Unknown multiple object tracking using 2D lidar and video camera

K. Kwak, J.-S. Kim, J. Min and Y.-W. Park

A multisensor fusion-based unknown object detection and tracking approach is presented. The approach consists of two new algorithms: (i) object segmentation by boundary detection and (ii) data association by integer programming. The proposed approach tightly combines the information from a single-line scanning lidar and a camera. The performance of the proposed algorithm is evaluated by comparing with a state-of-the-art method and is demonstrated with real datasets obtained from a moving platform.

Introduction: Object detection and tracking in an outdoor environment are important for many computer vision applications. The performance of approaches is influenced by the available sensors. A camera provides dense appearance information, but it does not give a sense of scale. A lidar gives accurate geometric information, but only at sparse locations, and much of the appearance information cannot be captured. Properly combining individual sensor measurements has a benefit because one sensor's data can provide complementary information to the other sensor.

In this Letter, we propose a new moving object detection and tracking approach by feature-level sensor fusion of a single-line scanning lidar and a camera. To achieve the goal, we have developed two key algorithms: (i) object segmentation by a vertical boundary detection that classifies whether an object boundary exists between two consecutive range measurements and corresponding images and (ii) data association that chooses the best matching pairs between multiple segmented data. The relative pose between the lidar and camera was estimated by Kwak *et al.* [1].

Object segmentation by boundary detection: We segment objects in a scene by detecting their boundaries. The boundary detection task is formulated as a classification problem that determines whether an object boundary exists between two consecutive range measurements, as shown in Fig. 1. We use a support vector machine (SVM) classifier. The SVM inputs are features derived from the projected lidar measurements and image patches surrounding the lidar measurements. By integrating the information of the heterogeneous sensors in the classifier input, the feature-level sensor fusion is achieved seamlessly.

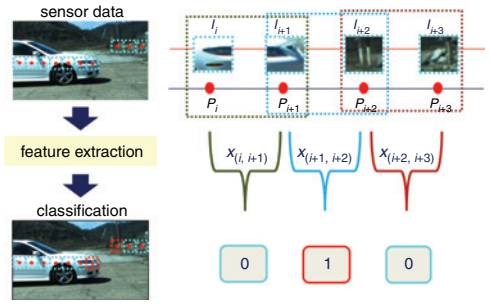

Fig. 1 *Overview of object segmentation by boundary detection (0: non-boundary and 1: boundary)*

Each lidar scanned point p_i ($i = 1, \ldots, n$) is represented as ($x_i = r_i \cos \phi_i$, $y_i = r_i \sin \phi_i$) in Cartesian coordinates from the raw measurements (r_i, ϕ_i) in polar coordinates. At an object boundary, the lidar measurements are likely to be discontinuous. It is represented as three features derived from two consecutive lidar measurements p_i and p_{i+1}: distance d_i between the points, range difference l_i of the points and surface orientation θ_i. The lidar features are defined as

$$d_i = \| p_i - p_{i+1} \|_2 \tag{1}$$

$$l_i = |r_i - r_{i+1}| \tag{2}$$

$$\theta_i = \arccos \frac{\boldsymbol{a} \cdot \boldsymbol{b}}{\| \boldsymbol{a} \| \| \boldsymbol{b} \|} \tag{3}$$

where $\boldsymbol{a}$ is the vector between p_i and p_{i+1}, and $\boldsymbol{b}$ is a vector between the mid-point of p_i and p_{i+1} and the sensor origin.

For image features, we consider two rectangular patches I_i and I_{i+1} corresponding to the projected lidar measurements p_i and p_{i+1}, respectively. Intuitively, if the image patches surrounding two lidar points belong to the same object, then the statistics of the image patches are more likely to be similar than if they belong to two different objects. It is represented as two features from two corresponding image patches: histogram intersection distance h_i of colour histograms of two patches I_i and I_{i+1}, and normalised sum of square differences e_i of intensity images [2]. The image features are defined as

$$h_i = 1 - \frac{\sum_B \min(H_i, H_{i+1})}{\min(|H_i|, |H_{i+1}|)} \tag{4}$$

where H_i is a colour histogram of I_i, $|H_i|$ is the magnitude of H_i and B is the number of bins in the histogram, and

$$e_i = \frac{1}{2} \frac{\sum_{\boldsymbol{x}} \left[(I_i(\boldsymbol{x}) - \mu_i) - (I_{i+1}(\boldsymbol{x}) - \mu_{i+1}) \right]^2}{\sigma_i \sigma_{i+1}} \tag{5}$$

where $\boldsymbol{x}$ is the pixel location of the $n \times m$ image patches and μ_i and σ_i are the mean and standard deviation of the intensity values of I_i.

Data association for multiple object tracking: We track multiple object clusters by a frame-by-frame data association. In the data association problem, the best matching pairs between the tracks and observations are estimated by solving an integer programming problem maximising the likelihood between the matching pairs. The relative likelihoods between the pairs are computed by the similarity values from the segmented lidar clusters and the corresponding image patches. In our approach, each of the segmented clusters composed of a set of lidar scan points is represented as a mixture of Gaussians considering the sensor noise [3]. To produce the probabilistic representation of each cluster, each lidar scan point is represented as a Gaussian point considering the uncertainties with respect to its scanning range and angle.

