A novel outdoor scene-understanding framework for unmanned ground vehicles with 3D laser scanners
Yan Zhuang, Guojian He, Huosheng Hu and Zhenwei Wu
Transactions of the Institute of Measurement and Control published online 14 July 2014
DOI: 10.1177/0142331214541140
The online version of this article can be found at:
http://tim.sagepub.com/content/early/2014/07/10/0142331214541140
Published by:
SAGE
http://www.sagepublications.com
On behalf of:
The Institute of Measurement and Control
Additional services and information for Transactions of the Institute of Measurement and Control can be found at:
Email Alerts: http://tim.sagepub.com/cgi/alerts
Subscriptions: http://tim.sagepub.com/subscriptions
Reprints: http://www.sagepub.com/journalsReprints.nav
Permissions: http://www.sagepub.com/journalsPermissions.nav
Citations: http://tim.sagepub.com/content/early/2014/07/10/0142331214541140.refs.html
>> OnlineFirst Version of Record - Jul 14, 2014
What is This?
A novel outdoor scene-understanding framework for unmanned ground vehicles with 3D laser scanners

Yan Zhuang¹,², Guojian He¹, Huosheng Hu³ and Zhenwei Wu²

Abstract
Outdoor scene understanding plays a key role for unmanned ground vehicles (UGVs) to navigate in complex urban environments. This paper presents a novel 3D scene-understanding framework for UGVs to handle uncertain and changing lighting conditions outdoors. A 2D bearing angle (BA) image is deployed to perform scene understanding so that the computational burden in the process of segmentation and classification of the 3D laser point cloud can be reduced. An improved super-pixel algorithm is used for fast 3D scene segmentation, and then the Gentle–Adaboost algorithm is utilized to perform super-pixel patch classification using the texture features of the Gray Level Co-occurrence Matrix. All false classification results in the uncertain super-pixel patches of BA images are transformed back to raw 3D laser points and a re-classification is conducted to refine the 3D scene understanding for UGVs. The results from a real laser dataset taken from a large-scale campus environment show the validity and robust performance of the proposed approach, in comparison with the results from Korea Advanced Institute of Science and Technology dataset.

Keywords
3D point clouds, laser scanner, unmanned ground vehicles, urban scene understanding

Introduction
Active environment perception and autonomous scene understanding are crucial tasks for mobile robots in navigating complex outdoor environments (Ponsa et al., 2011; Tan et al., 2013). A variety of vision systems and associated scene-understanding algorithms have been developed for such tasks (Floros and Leibe, 2012; Li et al., 2009). Recently, laser range finders have been gradually deployed to replace vision systems since they can provide a wider field of view and are not susceptible to different lighting conditions.

A new research branch in robotics society has emerged on how to effectively perform online outdoor scene understanding based on 3D laser data points (Behley et al., 2012; Choe et al., 2013). The traditional 3D laser-based scene-understanding system for an unmanned ground vehicle (UGV) directly uses the raw 3D laser point clouds. To perform discrete multi-label classification, Munoz et al. (2009a) adopted a functional gradient approach for learning the high-dimensional parameters of random fields. Although contextual reasoning techniques (e.g. Markov random fields) can show superior performance against local classifiers, their computational burden is often unacceptable for the real-time navigation task of UGVs.

An efficient learning of a random field with higher-order cliques was utilized by Munoz et al. (2009b) in a UGV platform to conduct 3D point cloud classification at 1/3 Hz for a map of 25 × 50 m. In order to reach better performance, e.g. 10 Hz, a two-class classification approach was proposed by Himmelsbach et al. (2009) using a 2.5D occupancy grid and point feature histograms. Several histogram-based descriptors were evaluated by Behley et al. (2012) for the classification of 3D laser scanning data on three datasets and they showed that the 3D histograms were a good choice for the classification in urban environments.

In this paper, a novel 3D scene-understanding framework is developed to label a 3D point cloud in a complex urban environment. A custom-built 3D laser scanning system, which is located in the middle front row of the UGV platform shown in Figure 1, is used to obtain the point clouds. In contrast, Figure 2 shows a closer look at this custom-built 3D laser scanner, which is realized by rotating a 2D laser range finder (SICK LMS 291 with a 180°/0.5° resolution) on a rotating platform. By classifying and labelling the 3D points of ground, vegetation, buildings, cars and trunks (trunks refer to the main stems of the trees, similarly hereinafter), the proposed approach enables a UGV to understand outdoor environments and performs autonomous navigation.

