Mono-Vision based Moving Object Detection using Semantic-Guided RANSAC

Songming Chen, Haixin Sun, Vincent Frémont

Mono-Vision based Moving Object Detection using Semantic-Guided RANSAC

Songming Chen[1], Haixin Sun[1], Vincent Frémont[1]

***Abstract*— This paper proposes a novel two-stage approach for detecting moving objects with a non-stationary monocular camera mounted on a vehicle. We formulate an innovative method called semantic-guided random sample consensus (Semantic-Guided RANSAC) to detect moving objects by semantic-geometric information fusion and integration. Firstly, semantic constraints from deep learning architecture (YOLO v4) are applied to predict the objects' location in the image frame. The fundamental matrix is then estimated robustly from two views through the sparse optical flow tracking with the help of semantic prior. Semantic-guided RANSAC is used to reject instance-level outliers which are actually moving objects based on the epipolar geometry and flow vector bound constraints. Experimental results on KITTI dataset reflect the effectiveness of our approach to identify moving objects in complex urban traffic scenes with the average precision above 0.82 for 4 sequences in the City category.**

I. Introduction

Vision-based traffic scene understanding is a complex yet indispensable task for the perception of autonomous vehicles. Typically, Moving Object Detection (MOD) is fundamental for high-level tasks such as obstacle avoidance in dynamic and uncertain environments. Identifying the dynamic objects also plays an important role in the vision based ego-motion estimation problem which usually has the assumption of static surroundings. Being able to recognize moving objects (cars, bicycles) and to obtain their states (stationary, non-stationary) can facilitate the safety of the autonomous vehicle operation.

Substantial research work has been devoted to the domain of moving object detection in recent years. Background subtraction [1] approach is widely applied to handle the MOD problem when image sequences are acquired from a static camera. However, for a moving camera, this approach cannot be directly utilized without additional constraints imposed. Because of the vehicle ego-motion, the object-motion and ego-motion are coupled together which makes the background subtraction non-trivial. In order to decouple and compensate for the ego-motion, the epipolar geometry [1] is commonly adopted for ego-motion estimation across two consecutive frames. Unfortunately, sparse feature-based state estimation may be unstable when the non-static feature points are chosen and incorporated in the estimation process. By convention, dynamic objects are regarded as outliers and a random sample consensus (RANSAC) [2] method is often applied to filter them out. However, this strategy fails to operate when the dynamic objects turn out to be the dominant components in the scene. Thus, effective moving object detection in a complex scene remains a critical issue to be solved for the perception of autonomous vehicles.

There are many challenges in developing a good moving object detector. It should be robust against aggressive ego-motion and be capable of tackling random moving objects motion flow. You Only Look Once (YOLO) [3] approach is a state-of-the-art, real-time object detection system where a deep neural network is passed only

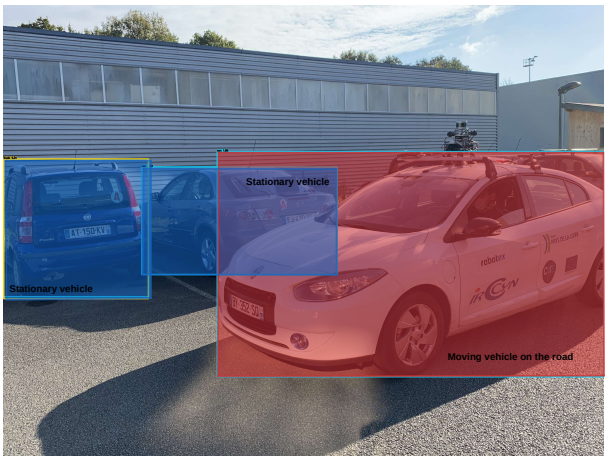


Fig. 1. Semantic-guided RANSAC scheme applied for moving object detection in the campus of Ecole Centrale de Nantes, blue color stands for stationary objects and red color represents non-stationary objects

