Structure Discovery in Multi-modal Data: a Region-based Approach

Alvaro Collet* Siddhartha S. Srinivasa† Martial Hebert*

*The Robotics Institute
Carnegie Mellon University
Pittsburgh, PA, USA
{acollet, hebert}@cs.cmu.edu

†Intel Labs Pittsburgh
4720 Forbes Ave. Suite 410
Pittsburgh, PA, USA
siddhartha.srinivasa@intel.com

***Abstract*—The ability of a perception system to discern what is important in a scene and what is not is an invaluable asset, with multiple applications in object recognition, people detection and SLAM, among others. In this paper, we aim to analyze all sensory data available to separate a scene into a few physically meaningful parts, which we term *structure*, while discarding background clutter. In particular, we consider the combination of image and range data, and base our decision in both appearance and 3D shape. Our main contribution is the development of a framework to perform scene segmentation that preserves physical objects using multi-modal data. We combine image and range data using a novel mid-level fusion technique based on the concept of regions that avoids any pixel-level correspondences between data sources. We associate groups of pixels with 3D points into multi-modal regions that we term *regionlets*, and measure the structure-ness of each regionlet using simple, bottom-up cues from image and range features. We show that the highest-ranked regionlets correspond to the most prominent objects in the scene. We verify the validity of our approach on 105 scenes of household environments.**

I. Introduction

Recent renewed interest in personal robotics [1]–[3] has produced key innovations in perception and planning, but has also revealed the complexity inherent to robots interacting with human environments. Such environments are dynamic, cluttered, and demonstrate an apparent lack of structure compared to industrial setups. These characteristics force robots to plan for large amounts of uncertainty, often requiring complete redesigns of current algorithms [4].

And yet, there *is* structure in the so-called unstructured environments, that humans are able to find. Perceptual grouping plays a powerful role in human visual perception [5]. To our eyes, homes, offices, parks and streets all show remarkable amounts of structure, such as walls, objects, trees, or cars.

In this work, our goal is to generate a scene segmentation, together with a ranking mechanism, such that the highest-ranking segments correspond to physical entities in the scene. We call this process *structure discovery*. We combine range and image data to compute perceptual cues such as concavities and discontinuities. These cues are then used to generate scene segmentations that preserve physical entities.

We believe that perceptual grouping has broad applications in robotic perception, from object recognition and people detection, to reliable features for localization and SLAM.

We are motivated by the results of Hoiem et al. [6] on scene interpretation from a single image. In [6], Hoiem et al. reason about occlusions as a way to segment a scene while preserving its 3D surfaces. We aim to solve a similar problem, but using a radically different approach. In single-image scene interpretation [7], most of the effort is spent in reconstructing qualitative 3D interpretations of the scene. We claim that, by adding range data and performing multi-modal sensing, we can recover more information from more complex scenes, using much simpler algorithms.

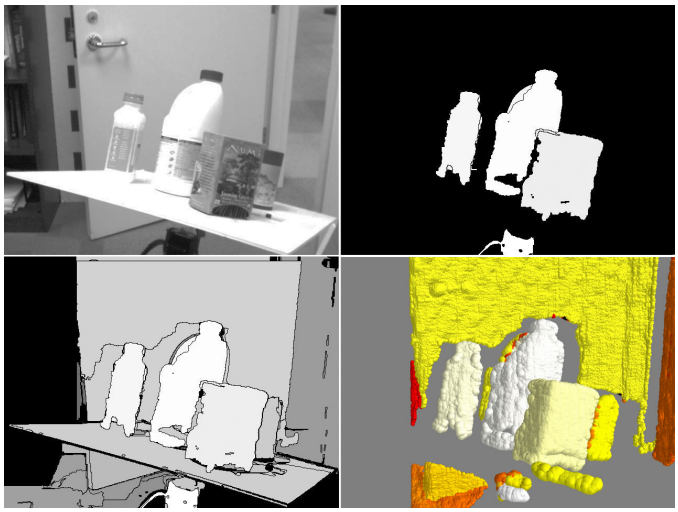

Fig. 1. Results of our algorithm for structure discovery applied to household items. (Top-left) Input image (range data is also an input, not shown). (Top-right) Highest ranked structures according to our algorithm. (Bottom-left) All structures found by our algorithm. (Bottom-right) 3D point cloud of all structures found by our algorithm, color-coded from white to red for easier visibility, being white the best score.There are 4 objects in the scene, 3 of which are detected as one of the highest ranked structures. The fourth object is detected correctly as a single entity but not one of the highest ranked due to being over 50% occluded.

Our main contribution is a structure discovery algorithm that exploits the availability of multi-modal data. We generate multiple segmentations [7], [8] of image and range data by varying the parameters of a standard segmentation algorithm (in our case, the Felzenszwalb-Huttenlocher graph-based segmentation [9]). While no single segmentation is completely correct, we hope that some segments in some of the segmentations are correct and contain a whole structured entity. We define a linkage step to relate segments in an image to the corresponding range measurements, and vice versa, to create multi-modal data regions. We term these multi-modal regions *regionlets*. We then compute region-wide features for each

regionlet, and aggregate them in a single energy function that measures the structure of each region. We show how simple features such as color consistency, continuity, alignment, and concavity work very well to identify potential structures.

As a key part of our approach, we develop a novel region-based image and range data fusion based on regionlets. Most of the literature focuses on low-level fusion with one-to-one correspondences to merge image and range data, e.g. [10]–[12]. We claim that by using regionlets instead, we can supersede low-level sensor fusion with a unified framework for one-to-one, one-to-many and many-to-many correspondences. Low-level fusion merges data at either the pixel or at the 3D point level. Pixel level fusion requires computing depth measurements for every pixel, that is, to generate depth images [10], [12]. If the range data source is a laser range finder, these algorithms often require super-resolution techniques [13] because images usually have higher resolution than their corresponding point clouds. Fusion at the 3D point level requires projecting the image out into the 3D point cloud to obtain colored 3D points, such as in [11]. The advantage of low-level fusion is that it is then easy to reuse existing image-based or point cloud-based algorithms to operate in the joint image-range data.

