2016 23rd International Conference on Pattern Recognition (ICPR)
Cancún Center, Cancún, México, December 4-8, 2016

Unknown Object Tracking in 360-Degree Camera Images

Ahmad Delforouzi, Seyed Amir Hossein Tabatabaei, Kimiaki Shirahama and Marcin Grzegorzek
Research Group for Pattern Recognition
University of Siegen, Germany
Emails: {ahmad.delforouzi, amir.tabatabaei, kimiaki.shirahama, marcin.grzegorzek}@uni-siegen.de

***Abstract*—In this paper, a method for unknown object tracking in output images from 360-degree cameras called Modified Training-Learning-Detection (MTLD) is presented. The proposed method is based on the recently introduced Training-Learning-Detection (TLD) scheme in the literature. The flaws of the TLD approach have been detected and significant modifications are proposed to enhance and to elaborate the scheme. Unlike TLD, MTLD is capable of detecting the unknown objects of interest in 360-degree images. According to the experimental results, the proposed method significantly outperforms the TLD method in terms of detection rate and implementation cost.**

I. Introduction

The 360-degree camera is a type of camera which provides 360-degree view. By a wider field of view, the required number of installed cameras can be significantly decreased in comparison with the case of commonly used normal cameras. The limited field of view for conventional cameras confines their applications as well. For example in the cases of areas like factories, big halls, or even in one single room, there are many blind spots when conventional cameras are in use. When the conventional security systems are replaced by 360-degree cameras, there will be a significant reduction of hardware costs, software license and maintenance costs. Thus the application of the new sensor of 360-degree cameras is growing in various areas such as robotics, car traffic control and intelligent surveillance systems.

Many methods have recently been developed to use similar types of cameras in various applications. Harabar et al. [1] utilized such an optic to automatically pilot a small traffic surveillance helicopter or a robot. Nayar et al. [2] produced a system which is able to detect activity in the monitored scene by fusing catadioptric sensors with PTZ cameras. In another application [3], a video surveillance system based on a catadioptric sensor for detecting and tracking objects in complex environments is presented. Human tracking for surveillance and security purposes is one of the important techniques which has attracted much attention in the state of the art (e.g., [4]). In some works [5, 6], color information has been used to analyze an object and resolve occlusion problems or to estimate the likelihood of objects matching for multiple human tracking in different cameras. In [7], a method based on the geometry and homography calibration has been introduced to overcome nonlinear spatial correspondences between the omnidirectional camera and the wide zoom range (PTZ) camera in object tracking application. Cui et al. [8] used background differencing and radial profile for object detection and tracking in dual camera systems. The nonuniform resolution of the omnidirectional camera and the corresponding calibration have been used in [9]. Multiple cues such as color, shape, and position are selected as human tracking features. In the images of fixed camera scenarios, the background is still. So by using background detection and subtraction, one can easily detect foreground objects. In other words, by looking for only moving objects in a more limited area, one can find the desired objects. There are many other methods for human detection and tracking [10–17] in literature. The human body is usually described by some simple shapes such as a circular shape for the top part and a cylindrical shape for other body parts. Thus, a very commonly used method is modeling the human appearance to 2d [10] or 3d shapes [11]. Some other methods are using offline-trained objects for the sake of human tracking [12].

II. Contribution of this paper

In this paper, the unknown object tracking in the images from the 360-degree camera is considered. Tracking of different objects e.g., vehicles, airplanes and humans by using of a unique system is very fascinating and various applications based on unknown object tracking can be designed to use in the industry. The problem of unknown object tracking in images is much more complicated than human tracking. Since no information about the desired object is available, model-based object tracking and also offline-trained object tracking are not applicable. Instead, other methods based on energy minimizing and/or online-training methods seem practical [15, 16]. The given contribution here is based on the recent online tracking approach called Learning-Training-Detection (TLD) [18]. The proposed method in this paper which is called Modified Learning-Training-Detection (MTLD) imposes important modifications on the TLD components to elaborately adapt it to solve the problem of unknown object tracking in 360-degree images. The problems of TLD and our modifications will be explained in the next sections.