* This work was supported by China Scholarship Council

[1] S. Chen, H. Sun and V. Frémont are with the Laboratoire des sciences du numérique de Nantes (LS2N), UMR 6004, at École Centrale de Nantes, 44321 Nantes, France. songming.chen@ec-nantes.fr, haixin.sun@ec-nantes.fr, vincent.fremont@ec-nantes.fr

once to the whole image frame. The YOLO network cuts the image into different segments and outputs instance-level bounding boxes which are weighted by the prediction scores. The YOLO network is able to predict the existence of objects in the scene. However, the state of an object remains unknown if its semantic label is not definitely static, such as pedestrian, bicycle or vehicle which are defined as movable objects in this paper. In Microsoft COCO dataset [4], definitely static (non-movable) objects are listed as traffic light, fire hydrant, stop sign, parking meter, bench and potted plant. The rest are classified as movable objects which need further information to solve the ambiguities of their state. For movable objects, we formulate a semantic-guided RANSAC algorithm to reject instance-level outliers which helps to discriminate truly moving objects from stationary ones.

The main contribution of this paper is to fuse the geometric and semantic information to segment the moving objects within the field of view. The RANSAC process is applied at a higher level abstraction to reject instance-level outliers which depends on the proportion of pixel-wise outliers in the bounding box. A moving object is extracted seamlessly from the semantic-guided RANSAC process and the computational complexity is reduced since sparse Shi-Tomasi corner features are used with semantic prior instead of all pixels in an image. Moreover, the fundamental matrix estimation process becomes more robust and efficient by taking the semantic prior into consideration. Without dedicated objects tracking and ego-motion estimation, our approach still achieves high precision and F-score on KITTI benchmarking sequences.

The remaining parts of this paper are organized as follows. Section II presents the recent literature regarding model-based and learning-based moving object detection with a non-stationary camera. The innovation and contribution of the semantic RANSAC approach is elaborated in Section III and validated with the KITTI benchmarks in Section IV. A concise conclusion and future work plan are given in Section V.

II. Related work

A. Geometric Constraint Based Detection

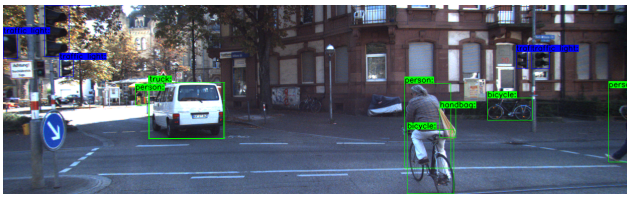
In [5], the multi-view geometry and the structure consistency constraints are combined to segment moving objects in the scene with a monocular camera. Such redundant constraints ensure the high detection precision in degraded circumstances. In [6], the challenge of estimating a vehicle's ego motion as well as the movements of dynamic objects at the same time is addressed based on projective factorization of the multiple-trajectory matrix. Stereo-vision based moving object detection methods are proposed in [7] [8] [9], where the motion likelihood of every pixel is calculated given the approximated ego-motion uncertainty and U-disparity map is built to characterize on-road obstacles. Color and depth hints are leveraged in the graph-cut framework for connected regions (moving objects) extraction. In [10], a Bayesian framework is applied to generate a probability value for each pixel, either being static or dynamic, according to the epipolar and focus of expansion constraints. The framework enables the system to detect moving objects with degenerate motion due to the flow vector bound constraint attached.

B. Deep Neural Network Based Detection

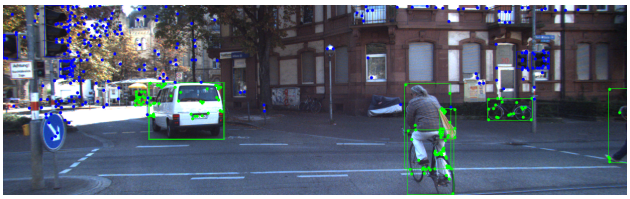
In [11], an unsupervised adversarial contextual model is proposed to detect dynamic objects in the image frame. The contextual information of the surroundings is fed for the neural network training to infer the optical flow in specific regions, meanwhile another network formats the context as uninformative as possible since the optical flow of a moving object is uncorrelated with the background. The term of moving objectness is introduced in [12], which represents the possibility that they belong to moving objects. Several prediction are firstly proposed using multiple figure-ground segmentations and then the proposals are ranked with the moving objectness criteria to identify moving objects. In [13], Neural-Guided RANSAC is applied to a wide range of computer vision tasks such as fundamental matrix estimation, horizon line estimation and camera re-localization. Different from differentiable RANSAC in [13], our two-stage approach (semantic prediction and geometric validation) is more flexible to add constraints to detect the objects with degenerate motion regardless of the ego-motion variation, without modifying or re-training the existing neural networks.

