Unsupervised 3D Category Discovery and Point Labeling from a Large Urban Environment

Quanshi Zhang[1], Xuan Song[1], Xiaowei Shao[1], Huijing Zhao[2] and Ryosuke Shibasaki[1]

***Abstract*— The building of an object-level knowledge base is the foundation of a new methodology for many perception tasks in artificial intelligence, and is an area that has received increasing attention in recent years. In this paper, we propose, for the first time, to mine category shape patterns directly from a large urban environment, thus constructing a category structure base. Conventionally, category patterns are learned from a large collection of object samples, but automatic object collection requires prior knowledge of category structures. To solve this chicken-and-egg problem, we learn shape patterns from raw segmentations, and then refine these segmentations based on the pattern knowledge. In the process, we solve two challenging problems of knowledge mining. First, as some categories have large intra-category structure variations, we design an entropy-based method to determine the structure variation for each category, in order to establish the correct range of sample collection. Second, because incorrect segmentation is unavoidable without prior knowledge, we propose a novel unsupervised method that uses a pattern competition strategy to identify and subtract shape patterns formed by incorrectly segmented objects. This ensures that shape patterns are meaningful at the object level. Experimental results demonstrated the effectiveness of the proposed method for category structure mining in a large urban environment.**

I. Introduction

Object-level visual knowledge is of great importance in many areas of artificial intelligence, such as computer vision, robotics, and data mining. Recently, increased attention has been paid to the mining of categorical visual knowledge, and this has led some researchers to attempt to find an efficient way of constructing a category knowledge base. The knowledge base can directly provide high-level knowledge for many visual tasks, such as object retrieval, recognition, segmentation, and tracking. In contrast to conventional learning of a specific model for each different task, the construction of a general and comprehensive category knowledge base may pave the way for a new methodology for advanced visual tasks.

Some pioneering work has mined category knowledge from images via unsupervised object discovery [15] and semi-supervised learning with image search engines [13], [14]. These image-based approaches mainly use bag-of-words models without encoding spatial structure information. However, in many cases, it is the 3D structure that dictates the function and thus the category of an object. Thus, learning a category structure base from a large 3D point cloud could present a practical alternative. However, the mining of category structures is hampered by the following chicken-and-egg conundrum:

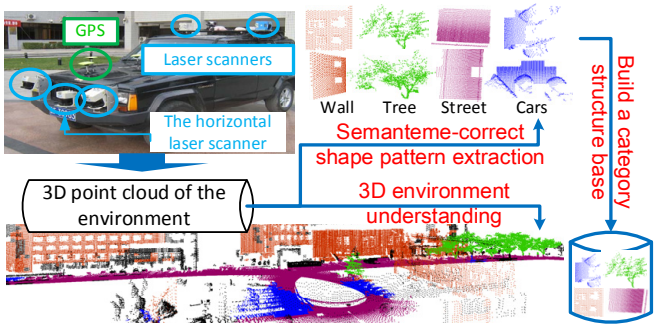

Fig. 1. Category structure knowledge mining, sample collection, and environment understanding are achieved in an unsupervised manner from the unlabeled 3D point cloud of a large urban environment.

1) In general, accurate point cloud segmentation and classification requires prior knowledge of the global structure of various objects to overcome challenges caused by the possible complex object structures and backgrounds, as well as intra-category structure variations. For example, objects with complex structures may be segmented into several distinct categories, connected objects may not have clear boundaries for segmentation, and objects in some categories may have relatively low structure similarities.

2) On the other hand, prior category knowledge needs to be learnt from a large collection of object samples in different categories, but accurate sample collection depends on the accuracy of the classification and segmentation.

Some previous approaches are of great significance in terms of supervised classification [20], [21], [22], [23], [24], [25], [26], [28] and unsupervised segmentation [30] of 3D point clouds based on local features. Other studies contribute to the learning of common structures of different categories from well-segmented 3D objects [29] or environments that are clear for segmentation [3], [5]. However, for the construction of a category structure base, we must *collect object samples in an unsupervised manner, and then learn all structural patterns of each category from the unlabeled 3D point cloud of a large and real environment*, which requires the simultaneous solution of the two interdependent issues mentioned above in a single framework.

Thus, we propose a two-step category-mining method consisting of repetitive shape pattern extraction and incorrectly segmented pattern subtraction. In the first step, we cluster the objects produced by raw segmentation into different repetitive shape patterns. We learn the common

[1] Quanshi Zhang, Xuan Song, Xiaowei Shao, and Ryosuke Shibasaki are with Center for Spatial Information Science, University of Tokyo. /zqs1022,songxuan,shaoxw,shiba/ at csis.u-tokyo.ac.jp

[2] Huijing Zhao is Key Laboratory of Machine Perception (MoE), Peking University. zhaohj at cis.pku.edu.cn

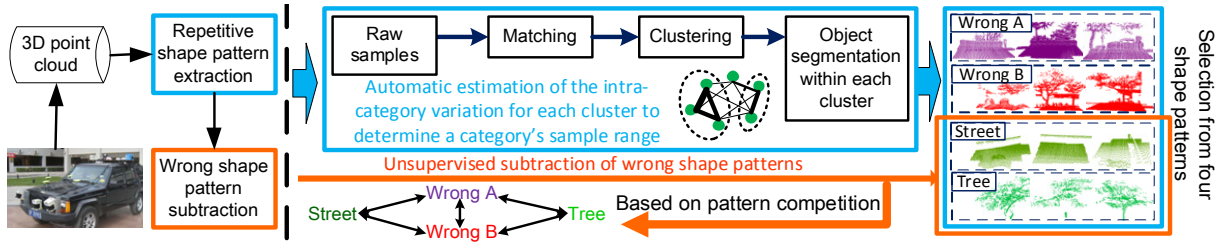

Fig. 2. Flowchart of our proposed framework showing the two main steps—repetitive shape extraction and unsupervised incorrect shape pattern subtraction. In the first step, we cut, match, and cluster object samples in the environment. Simultaneously, the minimum description length (MDL) principle is used to estimate the intra-category variation of each cluster. These clusters stand for the initial shape patterns, and object segmentation is then refined with shape pattern knowledge. In the second step, we use pattern competition in the descriptive area to identify and subtract incorrect patterns. Thus, we ensure that the extracted shape patterns are meaningful at the object level, which is a key issue for category mining. Here, two out of the four shape patterns on the right are subtracted, and the orange box contains the remaining two.

global structure of objects in each shape pattern, and further use this to refine the object segmentation within each cluster. During the clustering process, we must determine the range of each category to simultaneously avoid both over division of *loose* categories and low purity of *dense* categories. Thus, we utilize the minimum description length (MDL) principle to control the clustering process and provide a global solution to the estimation of intra-category shape variation.