This paper is organized as follows. The characteristics of 360-degree images are described in Section III. Section IV gives an overview of the TLD method. The MTLD method and its application on 360-degree images is explained in Section V. The experimental results, evaluation and comparisons are shown in Section VI. Section VII concludes the paper.

978-1-5090-4846-5/16/$31.00 ©2016 IEEE

Fig. 1: A 360-degree image

III. 360-DEGREE IMAGES

The samples of output images from a 360-degree camera are presented in this section. To generate the dataset [19], several video samples are used whose snapshots are shown in Fig. 1 and Fig. 2. The characteristics of video samples are as follows:

1) Natural in-plane rotation
2) Out-of-plane rotation
3) Moving camera
4) Various types of objects
5) High image resolution
6) Complex background

As shown, the 360-degree images are polar images in which the top region of a normal image has been transferred to the center of the corresponding polar image and other regions are arranged in the peripheral regions around the center. On the other hand, almost all video samples contain out-of-plane rotation which means the camera is looking at the diverse sides of the same object in different frames of each video. The used data-bank consists of some samples for both moving and fixed cameras wherein we are interested in the tracking of different objects e. g., human, airplane, car, motorcycle with diverse shapes and features. So in general, we need to develop an unknown object tracking method. Applying conventional tracking methods for unknown object tracking is very challenging. For example it would be very costly for the image size of 1500×1500 pixels as used in the experiments due to the total amount of processing load.

IV. AN OVERVIEW ON TRAINING-LEARNING-DETECTION

Training-Learning-Detection is an object tracking method which intends to track objects in normal videos when an unknown object is manually selected in the first frame. It uses online training to continuously detect the desired object and track it in successive frames. This method is composed of four modules including detection, tracking, learning and integration shortly described as follows.

1) The detection module looks for interesting objects in all frames independently. First, it segments the input image into all possible overlapping candidates with the same aspect ratio as of the desired object in the first frame. Then these candidates are fed to a hierarchical structure to be classified as desired object and or rejecting candidate.
2) The tracker estimates the current position of the object by using its position in the last frame. The tracking method is based on pyramidal Lucas-Kanade that uses Median-Flow for tracking [20]. However, in the TLD a failure detection [18] has been added to the median displacement measuring of the object's points.
3) The learning module uses two distinct parts of P and N experts to generate positive and negative examples respectively. P and N experts continuously bring new data for detector and tracker respectively. In other words P-experts identify just false negatives and N-experts identify only false positives [18].
4) The integrator compares the output of tracker and detector, then selects the final output by giving more priority to the detector.

The detector has a hierarchical framework to reject some of the candidates in each step. It is composed of three cascading classifiers as follows:

1) Variance Classifier: It rejects the non-object candidates (e.g., sky, street, snow).
2) Ensemble Classifier: It uses a set of pixel comparisons on a randomly chosen set of pixels [21].
3) Nearest Neighbor Classifier: It compares the similarity between remaining candidates and training data to a predefined threshold set to 0.65.

To evaluate the TLD method, we have applied it on 360-degree images. According to the experiment, the variance classifier is robust to rotation and can classify the candidates from 360-images. The ensemble classifier is sensitive to in-plane rotation and the nearest neighbor classifier is sensitive to both in-plane and out-of-plane rotation. In the next section, we improve the TLD method to be able to track the objects in 360-degree images.

V. MODIFIED TRAINING-LEARNING-DETECTION

A. Image Rectifying

Fig. 3 shows the building blocks of the proposed MTLD method. To overcome the problem of in-plane-rotation, a rectification transformation is engaged which converts the polar image of Fig. 1 to a normal image as shown in Fig. 4.To rectify the polar images we use the following equation:

