16th International Conference on Machine Vision Applications (MVA)
National Olympics Memorial Youth Center, Tokyo, Japan, May 27-31, 2019.

A Hierarchical Segmentation Approach with Convolution-Recursive Deep Learning for 3D Multi-Object Recognition under Partial Occlusion Conditions

Soma Boubou[1,2*] Tatsuo Narikiyo[1] Michihiro Kawanishi[1]
1. Toyota Technological Institute
2-12-1 Hisakata, Tenpaku, Nagoya 468-8511, JAPAN
boubou,n-tatsuo,kawa@toyota-ti.ac.jp

2. Omron Corporation
9-1 Kizugawadai, Kizugawa-City, Kyoto 619-0283, JAPAN
soma.boubou@omron.com

Abstract

Depth data based object recognition has recently emerged as a challenging research topic. In this work, we develop a novel approach to perform detection and recognition of occluded 3D objects. We propose a hierarchical segmentation algorithm in order to obtain the homogeneous sub-regions contained in each depth frame which in turn facilitates the recognition under severe occlusion conditions. Our proposal consists of three steps: the first step is to build a tree structure contains all key sub-surfaces visible in the depth frame with their intra-hierarchical relations. Thereafter, we draw a classification prediction for all nodes based on a combination of convolution and recursive neural networks. Finally, we employ the hierarchy scheme to refine the classification results. Our proposal obtained competitive results and proved to be invariant to objects scale, rotation, and viewpoint variations.

1 Introduction

Recent computer vision sensing technology facilitates acquiring 3D geometric data of the surrounding environment. Depth data acquisition with Kinect and Structure sensor become popular among researchers in computer vision domain. Thus depth data based researches tackle a wide range of computer vision topics including human action recognition [1], simultaneous localization and mapping (SLAM) [2] and object recognition [3, 4, 5, 6].

Due to the noisy nature of depth data, object recognition with depth data is considered as one of the challenging topics in computer vision. Improving the recognition performance with depth data is essential for the development of autonomous robots. By exploiting depth data for vision applications, the recognition system is provided with a useful source of extra information stored in the depth modality in order to solve the complex problem of general environment recognition. Depth data is invariant to illumination and color variations. Moreover, it provides geometric cues and allows simple straight forward foreground segmentation. However, although the-state-of-the-art research works show a promising recognition achievement with depth data, it is still far away from being perfect. If we look carefully to the object recognition achievements with RGB-D sensors, RGB-Depth split-out results presented in the literature shows that the main recognition achievement comes from the RGB data while the recognition with depth data is still a challenging task [7, 8]. In this work, we aim to enhance the recognition performance from depth-only data. This will facilitate object recognition with depth-only sensors such as Structure sensor by Occipital, Inc.

In case of occlusion, objects will be partially visible to the depth sensor. Hence, there is a chance that some/all local features of the targeted object are invisible which will cause a degradation in the recognition performance. To encounter this challenge, our approach is based on the fact that the 3D surface of an object is an aggregation of a set of contiguous key homogeneous sub-surfaces. Therefore, the recognition process of occluded objects is achieved through the detection and the classification of visible key sub-surfaces.

In this paper, we introduce a hierarchical approach in order to divide the 3D object surface into a set of homogeneous key sub-surfaces. Moreover, we employ convolutional-recursive deep learning model proposed by Socher et al., [9] to classify the key sub-surfaces obtained from raw depth images in the first step. Finally, We exploit the hierarchical relations obtained in the first step to perform a refinement of the classification results. Compared to other state-of-the-art 3D feature learning methods [4, 5, 10, 11], our approach is invariant to scale and viewpoint variation and could perform successful recognition under severe occlusion conditions.

The rest of the paper is organized as follows: related works are presented in the following section. Thereafter, we present the details of our approach in Section 3. The implementation and experimental results are presented in Section 4. Finally, we conclude in Section 5.