III. Presentation of the method

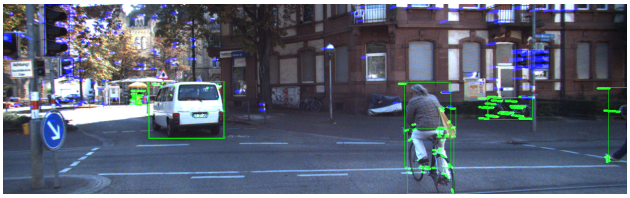
In Fig. 2, it is shown that the proposed framework starts with a yolo object detection module. Objects with static semantic labels such as traffic lights are directly classified as stationary. However, movable objects with the labels such as person, bicycle and car need further information to make the inference. Thus, Shi-Tomasi corner points [15] are extracted from the image and iteratively tracked using Lucas-Kanade optical flow. Feature points which belong to the static objects and background are utilized to estimate the fundamental matrix. Semantic-guided RANSAC takes full advantage of instance-level semantic segmentation and enables the



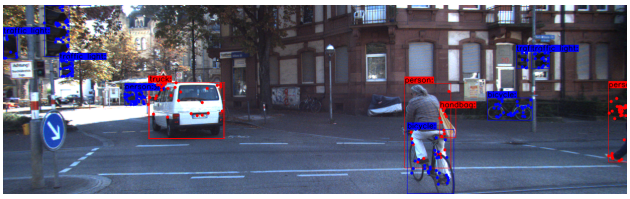
2.(a) Get the bounding boxes with semantic labels using yolo v4 detector



2.(b) Extract the Shi-Tomasi corner feature points, rendering in green for movable objects, rendering in blue for non-movable objects and background points



2.(c) Estimate the Fundamental matrix with the feature correspondences rendering in blue (pyramidal Kanade-Lucas-Tomasi tracker)



2.(d) Render the moving bounding box in red when the proportion of pixel-level outliers in the bounding box is over the threshold of 0.6

Fig. 2. Overview of the proposed semantic-guided RANSAC for moving objects detection. Blue, red, green colors stand for stationary, non-stationary and unknown objects respectively in kitti dataset [14]

fusion of semantic labels and geometric constraints for moving object detection. Combining semantic and geometric cues results in accurate moving object detection by checking the residual value of the epipolar constraint and flow vector bound for all suspicious points lying in the movable bounding boxes. Instead of training an end-to-end fashion neural network which outputs object existence and its state, a two-stage approach is taken. The yolo network output provides good semantic prior to predict the existence of the objects and then semantic-guided ransac decides the state of the objects based on the epipolar geometry and flow vector bound constraint which is detailed in Section III-C.

A. Semantic-Guided KLT Feature Tracking

The well-known Kanade–Lucas–Tomasi (KLT) [16] tracker leverages spatial intensity cues to guide the search for the corresponding features across two frames. In order to deal with large camera motion across frames, a pyramidal KLT tracker is implemented to allow for tracking points with large displacements between frames. Moreover, semantic label consistency and forward-backward flow consistency constraints are added into feature points tracking process to reduce the occurrence of mismatches due to occluded pixels and pixels with strong illumination changes.