$$x_r = G(y_p) \times cos(x_p), \quad y_r = G(y_p) \times sin(x_p) \qquad (1)$$

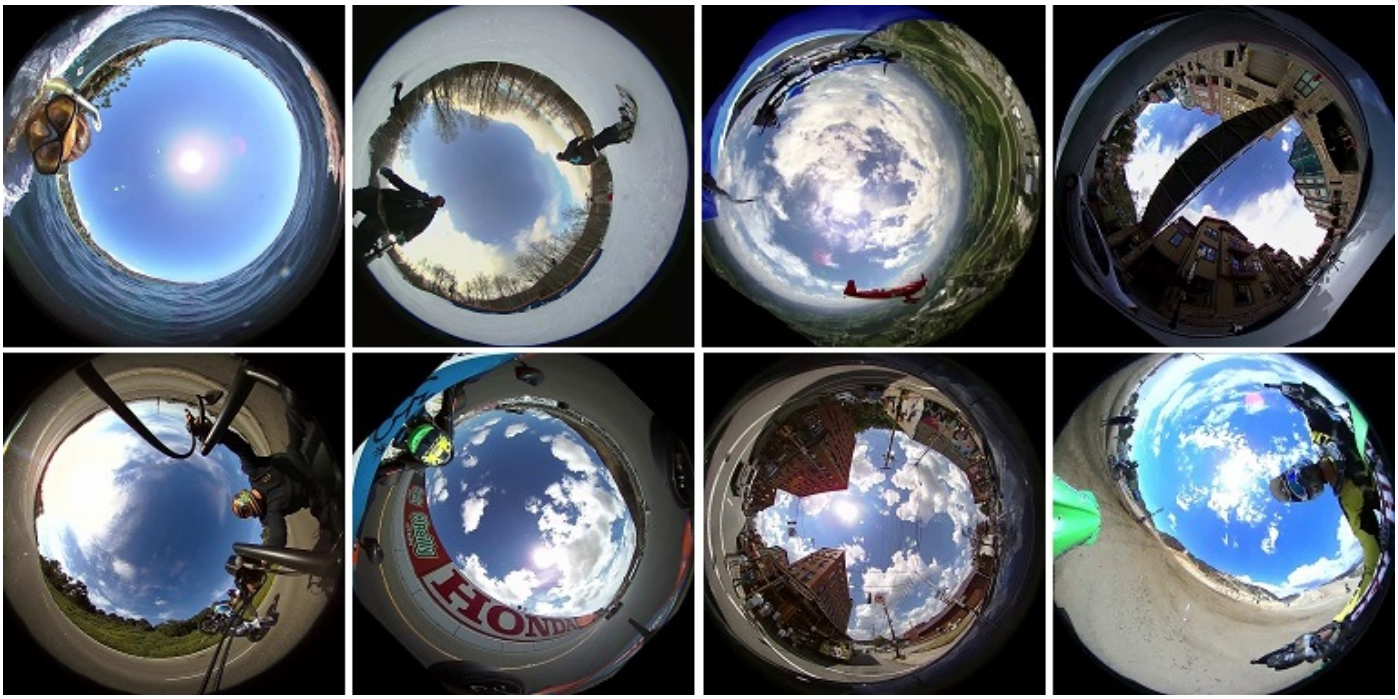

Fig. 2: 360-degree images

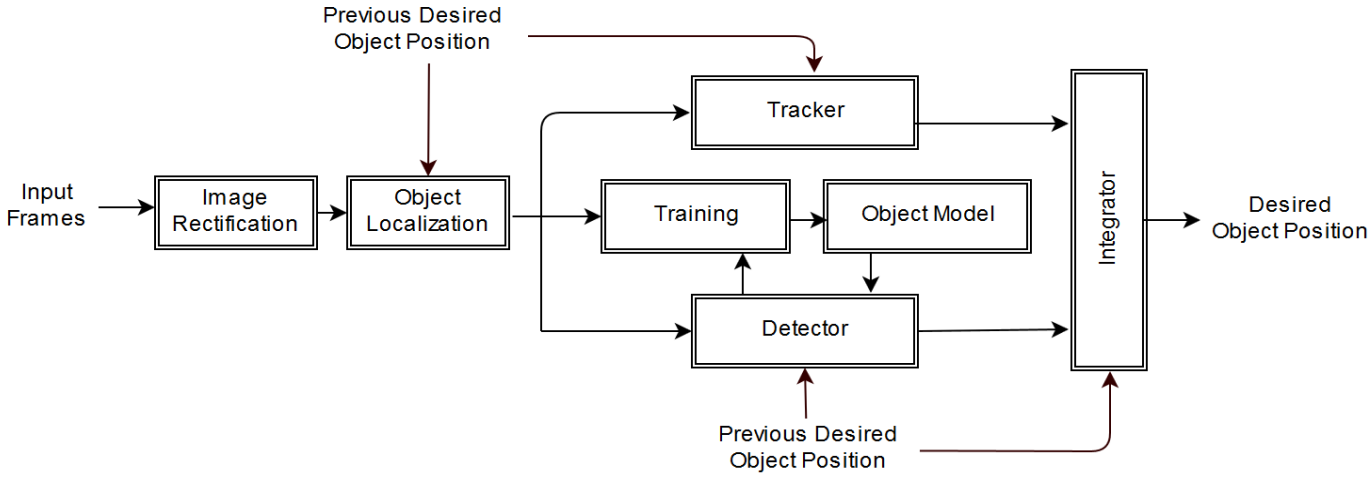

Fig. 3: Block diagram of the proposed MTLD method

where (x_p, y_p) and (x_r, y_r) are polar point and correspondent rectified point respectively and $G(.)$ is the stretching function over the Y axis. As a result, the direction of each object for a given frame is changed in a manner to become upright.

Fig. 4: Rectified image of Fig. 1

B. *Classifier Modifying*

By using image rectification in the MTLD approach, in-plane rotation is removed and the desired object could pass the ensemble classifier. However, in some cases, the desired object is still being rejected by the nearest neighbor classifier due to the out-of-plane rotation problem. To resolve this problem, the nearest neighbor classifier is modified in such a way to accept the desired object with out-of-plane rejection. This increases the acceptance rate of the nearest neighbor classifier and thus the one of whole system, but also the false positive rate increases. To tackle the latter concern, we have observed that many of objects are very far from their location in the previous frame. On the other hand, due to the high field of view of 360-degree

cameras, the movement of the desired object in both scenarios (fixed point camera and moving camera) is very limited. So we can easily ignore the remaining objects that are located far from the desired object. The block diagram of the proposed modified detector method is shown in Fig. 5 wherein a distance classifier is used to control the false positive rate caused by the modification of the nearest neighbor classifier. Thus, objects disqualified subjected to the constraint $\mid d_i - d_P \mid > 20$ are rejected wherein d_i and d_P denote the center of mass for candidate objects and last accepted object respectively.

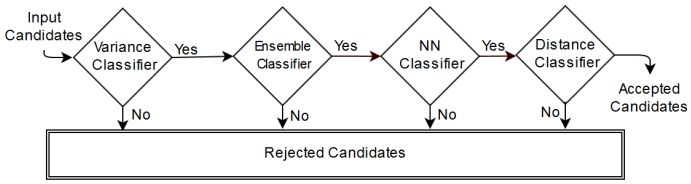

Fig. 5: Block diagram of the modified detector method

