Moving object detection from a point cloud using photometric and depth consistencies

Atsushi Takabe*, Hikari Takehara*, Norihiko Kawai*, Tomokazu Sato*,
Takashi Machida†, Satoru Nakanishi† and Naokazu Yokoya*
*Nara Institute of Science and Technology,
8916-5, Takayama-cho, Ikoma, Nara, Japan
Email: takabe.atsushi.sm7@is.naist.jp
†TOYOTA Central Research and Development Labs,
41-1, Yokomichi, Nagakute, Aichi, Japan

***Abstract*—3D models of outdoor environments have been used for several applications such as a virtual earth system and a vision-based vehicle safety system. 3D data for constructing such 3D models are often measured by an on-vehicle system equipped with laser rangefinders, cameras, and GPS/IMU. However, 3D data of moving objects on streets lead to inaccurate 3D models when modeling outdoor environments. To solve this problem, this paper proposes a moving object detection method for point clouds by minimizing an energy function based on photometric and depth consistencies assuming that input data consist of synchronized point clouds, images, and camera poses from a single sequence captured with a moving on-vehicle system.**

I. Introduction

3D models of outdoor environments have been used for several applications such as a virtual earth system [1] and a vision-based vehicle safety system [2]. Such 3D models are constructed by real-world measurement using an on-vehicle system equipped with laser rangefinders, cameras, and GPS/IMU [3]–[5]. Fig. 1 shows an example of an on-vehicle system equipped with sensors.

One problem here is that moving objects in the measured data lead to inaccurate 3D models and such 3D models cause undesirable artifacts in virtual views generated using the 3D models. For example, if we produce virtual views using view-dependent image-based rendering technique [6] from 3D models including moving objects, implausible textures often appear as shown in Fig. 2. To cope with this problem, a number of methods that detect or remove moving object regions in the data of outdoor environments have been proposed.

Kanatani *et al.* [7] detect 3D points on moving objects using photometric consistency between pixels obtained by projecting a 3D point onto omnidirectional images captured from different viewpoints. Since this method uses only photometric consistency, this method cannot detect points on moving objects whose luminance values are similar to those on background static objects. In addition, the applicable range is limited because the method is based on the assumption that the moving object exists on a road. Yan *et al.* [8] detect and remove moving objects from a single sequence data captured with a moving on-vehicle system. They detect moving objects by tracking sparse 3D points on the moving objects between a reference frame and its next frame. Although the accuracy

Fig. 1. On-vehicle system.

Fig. 2. Implausible image.

of the moving object detection largely depends on that of object tracking, this method often fails in tracking moving objects because tracking a sparse 3D point cloud captured with a laser rangefinder while moving is quite challenging. Huang *et al.* [9], Premebida *et al.* [10], and Spinello *et al.* [11] detect moving objects based on machine learning. They first detect moving objects from images and 3D point clouds, which are measured with LIDAR, using supervised object detection methods, and then integrate the detection results. Since these methods can detect only specified moving objects such as humans and vehicles, it is difficult to detect unknown moving objects.

This paper proposes a novel moving object detection method for 3D point clouds by minimizing an energy function based on photometric and depth consistencies assuming that input data consist of synchronized point clouds, images, and camera poses measured with a moving on-vehicle system from a single sequence. The proposed method can detect arbitrary moving objects in unconstrained regions without tracking moving objects.

II. Moving object detection from a point cloud

The proposed method detects 3D points on moving objects frame-by-frame using 3D point clouds, images, and camera poses of multiple frames measured by a moving on-vehicle system equipped with laser rangefinders, cameras, and GPS/IMU. Here, a 3D point cloud, an image and a camera pose are synchronously obtained in each frame. Since the 3D point cloud is obtained by raster-scanning a target environment in whole directions around the laser rangefinders, the relationship among neighboring 3D points is known.

In this study, assuming that luminance, position and shape of static objects are fixed during a short time period, we detect the 3D points on moving objects by minimizing an energy function based on photometric and depth consistencies between multiple frames using graph cuts [12]. In addition, to improve the detection accuracy, we design a function that evaluates the likelihood of moving objects from the relationship between manually labeled objects and their photometric and depth consistencies in measured data, and use it for the energy function. In the following sections, we describe the definition of the energy function, photometric consistency, depth consistency, and the function for evaluating the likelihood of moving objects.