Pyramidal KLT tracker can be applied to get the pair of matched feature points $(\boldsymbol{x}_1, \boldsymbol{x}_2)$ across frames, $\boldsymbol{x}_2 = KLT_{forward}(\boldsymbol{x}_1)$. Then the optical flow propagates backward to get the estimated initial feature point $\hat{\boldsymbol{x}}_1$, $\hat{\boldsymbol{x}}_1 = KLT_{backward}(\boldsymbol{x}_2)$. The forward-backward constraint is imposed to compute the euclidean distance $dist(\hat{\boldsymbol{x}}_1, \boldsymbol{x}_1)$ for matched points, and this metric is used to discard potentially erroneous feature matches when their discrepancy is over 2 pixels.

$$Status(\boldsymbol{x_1}, \boldsymbol{x_2}) = \begin{cases} valid & dist(\hat{\boldsymbol{x}}_1, \boldsymbol{x}_1) < 2 \\ invalid & dist(\hat{\boldsymbol{x}}_1, \boldsymbol{x}_1) > 2 \end{cases} \tag{1}$$

$$Status(\boldsymbol{x_1}, \boldsymbol{x_2}) = \begin{cases} valid & label(\boldsymbol{x}_1) == label(\boldsymbol{x}_2) \\ invalid & label(\boldsymbol{x}_1) != label(\boldsymbol{x}_2) \end{cases} \tag{2}$$

B. Semantic-Guided Fundamental Matrix Estimation

RANSAC [2] is an iterative method to estimate the underlying model parameters which meanwhile divides the input data into inliers and outliers. The main limitation of RANSAC is that, when large number of outliers are incorporated in the dataset, biased estimation output may be provided due to the limits of iteration times. Due to the increase of outliers ratio, RANSAC needs exponentially more iterations to reach the point with a outlier-free subset found, see Fig. 3. The expected number of iterations r to reach a certain probability p with a minimal outlier-free subset found is

$$r = \frac{\log(1-p)}{\log(1-w^N)} \tag{3}$$

where w is the fraction of inliers and N is the minimum number of samples needed for model estimation which should be eight pairs [1] of matching points for fundamental matrix estimation in our case. Higher inlier fraction is preferred since it helps to incorporate more correct correspondences in the consensus set and fewer iterations are needed get obtain the model parameters.

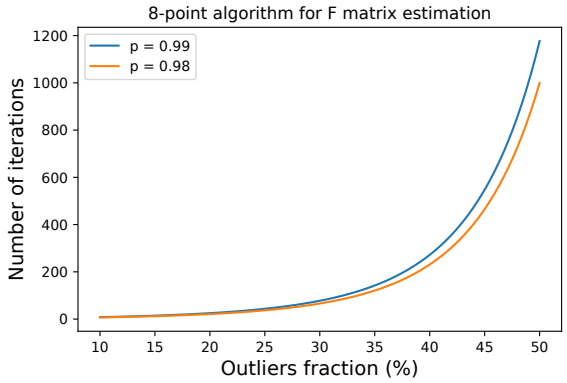


Fig. 3. RANSAC number of iterations for 8-point fundamental matrix estimation

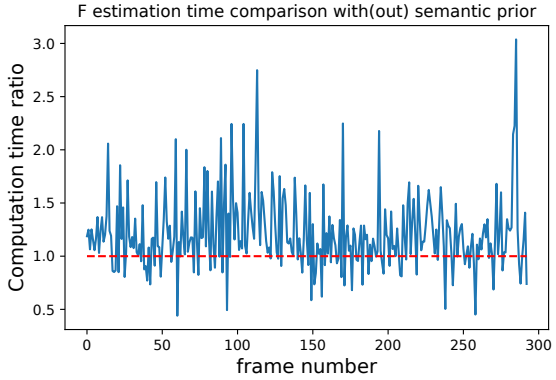


Fig. 4. The ratio of fundamental matrix estimation time without semantic prior to that with semantic prior

Semantic-guided fundamental matrix estimation makes use of semantic priors to guide the model fitting, which facilitates obtaining the outlier-free minimal subset. Static feature points belonging to background and objects with non-movable semantic labels such as traffic lights and traffic signs have higher priority than movable objects such as pedestrians and vehicles to be utilized for fundamental matrix estimation. Moreover, the semantic bounding box from yolo v4 output whose objectness score is lower than a threshold of 0.2 is suppressed and not taken into account. The semantic prior ultimately increases the fraction of inliers in the tentative pairs set and as a result, fundamental matrix estimation is more well conditioned which requires much less number iterations to converge, see Fig. 4, where the frames are taken from KITTI raw data 2011_09_26_*drive*_0056 sequence.