The decision making strategy in the integrator of the TLD scheme is very naive with room left for improvement. Regarding this matter, a comparison between the location of the detector output and tracker is made and the object which is closer to its location in the last frame is selected. So both the location of the object and its similarity to the accepted object in the previous frame are considered for output selection. It revives the role of the integrator in TLD which ignores the tracker for decision making. As a result, tracker and detector have equal chance to determine the final output. The experiments show that the acceptance rate for a desired object increases as well.

C. Restricting the Search Area

The proposed MTLD approach attempts to decrease the computational cost of the TLD method as well. For this sake, the object searching area in the MTLD method is restricted to a region around the position of the desired object in the previous frame. Because of the wide field of view in 360-degree images, even when the camera is moving, the objects in successive frames do not move swiftly. So the searching area can be limited without compromising the tracking performance rate.

D. Changing in the Detection Strategy

TLD uses the output of the tracker to train the detector. However, it is observed that it does not play an important role to update the detector. To increase the detection chance of the desired object, the current detected objects is included in the positive example set rather than the tracker output. This increases the chance of object detection especially in the case of out-of-plane rotation. According to the observation on various types of 360-images, a variation of the object size between two successive frames has been noticed. Thus, the object size variation in successive images is limited to $\pm 25\%$ in MTLD. The evacuation strategy for the object model has also been improved in MTLD. In the proposed strategy, the random selection has been substituted by a FIFO (i.e., first input first output) model. Using FIFO as an evacuation strategy will decrease the overfitting to the desired object in the first frame.