¹School of Control Science and Engineering, Dalian University of Technology, Dalian, China
²State Key Laboratory of Robotics, Shenyang Institute of Automation, Chinese Academy of Sciences, China
³School of Computer Science and Electronic Engineering, University of Essex, Colchester, UK

Corresponding author:
Yan Zhuang, School of Control Science and Engineering, Dalian University of Technology, Dalian, China.
Email: zhuang@dlut.edu.cn
Our work is focused on improving the performance of online outdoor scene understanding with the following features:

1) Instead of using the raw 3D laser point clouds, we use 2D bearing angle (BA) images to implement fast 3D scene segmentation and classification. Based on an improved super-pixel algorithm, the texture information in the 2D BA images can be effectively used. After image segmentation, the Gray Level Co-occurrence Matrix (GLCM) texture features are extracted from super-pixel patches in BA images, and then the Gentle-Adaboost algorithm is utilized to perform super-pixel patch classification with GLCM features. As there are exact correspondences between each laser point and each pixel in a BA image, the super-pixel patches classification results can be transformed and labelled in the corresponding 3D point clouds directly.

2) The BA-image-based segmentation and classification approach can label the 3D points with Ground (with elevation information), Vegetation (e.g. trees and shrubs) and the upper parts of Building (with semantic information) accurately. However, cars, trunk and lower parts of buildings are labelled Uncertain Objects. A re-classification approach is needed. All these uncertain super-pixel patches are transformed to raw 3D laser points, and an Agglomerative Hierarchical Clustering (AHC) algorithm is adopted to implement 3D laser data segmentation. Statistical and geometrical shape features are then extracted from these laser points so that Car, Trunk, Building and others can be labelled accurately.

The rest of this paper is organized as follows. In the next section, we describe the fast 3D point classification using 2D BA images, where accurate patch labels are generated together with some uncertain patches. Then we present the re-classification step for the 3D points of uncertain patches, where the refined scene-understanding results are obtained. The experimental data and the ultimate scene-understanding results are presented to show the feasibility and performance of the proposal approach. Finally, a brief concluding remark and future work are given.

3D point classification with 2D BA image

Transformation from 3D laser data to 2D BA image

The BA model was first proposed by Scaramuzza et al. (2007) to perform extrinsic calibration between a camera and a 3D laser scanner. In our early work, BA images were adopted to perform 3D-laser-based indoor place recognition and showed superior performance in representing details of the indoor scenes over traditional range images (Zhuang et al., 2013). In contrast, the large-scale urban scenes include various objects such as trees, shrubs, cars and buildings, and are more complicated than indoor scenes. However, BA image remains a better choice for urban scene understanding than range images, as the edges and contours of objects can be detected more effectively in BA images, as shown in Figure 3.

The algorithm for generating BA images can be found in our early work (Zhuang et al., 2013). There is an exact correspondence between laser scanning data and each pixel in a BA image, therefore the features extracted from both 2D images and 3D raw laser points will be used for scene segmentation and classification.

Super-pixel segmentation in 2D BA image

A BA image can be segmented into super-pixel patches using the Simple Linear Iterative Clustering (SLIC) algorithm proposed by Achanta et al. (2012), which is simpler, faster and more memory efficient than other super-pixel methods. SLIC adapts a k-means clustering to generate super-pixels, where a distance measure $D$ is applied to describe the distance between a pixel and a cluster centre. The distance measure $D$ consists of grey-scale and 2D position information so that SLIC can adhere to object boundaries in line with the grey-scale changing.
However, SLIC performs poorly in boundary adherence for BA images, as the grey-scale change for different objects in BA image is not as remarkable as a common vision image. As shown in Figure 4(a), SLIC cannot perform accurate super-pixel segmentation in some regions label by the blue rectangles, and some pixels of the car and the distant building are clustered into the same super-pixel patches despite the long distance between them in the real world. To solve this problem, 3D metric information is used in the standard SLIC algorithm to generate an improved one, namely 3DSLIC in our work, which can modify the distance measure $D$ in SLIC and be calculated as follows:

\[
D' = \begin{cases} 
D & (d_{3D} < d_{th}) \\
+\infty & (d_{3D} \geq d_{th}) 
\end{cases}
\]

where $D'$ is the new distance measure, $d_{3D}$ is the 3D distance between the pixel and the cluster centre, and $d_{th}$ is the threshold.