A. Definition of the energy function

The proposed method assigns *moving* or *static* label to each 3D point so as to minimize an energy function based on the likelihood of moving objects and the relationship among neighboring 3D points. Specifically, we define energy function E with respect to label X for target frame as follows:

$$E(X) = \sum_{v \in V} g_v(X_v) + \kappa \sum_{(u,v) \in A} h_{u,v}(X_u, X_v), \tag{1}$$

where g_v is an data term that measures the likelihood of the label X_v for a 3D point v, $h_{u,v}$ is an smoothness term between labels of two neighboring 3D points u and v, V is a set of measured 3D points in target frame, A is a set of pairs of two neighboring points (u, v), and κ is a weight to control the contribution of the second term versus the first term.

Data term g_v is defined on the basis of the likelihood of moving objects based on photometric consistency $M_{P,v}$ and the likelihood of moving objects based on depth consistency $M_{D,v}$ for 3D point v as follows:

$$g_v(X_v) = \begin{cases} (1 - M_{D,v}) + \alpha(1 - M_{P,v}) & (X_v : moving) \\ M_{D,v} + \alpha M_{P,v} & (X_v : static), \end{cases} \tag{2}$$

where the range of $M_{P,v}$ and $M_{D,v}$ is the closed interval $[0, 1]$ and α is a weight to control the contribution of $M_{P,v}$ versus $M_{D,v}$. When the likelihood of moving objects is large, label X_v tends to become *moving* so that g_v gets small, and vice versa. The likelihoods of moving objects $M_{P,v}$ and $M_{D,v}$ are described in the following section.

Smoothness term $h_{u,v}$ is defined on the basis of the difference between luminance values I_v^n and I_u^n (which are V in HSV color space) and the difference between depth values d_v^n and d_u^n at the projected positions of two neighboring 3D points u and v on target frame n as follows:

$$h_{u,v}(X_u, X_v) = \begin{cases} 0 & (X_v = X_u) \\ \frac{1}{|d_u^n - d_v^n| + |I_u^n - I_v^n| + \epsilon} & (X_v \neq X_u), \end{cases} \tag{3}$$

where ϵ is a positive constant value to make the denominator of the lower case in Eq. (3) non-zero. When different labels are assigned to two neighboring points, $h_{u,v}$ gets small if the differences of luminance values I_v^n and I_u^n and depth values d_v^n and d_u^n are large. Therefore, different labels tend to be allowed for the neighboring points in such a case.

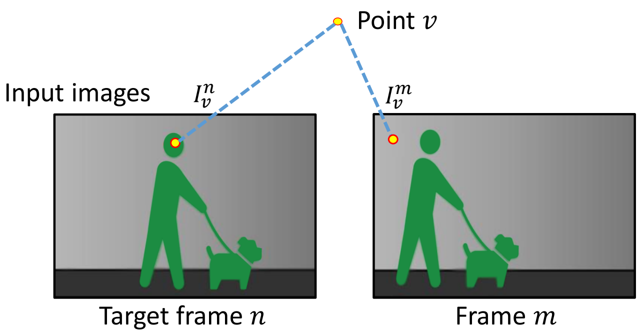


Fig. 3. Photometric consistency.

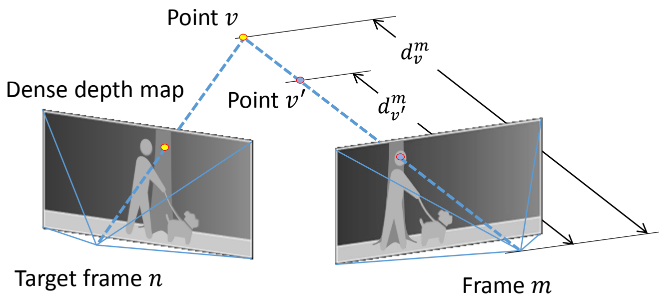


Fig. 4. The case of occlusion.