TABLE I: Evaluation results in terms of recall and precision

Video	FramesNo	TLD Recall/Precision	MTLD Recall/Precision
Snorkeling	683	0.038 / 1.0	0.84 / 0.87
Snowboard	248	0.516 / 0.516	0.67 / 0.67
Airplane	761	0.76 / 0.76	0.76 / 0.74
Street1	58	0.65 / 1.0	1.0 / 1.0
Pedestrians	129	0.64 / 0.64	0.86 / 0.86
Motor cycles	83	1.0 / 1.0	1.0 / 1.0
Car Racing	330	1.0 / 1.0	1.0 / 1.0
Street2	294	0.92 / 1.0	1.0 / 1.0
Motocross	250	1.0 / 1.0	1.0 / 1.0
Mean	-	0.6238 / 0.8749	0.8596 / 0.8615

VI. Experimental Results

To evaluate the proposed MTLD method, a set of 9 different video samples [19] with diverse frame numbers and various desired objects as shown in Fig. 1 and Fig. 2 is used. They have been captured in different scenes and the desirable objects for tracking include cars, motorcycles, pedestrians, the human head the human body and airplane. They have both in-plane and out-of-plane rotations. Our initial experiments showed that TLD cannot track objects in 360-degree images due to the lack of the rectification step. Thus, at first the input frames have been rectified and then inputted by TLD as done in MTLD modules. To evaluate the performance, the recall and precision variables have been used from the following equations.

$$Recall = \frac{T_P}{T_P + F_N}, \quad Precision = \frac{T_P}{T_P + F_P} \tag{2}$$

where T_P, F_N and F_P indicate true positive, false negative and false positive respectively. In this case, precision P is the number of true positives divided by the number of all responses and recall is the number of true positives divided by the number of object occurrences that should have been detected [18]. The proposed method is still sensitive to the background clutter. The results of the evaluation of the MTLD method and the TLD method in terms of recall and precision measures are listed in Tab. I.

According to the information of Tab. I, the proposed MTLD method outperforms the TLD method significantly. The mean recall rate has been improved by more than 20% while the precision rate stays in the same range as for the TLD method. Also, TLD is not successful in object tracking in the case of high rate of out-of-plane rotation in video samples of snowboard, street1, street2 and pedestrians. However, MTLD shows better results in these sample videos. Moreover, the TLD method fails to track the diver's head when he goes underwater. Therefore, TLD is very sensitive to environment changing, while the MTLD method can handle this variation. In the video of airplane, in some frames, the desired object goes to a region with a complex background which has some patterns similar to the object. Thus, neither MTLD nor TLD can track the object in those frames. The results of evaluation of recall measure for both methods of TLD and MTLD have been shown in a bar

TABLE II: Effect of each modified module on the recall rate of Snorkeling

Modules	Recall
TLD	0.038
TLD + NN	0.282
TLD + NN + Dis	0.761
TLD + NN + Dis + Int	0.770
TLD + NN + Dis + Int + T	0.821
TLD + NN + Dis + Int + T + FIFO	0.840

graph in Fig. 6. According to Fig. 6, MTLD has improved the recall variable for most of the video samples. For other video samples the recall value remains unchanged.

To evaluate individual effect of the each proposed modification in Section V, we applied each module on snorkeling. Because it has maximum recall difference between the MTLD and TLD. Restricting the search area does not have any effect on the recall rate and just increases the implementation speed. Let the NN threshold modification, Distance classifier, Integrator modification, changing the input source of the trainer and using the FIFO strategy for fulling the training queue be denoted by NN, Dis, Int, T and FIFO respectively. Tab. II shows the result of the above experiment.