Next, we wish to determine *how to detect incorrect segmentations without high-level category knowledge?* Strictly speaking, this is a high-level problem that requires the human object cognition; nevertheless, we can handle it in a complex but practical fashion. In the second step, our key contribution is the proposal of an unsupervised method to detect and subtract incorrect shape patterns. The descriptive areas of correct and incorrect patterns usually overlap in the environment, as the incorrect patterns mainly consist of partially segmented objects. Thus, based on the assumption that the correct patterns are those that are able to encode the environment with the fewest conflicts within their descriptive areas, we define an energy function to model the competitive relationship between different shape patterns in object representation. In this way, the detection of incorrect segmentations is converted into a global energy minimization problem (Fig. 2).

The main contributions of this paper are as follows. We propose, for the first time, unsupervised construction of a 3D category shape knowledge base as a starting point for many advanced visual tasks, and design an algorithm for category discovery and sample collection. When applied to the mining of category structures from a complex environment, our proposed method is the first to simultaneously deal with incorrect segmentations and considerable intra-category shape variation.

II. Related Work

Point cloud processing has developed rapidly in recent years. In this section, we discuss a wide range of related work to provide a better understanding of category structure mining.

Knowledge mining: The segmentation and the classification (point labeling) of 3D point clouds are two kinds of 3D environment understanding [20], [21], [22], [23], [24], [25], [26], [27], [28]. Munoz *et al.* [20], [28], Triebel *et al.* [21], and Anguelov *et al.* [27] made a breakthrough that they employed associative Markov networks (AMNs) with a max-margin strategy for supervised point cloud classification and segmentation. However, knowledge mining mainly requires learning to be conducted from unlabeled data in an unsupervised or semi-supervised manner. Besides, the mined knowledge should also represent the global structure of an object, rather than just meeting segmentation criteria based on local information.

Nevertheless, we also use our mined category structures to achieve scene understanding, and perform experiments to compare its performance with the classical supervised AMN-based point labeling.

Object-level global structure: A number of pioneering studies have contributed to the extraction of high or middle level structure knowledge. Hebert *et al.* [10] used some high-level shape assumptions to discover various structures in the environment, whilst Ruhnke *et al.* [8] learned a compact representation of a 3D environment based on Bayesian information criteria. Endres *et al.* [4] used latent Dirichlet allocation to discover 3D objects. These methods focused on part-level patterns, whereas we expected to extract global structures with the correct object-level semantemes.

Intra-category variation: Closer to our field of category structure mining, some approaches for the collection of 3D object samples have been proposed. Herbst *et al.* [1], [2] detected which objects had been moved across multiple depth images of the same scene, and Somanath *et al.* [9] detected the same objects appearing in different 3D scenes. Detry *et al.* [7] learned a general hierarchical object model from stereo data with clear edges. Category structure mining is not limited to the segmentation of objects with some specific shapes or recurrent objects, but also includes the discovery and learning of all shape patterns in each category, including any possible shape variations. From this viewpoint, the most closely related work involves repetitive shape detection [3], [5].

However, category structure mining must also be achieved in a more general environment, in which ground subtraction cannot be used for object segmentation, as many objects with irregular shapes could be connected to each other. In other words, *we have to solve the chicken-and-egg problem mentioned in Section I. Thus, we learn to correctly segment objects and subtract incorrect ones.* In our experiments, we

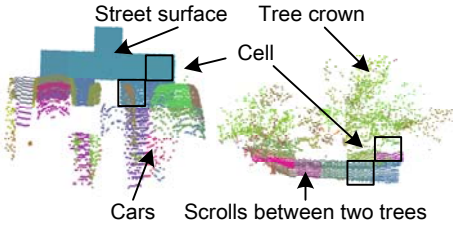

Fig. 3. Object samples and local features. There are two object samples—cars on the side of the street and trees with scrolls. Values in the red, green, and blue channels of the point color show {*linear, surface, block*}-*ness* of the local geometry, respectively.

realize the core idea of conventional repetitive shape detection in competing frameworks, and evaluate the performance of our method in a general urban environment.

III. Repetitive Shape Pattern Extraction

Repetitive shape pattern extraction is the first step of our system and achieved using the following processes. First, we search for object samples in all positions of the environment, and then apply 3D matching to calculate the object similarity. Based on the object similarity, we use hierarchical clustering to extract repetitive shape patterns in the environment. Finally, we refine the segmentation of object samples in each cluster based on the cluster's common shape.

This style of shape pattern extraction—combining matching and clustering—encounters two challenges in a real environment: 1) incorrect segmentation subtraction, and 2) intra-category variation determination without prior knowledge of category structures.

To tackle the first challenge, we use a cluster's shape pattern to refine the raw segmentation of its objects, and leave the detection and subtraction of incorrect shape patterns to be performed afterwards (see Section IV).

For the second challenge, it is necessary to estimate a specific intra-category shape variation to help determine the sample range of a given category. Otherwise, a category with a large shape variation could be incorrectly divided into several sub-categories, or a category with a small shape variation could be erroneously grouped with other objects. Thus, we use the MDL principle to incrementally estimate the intra-category structure variation during the clustering process.