Given N tracks at time t and M observations at time $t+1$, and considering a likelihood vector $\boldsymbol{c}$ of V matching hypotheses and a constraint matrix $\boldsymbol{A}$, selecting the best matching pairs between the tracks and the observations is equivalent to finding a subset of rows of $\boldsymbol{A}$ such that the sum of corresponding elements in $\boldsymbol{c}$ is maximised, under the constraint that no two rows share common non-zero entries. This can be posed as the following integer programming problem [4]:

$$\max_{\boldsymbol{x}} \boldsymbol{c}^{\mathrm{T}} \boldsymbol{x} \quad \text{s.t.} \quad \boldsymbol{A}^{\mathrm{T}} \boldsymbol{x} \leq \boldsymbol{b} \tag{6}$$

where $\boldsymbol{c}$ is a $V \times 1$ likelihood vector, $\boldsymbol{x}$ is a $V \times 1$ binary vector which is 1 if the vth row is in the solution, or 0 otherwise, $\boldsymbol{A}$ is a $V \times (N+M)$ matrix and $\boldsymbol{b}$ is a $V \times 1$ identity vector.

The likelihood vector $\boldsymbol{c}$ and the constraint matrix $\boldsymbol{A}$ are determined by the similarity measures S_{L} and S_{I} from both sensor data. Considering vth hypothesis by nth track and mth observation, its likelihood c_v is estimated as

$$c_v = S_L(Z_n, Z_m) S_I(I_n, I_m) \tag{7}$$

where I is the image patch corresponding to a segmented lidar cluster Z. The constraint matrix $\boldsymbol{A}$ whose row index corresponds to the vth hypothesis is built as

$$\boldsymbol{A}(v, i) = \begin{cases} 1, & \text{if } i = n \text{ or } i = N + m \\ 0, & \text{otherwise} \end{cases} \tag{8}$$

The similarity $S_{\mathrm{L}}(Z_n, Z_m)$ of the two segmented lidar clusters Z_n and Z_m is estimated with the two probability distributions. Let the probability distribution of Z_n be $\boldsymbol{P}_n \sim \mathcal{N}(\mu_n, \Sigma_n)$ and that of Z_m be $\boldsymbol{P}_m \sim \mathcal{N}(\mu_m, \Sigma_m)$. The similarity between $\boldsymbol{P}_n$ and $\boldsymbol{P}_m$ is measured by the Bhattacharyya distance (D_{Bhatt}). Once we compute the similarity, the likelihood S_{L} for matching Z_m to Z_n is estimated by the following exponential function:

$$S_L(Z_n, Z_m) = \exp(-D_{\mathrm{Bhatt}}) \tag{9}$$

The image similarity is measured with sub-image regions I_n and I_m in the object region of interests (ROIs) which are cropped from the vertical boundary detection. Since a single-line scanning lidar measures depth of a stripe line of an object, we define the sub-image region to be a rectangular region. The width of the rectangular region is the same as the object ROIs, whereas its height is measured using a predefined size onto the object ROIs centred on the projection of lidar measurements.

Considering the normalised colour histograms H_n and H_m extracted from I_n and I_m, respectively, the image similarity S_I is estimated by the Bhattacharyya measure between distributions of H_n and H_m. However, the Bhattacharyya measure is a divergence-type measure between distributions and does not impose a metric structure. We use the modification of the Bhattacharyya measure proposed by Comaniciu *et al.* [5]. This measure represents a metric distance between two distributions as

$$S_I(I_n, I_m) = \sqrt{1 - \sum_{k=1}^{K} \sqrt{(H_n(k), H_m(k))}} \quad (10)$$

where K is the number of bins.

Experimental results: Fig. 2 shows the experimental results: (Fig. 2*a* is the) effect of sensor modality and (Fig. 2*b* is a) comparison with the adaptive breakpoint detector (ABD) algorithm described by Borges and Aldon [6]. As can be seen from Fig. 2*a*, the lidar features are more discriminative than the image-based features. This agrees with the intuition that the depth discontinuity from two-dimensional (2D) laser range measurements is a more reliable indicator of object boundaries than the difference in image intensity or texture. However, the combination of lidar and image-based features performs better than features from either modality individually. This result suggests that the information in the lidar and image features is not entirely redundant. Fig. 2*b* shows that our algorithm outperforms the ABD algorithm for all threshold values. We also manually extracted two subsets of the full data that consisted entirely of easy (sample 1: cars, walls and buildings) and challenging cases (sample 2: bushes, occlusion and transparent surfaces). The performance of both algorithms on these challenging datasets is worse than on the full dataset, but our algorithm still outperforms the ABD method.