However, ease of use notwithstanding, individual pixels and 3D points have no meaning in isolation, and their union in a single colored 3D point does not convey any information about a scene either. It is when pixels are grouped together (a minimum of approximately 32×32 pixels for human object recognition [14]) that information can be extracted. We use this rationale to define regionlets as semantically equivalent regions in both image and range data. The smallest regionlet possible is a correspondence between one pixel and one 3D point, thus being equivalent to low-level sensor fusion and compatible with the existing literature. By using larger image and range data regions, we can compute more powerful features for the different data sources and merge information at a higher level. An additional advantage is that we can work with the native resolution of each data source; the data sources may have different resolutions and even non-linear densities (e.g. a rotating laser range finder) with no extra overhead. Related work in the mid-level sensor fusion literature includes feature-based approaches for a variety of sensors (e.g. [15], [16]) and region-based approaches for multiple images [17], but to the best of our knowledge no previous work has been published in region-based fusion of image and range data.

We apply our Structure Discovery algorithm to discover objects in indoor environments. The current state of the art in this area is to use range data, tuned to exploit domain knowledge of the geometry of the scene [18]. The use of multiple constraints simplifies the solution, but also limits its application to finding objects on top of a nearby table. Image-based object discovery techniques are more focused on larger objects and natural scenes [19], [20], and are outperformed by the range-based approach in indoor environments. In the experiments section, we show that our algorithm performs as well as the state of the art in finding objects on top of a table, without any limiting constraints about the scene. Algorithms optimized for particular scene geometries work poorly when their assumptions about the scene are violated (e.g. Fig. 1, in which the table is not visible enough). The greater generality of our algorithm is showcased in a second set of experiments, in which we discover objects in generic indoor scenes that the current state of the art is unable to process.

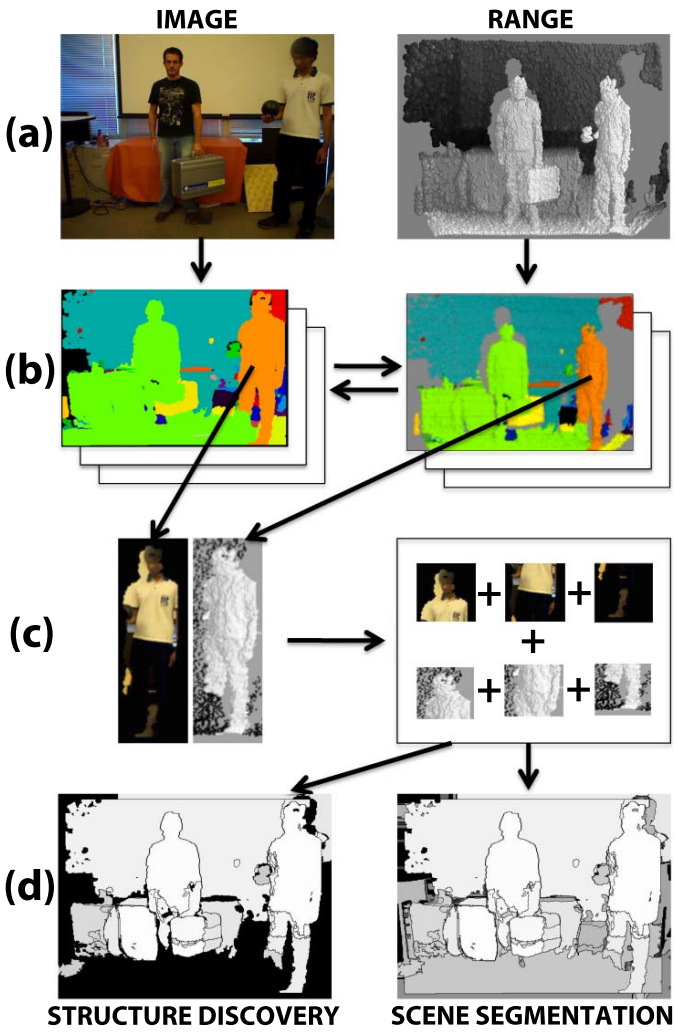

Fig. 2. Method overview. (a) Input data, image + range data. (b) Initial segmentations on each data source. Segmentations are projected to second data source. (c) Regionlets and sub-regionlets are generated from image, range data segmentations. Structure-ness score is computed from features on regionlets. (d) Output: structure discovery and scene segmentation.

II. Method overview

Given a single image and a single range scan of a scene, our goal is to automatically discover entities with structure, mainly in terms of appearance, shape smoothness and continuity. Our algorithm is summarized in Fig. 2. The result is a scene segmentation of the different structures that compose a scene, ranked from most to least structured.

We organize our method in four different steps, which we describe in detail in the following sections. Fig. 2.(a) shows an example of the input we receive from our sensors for a particular scene: an image, and a point cloud or depth image. Fig. 2.(b) shows example segmentations from each data source. Each of the segments in each of the segmentations is considered a hypothesis for a potential structure; our work,

therefore, is to analyze each of the segments and rank them in terms of their structure-ness, according to the features defined below.