A. Object sampling and features

A brute-force search is employed to detect object samples in the environment. These samples are then segmented by a cylinder of height s and radius $s/2$. The point cloud within the cylinder is considered as the raw object segmentation. The environment is divided into local cells, from which local features are extracted. Object samples are also represented at the cell level. Considering the noise and point density, cells are defined as cubes with an edge length of $1.25m$ in order to obtain reliable local features. An object sample is represented by the local features and 3D coordinates of a set of cells. We use three geometric features inspired by the spectral analysis of point clouds [19], [12]. Given the point cloud within a cell, we define $\lambda_1 \geq \lambda_2 \geq \lambda_3$ to be the eigenvalues of the scatter matrix over these 3D points. We use $\{f_1^*=\lambda_1/\max\{\sqrt{\lambda_2\lambda_3},\varepsilon\}, f_2^*=\max\{\lambda_2/\max\{\sqrt{\lambda_1\lambda_3},\varepsilon\}, \sqrt{\lambda_1\lambda_3}/\max\{\lambda_2,\varepsilon\}\}, f_3^*=\sqrt{\lambda_1\lambda_2}/\max\{\lambda_3,\varepsilon\}\}$ to measure the relative {*linear, surface, block*}-*ness* of the local geometry. ε is set to 0.1 to prevent too large values of $\mathbf{f}^*$. The features are normalized $\mathbf{f}=\mathbf{f}^*/\sqrt[3]{f_1^* f_2^* f_3^*}$, and then each dimension is normalized to the same variation. Fig. 3 shows the object samples and their local features.

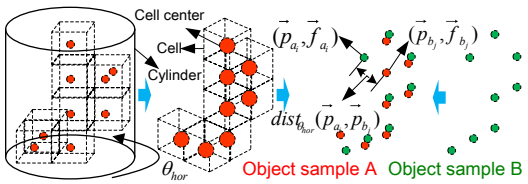

Fig. 4. Cell-level 3D matching. In this figure, we use cell centers to represent the object. For clarity, we picture a smaller cylinder and a simpler object than in the real parameter settings.

B. 3D matching and similarity graph

To make a full use of spatial information, we use 3D matching to calculate the structure similarity between two samples. The matching uses both global positions and local features. We can ignore translations in matching and only consider horizontal rotations, as object samples are searched in all possible positions and assumed to *stand* on the ground.

Given two object samples A and B with their cell sets $\{a_i\}$ and $\{b_j\}$, $i=1,2,...,n_A, j=1,2,...,n_B$, their sets of local features and spatial coordinates are denoted as $\{\mathbf{f}_{a_i}\}$, $\{\mathbf{f}_{b_j}\}$ and $\{\mathbf{p}_{a_i}\}$, $\{\mathbf{p}_{b_j}\}$, respectively. Fig. 4 illustrates some variables for clarity. When A and B are matched under horizontal rotation θ_{hor}, the matching probability of two cells a_i and b_j is calculated as follows:

$$P(\mathbf{p}_{a_i},\mathbf{p}_{b_j},\mathbf{f}_{a_i},\mathbf{f}_{b_j}|\theta_{hor}) = P(\mathbf{p}_{a_i},\mathbf{p}_{b_j}|\theta_{hor})P(\mathbf{f}_{a_i},\mathbf{f}_{b_j}) \quad (1)$$

$$P(\mathbf{p}_{a_i},\mathbf{p}_{b_j}|\theta_{hor}) \sim \mathcal{N}(dist_{\theta_{hor}}(\mathbf{p}_{a_i},\mathbf{p}_{b_j})|\mu=0,\sigma^2) \quad (2)$$

$$P(\mathbf{f}_{a_i},\mathbf{f}_{b_j}) \sim \beta\mathcal{N}(\mathbf{f}_{a_i}-\mathbf{f}_{b_j},|\mu=0,\Sigma) \quad (3)$$

where, $dist_{\theta_{hor}}(\mathbf{p}_{a_i},\mathbf{p}_{b_j})$ indicates the distance between a_i and b_j with rotation θ_{hor}; Σ is the covariance matrix of local features, σ^2 is the variance of position distances, and β is a weighting for local features ($\beta=1$, here).

The probability that a cell a_i in A is well matched to B, is calculated as:

$$\begin{aligned} P(a_i,B|\theta_{hor}) &= \max_{b_j} P(\mathbf{p}_{a_i},\mathbf{p}_{b_j},\mathbf{f}_{a_i},\mathbf{f}_{b_j}|\theta_{hor}) \\ &\approx \textstyle\sum_j P(\mathbf{p}_{a_i},\mathbf{p}_{b_j},\mathbf{f}_{a_i},\mathbf{f}_{b_j}|\theta_{hor}) \end{aligned} \quad (4)$$

The approximation is based on the fact that the probability of a_i matching to its best matching cell in B is usually far greater than the probability matching to other cells in B. Thus, the similarity between A and B is the maximum of the following value:

$$\max_{\theta_{hor}} \frac{\sum_i P(a_i,B|\theta_{hor}) + \sum_j P(b_j,A|\theta_{hor})}{n_A+n_B} \quad (5)$$

The minimization problem can be solved via gradient descent methods with some initial pose estimations, or just exhaustive search.

C. MDL-based hierarchical graph clustering

Repetitive shape patterns are extracted via hierarchical graph clustering. The MDL principle is used to provide a global solution to the intra-category shape-variation estimation problem for an accurate boundary of each cluster. Let $G = (V, E)$ be an undirected and weighted similarity graph. Each vertex $v_i \in V$ denotes an object sample, and each edge $e_{ij} \in E$ indicates the sample similarity that is defined by (5). The total description length provides a global penalty of the current clustering status, and is formulated based on [17] as follows:

$$L(V,C) = L(V|C) + L(C) \tag{6}$$

where, $C = \{c_j\}$ is the current cluster set; $L(C)$ is the description length of the cluster division; $L(V|C)$ represents the residual uncertainty of object samples due to their intra-cluster variations.

$$L(C) = -\sum_i P_i \log P_i \tag{7}$$

$$L(V|C) = -\alpha \sum_i P_i \log D_i \tag{8}$$

where, $P_i = \|c_i\|/\|V\|$ is the probability of observing cluster c_i; $D_i = \sum_{e_{jk} \in E: v_j, v_k \in c_i} e_{jk} / \|\{e_{jk} \in E | v_j, v_k \in c_i\}\|$ is the cluster density of c_i and measures the intra-cluster variation; α is a weighting that connects the cluster density and the cluster uncertainty ($\alpha = 1.5$, here).