B. Photometric consistency

As illustrated in Fig. 3, in the case of projecting 3D point v on a moving object measured in target frame n onto images of different frames, luminance values of the projected positions often vary because different objects often exist on the projected positions. Therefore, the proposed method projects 3D points of target frame n onto images of multiple frames m ($n - N \leq m \leq n + N$), and evaluates the likelihood of moving objects on the basis of the differences of luminance values.

Specifically, we calculate the likelihood of moving objects based on photometric consistency $M_{P,v}$ for 3D point v as follows:

$$M_{P,v} = F_P\left(\max_m(I_{err}(m, v))\right), \tag{4}$$

where $I_{err}(m, v)$ is defined using luminance values I_v^n and I_v^m at the projected positions of 3D point v on target frame n and another frame m, respectively, as follows:

$$I_{err}(m, v) = \begin{cases} 0 & (d_v^m - d_{v'}^m > T) \\ |I_v^n - I_v^m| & (otherwise). \end{cases} \tag{5}$$

If 3D point v measured in target frame n is occluded in different frame m, a different 3D point v' is observed at the projected position of 3D point v on frame m. This means that the projected positions of v and v' on frame m are the same. In this case, point depth d_v^m at 3D point v on frame m is larger than the observed depth $d_{v'}^m$ at the projected position of 3D point v as illustrated in Fig. 4. Thus, we assume that a 3D point v is occluded in frame m if $d_v^m - d_{v'}^m > T$ is satisfied, where $T(\geq 0)$ is a threshold, and we set luminance error $I_{err}(m, v)$ as zero in this case. The function $F_P(\cdot)$ $(0 \leq F_P(\cdot) \leq 1)$ in Eq. (4) evaluates the likelihood of moving objects on the basis of the relationship between manually labeled objects and their

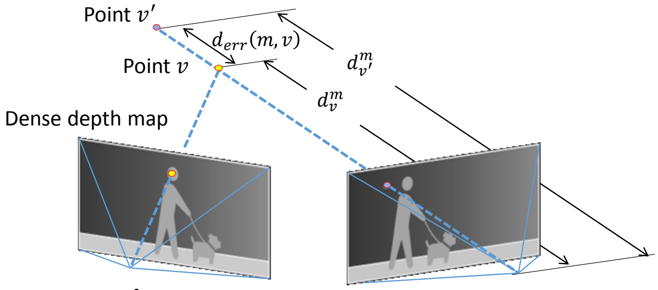


Fig. 5. Depth consistency.

photometric consistencies in measured data. We describe the details of function F_P in Section II-D.

C. Depth consistency

As with the photometric consistency in the previous section, in the case of projecting 3D point v on a moving object measured in target frame n onto depth maps of a different frame m, an observed depth $d_{v'}^{m}$ at the projected position of 3D point v on frame m differs from d_{v}^{m} because different objects exist on the projected positions as illustrated in Fig. 5. Accordingly, the proposed method projects 3D points in target frame n onto depth maps of multiple frames m, and evaluates the likelihood of moving objects on the basis of the difference of depth values. It should be noted here that we generate a dense depth map in each frame by linearly interpolating depth values using connected neighboring 3D points on the depth map as shown in Fig. 6 because depth maps generated from sparse 3D points have missing values.

Specifically, we calculate the likelihood of moving objects based on depth consistency $M_{D,v}$ as follows:

$$M_{D,v} = F_D\left(\max_{m}(d_{err}(m,v))\right), \tag{6}$$

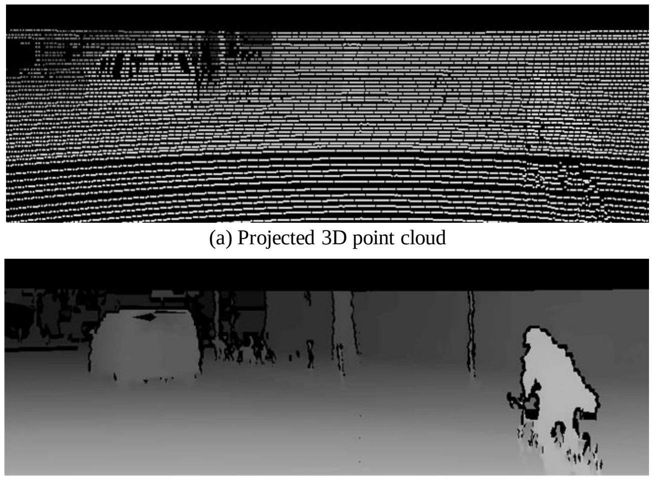
where $d_{err}(m,v)$ is defined using depth d_{v}^{m} and observed depth $d_{v'}^{m}$ at the projected position of 3D point v onto frame m as follows:

$$d_{err}(m,v) = \begin{cases} 0 & (d_{v}^{m} - d_{v'}^{m} > T) \\ |d_{v'}^{m} - d_{v}^{m}| & (otherwise). \end{cases} \tag{7}$$