Once the initial segmentations in each data source are generated, we need to compute a multi-modal region, a regionlet, from each segment in each segmentation. Segments from each data source are associated with data points from the other data source through a projection scheme. Each resulting regionlet contains a set of pixels and a set of 3D points, but no individual pixel to 3D point association is made. In order to rank the different hypotheses, we develop a hierarchical scheme: each regionlet is split into sub-regionlets to evaluate both local (at the sub-regionlet level) and global (at the regionlet level) consistency, in a multi-resolution grid similar to segmentation trees [21], [22]. We then calculate each regionlet score by evaluating image features, range features and mixed features at each level, and produce a ranking of regionlets based on their score (Fig. 2.(c)). Finally, we assign each pixel and 3D point to the highest-ranked regionlet that contains it, thus producing a segmentation of the scene based on structure-ness, as shown in Fig. 2.(d). If we wish to retrieve only the most structured regionlets, we only need to choose the few highest ranked regionlets.

III. Hypothesis generation

Generating likely hypotheses to evaluate their structure-ness is a hard problem. Ideally, we would like to try all possible segmentations from a scene and rank them, keeping only those with highest score. Unfortunately, such a procedure is infeasible. Following [7], [8], we generate a small number of segmentations with a well-known segmentation algorithm [9] as a representative sample of the set of all possible segmentations. Our assumption is that none of the segmentations are correct as a whole, but that *some* segments in *some* segmentations will be correct and contain entities with consistent structure. It is important to note that, since we do not rely on the full segmentation to be correct, the particular choice of a segmentation algorithm is not that critical. We choose [9] because it is fast and produces reasonable results. We perform this procedure independently for the image and range data. For the image segmentations, we use the difference between pixel colors in RGB space as a similarity measure, and generate multiple segmentations by progressively increasing the threshold k, typically from $k = 50$ to $k = 500$ in increments of 50. For the range data segmentations, we define similarity as the Euclidean distance between the eigenvalue-based features in [23], computed in a neighborhood of up to 20 nearest neighbors. To generate multiple segmentations, we progressively increase the threshold k from $k = 1$ to $k = 10$ in increments of 1.

IV. Generating Regionlets

One of the main contributions of our work is the mid-level fusion of image and range data. We use regionlets as our elementary processing units, which we create from image and range segments. Each regionlet

$$R_i = \{\mathbf{I}_i, \mathbf{p}_i, \mathbf{P}_i, \mathcal{A}(R_i)\} \quad (1)$$

is defined by a set of N image pixels $\mathbf{I}_i$ and their corresponding pixel positions $\mathbf{p}_i$, a set of M 3D points $\mathbf{P}_i$, and a set of adjacencies $\mathcal{A}(R_i)$. Hierarchies of regionlets are denoted by $R_i^{(k)}$ for regionlet R_i in the k-th level of the hierarchy. In this case, the adjacencies of a regionlet $R^{(k)}$ can be parents, neighbors and children of $R^{(k)}$ depending on whether they are from level $k-1$, k or $k+1$ in the hierarchy, respectively. In this work, we use a two-layer hierarchical scheme. For simplicity, we refer to the top-level regionlets $R^{(0)}$ as simply *regionlets* R, and the bottom-level regionlets $R^{(1)}$ as *sub-regionlets* r.

A. Associating 3D to image segments

Given an image segmentation, we must compute the 3D range measurements associated with each segment in the image. Assuming that camera and 3D sensor are calibrated, the 3D data are projected into the image and associated with the image segments into which they are projected.

B. Associating image pixels to 3D segments

To generate a regionlet given 3D segments, we must associate a set of pixels in the image with each 3D segment. We compute which image pixels correspond to a range data segment via Z-buffering [24]. We sort all range measurements according to their depth with respect to the camera. Starting with the points that are furthest away, each 3D point P_j paints a circle of pixels around its image projection $\hat{p}_j$ with the regionlet ID that corresponds to P_j. 3D points closer to the camera paint over 3D points further away. As a result, we obtain a set of image pixels associated with a range data segment. In order to account for differences in range data density, we adapt the radii of the circles painted according to the density of each region.

C. Generating sub-regionlets

In order to calculate a score for each regionlet and keep computations tractable, we separate each top-level regionlet into sub-regionlets, akin to the use of super-pixels from an image over-segmentation as elementary units of processing. In our work, we consider three choices to generate sub-regionlets: super-pixels (over-segmentation in image space), super-points (over-segmentation in 3D space), and fixed shape and size sub-regionlets. One important drawback of both super-pixels and super-points is that their shapes are often elongated and contain many twists and turns. This effect often results in very narrow sub-regionlets, with a width of only a few data points. Computing any kind of 3D features (e.g. surface normals) is unreliable in such data, so we choose instead the safer approach of sub-regionlets of non-overlapping shapes that maximize the visible surface area, in order to compute both 2D and 3D features reliably. In particular, we use squares of side $1/12$ of the vertical resolution of the image (e.g. 40×40 pixels in a 640×480 pixels image).

To generate sub-regionlets r_{ij}, we first compute a bounding box around a given regionlet R_i in the image domain. We then separate the bounding box in squares and compute the range measurements associated with them using the method described in Section IV-B. Sub-regionlets with less than a certain number of pixels and range measurements are considered insufficient and discarded.

TABLE I
UNARY AND PAIRWISE TERMS, AND APPLICATION TO REGIONLETS AND SUB-REGIONLETS.

	Unary	Pairwise	Regionlet	Sub-regionlet
Appearance Model	✓	✗	✓	✗
Shape Model	✓	✗	✓	✗
Self-Continuity	✓	✗	✓	✓
Contour Compactness	✓	✗	✓	✓
Pair-Continuity	✗	✓	✗	✓
Verticality	✓	✗	✓	✓
Concavity	✗	✓	✗	✓
Projection	✗	✓	✗	✓
Alignment	✗	✓	✗	✓
Color histogram	✗	✓	✗	✓
Surface compatibility	✗	✓	✗	✓

V. STRUCTURE DISCOVERY AS REGIONLET SCORING

Once the multiple segmentations are computed and regionlets generated, we assume we have populated our hypothesis space with all potential objects we want to discover. We need now to score all regionlets in terms of their structure-ness and rank them from best to worst. In this section, we develop a framework to evaluate the structure-ness of regionlets, and define the different features we use in each data modality.