The term $L(C)$ encodes the prior penalty of each cluster, as small clusters have large observation uncertainties in the environment. The term $L(V|C)$ assigns the cluster with large intra-cluster variation with large penalty. The minimization of $L(V,C)$ balances these two terms and estimates the suitable intra-category variation via a global optimization.

Before clustering, each object sample is initialized as a cluster. Then, we achieve the clustering via the gradient descent method. In each subsequent step, the pair of clusters (c_i, c_j) with the steepest descent of description length is merged, as an approximate solution of description length minimization.

$$\max_{c_i, cj} \frac{L(V,C) - L(V, (C \cup \{c_i \cup cj\}) \backslash \{c_i, cj\})}{\|c_i\| + \|c_j\|} \tag{9}$$

From all the learned clusters, we select ones with a size no less than τ as reliable clusters ($\tau = 30$, here) for further processing.

D. Object segmentation

Each cluster stands for a specific shape pattern in the environment, and we use this top-down information to segment objects in each cluster (Fig. 5). We match each pair of object samples in a cluster, and the most frequently matched parts of each object sample are taken as the true body of the object.

Given sample A with its cells $\{a_i\}$ $(i = 1, 2, ..., n_A)$ and sample B, their matching rotation θ_{hor} is calculated by (5). $P(a_i, B|\theta_{hor})$ indicates the probability that a_i is well matched to B under rotation θ_{hor} calculated by (4). If $P(a_i, B|\theta_{hor})$ is greater than a threshold (0.02 here), then a_i is considered well matched to B. Those cells that cannot well match enough

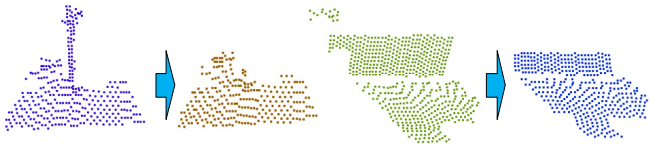

Fig. 5. Object segmentation. Here, the pole (left) and the wall (right) on the side of the street are removed, according to the common structure of the *street* pattern.

object samples (at least 60% of the samples in the cluster) will be considered as a part of the background and removed from A.

IV. Incorrect Pattern Subtraction by Pattern Competition

In a noisy and complex environment, the techniques of ground subtraction, plane detection [3], [5], and hierarchical segmentation [11] cannot be reliably used as preprocessing to provide object candidates. Consequently, the extraction of incorrect shape patterns cannot be avoided. For example, a tree crown and trunk could be divided into different objects; the street with cars parking along the side may share the same shape pattern with various small objects on the ground. Thus, we propose an unsupervised method to subtract incorrect patterns as a key step between conventional repetitive shape pattern extraction and the category mining task, thus ensuring correct object-level semantemes (Fig. 7).

We first define the descriptive area of a shape pattern as follows: a cell x_i in the environment is described by a shape pattern c_j if and only if x_i is contained by an object sample of c_j. We write this descriptive relationship as $x_i \Rightarrow c_j$.

Obviously, each cell can be described by multiple shape patterns; alternatively we can say that different shape patterns are competing for their descriptive area in the environment, as shown in Fig. 6. Compared to incorrect shape patterns, correct patterns describe different object categories more clearly, with fewer conflicts in their descriptive area. Thus, although the problem of incorrect pattern subtraction cannot strictly be solved in an unsupervised way, pattern competition provides the following approximate solution to this cognition-level problem.

Our goal is to select a set of shape patterns to represent objects in the environment. These shape patterns should satisfy the criteria that: 1) they can successfully describe most cells in the environment; and 2) only one, or a limited number of shape patterns, describes each cell. We encode these two requirements in an energy function.

Given a set of cells $X = \{x_i\}$ and a set of shape patterns $C = \{c_j\}$, the penalty for a cell $x_i \in X$ that is described by n_i shape patterns, is defined as follows:

$$E(x_i|C) = \begin{cases} \log n_i & n_i > 0 \\ \lambda & n_i = 0 \end{cases} \tag{10}$$

$$n_i = \|\{c_j \in C | x_i \Rightarrow c_j\}\| \tag{11}$$

where, a large penalty λ ($\lambda = 1$, here) is assigned if x_i cannot be described by any shape patterns in C, preventing excessive subtraction of shape patterns.

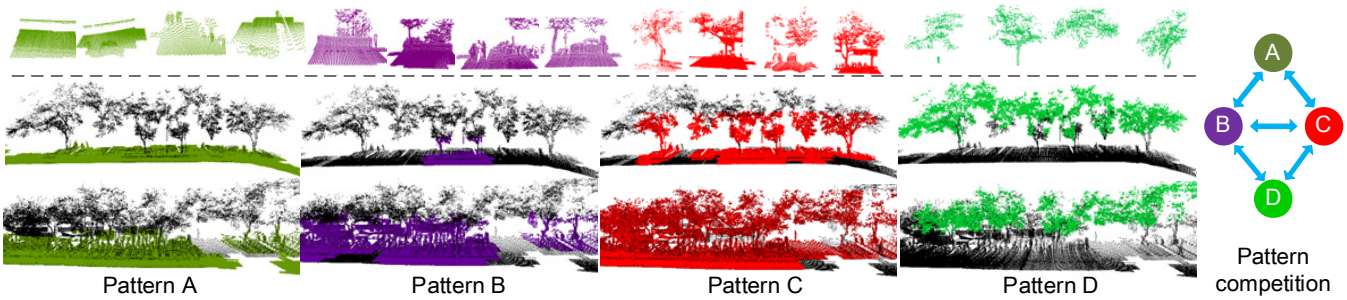

Fig. 6. Competitive relationship in the descriptive area between shape patterns. A cell is described by a shape pattern, iff the cell is contained by an object sample of this pattern. Object samples of four different shape patterns are shown in the first row. Pattern A and Pattern D describe the street and the tree, which are clear in semanteme, but Pattern B and Pattern C are incorrect patterns. The main competitive relationship between the four patterns is shown on the right.