2 Related work

There are various limitations within existing techniques for 3D object recognition in terms of efficiency, robustness, clutter, occlusion and the discrimination capability of the feature representation. In this section, we will start by presenting a small taxonomy of the popular approaches which are pursued to perform the task of 3D objects recognition under occlusion.

There are two approaches for partially occluded object recognition. The first one is based on local features recognition such as SHOT [12] or TriSI [13] and, the second approach, based on frame segmentation, is more popular in object recognition from 2D images such as superpixel [14].

*At the time of conducting the research, Soma Boubou was a post-doctoral researcher at Toyota Technological Institute. Currently, he serves as a permanent expert researcher at Omron Corporation.

© 2019 MVA Organization

On the other hand, 3D objects segmentations is classified into two types, namely surface-type [15, 16] and part-type [17] segmentations. In this work, we follow a surface-type segmentation approach. In order to cluster nonplanar regions in surface-type based segmentation, cluster representatives and discriminatory criteria are essential. Thus, primitives surfaces such as spheres, cylinders and, cones are employed by several works in order to find the best possible fitting primitive in least squares sense. A more straightforward approach to cluster non-planar surfaces is simply to measure the differences in normal direction or, in dihedral angles between mesh elements or depth frame pixels [3, 5]. Depending on the tolerance of this difference, planar and several types curved parts can be segmented.

Segmentation techniques are classified into five types, single or multiple source region grow [18], hierarchical clustering [19], iterative clustering [20], and spectral analysis [21]. Since the number of regions depends heavily on the choice of initial seeds, searching for the local optimum of each region separately in region grow techniques may in some cases create unsatisfactory global results. Furthermore, there are times when a hierarchical segmentation structure is beneficial for specific applications. Hierarchical clustering, while still a greedy approach, can be seen as global-greedy because it always chooses the best merging operation for all clusters and does not concentrates on growing one.

Similar to region-growing, the difference between various hierarchical clustering algorithms lies mainly in the merging criteria and the priority of elements in the queue. According to the type of implicit approach employed for segmentation, it can be classified into three types: Boundaries construction, Top-down approach, and Inferring approach. While previous approaches produce an unknown number of clusters, such number is given as a priori to iterative clustering approach. The approach iteratively employs k-means algorithm to assign all segments into the given number of clusters defined by its pre-known representatives. There are several issues concerning iterative clustering such as convergence and choice of initial representatives.

Finally, spectral graph theory demonstrates the relationship between the combinational characteristics of a graph and the algebraic properties of its Laplacian. Due to its high computational cost, the surface must be pre-segmented into smaller balanced-in-size sub-surfaces each of which must be treated separately. Moreover, edges must be minimized in order to reduce the visual effects.

On the other hand, there are several approaches employed to achieve recognition with 3D data. Recognition based on HOT curves such as the one proposed by Joshi *et al.* [22] relies on the accurate localization of inflection points, which in turn are sensitive to noise. A novel recognition algorithm based on spin image representation is proposed by Johnson and Hebert [23]. However, the proposed representation is sensitive to the resolution and sampling of the models. Spin images map a 3D surface into a 2D histogram [24] with a low discrimination capability which leads to many ambiguous matches. Carmichael *et al.* [25] proposed an improvement to overcome the sensitivity of the spin images regarding resolution. However, problems such as of the low discrimination capability of the feature representation and the inefficiency of the algorithm remain unsolved.

Socher et al., [9] proposed a recognition algorithm for RGB-D data based on convolution-recursive deep neural network (CNN-RNN) model. Using a CNN with multi-fixed RNN tree structure. Socher et al., demonstrated that RNNs with random weights can produce high-quality features and the model performance could be improved by increasing the number of features. Authors in [9] used a different tree structure for each input and trained the RNNs with back-propagation through a structure presented in [26]. More recently, Boubou et al. [4] proposed a recognition system based on extreme learning machines with a local receptive field. The approach demonstrates a competitive recognition performance with a short computational time. The proposed approach in [4] is proved to more robust regarding objects rotation and viewpoint variation.