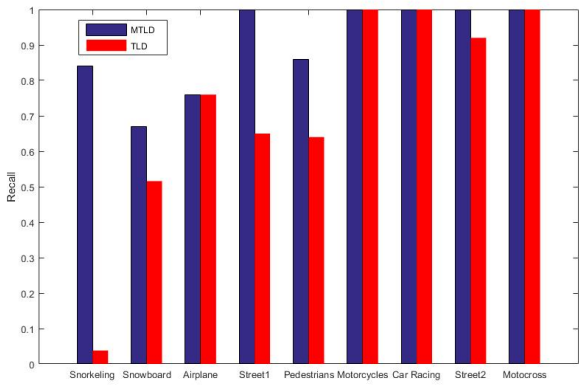

Fig. 6: Results of recall measures for MTLD and TLD

In another experiment, we measured the average computational time per frame for both MTLD and TLD. The comparison result has been shown in Fig. 7. As indicated, the implementation speed has been efficiently increased in MTLD for all video samples except for the snorkeling video. Since TLD mostly rejects all candidates of snorkeling in the first steps of the detector, the candidates do not pass through all steps of the detector. Thus, some modules of TLD are not involved in most frames of this video and therefore the total time consumption is limited. Finally, Fig. 8 and Fig. 9 show the whole image and area of interest for the searching region for the airplane video sample respectively. According to the figures, MTLD searches the desired object in more limited area than TLD. By searching the desired object in a limited area of the image, the number of candidates is dramatically reduced. So MTLD tracks objects in a lower period as shown in Fig. 7.

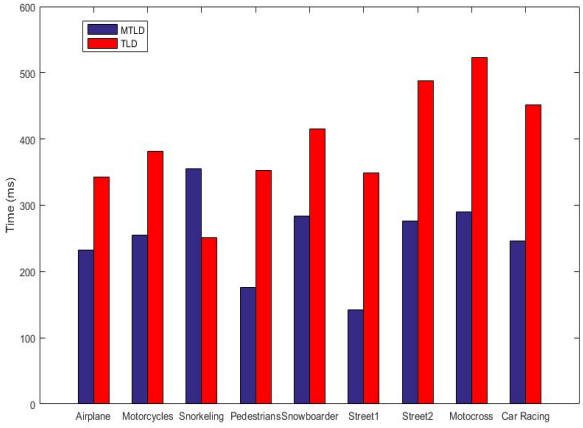

Fig. 7: Computation time comparison

VII. Conclusion

In this paper, we proposed an efficient method to track unknown objects in 360-degree images. We improved a state-of-the-art method of TLD to overcome the tracking problems in the challenging conditions of 360-degree images. The resolution of our 360-degree images is much higher than the TLD dataset, thus we suggested to restrict the searching area to decrease the computation load in the MTLD method. The experimental results show that our method outperforms the state-of-the-art method of TLD. This method can track objects even when they have out-of-plane rotation and varying environment. However, like TLD, our method is fragile to complex backgrounds when the desired object is similar to a part of the background. Therefore, our future work will address this concern.

Acknowledgment

Research and development activities leading to this article have been supported by the German Federal Ministry of Education and Research within the project “Cognitive Village: Adaptively Learning Technical Support System for Elderly” (Grant Number: 16SV7223K).

References

[1] S. Hrabar and G. S. Sukhatme, “Omnidirectional vision for an autonomous helicopter,” in *Robotics and Automation, 2003. Proceedings. ICRA ’03. IEEE International Conference on*, vol. 1, pp. 558–563 vol.1, Sep. 2003.

[2] S. Nayar and T. Boult, “Omnidirectional vision systems: 1998 pi report,” 1998.

[3] T. E. Boult, R. J. Micheals, X. Gao, and M. Eckmann, “Into the woods: visual surveillance of noncooperative and camouflaged targets in complex outdoor settings,” *Proceedings of the IEEE*, vol. 89, pp. 1382–1402, Oct. 2001.

[4] D. T. Lin and K. Y. Huang, “Collaborative pedestrian tracking with multiple cameras: Data fusion and visual-

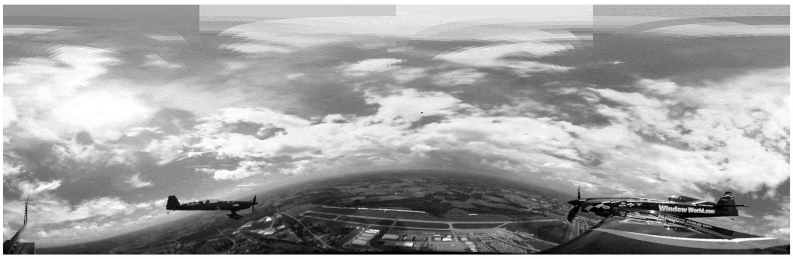

Fig. 8: whole image searching in TLD

Fig. 9: Limited searching area in MTLD

ization," in *Neural Networks (IJCNN), The 2010 International Joint Conference on*, pp. 1–8, July 2010.