We evaluate regionlets at both the global level, i.e. considering each region as a whole, and at the local level via sub-regionlets. Therefore, the energy function we define is comprised of unary terms at both the local and global level, and of pairwise terms at the local level.

Each regionlet R_i is composed of sub-regionlets r_{ij}. The set $\mathbf{r}$ of all sub-regionlets within a regionlet R is expressed as the *children* $\mathcal{C}$ of R, i.e., $\mathbf{r} = \mathcal{C}(R)$. Each sub-regionlet r_{ij} is characterized by a set of RGB values $\mathbf{I}_{ij}$, a set of 3D points $\mathbf{P}_{ij}$ and a set of neighbors $\mathbf{N}(r_{ij})$. In particular, each sub-regionlet r_{ij} is 4-connected to the adjacent sub-regionlets.

The final structure-ness score $\mathrm{St}(\cdot)$ for regionlet R is

$$\mathrm{St}(R;\Theta) = \phi(R;\Theta) \prod_{r\in\mathcal{C}(R)} \mathrm{St}(r;\Theta), \tag{2}$$

where

$$\mathrm{St}(r;\Theta) = \phi(r;\Theta) \prod_{r_j\in\mathbf{N}(r)} \Phi(r,r_i;\Theta). \tag{3}$$

The structure of this energy function is similar to a MRF [25], although for the task of region scoring we do not need to perform any inference on this function. It is important to mention that, for stability purposes, the log-linear version $\log(\mathrm{St}(\cdot))$ is preferred over the score $\mathrm{St}(\cdot)$.

The features we explain in this section are inspired from previous work from [6], [23], [26]–[28]. It is important to mention that, while the features we present work well in practice, they are just an example for the implementation of our structure discovery framework. Any other image-based, 3D-based, or mixed feature f to be computed in a group of data points such that $f(R_i) \in [0,1]$ is a potential feature to be used in our framework.

A. Unary terms

The unary terms of regionlets $\phi(R)$ and sub-regionlets $\phi(r)$ are computed as the interaction between the different features described in Table I. For a given set of features $\mathbf{f}$ and parameters Θ, the unary term

$$\phi(\cdot;\Theta) = \frac{1}{\sum_i w_i} \sum_{f_i\in\mathbf{f}} w_i f_i(\cdot;\Theta), \tag{4}$$

where the weights w_i account for the different importance of individual features when searching for particular types of structure. In order to be able to compare the different terms, we normalize all features (unless otherwise noted) to have a range $[0,1]$, where 0 is the worst and 1 the best score a feature can achieve.

1) Appearance model: An appearance model is used to search for structure with particular visual properties. The appearance model $\mathrm{App}(\cdot)$ can be any image-based likelihood function that returns the confidence value in the range $[0,1]$ of an image segment given model parameters Θ. For an extensive discussion and more details on appearance models for color and grayscale segmentation, see [28]. Examples of useful appearance models are:

- **Single color distribution**, to model non-textured, smooth regions.
- **Mixture of gaussians**, useful to model textured color regions.
- **Global color or intensity histograms**.

2) 3D shape model: A 3D shape model is defined equivalently to the image-based appearance model in Section V-A1. The shape model $\mathrm{Sh}(\cdot)$ specifies which types of structures are desired. Useful shape priors include:

- **Planar shape**. A planar approximation $\hat{\mathbf{P}}$ of a set of 3D points $\mathbf{P}$ can be easily computed via PCA analysis or RANSAC. Then $\mathrm{Sh}(R;\Theta) \in [0,1]$ is a confidence measure of how planar R is.
- **Size**. It is not uncommon for structures within the same scene to have vastly different scales, and we can specify surface or volume constraints on structures to prioritize which ones should be preferred.
- **Scale**. An alternative to a fixed size constraint is to define a size dependent on distance. This way, we can discover small structures near the camera, and larger structures further away from the camera.

3) Self-continuity: The self-continuity feature measures abrupt changes in depth within a regionlet or sub-regionlet, which are often representative of discontinuities and boundaries between objects. Finding discontinuities in unstructured point clouds is a hard problem and multiple algorithms have been developed for this purpose [29]. In our framework, we simplify this problem by accounting for the implicit ordering given by the image data. We construct a grid of control points in the image domain, and these control points are associated with the 3D points in the regionlet with minimal reprojection error.

In a sub-regionlet, the control points are equally spaced to form a grid of 64 control points. In a regionlet R_i, the control points are the centers of each sub-regionlet r_{ij}, and their connectivity maps that of the sub-regionlets. Once the Euclidean distances d_k between connected 3D control points have been computed, we define the continuity score $\mathrm{Cont}(\cdot;\Theta)$ as

$$\text{Cont}(\cdot;\Theta) = \exp\left(-\frac{1}{w_{cont}^2}\frac{1}{|\bar{P}|}\max_k d_k\right), \quad (5)$$

where $|\bar{P}|$ is the average distance from regionlet R to the camera, which is necessary in order to reliably compare regionlet scores from different depths.

4) Verticality: Structures facing the camera and range sensor, i.e., parallel to the image plane, are more desirable than structures almost perpendicular to the image plane, as it is less reliable to estimate shape and appearance parameters on structures under heavy projective distortion. In addition, it is more complicated to use discovered structures under heavy projective distortion in further tasks, such as object modeling or object recognition. Therefore, we implement verticality as a feature in our regionlet-scoring framework.