As with the photometric consistency, if $|d_{v}^{m} - d_{v'}^{m}| > T$ is satisfied, *i.e.*, a occlusion occurs, we set depth error $d_{err}(m,v)$ as zero. The function $F_D(\cdot)$ $(0 \leq F_D(\cdot) \leq 1)$ in Eq. (6) evaluates the likelihood of moving objects on the basis of relationship between manually labeled objects and their depth consistencies in measured data. We describe the details of function F_D in the next section.

D. Definition of function for evaluating the likelihood of moving objects

In order to improve the detection accuracy, we design F_P and F_D that evaluate the likelihood of moving objects from the relationship between manually labeled objects and their photometric and depth errors, which are calculated by Eq. (5) and Eq. (7), respectively.



(a) Projected 3D point cloud

(b) Dense depth map

Fig. 6. Interpolation of depth map.

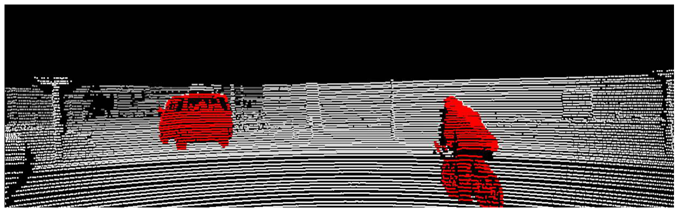


Fig. 7. Manually labeled moving objects.

Specifically, we first manually assign *moving* or *static* label to measured 3D points in order to make the ground truth as shown in Fig. 7. Next, we compute histograms of the luminance and depth error for *moving* and *static* labels, respectively. Finally, we calculate ratios of the number of *moving* labels to the entire number for each bin in the histograms, and store the ratios into lookup tables that are used as the functions of the likelihood of moving objects F_P and F_D.

III. Experiments

To verify the effectiveness of the proposed method, we detected 3D points on moving objects using a single sequence of 3D point clouds, images, and camera poses that are synchronously acquired with a moving on-vehicle system, and performed quantitative evaluations.

A. Experimental conditions

We applied the proposed method to the public data sets from KITTI [13], which are captured by a moving on-vehicle system. These data sets consist of synchronized point clouds measured by an omnidirectional laser rangefinder (Velodyne HDL-64E), images captured by a camera (Point Grey Flea 2 (FL2-14S3C-C)), and camera poses measured by a GPS/IMU (OXTS RT 3003). The specifications of the sensors are shown in Table I. First, we gave the ground truth to 10 frames in data set A shown in Fig. 8 to calculate the function for evaluating the likelihood of moving objects. Then, we applied the proposed method with the calculated function to respective 20 frames in data sets B, C, and D (shown in Figs.9 - 11) to detect

TABLE I
SPECIFICATIONS OF SENSORS.

Velodyne HDL-64E	
Data acquisition	10Hz
Acquisition points	About 130,000 points
Angular resolution	0.09°
Range of the declination angle	0°-26.8°
Measurement error of the distance	$\pm 20mm$
Range of the measurement	$\leq 120m$
Point Grey Flea 2	
Data acquisition	10Hz
Resolution(pixel)	1242×375
OXTS RT 3003	
Data acquisition	10Hz
Resolution of GPS·IMU	0.02m / 0.1°

Fig. 8. Example image in data set A.

Fig. 9. Example image in data set B.

Fig. 10. Example image in data set C.

Fig. 11. Example image in data set D.

points on the moving objects. In addition, this experiment used 11 frames ($N = 5$) for calculating photometric and depth consistencies and used $\kappa = 12$ in Eq. (1) and $\alpha = 2.5$ in Eq. (2). In the following sections, we explain the calculation of the function for evaluating the likelihood of moving objects, the detection results and quantitative evaluation.