The energy function is defined as follows:

$$E(X|C) = \sum_i E(x_i|C) \tag{12}$$

The whole task is converted to the selection of a subset of shape patterns, $C_{sub} \subseteq C$, that minimizes the energy function as follows:

$$\arg\min_{C_{sub} \subseteq C} E(X|C_{sub}) \tag{13}$$

We use a greedy strategy to obtain an approximate solution to the energy minimization problem. Initially, C_{sub} is set as C. In each subsequent step, we select and remove a shape pattern from C_{sub}, in order to reduce the system energy in the direction of the steepest descent, as follows:

$$\arg\max_{c_i} \frac{E(X|C_{sub}) - E(X|C_{sub} \backslash \{c_i\})}{\|\{x_j \in X | x_j \Rightarrow c_i\}\|} \tag{14}$$

The energy decrease is normalized by the descriptive range of a shape pattern in order to obtain the energy gradient. This iterative process stops when the energy cannot be reduced by the removal of any more shape patterns. Fig. 7 illustrates the improvement in the accuracy of shape patterns.

V. Experiments

Intelligent vehicle and 3D point clouds: We develop an intelligent vehicle system to collect 3D data. The vehicle is equipped with five single-row laser scanners to profile its surroundings in different directions, as shown in Fig. 1. A global positioning system (GPS) and an inertial measurement unit (IMU) are mounted on the vehicle. A localization module is developed by fusing the GPS/IMU navigation unit with a horizontal laser scanner. In this module, the localization problem is formulated as a simultaneous localization and mapping system with moving object detection and tracking [16], thus ensuring both the global and local accuracy of the vehicle's pose estimation. The 3D representation of the environment can be obtained by geo-referencing the local range data from four slant laser scanners in a global coordinate system, given the vehicle's pose and the geometric parameters of the slanted laser scanners.

The proposed category structure mining requires that each category contains enough objects to form a shape pattern, so the environment must be quite large and contain various objects. Thus, our intelligent vehicle collects 3D point cloud data in a large urban environment. The size of the whole point cloud in our experiment is $300m \times 400m$. Within the environment, the smooth street has a small variation in shape, whereas the wall has a larger shape variation. The wall has a high noise level due to its distance from the intelligent vehicle. The wall also has various shapes, as it may be occluded by tree branches. Trees have the largest shape variation, due to their varied structures.

Results: Our proposed system, which combines MDL-based intra-category variation estimation and incorrect pattern subtraction, generates 10 shape patterns: 5 *wall* patterns, 1 *tree* pattern, 2 *street* patterns and 2 *cars* patterns. Some object samples of each pattern are shown in Fig. 8.

A. Comparison with repetitive shape pattern extraction

As discussed in Section II, the main comparable approaches close to the spirit of category mining are the repetitive shape detection [3], [5]. They clustered directly segmented objects into shape patterns, and we realize their core idea in a competing framework for comparison. Considering these methods' difficulties in a complex environment, we convert their segmentation and matching techniques into our style to make it fit our noisy and sparse data, and the competing framework uses average-linkage hierarchial graph clustering instead of our MDL-based clustering. The proposed system obtains 47 clusters with sizes larger than τ before pattern subtraction, so for comparison, the stopping criterion for the average-linkage clustering is set as that 47 clusters with sizes larger than τ have been obtained. Moreover, we further design two frameworks in between—one is made by replacing the average-linkage clustering in the competing framework with the MDL-based clustering, and the other by adding incorrect pattern extraction to the competing framework. Thus, each module's performance in the proposed method can be evaluated sequentially.

As category mining can be considered as a combination of object detection and segmentation within a single framework, we need to evaluate object detection and segmentation simultaneously, rather than sequentially. Moreover, in contrast to conventional sample-based model learning and testing, category mining learns directly from the point cloud of a large environment, not from a set of depth images as samples. In the environment, some objects are uncountable, such as the wall and the street, so it is impossible to apply a sample-based evaluation.

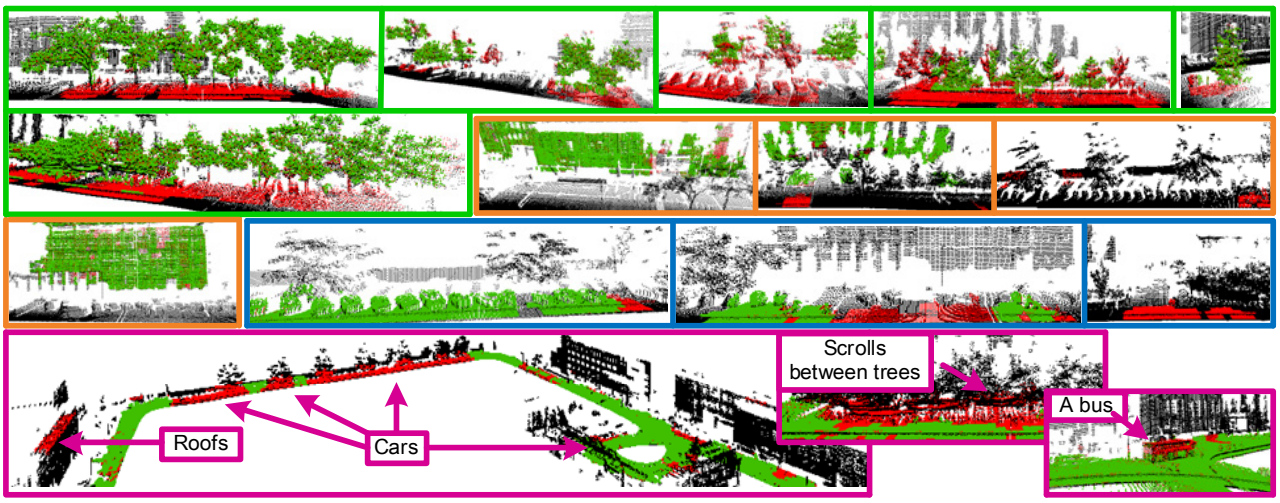

Fig. 7. Effects of incorrect shape pattern subtraction. The extracted shape patterns are assigned to different categories (see Section V-A for details). The descriptive areas of shape patterns in the same category are shown in these sub-figures. The color of the sub-figure box indicates its category—*wall* (orange), *tree* (green), *street* (purple) or *cars* (blue). In each sub-figure, the descriptive areas of the different shape patterns overlap. The subtraction of incorrect shape patterns reduces the descriptive area of all the shape patterns (red and green) to that of the *correct* shape patterns (green) that are better matched to human cognition.