In this work, we present a novel hierarchical segmentation approach in order to extract key sub-surfaces and build a robust recognition system targeting occluded objects based on convolution-recursive deep learning approach. Thereafter, our approach finally refines the classification results based on the hierarchy structure knowledge built in the first step. Our experiments show that our proposed approach performance overcome the one shown by CNN-RNN approach and ELM-LRF.

3 METHODOLOGY

3.1 Segmentation

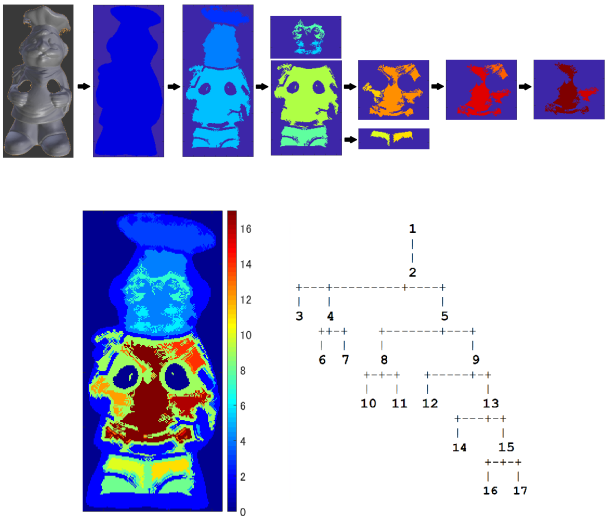
Briefly, our hierarchical clustering proposal is a subsurface-type based top-down approach aiming to construct a data tree from the depth input frame. Starting from a root which represents the input depth frame, a which is subject to segmentation process in order to partition the root into two (or more) nodes. This process continues iteratively for each of the tree nodes until a certain desired node size threshold is met. Each partitioning process is achieved by finding the best boundary fit between two sub-surfaces. Algorithm 1 explicates tree construction mechanism.

Algorithm 1 Hierarchical segmentation

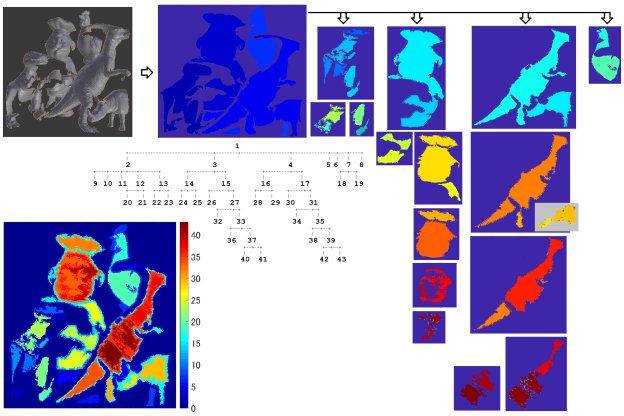
Require: train/test depth frame

1: Create a root set $S_{l=1}$ including all segmented sub-surfaces ▷ Depth based threshold is used
2: Insert $S_{l=1}$ to a priority queue $Q_{l=1}$
3:
4: **while** tree leaves can be split **do**
5: **for** `size of` Q_l **do**
6: Get the top set S_{li} from Q_l
7:
8: **if** S_{li} can be split & $A_{li} > A_{min}$ **then**
9: Split S_{li} into a set Sj ▷ A gradual threshold's linear relaxation is implemented
10: Insert all Sj to Q_{l+1}
11: Move to the next level $l+1$

In details, let us consider a raw depth frame which is represented as the tree root with a single element set $S_l = \{N_{ind}\}$ where $l, ind \in [0, \infty[$ are natural numbers represents the level of the corresponding sets and, the index of the tree node N_{ind} respectively. A depth based threshold is used (c.f., algorithm 1 line 1), in order to segment the background and spilt the rest of the tree root frame into an ordered set of j objects or sub-objects $Q_l = \{N_{ind}\}$ with corresponding surface sizes $A_{li} > A_{min}$, that is $l = 1$, $ind \in [2, j+1]$ and j represents the number of children nodes. Each element of Q_l will be tested for possible splitting.