Equation (1) indicates that a pixel will not belong to a cluster when the 3D distance between the pixel and the cluster centre is larger than the threshold $d_{th}$ that represents the average minimum 3D distance between different objects in the real world. In our work, $d_{th} = 3$ m and this proves to be the best choice by experimental results. As shown in Figure 4(b), the segmentation results in the blue rectangles are more accurate than the corresponding ones in Figure 4(a). It is clear that by adding 3D metric information to the distance measure, the 3DSLIC algorithm can perform more accurate super-pixel patches.

**GLCM feature extraction and classification**

After super-pixel segmentation in BA images, each super-pixel patch is considered a basic unit in our classification algorithm. In most situations, every pixel in a patch will belong to one part of an object in the scene, and these super-pixel patches will be classified into Ground, Vegetation, Building and Uncertain Object with a Boosting classifier. For the classification procedure, we select the GLCM texture feature presented by Haralick et al. (1973) as the feature descriptor of a super-pixel patch, which has been proven effective for image texture analysis.

GLCM is an estimate of the second-order joint probability, $P_{\delta}(i,j)$ of the intensity values of two pixels ($i$ and $j$), a distance $\delta$ apart along a given direction $\theta$, i.e. the probability that $i$ and $j$ have the same intensity. This joint probability takes the form of a square array $P_{\delta}$, with row and column...
dimensions equal to the number of discrete grey levels (intensities) in the image zone (Bharati et al., 2004). With measurement at multiple distances and in multiple directions, GLCM descriptors can capture complex textural differences in an image zone.

In this paper, we set the distance \( d = 1 \). The GLCM for the selected image super-pixel is calculated in directions \( (\theta = 0^\circ, 45^\circ, 90^\circ, 135^\circ) \) at 16 discrete grey levels, resulting in four \( 16 \times 16 \) GLCM matrices for each super-pixel. Finally, four GLCM descriptors (contrast, energy, homogeneity and correlation) are calculated for each of the four GLCM matrices, resulting in 16 descriptors for each super-pixel (Bharati et al., 2004). All these parameters are chosen by practical testing and are proved a good choice for representing texture features in our BA image.

In practical situations, to distinguish the Ground from other classes, 3D elevation information is added into GLCM texture features, resulting in a feature vector of 17 dimensions for each super-pixel. There are different boosting variants for multi-class generalizations such as AdaBoost.M1, AdaBoost.M2, AdaBoost.MH and AdaBoost.MR (Schapire, 2003), or by using weak classifiers with multiple classification ability (CARTs, C4.5, etc.), two-class classification methods (Real-AdaBoost, Gentle-AdaBoost, etc.) can be directly extended for solving multi-classification. In our work, Gentle-AdaBoost with CARTs as basic classifiers is adopted to deal with the multi-classification problem due to its excellent noise immunity (Lienhart et al., 2003). Finally, fast 3D point classification is implemented and some results are shown in Figure 5. The super-pixel patches of BA images and the corresponding 3D points are classified into four classes: Vegetation, Ground, Building and Uncertain Object.

**Refined scene understanding**

This section introduces the re-classification procedure that transforms uncertain super-pixel patches to raw 3D laser points, and then classifies them into cars, trunk, lower parts of buildings and others. The procedure mainly covers two steps: re-segmentation and re-classification.

**Re-segmentation**

In order to obtain refined scene-understanding results, the 3D points of uncertain patches are firstly re-segmented into

---

**Figure 4.** (a) Simple Linear Iterative Clustering (SLIC)-based super-pixel segmentation results in a bearing angle (BA) image; (b) improved 3DSLIC-based super-pixel segmentation results in the same BA image.
precise objects using the AHC algorithm proposed by Johnson (1967). AHC iteratively agglomerates two nearest clusters into a new one by using the distance measure $d_{AHC}$ until only one cluster is remained or a determined cluster number $c$ is reached, which results in a hierarchical structure along the clustering process. Murtagh and Contreras (2012)