We compute the verticality score as a ratio between the area projected in the image and the maximum area spanned by the set of 3D points $\mathbf{P}$ of regionlet R. We approximate the computation of the area of a regionlet by the area of its bounding box along the directions of maximum variation.

5) Contour compactness: Object boundaries in the real world are usually smooth and contain few jagged edges. This fact has been used in the image segmentation and sensor fusion literature to produce smoother segmentations of objects (e.g. [26]). We use the same definition as [26] for a Contour Compactness $\text{CC}(\cdot)$ feature to encourage smooth edges. In particular, we measure the ratio of regionlet area to perimeter length, both in the image domain, normalized so that $\text{CC}(\cdot) \in [0,1]$.

B. Pairwise terms

Pairwise terms $\Phi(\cdot,\cdot)$ measure the interactions between neighboring sub-regionlets, in order to compute the likelihood that two sub-regionlets belong to the same structure.

For a given set of features $\mathbf{f}$ and parameters Θ, the pairwise term

$$\Phi(r_j, r_k; \Theta) = \frac{1}{\sum_i w_i}\sum_{f_i \in \mathbf{f}} w_i f_i(r_j, r_k; \Theta), \quad (6)$$

where the weights w_i account for the different importance of individual features when searching for particular types of structure. As with the Unary terms, we normalize all features (unless otherwise noted) to $\Phi(r_j, r_k; \Theta) \in [0,1]$.

1) Pairwise continuity: The pairwise continuity feature measures abrupt changes in depth and discontinuities between two sub-regionlets. We follow a similar approach to Section V-A3, and reuse the same set of control points $\mathbf{P}_i^c \in r_i$ and $\mathbf{P}_j^c \in r_j$. We compute the pairwise continuity score as the average between the M minimal distances between $\mathbf{P}_i^c$ and $\mathbf{P}_j^c$, normalized as in Eq. 5.

2) Concavity: Studies in human perception have shown that concavities are one of the major cues in the human visual system to segment a scene into parts [30], and have been used successfully in 3D segmentation and mesh decomposition [27]. In our work, we compute the concavity between two sub-regionlets as the difference in orientation α_{ij} between their surface normals pointing towards the camera.

TABLE II
OBJECT DISCOVERY AND NUMBER OF OBJECTS PER SCENE.

	Single	Single, Non-transp.	Multiple	Total
Structure Discovery	52.7%	76%	61.2%	59.9%
Rusu *et al* [18]	38.8%	56%	66.8%	62.5%

Given the importance of concavities in 3D segmentation, we can enforce a strong penalty on concave unions and reward convex unions by using

$$\text{Cv}(r_i, r_j) = \sin\alpha_{ij}. \quad (7)$$

In this case, $\text{Cv}(\cdot,\cdot) \in [-1,1]$.

3) Projection: The Projection feature captures information about the smoothness of a surface, by computing the projection error of a planar approximation of a sub-regionlet onto its neighbor. This way, surfaces with small variation score high, since both sub-regionlets have similar global properties, while different shapes and orientations achieve a low score. Let $\hat{\mathbf{P}}_i^j$ be the projection of the 3D points $\mathbf{P}_i \in r_i$ onto r_j. The projection score $\text{Proj}(\cdot,\cdot;\Theta)$ is then

$$\text{Proj}(r_i, r_j; \Theta) = \exp\left(-\frac{1}{w_{\text{proj}}}\frac{1}{N}\sum_{k=1}^{N}\|E_k\|_2\right) \quad (8)$$

$$\mathbf{E} = \min(\mathbf{P}_i - \hat{\mathbf{P}}_i^j, \mathbf{P}_j - \hat{\mathbf{P}}_j^i) \quad (9)$$

4) Alignment: The Alignment feature grades the depth alignment of parallel surfaces. Despite this fact being partially captured by the pairwise continuity feature, we enforce the alignment of surfaces with a more robust feature that operates on average range measurements from regionlets, and not individual distances between points. For this task we re-use the Projection score from Section V-B3, but we measure the projection error of the vector difference of means, i.e., $E = \mu_i - \mu_j$, where μ_i, μ_j are the average of the sets of 3D points $\mathbf{P}_i \in r_i, \mathbf{P}_j \in r_j$.

5) Color histogram: This image-only feature captures the similarity in terms of appearance between sub-regionlets. Following [6], we compute the distance between the color histogram (in $8\times8\times8$ bins) of each individual sub-regionlet compared to the histogram union of the two sub-regionlets, normalized so that $\text{ColorHist}(r_i, r_j; \Theta) \in [0,1]$.

6) Surface compatibility: The surface compatibility feature is based on the 3D features described in [23] of linear-ness l, planar-ness p and scatter-ness s, computed from the relative weights of the eigenvalues of the range data. We hypothesize that physical objects do not have abrupt changes in their surface properties. In other words, we assume that two planar surfaces are more likely to be parts of the same physical entity than a planar and a spherical surface, or that two curvy surfaces are more likely to be the same entity than a curvy surface and a plane. A simple way of encode this information is to compute the Euclidean distance between the two vectors $V_i = [l_i, p_i, s_i]$ and $V_j = [l_j, p_j, s_j]$ from r_i and r_j.

VI. EXPERIMENTS

Our goal in the experiments section is to compare our algorithm to the state of the art in object discovery in indoor scenes. We want to show that our algorithm performs as

Fig. 3. Examples of our structure discovery algorithm applied to household objects, in the Objects Pan-tilt Database. (Top row) Input images. (Bottom row) Top 10% ranked regionlets for each scene, color-coded from white to black, being white the highest ranked regionlet.

well as a specialized algorithm carefully optimized for the particular scene geometry of objects on top of a table. We also want to show that that our algorithm generalizes better to generic indoor scenes, because we do not enforce any limiting constraints on the scene geometry. To that end, the first set of experiments focuses on the discovery of common household objects on top of a table, while the second one focuses on the discovery of larger structures such as people, tables or walls. Each dataset has been gathered with a different source of depth information (Projected Textured Stereo [31], and an RGBD camera) to test the performance of our algorithm with different data sources.