TABLE I

THE NMI VALUES OF SHAPE PATTERNS

Method	NMI value
Proposed method	0.629
Competing method with MDL-based clustering	0.341
Competing method with incorrect-pattern subtraction	0.320
Original competing method[3],[5]	0.074

Therefore, motivated by evaluation methods of object segmentation, we choose cell-level evaluation instead. For this purpose, cells in the environment are manually labeled as four categories: *wall*, *tree*, *street* and *cars* as the ground truth. Some small fragments are unclear in object-level semanteme, and not labeled nor used in evaluation. A shape pattern is assigned the same label as most of its describable cells.

We use multiple evaluation indexes in our experiment. The normalized mutual information (NMI) is a general evaluation of shape patterns. A high NMI value indicates that (1) the composition of the cells described by each shape pattern has high purity, and (2) the number of shape patterns for one object category is small. For detailed comparison, we also utilize the purity, shape pattern number, detection rate and error rate to evaluate the result from different perspectives. All the evaluation indexes are calculated at the cell level.

NMI: m_i^k ($k \in \{wall, tree, street, cars\}$) denotes the number of cells in the i^{th} pattern marked with label k. The NMI value is calculated as follows:

$$NMI = \sum_i \sum_k \frac{m_i^k}{N} \log \frac{N m_i^k}{M^k M_i} \tag{15}$$

$$M_i = \sum_k m_i^k \quad M^k = \sum_i m_i^k \quad N = \sum_k M^k \tag{16}$$

Table I lists the NMI values of the proposed method and three competing frameworks. From this table, we can conclude that the NMI value is increased by the incorrect pattern subtraction and the use of the MDL principle in clustering.

TABLE II

THE SHAPE PATTERN NUMBER AND THE PURITY OF DIFFERENT OBJECT CATEGORIES

Shape pattern number (upper) / Purity (lower)				
The method	Wall	Tree	Street	Cars
The proposed method	5 0.985	1 0.974	2 0.860	2 0.635
The competing method with MDL-based clustering	12 0.974	15 0.850	15 0.746	5 0.588
The competing method with incorrect pattern subtraction	1 1.000	0 none	4 0.857	1 0.629
The original competing method[3],[5]	1 1.000	0 none	40 0.848	6 0.589

Purity, detection rate, and error rate: The purity of a shape pattern $Purity_i$ is calculated as follows:

$$Purity_i = \max_k m_i^k / M_i \tag{17}$$

We show the shape pattern purity in Fig. 8. The purity of a category is calculated as follows:

$$Purity^k = \frac{\sum_{i:\forall j, m_i^k \geq m_i^j} \max_k m_i^k}{\sum_{i:\forall j, m_i^k \geq m_i^j} M_i} \tag{18}$$

Table II lists shape pattern numbers and purities of different categories learned in the proposed method and three competing frameworks. As the cars are connected by the street in object samples of *cars* patterns, the *cars* category has relatively low purity.

We can consider category mining as a category detection task: for a cell $x_i \Rightarrow c_j$, if shape pattern c_j belongs to category k, we say that x_i can be detected as category k. The same cell may be detected as different categories. The category detection result is shown in Fig. 8. The detection rate of a category is the percentage of cells that are correctly detected, out of all the cells that are manually labeled as this category in the ground truth. The error rate of a category is the percentage of cells incorrectly detected, out of all the cells