C. Semantic-Guided Instance-Level Outlier Rejection

Considering pairs of matched points belonging to the background and static objects, the fundamental matrix is robustly estimated with the 8-point algorithm in Section III-B. And given a pair of matched points $(\boldsymbol{p}_n, \boldsymbol{p}_{n+1})$ lying in the movable object, geometric constraint can be leveraged to distinguish the truly dynamic objects from the static ones. Fundamental matrix maps the point $\boldsymbol{p}_{n(n+1)}$ to its corresponding epipolar line $\boldsymbol{l}_{n+1(n)}$ as $\boldsymbol{l}_{n+1(n)} \sim \boldsymbol{F}\boldsymbol{p}_{n(n+1)}$ across two frames, where $\sim$ represents an up-to-scale equality. Then it is possible to calculate the epipolar geometry residual $r_{\boldsymbol{F}}$ for matched points to implement outlier rejection based on point-to-line distance d_{p2l} in the image.

$$r_{\boldsymbol{F}} = \max\{d_{p2l}(\boldsymbol{p}_n, \boldsymbol{l}_n), d_{p2l}(\boldsymbol{p}_{n+1}, \boldsymbol{l}_{n+1})\} \tag{4}$$

However, when a 3D point in the scene appears on the epipolar plane which is constructed by the point itself the camera center in the previous and current frames, the perspective projection of the moving point always stays on the corresponding epipolar line. In this case, null epipolar residual does not represent the point is static. So the epipolar constraint is not capable to detect such moving points with degenerate motion. Thus, another constraint of Flow Vector Bound (FVB) [10] is additionally imposed to find the bound of parallax range $[d_{min}, d_{max}]$ for static and background points in the scene. Given images captured from a pin-hole camera, pixel-wise displacement d^i for the feature point $\boldsymbol{p}^i = (u_i, v_i)$ which has the depth value z can be obtained with the equation:

$$\begin{aligned} \boldsymbol{p}^i_{n+1} - \boldsymbol{K}\boldsymbol{R}_{n:n+1}\boldsymbol{K}^{-1}\boldsymbol{p}^i_n &= \tfrac{1}{z}\boldsymbol{K}\boldsymbol{t}_{n:n+1} \\ d^i &= \left|\boldsymbol{p}^i_{n+1} - \boldsymbol{K}\boldsymbol{R}_{n:n+1}\boldsymbol{K}^{-1}\boldsymbol{p}^i_n\right| \end{aligned} \tag{5}$$

where $\boldsymbol{K}$, $\boldsymbol{R}_{n:n+1}$ and $\boldsymbol{t}_{n:n+1}$ are the camera intrinsics, rotation and translation from timestamp n to $n+1$ respectively. Any point with a parallax value d^i which is not in the range $[d_{min}, d_{max}]$ will be also set as an outlier and rejected. Then, the ambiguous movable object can be classified as a truly dynamic (static) object if there are more than 8 feature points lying on the object and the proportion of pixel-level outliers is above (below) the threshold of 0.6. If the number of feature points inside the bounding box is less than 8 (minimum number for independent fundamental matrix estimation), the state of the object is set as unknown and waiting for further information to make the decision. Moreover it needs to be noted that epipolar geometry constraint only works for the moving camera mounted on the vehicle. When the camera does not move, the epipolar line can not be defined. In this situation, FVB constraint

complements the epipolar geometry constraint to detect moving objects in the scene.

Algorithm 1 Semantic-Guided Instance-level Outlier Rejection

Input: Corresponding feature points in two consecutive frames

Output: Segmented moving bounding boxes in the scene frame

1: ▶ Extract the background and static feature points in the scene by excluding the feature points with movable semantic labels
2: ▶ Apply the 8-point algorithm to estimate $\boldsymbol{F}$ with the static and background feature points
3: ▶ Check how well $\boldsymbol{F}$ matches feature points in the bounding boxes with movable labels using Eq. 4
4: ▶ Impose FVB constraint to detect feature points on objects with degenerate motion using Eq. 5
5: ▶ Determine the movable object as the truly dynamic object if there are enough feature points on the object and the proportion of pixel-level feature point outliers is above the threshold of 0.6