A. Our implementation

In our implementation of the Structure Discovery algorithm, we use the following constants:

- Appearance/shape model: we use a simple maximality prior instead of a full appearance/shape model, in which we encourage the creation of large regionlets. In particular, we use $App(R;\Theta) = \log(|\mathbf{I}|)$, where $|\mathbf{I}|$ is the total number of pixels of regionlet R. We use a logarithm to avoid this feature from overpowering all others.
- Normalization weights: all weights used for normalization purposes, i.e. inside the exponentials, are set according to the uncertainty/noise characteristics of each sensor. For the stereo data, these are set so an average error of 1.5 cm at 1 m outputs a feature score of 0.5. The RGBD camera has an average uncertainty of 3%, so we set the weights so that an average error of 3% outputs a feature score of 0.5. This normalization scheme is described in more detail in [4].
- Feature weights: we do not commit to any feature being more powerful than others, so all weights are set to 1.

All our parameters are kept constant throughout our experiments to demonstrate the generality of our algorithm in discovering structure.

B. Baseline algorithm

The baseline algorithm we compare against is the 3D object segmentation algorithm from Radu Rusu's Point Cloud Library [18] (Willow Garage). This algorithm exploits domain knowledge of the geometry of common household environments; it is specialized in finding objects on planar surfaces, by first performing a plane-fitting procedure and then clustering groups of 3D points that lie on top of the plane. It is a simple method that performs remarkably well as long as the plane-fitting procedure is able to find valid planes, and it has been showcased on multiple occasions in most demonstrations of the PR2 Personal Robot.

C. Results

In this first experiment, we use a subset of 90 scenes extracted from Willow Garage's "Objects Pan-tilt Database" [32] (some examples shown in Fig. 3). This dataset contains different sets of household objects attached to a pan-tilt unit that is moved around, so that multiple views of the objects are available. Images from this dataset are grayscale, with a resolution of 640×480 pixels. The depth information is computed using Projected Textured Stereo [31], a technique that extracts dense depth maps from images through the projection of a structured texture on the scene. The objects used in our evaluation have different shapes and appearances, and include a Gillette Shaving Cream, a Kleenex Cube, Mop'n'Glow floor cleaner, a soda can and a milk carton, among others. Some scenes contain multiple objects (up to 5) in different levels of occlusion and some contain single objects, for a total of 223 object instances in 90 scenes. Some examples of scenes and our segmentations are shown in Fig. 3.

In our evaluation, we ground truth each scene with a bounding box around each object, and use the PASCAL criteria [33] bounding box evaluation to identify which objects are correctly discovered.

Our results are shown in Table II. Both our algorithm (Structure Discovery, in Table II) and the baseline from Rusu *et al* [18] perform similarly well, despite using vastly different approaches. It is interesting to note that our algorithm tends to

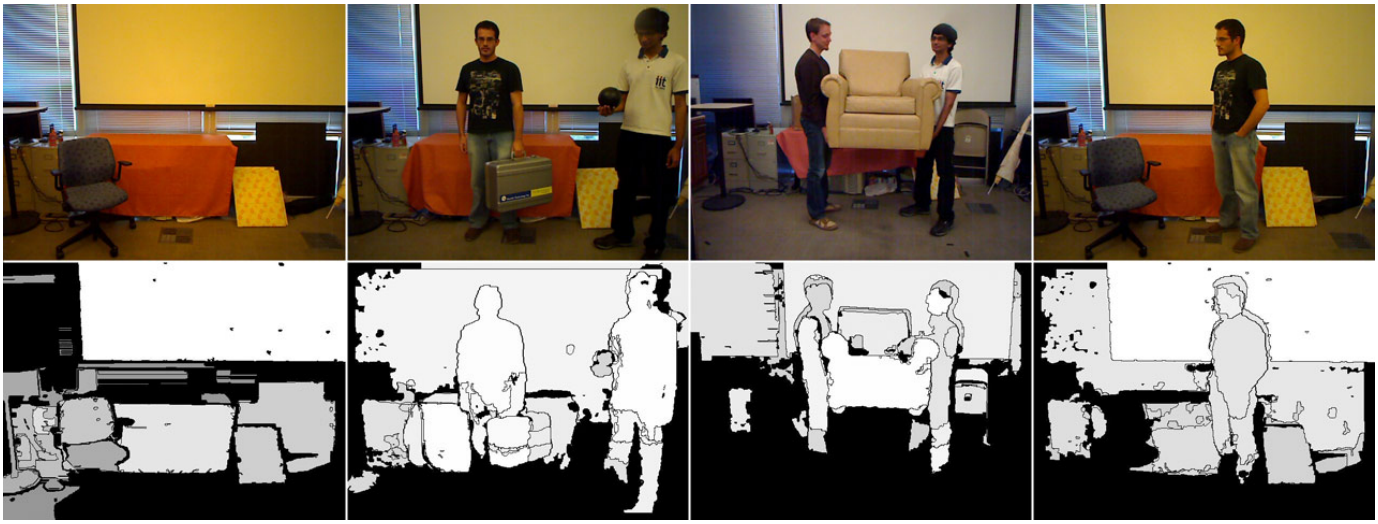

Fig. 4. Examples of our structure discovery algorithm applied to household scenes, in the Household Dataset. (Top row) Input images. (Bottom row) Top 10% ranked regionlets for each scene.

discover objects more reliably in scenes with little clutter. The explanation for this tendency is that our algorithm is designed to detect prominent structures in the scene, which is very often correlated to the size of the structures in the image. In some of the highly cluttered scenes, the objects seldom span more than a single sub-regionlet, and contain little 3D information. Under these circumstances, the algorithm from Rusu *et al* is a more convenient choice.