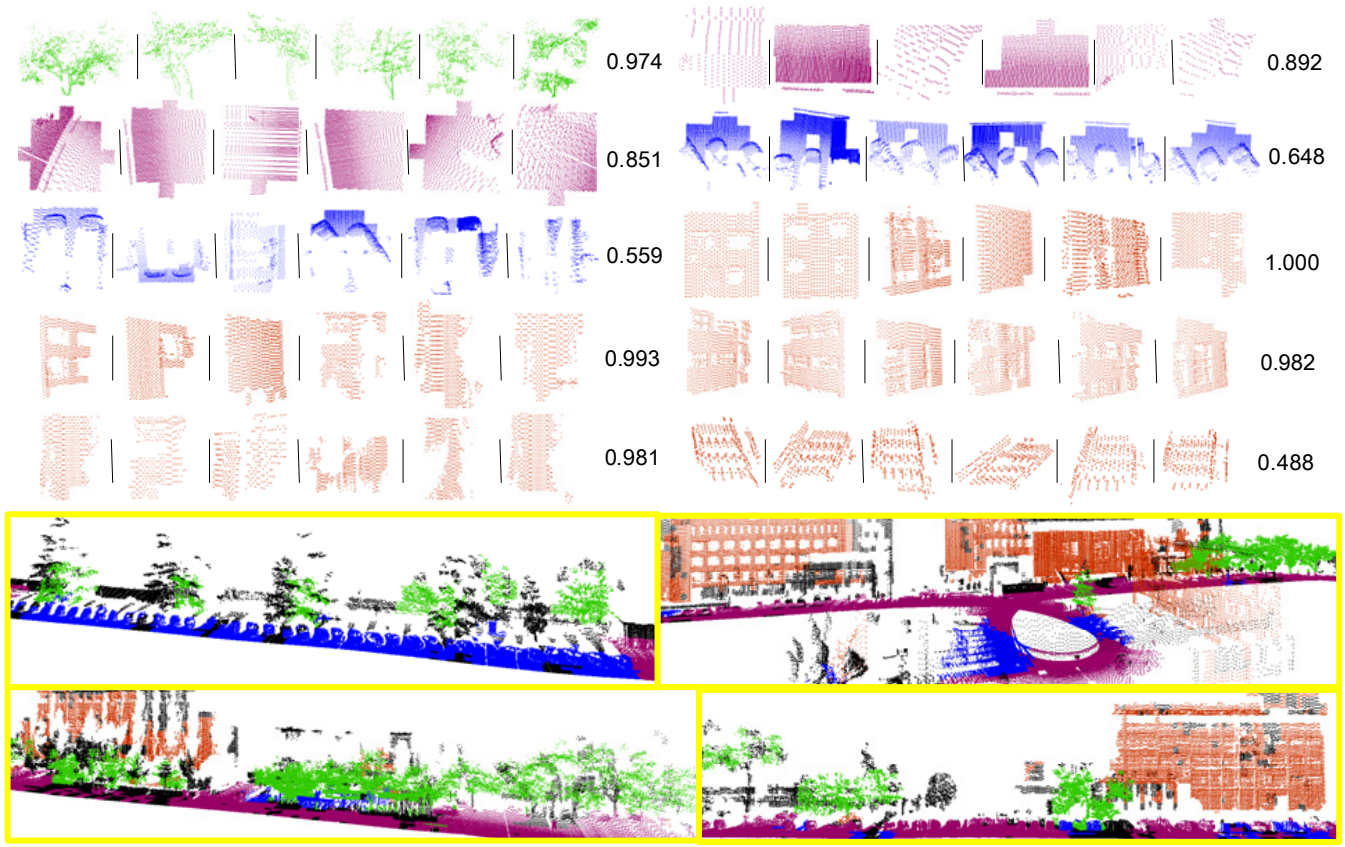

Fig. 8. Collected samples and environment understanding. Different colors indicate different categories—*wall* (orange), *tree* (green), *street* (purple) and *cars* (blue). Object samples of each shape pattern are shown with the pattern's purity (defined in (17)) on the right. In fact, the last *wall* pattern describes slanted roofs of buildings, so its shape is different from the ordinary wall. The environment is represented by cell-level category detection. For more results, please see our demo in the supplementary materials.

not labeled as this category. Table III shows the detection rate and error rate of each category.

The MDL-based graph clustering assigns different shape variations for different object categories, while the average-linkage graph clustering is prone to get clusters with high density in the similarity graph (small shape variation). Thus, it is difficult for average-linkage graph clustering to obtain clusters with large shape variations in a complex environment. As a result, the original competing framework (with average-linkage graph clustering) yields 40 *street* clusters, 6 *cars* clusters, only 1 *wall* cluster, and no *tree* cluster. This is because *street* has a small shape variation, *cars* has a moderate shape variation, whereas *wall* and *tree* have a large shape variations, as discussed in the beginning of Section V.

In contrast, the MDL principle assigns suitable shape variations for different object categories, so the competing framework with MDL-based clustering yields a more balanced result: 12 *wall* clusters, 15 *tree* clusters, 15 *street* clusters and 5 *cars* clusters. Moreover, for the large-shape-variation categories such as *wall* and *cars*, the MDL-based graph clustering leads to a much higher detection rate than the average-linkage graph clustering.

Incorrect pattern subtraction selects a set of semanteme-correct shape patterns to describe objects in the environment. The NMI value, as a general evaluation of shape patterns, proves the significant contribution of the incorrect pattern subtraction (Table I). More specifically, the incorrect pattern subtraction greatly decreases the shape pattern number and slightly increases the purity of each object category (Table II). The error rate for the category detection is greatly reduced by incorrect pattern subtraction at the cost of a slight decrease in the detection rate (Table III).

B. Comparison with AMN-based point cloud classification

AMNs have demonstrated a superior performance in point cloud classification [20], [21], [28] and segmentation [27] in recent years. Although not designed for category structure mining, and despite their requirement to learn a max-margin classifier from a large number of training data, we compare AMNs with our system from the perspective of environment understanding.

Thus, we compare the detection rate and error rate of supervised AMN-based classification [20], [28] with our category detection based on the mined category structure knowledge, as shown in Table III. AMNs are trained to classify the *wall*, *tree*, *street*, *cars*, and *unlabeled* categories. The *unlabeled* category mainly consists of small fragments due to data sparsity and other objects, such as buses. These are unclear in the object-level semanteme, and thus not used in the previous evaluation. Therefore, following the same criterion, neither the unlabeled data nor the point cloud

TABLE III

THE DETECTION RATE AND THE ERROR RATE OF DIFFERENT OBJECT CATEGORIES

Detection rate (upper) / Error rate (lower)				
The method	Wall	Tree	Street	Cars
The proposed method	0.647 0.004	0.646 0.008	0.821 0.070	0.587 0.045
The competing method with MDL-based clustering	0.768 0.016	0.886 0.179	0.996 0.217	0.672 0.058
The competing method with incorrect-pattern subtraction	0.164 0.000	none none	0.832 0.079	0.356 0.025
The original competing method[3],[5]	0.164 0.000	none none	0.813 0.084	0.530 0.062
AMN-based classification[20],[28] 300 cliques for training	0.621 0.049	0.692 0.029	0.157 0.010	0.050 0.022
AMN-based classification[20],[28] 900 cliques for training	0.617 0.019	0.729 0.020	0.343 0.019	0.241 0.017
AMN-based classification[20],[28] 2700 cliques for training	0.582 0.016	0.779 0.021	0.501 0.023	0.304 0.020

cliques (see [20], [28]) classified as *unlabeled* by the AMN are used in the evaluation.

The max-margin strategy allows the AMN to operate as a powerful classifier, but it only uses features extracted from a single local clique to determine the overall category of the clique. Thus, in many cases, normal objects are classified as *unlabeled* (object fragments).

VI. Discussion and Conclusions

We have successfully developed an algorithm specially for category pattern mining in a large urban environment. Experiments show its superior performance in an environment that has objects with different intra-category shape variations. In addition, we propose a global solution to the incorrect segmentation problem for complex and noisy objects.

In this study, we only cut object samples of a fixed scale from the environment, as most of the objects are on this scale. In future research, we intend to apply our approach to different 3D environment data using multiple scales for object sampling. Without color information, object segmentation based on the statistical common shape within a cluster cannot provide as clear an object boundary as image segmentation. Thus, we will add color information in future research.

ACKNOWLEDGMENT

This work was supported by Microsoft Research, a Grant-in-Aid for Young Scientists (23700192) of Japans Ministry of Education, Culture, Sports, Science, and Technology (MEST), and Grant of Japans Ministry of Land, Infrastructure, Transport and Tourism (MLIT).