IV. EXPERIMENTAL RESULTS

A. Evaluation Metrics

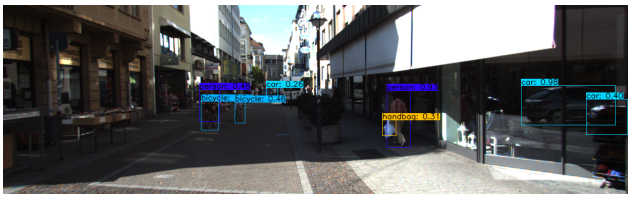
The KITTI dataset [14] contains image sequences recorded in urban and highway environments. In the category of raw data, 2D bounding boxes tracklets of moving objects are provided for several sequences. Our system is evaluated at the bounding box level with the metrics of Precision and F-score defined as:

$$P = \frac{t_p}{t_p + f_p}\ ,\ F = \frac{2t_p}{2t_p + f_p + f_n} \tag{6}$$

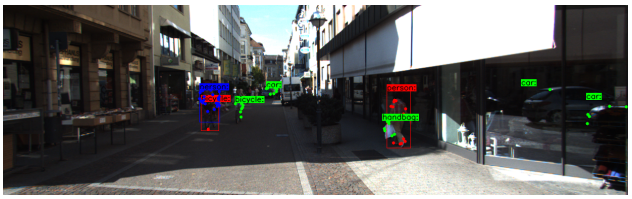
with t_p, f_p and f_n represent true positive detection, false positive misdetection and false negative alarm successively. Evaluation results are obtained on a laptop PC with an Intel i7-9750H CPU and 32GB of RAM.

B. Result Analysis

Fig. 5. Flow vector bound constraint for detecting moving objects with degenerate motion with the frame from KITTI raw data sequence 05



6.(a) YOLO network wrongly classifies reflections in the mirror as cars with high confidence



6.(b) States of cars in the mirror are set as unknown (green) in our pipeline due to lack of consistently matched feature points for decision-making

Fig. 6. Result analysis for false alarms due to mirror reflection with the frame from KITTI raw data sequence 71

In order to highlight the advantage of our proposed approach which fuses semantic and geometric information, the method presented in [8] and [9] which utilize stereo-vision without considering semantic clues are chosen as the baseline. Four typical heavy traffic scene sequences are selected to evaluate our system and a short demonstration video can be accessed with the attached link[1]. Tab. I illustrates the quantitative results for the comparison purpose. From Tab. I, it is shown that the precision of moving object detection has been greatly improved due to the semantic cues involvement. Taking the semantic information into account increases the true positive detection rate and decreases the false negative alarm rate. Moreover, with RTX 2070 GPU acceleration, our system can run at the speed of 30FPS, which achieves real-time performance. It is superior to the approach in [8] and which takes more than 0.2 second to estimate the ego-motion along with its uncertainty for each single frame. In our method, the ego-motion is implicitly integrated in the epipolar geometry point-to-line residuals and the sparse feature points optical flow ensures the efficiency of the whole pipeline. Fig. 5 demonstrates that, the FVB constraint effectively helps to detect moving objects with degenerate motion. At the same time, the FVB constraint also applies when the ego-motion is null, see the video presentation, which makes our system robust against the ego-motion variation. And it is presented in Fig. 6 that, the false

[1]https://uncloud.univ-nantes.fr/index.php/s/KLzoSQqPdFnt5fZ

TABLE I
COMPARISON OF MOD ACCURACY

Methods	Metrics	05	11	51	56
Our approach	Precision	**0.700**	**0.868**	**0.856**	**0.885**
	F-Score	0.762	**0.878**	0.773	**0.798**
Approach in [8]	Precision	0.690	0.696	0.680	0.768
	F-Score	**0.780**	0.792	**0.799**	0.777
Approach in [9]	Precision	0.383	0.675	0.556	0.510
	F-Score	0.513	0.770	0.706	0.664