B. Calculation of the function for evaluating the likelihood of moving objects

Figs. 12 and 13 show the histograms of the luminance and depth error and the likelihood of moving objects based

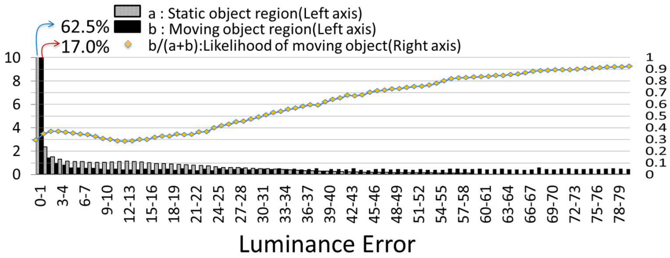


Fig. 12. Histogram of the luminance error and the likelihood of moving objects based on photometric consistency.

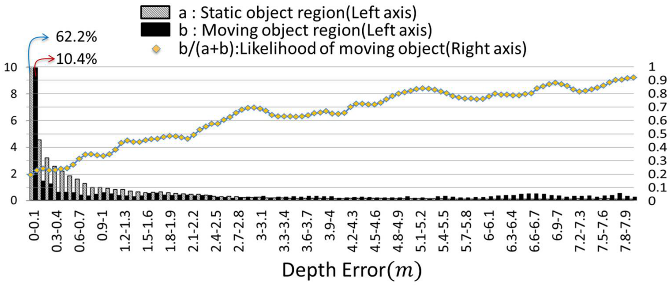


Fig. 13. Histogram of the depth error and the likelihood of moving objects based on depth consistency.

on photometric and depth consistencies for the data set A, respectively. In the histograms of the luminance error and depth error, we divide the ranges of luminance error (0 - 255) and depth error by the intervals of 1 and 0.1 m, respectively. The functions F_P and F_D for evaluating the likelihood of moving objects based on photometric and depth consistency were determined by smoothing the likelihood along the error-axis using a Gaussian filter. In addition, in the functions, we set the likelihood of moving objects as 0.9 when the luminance error is over 80, and we set the likelihood of moving objects as 0.9 when the depth error is over 8.0 m.

C. Detection results and quantitative evaluation

This section compares the results of moving object detection by the proposed method with those by the method using only photometric information for energy function E (referred to as photometric method) and the method using only depth information for energy function E (referred to as depth method) for data sets B, C and D to verify the effectiveness of the proposed method. Here, we quantitatively evaluate the results using two benchmarks TPR (True Positive Rate) and ACC (Accuracy). Specifically, given TP (True Positive) that is the number of points that exist on the moving object and are judged as points on moving objects correctly and FN (False Negative) that is the number of points that exist on moving objects but are judged as points on static objects incorrectly, TPR (which also means the detection rate of moving objects) is defined as follow:

$$\mathrm{TPR} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}. \tag{8}$$

In addition, given TN (True Negative) that is the number of points that exist on static objects and are judged as points