Figure 5. The fast 3D point classification results for two urban scenes. Colour code: Vegetation (green), Ground (grey), Building (orange), Uncertain Object (red), background (black): (a) bearing angle (BA) image and the corresponding 3DSLIC-based super-pixel segmentation results; (b) the classification result for super-pixel patches; (c) the classification result in the corresponding 3D points.
summarized seven hierarchical clustering algorithms distinguished by the different distance measure \( d_{AHC} \), among which Centroid is selected in our AHC algorithm for its simplicity in computation. In real-world situations, Centroid uses \( d_{AHC} = \| \mathbf{g}_i - \mathbf{g}_j \| \) to describe the dissimilarity between clusters \( G_i \) and \( G_j \), where \( \mathbf{g}_i \) is a vector that represents either an initial point or a cluster centre.

The traditional AHC segmentation algorithm has two problems: the high computational complexity and the difficulty in determining the final cluster number \( c \). Here, 3D voxels with suitable size are adopted as the initial clusters instead of raw 3D points to improve computational efficiency. For the second problem, the termination condition of clustering iterations is modified as \( d_{AHC-min} \leq d_{AHC-th} \) instead of reaching the determined cluster number \( c \). Where \( d_{AHC-min} \) is the distance for two nearest clusters of current iteration and \( d_{AHC-th} \) is the threshold. The final AHC algorithm applied in our re-segmentation step is described in Algorithm 1:

**Algorithm 1: AHC segmentation method**

**Inputs:** 3D voxels \( V_i, i = 1, \ldots, n \)

**Output:** AHC clusters \( \{ G_i \}, i = 1, \ldots, c \)

**Initialize** \( d_{AHC-min} = 0, G = V_i, i = 1, \ldots, n, c = n \)

for \( c > 1 \) and \( d_{AHC-min} < d_{AHC-th} \) do

Compute the smallest distance \( d_{AHC-min} \) and the corresponding cluster \( G_i \) and \( G_j \), agglomerate \( G_i \) and \( G_j \);

\( c = c - 1 \);

end for

return cluster \( \{ G_i \}, i = 1, \ldots, c \)

In our work, the size of 3D spatial voxels is set to be \( 0.4 \times 0.4 \times 0.4 \) m for re-segmentation. The threshold \( d_{AHC-th} \) for the termination of clustering iterations is a key parameter in the clustering process. By using Centroid \( (d_{AHC} = \| \mathbf{g}_i - \mathbf{g}_j \| ) \), \( d_{AHC-th} \) indicates the average minimum distance of the object centres in the real world, which is determined as 3.0 m (corresponding to the value \( d_{th} \) above). Some
Re-segmentation results are shown in Figure 6 (a) by rendering different clusters with different colours.

Re-classification

Aijazi et al. (2013) summarized two kinds of classification methods for 3D point classification: discriminate models such as conditional random field (CRF) and feature-based algorithms such as point feature histograms. As there are significant differences in the size and shape for the 3D objects in our re-classification tasks, we still use feature-based local classifier to discriminate buildings, cars, trunks and others. In this step, the geometrical shape features are extracted from 3D points of each segmented object.

The detailed features can be defined as:

- **Bounding box**: It describes the size of an object, which is large for buildings, medium for cars and small for trunk or others. The bounding box is represented with \( l_b \) and \( h \), where \( l_b \) is the length for the longest diagonal of the box and \( h \) is the height of the box (Figure 7a).

- **Point distribution along height**: The points of an object are evenly divided into three parts along height. The points are sparse in the top and bottom but dense in the middle part for cars, even for trunks, and erratic for buildings and others. If the point number in each part is \( n_1 \), \( n_2 \) and \( n_3 \) from bottom to top, the distribution features are \( d_1 = \frac{n_1}{n_2} \) and \( d_2 = \frac{n_2}{n_3} \) (Figure 7b).

With the help of geometrical shape features introduced above, 3D points of the re-segmented objects are classified into Building, Car, Trunk and others (Figure 6b). The final refined scene-understanding results implemented by both fast 3D point classification approach and re-classification approach are shown in Figure 6(c).