References

[1] Evan Herbst, Peter Henry, Xiaofeng Ren, and Dieter Fox. Toward Object Discovery and Modeling via 3D Scene Comparison. In *ICRA*, 2011.

[2] Evan Herbst, Xiaofeng Ren, and Dieter Fox. RGB-D Object Discovery via Multi-Scene Analysis. In *IROS*, 2011.

[3] Jiwon Shin, Rudolph Triebel, and Roland Siegwart. Unsupervised Discovery of Repetitive Objects. In *ICRA*, 2010.

[4] Felix Endres, Christian Plagemann, Cyrill Stachniss, and Wolfram Burgard. Unsupervised Discovery of Object Classes from Range Data using Latent Dirichlet Allocation. In *Robotics: Science and Systems*, Seattle, WA, USA, 2009.

[5] Michael Ruhnke, Bastian Steder, Giorgio Grisetti, and Wolfram Burgard. Unsupervised Learning of 3D Object Models from Partial Views. In *ICRA*, 2009.

[6] Jens Behley, Kristian Kersting, Dirk Schulz, Volker Steinhage and Armin B. Cremers. Learning to Hash Logistic Regression for Fast 3D Scan Point Classification. In *IROS*, 2010.

[7] Renaud Detry, Nicolas Pugeault, and Justus H. Piater. A Probabilistic Framework for 3D Visual Object Representation. In *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(10):1790–1803, 2009.

[8] Michael Ruhnke, Bastian Steder, Giorgio Grisetti, and Wolfram Burgard. Unsupervised Learning of Compact 3D Models Based on the Detection of Recurrent Structures. In IROS, 2010.

[9] Gowri Somanath, Rohith MV, Dmitris Metaxas, and Chandra Kambhamettu. D-Clutter: Building object model library from unsupervised segmentation of cluttered scenes. In *CVPR*, 2009.

[10] Alvaro Collet, Siddhartha S. Srinivasay, and Martial Hebert, Structure Discovery in Multi-modal Data: a Region-based Approach. In *ICRA*, 2011.

[11] Frank Moosmann, and Miro Sauerland. Unsupervised discovery of object classes in 3D outdoor scenarios. In *ICCV Workshops*, 2011.

[12] Jean-Francois Lalonde, Nicolas Vandapel, Daniel F. Huber, and Martial Hebert. Natural Terrain Classification using Three-Dimensional Ladar Data for Ground Robot Mobility. In *Journal of Field Robotics*, 23(1), 839–861, 2006.

[13] Li-Jia Li, Gang Wang, and Li Fei-Fei. OPTIMOL: automatic Online Picture collecTion via Incremental MOdel Learning. In *International Journal of Computer Vision*, 88(2):147–154, 2010.

[14] Sudheendra Vijayanarasimhan, and Kristen Grauman. Large-Scale Live Active Learning: Training Object Detectors with Crawled Data and Crowds, In *CVPR*, 2011.

[15] Tinne Tuytelaars, Christoph H. Lampert, Matthew B. Blaschko, and Wray Buntine. Unsupervised Object Discovery: A Comparison In *International Journal on Computer Vision*, 88(2):284–302, 2010.

[16] Huijing Zhao, M. Chiba, Ryosuke Shibasaki, Xiaowei Shao, Jinshi Cui, and Hongbin Zha. SLAM in a Dynamic Large Outdoor Environment using a Laser Scanner. In *ICRA*, 2008.

[17] Petri Kontkanen, Petri Myllymaki, Wray Buntine, Jorma Rissanen, and Henry Tirri. An MDL Framework for Data Clustering. In *Advances in Minimum Description Length: Theory and Applications*, 2005.

[18] Francis Crick. Astonishing Hypothesis: The Scientific Search for the Soul. Scribner reprint edition. ISBN 0-684-80158-2, 1995.

[19] G. Medioni, M.-S. Lee, and C.-K. Tang. A Computational Framework for Segmentation and Grouping. Elsevier, 2000.

[20] Daniel Munoz, Nicolas Vandapel, and Martial Hebert. Onboard Contextual Classification of 3-D Point Clouds with Learned High-order Markov Random Fields. In *ICRA*, 2009.

[21] Rudolph Triebel, Kristian Kersting, and Wolfram Burgard. Robust 3D Scan Point Classification using Associative Markov Networks. In *ICRA*, 2006.

[22] Huijing Zhao, Yiming Liu, Xiaolong Zhu, Yipu Zhao, and Hongbin Zha. Scene Understanding in a Large Dynamic Environment through a Laser-based Sensing. In *ICRA*, 2010.

[23] Aleksey Golovinskiy, Vladimir G. Kim, and Thomas Funkhouser. Shape-based Recognition of 3D Point Clouds in Urban Environments. In *ICCV*, 2009.

[24] Christian Wojek, Stefan Roth, Konrad Schindler, and Bernt Schiele. Monocular 3D Scene Modeling and Inference: Understanding Multi-Object Traffic Scenes. In *ECCV*, 2010.

[25] Ingmar Posner, Mark Cummins, and Paul Newman. A Generative Framework for Fast Urban Labeling Using Spatial And Temporal Context. In *Autonomous Robots*, 26(2–3), 153–170, 2009.

[26] Klaas Klasing, Dirk Wollherr, and Martin Buss. Realtime segmentation of range data using continuous nearest neighbors. In *ICRA*, 2009.

[27] Dragomir Anguelov, Ben Taskary, Vassil Chatalbashev, Daphne Koller, Dinkar Gupta, Geremy Heitz, and Andrew Ng. Discriminative learning of markov random fields for segmentation of 3d scan data. In *CVPR*, 2005.

[28] Daniel Munoz, J. Andrew Bagnell, Nicolas Vandapel, and Martial Hebert. Contextual Classification with Functional Max-Margin Markov Networks. In *CVPR*, 2009.

[29] Kevin Lai, Liefeng Bo, Xiaofeng Ren, and Dieter Fox A Large-Scale Hierarchical Multi-View RGB-D Object Dataset. In *ICRA*, 2011.

[30] Klaas Klasing, Dirk Wollherr, and Martin Buss. A Clustering Method for Efficient Segmentation of 3D Laser Data. In *ICRA*, 2008.