For the single object experiments, we split our results between "Single" and "Single, non-transparent" because neither algorithm is able to discover a transparent wine glass in the dataset, mainly because of a lack of consistent range data in its surface, as the Projected Texture Stereo algorithm fails to recover stereo data from it. In the single object experiments, our algorithm outperforms the baseline by 20%. This difference, however, is a bit misleading, since both algorithms perform equally well on all objects but one (the Mop'n'Glow bottle). In a general setting, we believe that these two algorithms perform very similarly, as long as there is a planar surface for the baseline algorithm to detect.

The second experiment we conduct is the discovery of larger entities in indoor environments, such as walls, furniture and people. In order to do so, we gathered a dataset of 15 indoor scenes, which we call the Household Dataset. The image/range sensor is an RGBD camera that outputs 640×480 resolution images with associated depth for every pixel. We downsample the range data to one third of the original resolution to verify that we do not require a pixel-level fusion of data sources. We have annotated and produced bounding boxes for four types of structure: wall, person, furniture, and "other". The label "other" is for other prominent objects in the scene, such as a backpack, a painting, a cardboard box or a suitcase.

For this experiment we are unfortunately unable to provide any results from the baseline system, as it is optimized for a much shorter range than these scenes, and does not return any detections in these scenarios.

Results from this dataset are shown on Table III, and example scenes and their highest ranked regionlets are shown in Fig. 4. Analyzing the results, we see that large, simple structures such as walls are discovered very reliably, while more complex entities such as people are often missed, in particular their legs. If we focus on the upper body of a person, we find a twofold increase in performance, as a person's upper body is often larger and has less variability.

TABLE III
STRUCTURE DISCOVERY ON INDOORS SCENES. RESULTS FROM RUSU *et al.* SHOW NO DETECTIONS AND ARE OMITTED FROM THE TABLE.

	Walls	Person	Person (torso)	Furniture	Other
Total Present	12	14	14	11	13
Found (%)	100%	35.7%	78.5%	63.6%	61.5%

VII. DISCUSSION

We have presented and validated an algorithm to perform structure discovery from multi-modal data using a novel region-based approach. We have demonstrated that our algorithm is able to discover common household objects with similar accuracy than specialized 3D object segmentation algorithms, without the need to rely on any rigid assumptions about the scene structure, such as the presence of a visible planar surface in which objects are placed.

Interestingly, the results from both algorithms are almost complementary in the kind of scenes they perform best. Our algorithm discovers objects best when the objects' largest faces are parallel to the image plane, as the image and range sensors capture more information about them. On the other hand, [18] performs best when the table is the most prominent part of the scene, e.g. seen from a high viewpoint. An interesting follow-up work to this algorithm would be a higher-level reasoning about the interpretation of a scene; in the case of an overhead picture, the detection of a planar surface could lead to a closer inspection of entities on top of it.

A close analysis on the limitations of our algorithm shows some important conclusions and areas of improvement. We have found that a bad performance of our algorithm is often tied to the bad performance of the initial segmentations. On multiple occasions where an entity is missed, the reason is that none of the initial segmentations was able to capture that

entity in its entirety in a single segment. When this happens, our assumptions do not hold and thus our algorithm cannot generate the expected results. In addition, while we discover structures such as walls with high accuracy, our structure model is sometimes not flexible enough to handle the large variations of more complex structures such as people or some furniture, as they are seldom segmented in their entirety. A possible solution to this issue would be the addition of a third regionlet layer on top of regionlets and sub-regionlets, and compute these as a combination of multiple regionlets via e.g. split and merge techniques [28].

We are confident that the introduction of regionlets as the minimal processing units is an important step in image-range data sensor fusion. Regionlets supersede and integrate low- and mid-level fusion in one framework, extracting the advantages from both approaches. Regionlets are compatible with existing algorithms that require one-to-one correspondences, while adding an extra layer of abstraction that may lead to more sophisticated perception tasks using mixed image and range data.

We believe that this structure discovery algorithm opens many interesting possibilities for future work. A higher level scene interpretation that reasons about the relationships between discovered structures could lead to massive improvements in the quality and reliability of the objects we discover, and enable the unsupervised learning of object models for object/category recognition. A related and also interesting direction to follow is the reutilization of our algorithm as a category recognition algorithm from multi-modal data. The exploration of different shape and appearance models, coupled with a per-class feature selection scheme may lead to reliable category recognition for robotics.

VIII. Acknowledgments

This material is based upon work partially supported by the National Science Foundation under Grant No. EEC-0540865. Alvaro Collet is partially supported by a Caja Madrid fellowship. Special thanks to Christopher G. Atkeson for his insightful comments and discussions.

References

[1] S. S. Srinivasa, D. Ferguson, C. J. Helfrich, D. Berenson, A. Collet, R. Diankov, G. Gallagher, G. Hollinger, J. Kuffner, and M. V. Weghe, "HERB: a home exploring robotic butler," *Autonomous Robots*, vol. 28, no. 1, pp. 5–20, 2010.

[2] WillowGarage, "The Personal Robot Project," 2008. [Online]. Available: http://www.willowgarage.com

[3] H. Nguyen, C. Anderson, A. Trevor, A. Jain, Z. Xu, and C. Kemp, "El-E: An Assistive Robot that Fetches Objects from Flat Surfaces," in *IEEE Proceedings of Human Robot Interaction, The Robotics Helpers Workshop*. IEEE, 2008.

[4] A. Collet, M. Martinez, and S. S. Srinivasa, "The MOPED framework: Object Recognition and Pose Estimation for Manipulation," *To appear in The International Journal of Robotics Research*, 2011.