alarms due to mirror reflection are set as unknown states because of the minimum feature number constraint imposed. Compared with the end-to-end moving object detection approach, our two-stage method is more flexible to add constraints without modifying or retraining the existing neural networks. Despite these advantages, our framework also has some drawbacks. It does not perform very well in some certain scenario where the objects are almost out of the range of detection. In this situation, they appear to be very small and there are not enough feature points on them for decision-making. Moreover, when the static objects are getting close to the vehicle due to ego-motion, false alarms will be raised if their parallax across frames exceed the flow vector bound for the current frame. Besides, object mutual occlusion might also cause the false alarm occurrence when overlapping bounding boxes share the feature points for outlier rejection. Indeed, in these situations, the detection precision degrades. However, in practice false alarms do not have fatal impact for the autonomous driving and misdetection of moving objects in the scene is not that critical when the objects are far away from the vehicle.

V. Conclusion and future work

In this paper, an effective approach which leverages both semantic and geometric cues is proposed to segment moving objects with a non-stationary monocular camera mounted on a vehicle. According to the experimental results, the proposed method provides accurate moving object detection in complex urban environments which achieves high precision (above 0.82) on KITTI benchmarking sequences. Despite many advantages of our proposed method, our system relies on the deep learning based objectness prediction. In the future, additional geometric constraints incorporating depth variance information will be employed to actively cluster high probability moving points in order to reduce the false alarm and misdetection effect.

References

[1] R. I. Hartley and A. Zisserman, *Multiple View Geometry in Computer Vision*, 2nd ed. Cambridge University Press, ISBN: 0521540518, 2004.

[2] M. A. Fischler and R. C. Bolles, "Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography," *Communications of the ACM*, vol. 24, no. 6, pp. 381–395, 1981.

[3] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 779–788.

[4] 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 *Proceedings of European Conference on Computer Vision (ECCV)*, 2014, pp. 740–755.

[5] V. Frémont, S. A. R. Florez, and B. Wang, "Mono-vision based moving object detection in complex traffic scenes," in *Proceedings of IEEE Intelligent Vehicles Symposium (IV)*. IEEE, 2017, pp. 1078–1084.

[6] R. Sabzevari and D. Scaramuzza, "Multi-body motion estimation from monocular vehicle-mounted cameras," *IEEE Transactions on Robotics*, vol. 32, no. 3, pp. 638–651, 2016.

[7] B. Wang, S. A. R. Florez, and V. Frémont, "Multiple obstacle detection and tracking using stereo vision: application and analysis," in *Proceedings of International Conference on Control Automation Robotics & Vision (ICARCV)*, 2014, pp. 1074–1079.

[8] D. Zhou, V. Frémont, B. Quost, Y. Dai, and H. Li, "Moving object detection and segmentation in urban environments from a moving platform," *Image and Vision Computing*, vol. 68, pp. 76–87, 2017.

[9] A. Talukder and L. Matthies, "Real-time detection of moving objects from moving vehicles using dense stereo and optical flow," in *Proceedings of IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2004, pp. 3718–3725.

[10] A. Kundu, K. M. Krishna, and J. Sivaswamy, "Moving object detection by multi-view geometric techniques from a single camera mounted robot," in *Proceedings of IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2009, pp. 4306–4312.

[11] Y. Yang, A. Loquercio, D. Scaramuzza, and S. Soatto, "Unsupervised moving object detection via contextual information separation," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 879–888.

[12] K. Fragkiadaki, P. Arbelaez, P. Felsen, and J. Malik, "Learning to segment moving objects in videos," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 4083–4090.

[13] E. Brachmann and C. Rother, "Neural-guided ransac: Learning where to sample model hypotheses," in *Proceedings of IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019, pp. 4322–4331.

[14] A. Geiger, P. Lenz, C. Stiller, and R. Urtasun, "Vision meets robotics: The kitti dataset," *The International Journal of Robotics Research*, vol. 32, no. 11, pp. 1231–1237, 2013.

[15] J. Shi and C. Tomasi, "Good features to track," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 1994, pp. 593–600.

[16] B. D. Lucas, T. Kanade, *et al.*, "An iterative image registration technique with an application to stereo vision," in *Proceedings of the 7th international joint conference on Artificial intelligence*, 1981, pp. 24–28.