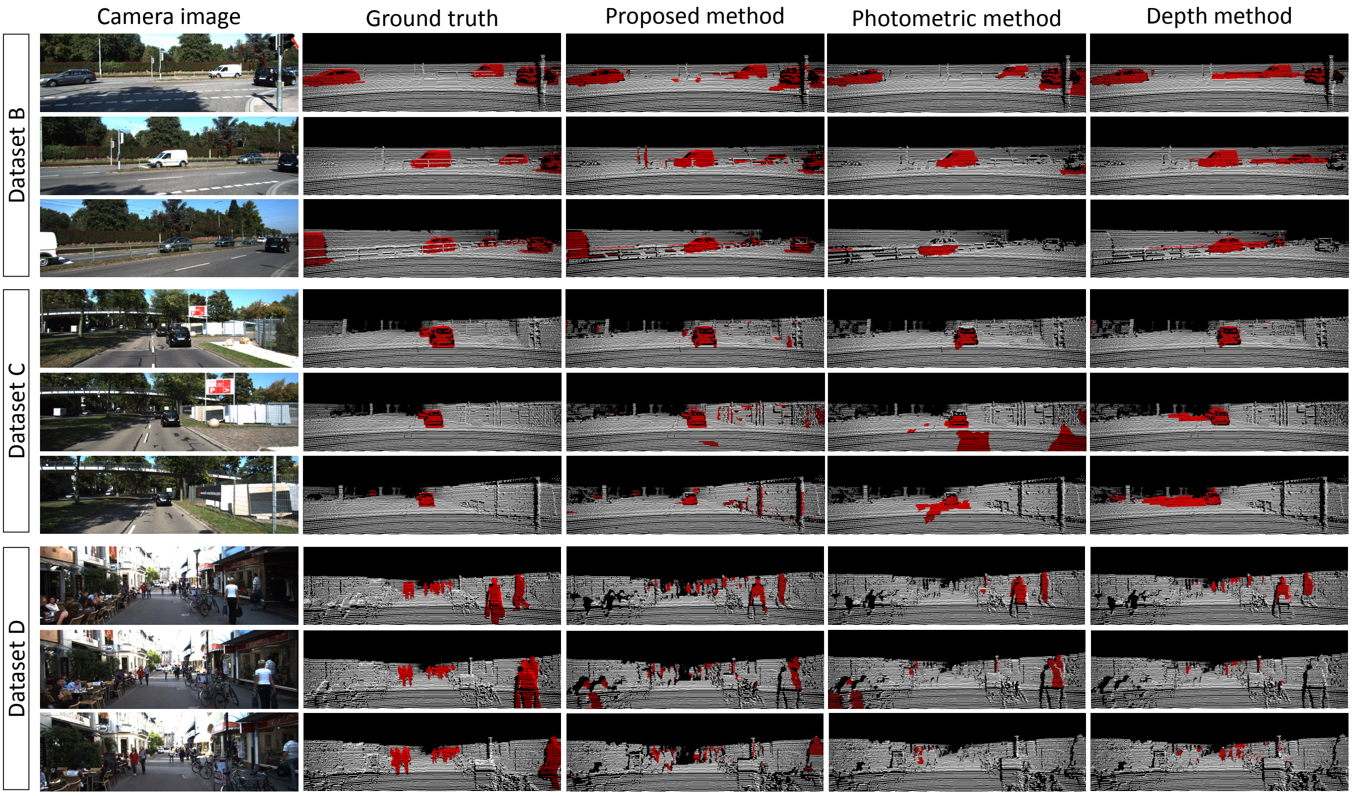


Fig. 14. Detection results for data set B, C, and D using the proposed method, photometric method, and depth method.

on static objects correctly and FP (False Positive) that is the number of points that exist on moving objects but are judged as points on static objects incorrectly, the accuracy (ACC) is defined as follow:

$$\mathrm{ACC} = \frac{\mathrm{TP}+\mathrm{TN}}{\mathrm{TP}+\mathrm{TN}+\mathrm{FP}+\mathrm{FN}}. \tag{9}$$

Fig. 14 shows ground truths, which given by manually assigning *moving* or *static* label to measured 3D points, and the detection results for data sets B, C and D obtained by the proposed, photometric and depth methods. Fig. 15 shows the quantitative evaluations for these three methods. Symbol "*" in Fig. 15 means that a significant difference was recognized using the t-test with a 5% significant level. In the experiment for data set B, as shown in Fig. 15(a), TPR of the proposed method is better than that of photometric method, and TPR and ACC of the proposed method are better than those of depth method. We also show the comparison of the results for the vehicle which moves slowly in the depth direction in data set B in Fig. 16. While it was difficult for the methods using only photometric consistency or depth consistency to detect the slowly moving objects in the depth direction because the luminance and depth error may be small, the proposed method could obtain better results by considering both the photometric and depth consistencies. In the experiment for data set C, as shown in Fig. 15(b), TPR and ACC of the proposed method are better than those of other methods. This is because FN of the proposed method is less than those of other methods and it is difficult for other methods to correctly divide the region around a boundary of the moving object in which the difference of luminance values or depth values is small. However, we confirmed false detections of the points on the ground near the camera in the result by the proposed method as shown at the 5th row and 3rd column in Fig. 14. We consider that this is because of the automatic exposure adjustment function of the camera in the sunny condition. In the experiment for data set D, as shown in Fig. 15(c), while TPR of the proposed method is better than those of other methods, TPRs of all the methods are low. This is because the camera and moving objects moves slowly as shown at the 7th-9th row and 1st column in Fig. 14. We confirmed that it is difficult to detect points on the objects that move slowly when the camera also moves slowly because the background of the moving objects is continuously occluded by the moving objects and the background is seldom observed.