Experimental results

**3D laser data acquisition with UGV platform**

The 3D laser dataset used in our experiment is acquired with the custom-built 3D laser scanner shown in Figure 2, which is mounted on our Smart-Cruiser UGV platform (Figure 1). In total, 60 urban scenes in our Dalian University of Technology (DUT) campus are scanned in a stop-and-go way along the selected path as shown in Figure 8. Each scene has nearly 80,000 points, which are hand-labelled Vegetation, Ground, Building, Trunk, Car and Others. All the labelled scenes are divided into a training part (60%) and a testing part (40%) for our scene-understanding experiment. The details of the DUT campus dataset are presented in Table 1.
3D scene-understanding results

The classifiers in the fast classification step and the re-classification step are both trained with the training data set and Gentle-Adaboost algorithm. The trained classifiers are then tested on the testing dataset for 3D scene understanding.

Figure 9 shows the experimental results of eight 3D scene understanding, which are randomly selected from our experiments. The details of the classification performances are given in Table 2, where rows are the ground truth labels and columns are the predicted labels. The overall classification rate is about 0.845 and the scene-understanding framework performs well overall with better or about equal precision and recall rates for the five classes.

Timing

The runtime analysis is carried out on a PC with a Core2 Duo 2.5-GHz CPU, and the time cost result of 12 scenes randomly selected from the testing dataset is given in Figure 10. Every step in our scene-understanding approach is tested for the time cost. We need no more than 50 ms to calculate the BA image. In the process of super-pixel segmentation, 300–380 ms is used to perform our 3DSLIC super-pixel segmentation algorithm. The GLCM feature extraction and classification with AdaBoost took 310–360 ms, which is one of the largest time consumption steps in our 3D scene understanding. Finally, the re-classification step takes about 60–100 ms to complete the refined scene understanding. The whole framework takes 730–850 ms and an average of 798 ms to complete the understanding work of an urban scene. The low time consumption ensures that the proposed framework can be implemented online.

Discussion

In order to demonstrate the proposed method working well in other laser datasets, we use three 3D point cloud data sets (KAIST 1, KAIST 2, KAIST 3) provided by Professor Myung Jin Chung at the Korea Advanced Institute of Science and Technology (KAIST) to test the effectiveness of our method. As introduced in Choe et al. (2013), the KAIST dataset was acquired by two vertical laser range finders in the KAIST campus within a 690 × 680-m area. There are in total about 2.88 million points in the KAIST 1–3 datasets.

In our testing experiments, we divided three large-scale outdoor scenes into a series of local 3D scenes, and each local scene is composed of 250 groups of 2D laser scanning data (250 × 361 = 90 250 points in each local scene). Therefore, there are 22, 12 and 23 local scenes in KAIST 1, KAIST 2 and KAIST 3, respectively. We conducted a series of 3D scene-understanding experiments using the new KAIST datasets and three of them are randomly selected and shown in Figure 11. Moreover, Figure 12 shows the entire 3D scene reconstruction and understanding results using dataset KAIST 1 and KAIST 2, while Figure 13 shows 3D scene reconstruction and understanding result using dataset KAIST 3. It should be noted that all of these laser data in KAIST 1–3 are never used as training data to train the classifier. They are only used as testing data in these new experiments.

Table 1. Details of the Dalian University of Technology campus dataset.

<table>
<thead>
<tr>
<th>Number of scenes</th>
<th>Number of points</th>
<th>Vegetation (%)</th>
<th>Ground (%)</th>
<th>Building (%)</th>
<th>Car (%)</th>
<th>Trunk (%)</th>
<th>Others (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Training set</td>
<td>36</td>
<td>3 024 000</td>
<td>34.2</td>
<td>24.8</td>
<td>27.8</td>
<td>1.5</td>
<td>3.2</td>
</tr>
<tr>
<td>Testing set</td>
<td>24</td>
<td>2 016 000</td>
<td>46.8</td>
<td>23.3</td>
<td>13.4</td>
<td>2.7</td>
<td>7.7</td>
</tr>
<tr>
<td>Total</td>
<td>60</td>
<td>5 040 000</td>
<td>39.3</td>
<td>24.2</td>
<td>22.0</td>
<td>2.0</td>
<td>3.0</td>
</tr>
</tbody>
</table>

Table 2. Scene-understanding results.