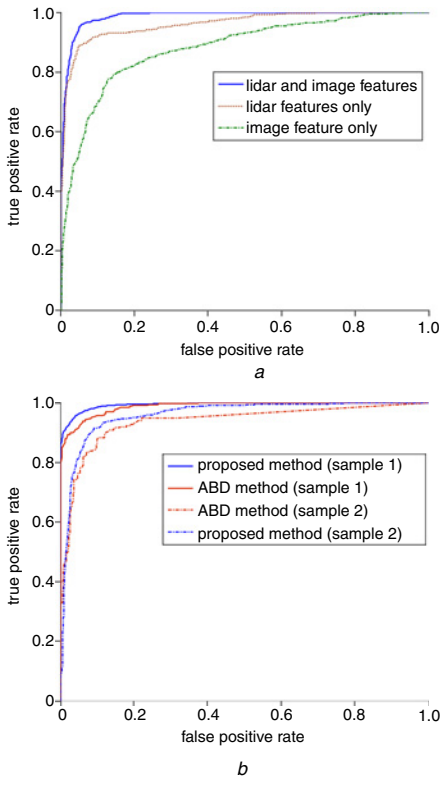

Fig. 2 *Performance evaluations of object boundary detection method*
a Threshold averaging ROC curves
b Comparison with ABD method

Fig. 3 shows the tracking result by the proposed data association algorithm on a moving platform. To do this experiment, we used sensor data from a lidar and a camera mounted on the right side of the platform. Fig. 3*a* describes the initial measurements from both sensors that consist of three moving objects (red circles) and one stationary object (a blue circle). Each object image is cropped from a region estimated by the boundary detection algorithm. As shown in Figs. 3*b* and *c*, the proposed data association algorithm tracked clustered objects correctly. When we registered the lidar measurements based on the data association result, the moving objects were filtered out and individually tracked for several frames.

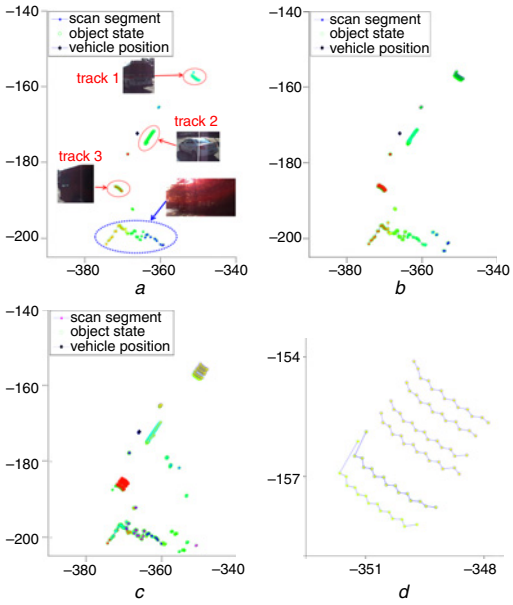

Fig. 3 *Data association results in moving platform (best viewed in colour)*
a Initial sensor data from lidar and camera
b Data association result between t and $t+1$ frames (blue: track 1, cyan: track 2 and red: track 3)
c Registered lidar measurements for six frames
d Zoomed-in result of track 1 (blue)

Conclusion: We have presented a new unknown moving object detection and tracking method that consists of the object boundary detection by classification and the data association by integer programming. The approach is achieved by tightly combining the two sets of sensor information from a single-line scanning lidar and a video camera. Through the experiments, we have proved the benefit of sensor fusion for the object boundary detection and data association.

2 February 2014
doi: 10.1049/el.2014.0355
One or more of the Figures in this Letter are available in colour online.

K. Kwak, J. Min and Y.-W. Park (*Agency for Defense Development, Bukyuseongdaero 488gil, Yuseong, Daejeon 305-152, Republic of Korea*)

J.-S. Kim (*Korea Institute of Science and Technology, Hwarangno 14gil 5, Seongbuk, Seoul 136-791, Republic of Korea*)

E-mail: junsik.kim@kist.re.kr

References

1 Kwak, K., Huber, D., Badino, H., and Kanade, T.: 'Extrinsic calibration of a single line scanning lidar and a camera'. IEEE Conf. Intelligent Robots and Systems, San Francisco, CA, USA, September 2011
2 Criminisi, A., Blake, A., and Rother, C.: 'Efficient dense stereo with occlusions for new view-synthesis by four-state dynamic programming', *Int. J. Comput. Vis.*, 2007, **71**, pp. 89–110
3 Katz, R., Nieto, J., and Nebot, E.: 'Probabilistic scheme for laser based motion detection'. IEEE Conf. Intelligent Robots and Systems, Nice, France, September 2008
4 Li, K., Chen, M., and Kanade, T.: 'Cell population tracking and lineage construction with spatiotemporal context'. Int. Conf. Medical Image Computing and Computer-Assisted Intervention, Brisbane, Australia, October 2007, pp. 295–302
5 Comaniciu, D., Ramesh, V., and Meer, P.: 'Kernel-based object tracking', *IEEE Trans. Pattern Anal. Mach. Intell.*, 2003, **25**, pp. 564–575
6 Borges, G.A., and Aldon, M.-J.: 'Line extraction in 2D range images for mobile robotics', *J. Intell. Robot. Syst.*, 2004, **40**, pp. 267–297