[5] M. Wertheimer, *Laws of organization in perceptual forms*. Harcourt, Brace and Company, 1938, pp. 71–88.

[6] D. Hoiem, A. N. Stein, A. A. Efros, and M. Hebert, "Recovering Occlusion Boundaries from a Single Image," in *IEEE ICCV*. IEEE, Oct. 2007, pp. 1–8.

[7] D. Hoiem, A. Efros, and M. Hebert, "Geometric Context from a Single Image," in *IEEE ICCV*. IEEE, 2005, pp. 654–661.

[8] J. Sivic, B. Russell, A. Efros, A. Zisserman, and W. Freeman, "Discovering Objects and their Localization in Images," in *IEEE ICCV*. IEEE, 2005, pp. 370–377.

[9] P. Felzenszwalb and D. Huttenlocher, "Efficient Graph-Based Image Segmentation," *International Journal of Computer Vision*, vol. 59, no. 2, 2004.

[10] M. Bjorkman and D. Kragic, "Active 3D scene segmentation and detection of unknown objects," in *IEEE ICRA*. IEEE, 2010, pp. 3114–3120.

[11] I. Posner, D. Schroeter, and P. Newman, "Online generation of scene descriptions in urban environments," *Robotics and Autonomous Systems*, vol. 56, no. 11, pp. 901–914, 2008.

[12] S. Gould, P. Baumstarck, M. Quigley, A. Y. Ng, and D. Koller, "Integrating Visual and Range Data for Robotic Object Detection," in *ECCV Workshop on Multi-camera and Multi-modal Sensor Fusion Algorithms and Applications*, 2008.

[13] J. Diebel and S. Thrun, "An application of markov random fields to range sensing," in *NIPS*, vol. 18, 2005, p. 291.

[14] A. Torralba, R. Fergus, and W. T. Freeman, "80 Million Tiny Images: a Large Data Set for Nonparametric Object and Scene Recognition." *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 30, no. 11, pp. 1958–1970, Nov. 2008.

[15] A. Gunatilaka and B. Baertlein, "Feature-level and decision-level fusion of noncoincidently sampled sensors for land mine detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 23, no. 6, pp. 577–589, Jun. 2001.

[16] V. Sharma and J. Davis, "Feature-level Fusion for Object Segmentation using Mutual Information," in *IEEE CVPR Workshops*. IEEE, 2006, pp. 139–147.

[17] J. J. Lewis, R. J. O'Callaghan, S. G. Nikolov, D. R. Bull, and C. N. Canagarajah, "Pixel- and region-based image fusion with complex wavelets," *Information Fusion*, vol. 8, no. 2, pp. 119–130, Apr. 2007.

[18] R. B. Rusu, "Point Cloud Library," 2010. [Online]. Available: http://www.ros.org/wiki/pcl

[19] I. Endres and D. Hoiem, "Category Independent Object Proposals," *ECCV*, 2010.

[20] B. Alexe, T. Deselaers, and V. Ferrari, "What is an object?" in *IEEE CVPR*. IEEE, Jun. 2010, pp. 73–80.

[21] E. Borenstein, E. Sharon, and S. Ullman, "Combining Top-Down and Bottom-Up Segmentation," in *IEEE CVPR Workshops*. IEEE, 2004.

[22] E. Sharon, A. Brandt, and R. Basri, "Segmentation and boundary detection using multiscale intensity measurements," in *IEEE CVPR*. IEEE, 2001, pp. 469–476.

[23] J.-F. Lalonde, N. Vandapel, D. F. Huber, and M. Hebert, "Natural terrain classification using three-dimensional ladar data for ground robot mobility," *Journal of Field Robotics*, vol. 23, no. 10, pp. 839–861, Oct. 2006.

[24] M. Wand, M. Fischer, I. Peter, F. Meyer, and W. Straser, "The randomized z-buffer algorithm: interactive rendering of highly complex scenes," in *SIGGRAPH*, 2001, pp. 361–370.

[25] F. Huang and Y. Ogata, "Generalized Pseudo-Likelihood Estimates for Markov Random Fields on Lattice," *Annals of the Institute of Statistical Mathematics*, vol. 54, no. 1, pp. 1–18, Mar. 2002.

[26] C. C. Chu and J. K. Aggarwal, "The Integration of Image Segmentation Maps using Region and Edge Information," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 15, no. 12, p. 1241, 1993.

[27] S. Katz and A. Tal, "Hierarchical mesh decomposition using fuzzy clustering and cuts," *ACM Transactions on Graphics*, vol. 22, no. 3, p. 954, Jul. 2003.

[28] Z. Tu and S.-C. Zhu, "Image Segmentation by Data-Driven Markov Chain Monte Carlo," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 5, pp. 657–673, May 2002.

[29] P. Tang, D. Huber, and B. Akinci, "A Comparative Analysis of Depth-Discontinuity and Mixed-Pixel Detection Algorithms," in *International Conference on 3-D Digital Imaging and Modeling*. Ieee, Aug. 2007, pp. 29–38.

[30] D. Hoffman and W. Richards, "Parts of recognition," *Cognition*, vol. 18, no. 1-3, pp. 65–96, Dec. 1984.

[31] K. Konolige, "Projected Texture Stereo," in *IEEE ICRA*. Anchorage, AK: IEEE, 2010, pp. 148–155.

[32] WillowGarage, "Objects Pan-tilt Database," 2010. [Online]. Available: http://vault.willowgarage.com/wgdata1/vol1/objects_pantilt_database/

[33] M. Everingham, L. Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The Pascal Visual Object Classes (VOC) Challenge," *International Journal of Computer Vision*, vol. 88, no. 2, pp. 303–338, 2010.