<table>
<thead>
<tr>
<th>Vegetation</th>
<th>Ground</th>
<th>Building</th>
<th>Car</th>
<th>Trunk</th>
<th>Others</th>
<th>Recall</th>
</tr>
</thead>
<tbody>
<tr>
<td>798 091</td>
<td>25 647</td>
<td>55 543</td>
<td>1204</td>
<td>3919</td>
<td>59 081</td>
<td>0.846</td>
</tr>
<tr>
<td>1138</td>
<td>265 940</td>
<td>1499</td>
<td>60</td>
<td>548</td>
<td>955</td>
<td>0.984</td>
</tr>
<tr>
<td>39 127</td>
<td>15 714</td>
<td>401 091</td>
<td>2614</td>
<td>6501</td>
<td>4677</td>
<td>0.854</td>
</tr>
<tr>
<td>2044</td>
<td>9202</td>
<td>9896</td>
<td>30 500</td>
<td>1542</td>
<td>1246</td>
<td>0.560</td>
</tr>
<tr>
<td>16 818</td>
<td>6360</td>
<td>10 880</td>
<td>216</td>
<td>19 233</td>
<td>922</td>
<td>0.353</td>
</tr>
<tr>
<td>21 313</td>
<td>34 507</td>
<td>12 343</td>
<td>902</td>
<td>626</td>
<td>154 082</td>
<td>0.689</td>
</tr>
<tr>
<td>0.908</td>
<td>0.744</td>
<td>0.816</td>
<td>0.859</td>
<td>0.594</td>
<td>0.697</td>
<td></td>
</tr>
</tbody>
</table>
The computational cost for our method is lower than the classification method used in Choe et al. (2013). In our work, the total time costs (Intel Core2 Duo 2.5-GHz processor, 2GB RAM, Windows 7, C++) for KAIST 1, KAIST 2 and KAIST 3 are 19.5, 10.1 and 18.6 s, respectively. However, in the work of Choe et al. (2013), the time costs (Intel Core2 Duo 3.16-GHz processor, 4GB RAM, Windows 7, MATLAB) for parts of dataset KAIST 1, KAIST 2 and KAIST 3 are 52.9, 56.8 and 100.5 s, respectively. Although the method proposed by Choe et al. (2013) was implemented

Figure 9. Eight 3D scene-understanding results in our experiments are selected and given randomly. Colour code: Vegetation (green), Building (orange), Ground (grey), Car (yellow), Trunk (pink), Others (blue).
Figure 10. Time cost of our scene-understanding method (12 scenes randomly selected from the testing dataset).

Figure 11. Three 3D scene-understanding results are randomly selected using the new KAIST dataset. Colour code: Vegetation (green), Building (orange), Ground (grey), Car (yellow).

Figure 12. The entire 3D scene reconstruction and understanding results using dataset KAIST 1 and KAIST 2.

Figure 13. The entire 3D scene reconstruction and understanding results using dataset KAIST 3.
in MATLAB and may be improved via implementation in C++, our method can still run with lower computational cost since a 2D BA image is deployed to perform fast segmentation and classification in 3D point clouds.

**Conclusion and future work**

This paper addresses how to accomplish urban scene understanding using 3D laser scanning data. A 2D BA image is deployed to perform scene understanding. Two novel approaches are proposed for 3D laser-point segmentation and classification. Firstly, an improved super-pixel algorithm and a Gentle–Adaboost algorithm are adopted to perform fast outdoor scene segmentation and super-pixels patches classification of BA images, respectively. Secondly, a re-classification approach is deployed to refine 3D scene understanding for UGVs in order to eliminate some false classification results that are labelled as the uncertain objects in BA images. Experimental results are given to show the validity and robustness of the proposed framework.

In the future work, we plan to improve the performance of on-the-fly urban scene understanding with the 3D point clouds captured by multiple 2D lasers at careful time synchronization. Furthermore, some real-time classification algorithms will be utilized to obtain good accuracy and robustness in real-world applications.

**Acknowledgements**

The authors would also like to thank Mr Yungeun Choe and Professor Myung Jin Chung at Korea Advanced Institute of Science and Technology for providing the 3D point cloud dataset KAIST.

**Conflict of interest**

The authors declare that there is no conflict of interest.

**Funding**

This work was supported in part by the National Natural Science Foundation of China (grant numbers 61375088 and 61035005) and the State Key Laboratory of Robotics (grant number 2013-O07).

**References**