IV. Conclusion

In this paper, we have proposed a moving object detection method for point clouds by minimizing an energy function based on photometric and depth consistencies assuming that input data consist of synchronized point clouds, images, and camera poses from a single sequence captured with a moving on-vehicle system. In experiments, we confirmed that the proposed method could obtain significantly better evaluation results than baseline methods considering only photometric or depth consistency for various scenes. However, we also

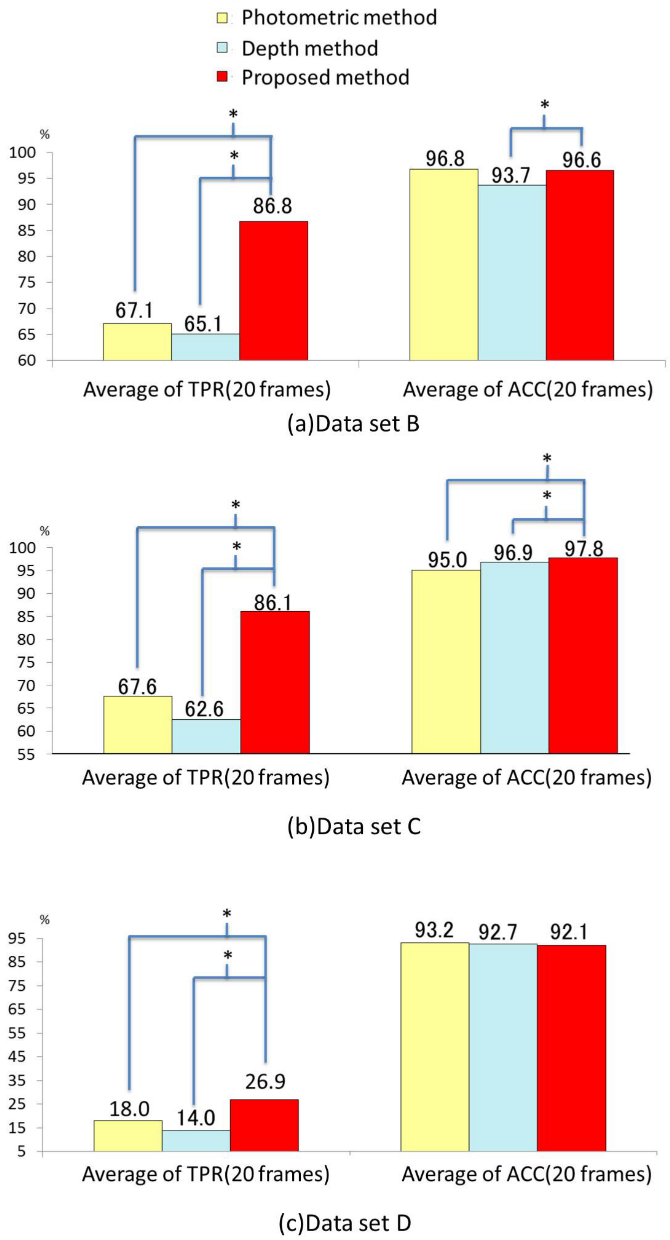


Fig. 15. Comparison of the results.

confirmed that it is difficult for the proposed method to detect moving objects in such scenes as those that include luminance changes due to automatic brightness adjustment of cameras. In future work, we will compensate for changes in the brightness by the automatic brightness adjustment of cameras to improve moving object detection.

ACKNOWLEDGMENT

This work was supported in part by JSPS KAKENHI No. 26330193.

REFERENCES

[1] Fumio Okura, Masayuki Kanbara, and Naokazu Yokoya. "Mixed-reality world exploration using image-based rendering." ACM Journal on Computing and Cultural Heritage, Vol. 8, No. 2, pp. 9:1-9:26, 2015.