[5] T. Fukuda, T. Suzuki, F. Kobayashi, F. Arai, Y. Hasegawa, and M. Negi, "Seamless tracking system with multiple cameras," in *Industrial Electronics Society, 2000. IECON 2000. 26th Annual Confjerence of the IEEE*, vol. 2, pp. 1249–1254 vol.2, 2000.

[6] L. Zhu, J. N. Hwang, and H. Y. Cheng, "Tracking of multiple objects across multiple cameras with overlapping and non-overlapping views," in *Circuits and Systems, 2009. ISCAS 2009. IEEE International Symposium on*, pp. 1056–1060, May 2009.

[7] C. H. Chen, Y. Yao, D. Page, B. Abidi, A. Koschan, and M. Abidi, "Heterogeneous fusion of omnidirectional and ptz cameras for multiple object tracking," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 18, pp. 1052–1063, Aug. 2008.

[8] Y. Cui, S. Samarasckera, Q. Huang, and M. Greiffenhagen, "Indoor monitoring via the collaboration between a peripheral sensor and a foveal sensor," in *Visual Surveillance, 1998. Proceedings., 1998 IEEE Workshop on*, pp. 2–9, Jan. 1998.

[9] G. Scotti, L. Marcenaro, C. Coelho, F. Selvaggi, and C. S. Regazzoni, "A novel dual camera intelligent sensor for high definition 360 degrees surveillance," in *Intelligent Distributed Surveilliance Systems, IEE*, pp. 26–30, Feb. 2004.

[10] Z. Lin, L. S. Davis, D. Doermann, and D. DeMenthon, "Hierarchical part-template matching for human detection and segmentation," in *Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on*, pp. 1–8, Oct. 2007.

[11] L. Wang and N. H. C. Yung, "Three-dimensional model-based human detection in crowded scenes," *IEEE Transactions on Intelligent Transportation Systems*, vol. 13, pp. 691–703, June 2012.

[12] B. Leibe, E. Seemann, and B. Schiele, "Pedestrian detection in crowded scenes," in *Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05) - Volume 1 - Volume 01*, CVPR '05, (Washington, DC, USA), pp. 878–885, IEEE Computer Society, 2005.

[13] L. Wang, N. H. C. Yung, and L. Xu, "Multiple-human tracking by iterative data association and detection update," *IEEE Transactions on Intelligent Transportation Systems*, vol. 15, pp. 1886–1899, Oct. 2014.

[14] L. Wang and N. H. C. Yung, "Extraction of moving objects from their background based on multiple adaptive thresholds and boundary evaluation," *IEEE Transactions on Intelligent Transportation Systems*, vol. 11, pp. 40–51, Mar. 2010.

[15] A. Milan, K. Schindler, and S. Roth, "Multi-target tracking by discrete-continuous energy minimization," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. PP, no. 99, pp. 1–1, 2015.

[16] A. Milan, S. Roth, and K. Schindler, "Continuous energy minimization for multitarget tracking," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 36, no. 1, pp. 58–72, 2014.

[17] B. Wu and R. Nevatia, "Detection and tracking of multiple, partially occluded humans by bayesian combination of edgelet based part detectors," *International Journal of Computer Vision*, vol. 75, no. 2, pp. 247–266, 2007.

[18] Z. Kalal, K. Mikolajczyk, and J. Matas, "Tracking-learning-detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 7, pp. 1409–1422, 2012.

[19] https://360fly.com/videos, last visit: Sep. 2015.

[20] J. yves Bouguet, "Pyramidal implementation of the lucas kanade feature tracker," *Intel Corporation, Microprocessor Research Labs*, 2000.

[21] M. Ozuysal, P. Fua, and V. Lepetit, "Fast keypoint recognition in ten lines of code," in *Proceedings of the 2007 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'07)*, pp. 1–8, June 2007.