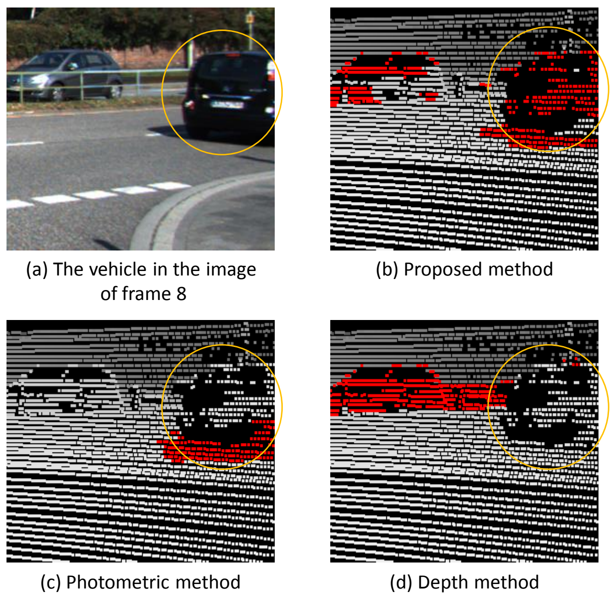


Fig. 16. Results for vehicle moving slowly to the depth direction in data set B.

[2] Akitaka Oko, Tomokazu Sato, Hideyuki Kume, Takashi Machida, and Naokazu Yokoya. "Evaluation of image processing algorithms on vehicle safety system based on free-viewpoint image rendering." In Proceedings of IEEE Intelligent Vehicles Symposium, pp. 706-711, 2014.

[3] Sabry F. El-Hakim, Claus Brenner, and Gerhard Roth, A multi-sensor approach to creating accurate virtual environments, Journal of Photogrammetry & Remote Sensing, Vol.53, pp.379-391, 1998.

[4] Toshihiro Asai, Masayuki Kanbara, and Naokazu Yokoya. "Data acquiring support system using recommendation degree map for 3D outdoor modeling." In Proceedings of Electronic Imaging, pp.64910H-1-64910H-8, 2007.

[5] Gordon Petrie. "Mobile mapping systems: An introduction to the technology." International Journal of GeoInformatics, Vol. 13, No. 1, pp. 32-43, 2010.

[6] Paul Debevec, Yizhou Yu, and George Borshukov. "Efficient view-dependent image-based rendering with projective texture-mapping." Proc. Eurographics Rendering Workshop, pp. 105-116, 1998.

[7] Tsunetake Kanatani, Hideyuki Kume, Takafumi Taketomi, Tomokazu Sato,and Naokazu Yokoya. "Removal of moving objects from point cloud data for 3D modeling of outdoor environments." IIEEJ Transactions on Image Electronics and Visual Computing, Vol. 3, No. 1, pp.54-62, 2015.

[8] Jizhou Yan, Dongdong Chen, Heesoo Myeong, Takaaki Shiratori, and Yi Ma. "Automatic extraction of moving objects from image and LIDAR sequences." In Proceedings of International Conference on 3D Vision, pp.673-680, 2014.

[9] Lili Huang and Matthew Barth. "Tightly-coupled LIDAR and computer vision integration for vehicle detection." In Proceedings of Intelligent Vehicles Symposium, pp.604-609, 2009.

[10] Cristiano Premebida, Oswaldo Ludwig, and Urbano Nunes. "LIDAR and vision-based pedestrian detection system." Journal of Field Robotics, Vol. 26, No. 9, pp. 696-711, 2009.

[11] Luciano Spinello, Rudolph Triebel, and Roland Siegwart. "Multiclass multimodal detection and tracking in urban environments". International Journal of Robotics Research, Vol. 29, No. 12, pp. 1498-1515, 2010.

[12] Yuri Boykov, Olga Veksler, and Ramin Zabih. "Fast approximate energy minimization via graph cuts." IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 23, No. 11, pp. 1222-1239, 2001.

[13] Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun. "Vision meets robotics: The kitti dataset." International Journal of Robotics Research, Vol. 32, No. 11, pp.1231-1237, 2013.