(a) A training frame that contains one training view of a 3D object sample. Since the training frame has one object, root has only one child representing the training object.



(b) Since the testing frame contains multi-objects, root has several children representing the extracted objects/sub-objects

Figure 1. Illustrates two tree structures obtained from 3D depth data using our proposed hierarchical segmentation for training and testing frames. Seeking to simplify the tree presentation, form now on we will represent segmentation tree structure with a single RGB image (left-bottom corner) where each color represents one node according to its index number.

A dihedral angles based threshold (c.f., algorithm 1 line 9) is gradually linearly relaxed from a maximum value until the realization of node split. Resulting children will be added to a priority set Q_{l+1}. All leaves will be iteratively split until the new split surface size does not satisfy the minimum split size A_{min} . Figure 1 illustrates two tree structures obtained from 3D data using our proposed hierarchical segmentation for training and testing frames. Seeking for presentation simplification, final trees are presented as one RGB figure where each node is presented with unique color according to the node self-index number.

3.2 Recognition Model

In this step, we consider each node of the segmentation tree as an isolated train/test sample. In other words, instead of relying on frames as training samples, our train set is a collection of all nodes extracted from training frames. Thereafter, we apply the recognition model on each node of the test frame in order to draw a classification prediction and a prediction confidence for each node separately.

In order to learn CNN-RNN model, we follow the procedure described by Socher et al., [9]. Each modality is given to a single convolutional neural network layer (CNN) [27] which provides a useful translational invariance of low level features such as edges. Moreover, CNN allows parts of an object to be deformable to some extent. First, random patches are extracted from depth data and a k-means algorithm is used to cluster those pre-processed patches. The single layer CNN which consists of a convolution layer, a rectification and local contrast normalization (LCN), is used to convolute each tree node which is previously resized to a square dimension d_i pixels with K square filters with size d_p. K filter responses with size $d_i - d_p + 1$ are average pooled with a pooling filter of size d_l and a stride of size s. The final output of CNN layer regarding one input node will be a 3D matrix $X \in \mathbb{R}^{K \times r \times r}$, where $r = (d_i - d_l)/s + 1$ is the size of pooling response.

Pooled filter responses $X \in \mathbb{R}^{K \times r \times r}$ are given to a recursive neural network (RNN) [28] which is designed to learn hierarchical features. Here, RNNs project the inputs into a lower dimensional space through multiple layers with tied weights and non-linearities. Thereafter, we merge adjacent $b \times b$ block of k-dimensional columns of X 3D input matrix according to equation 1, where $W \in \mathbb{R}^{K \times b^2 K}$ is a random weights matrix and, f is a non-linearity function.

$$P = f\left(W \begin{bmatrix} x_1 \\ \vdots \\ x_{b^2} \end{bmatrix}\right) \tag{1}$$

Applying equation 1 to all $b \times b$ blocks of X will produce a new matrix with a size equal to $K \times (r/b) \times (r/b)$. The procedure is iteratively applied until reaching a single column with K dimension. The classification model shows that multi-RNNs are more effective compared with a single RNN structure. In order to employ multi-RNNs, the system generates n_R random weighting matrices $\{W_i\}$ equal to the number of RNNs. As a result, each input depth image will generate n_R K-dimensional vectors which are then given to softmax classifier.

3.3 Classification Refinement

Depends on the minimum sub-surface size A_{min} threshold which we set to be 500 pixels, a large number of tree nodes represent a small sub-surfaces area down to 25×20 pixels which is a reason for misclassification due to insufficient features found in such small areas. Now, increasing the minimum sub-surface size A_{min} is actually a straightforward solution. Unfortunately, it has a major drawback that is in reality due to occlusion and other factors related to depth sensors capability, test frames often contain a plenty of small sub-objects some of which are 150 pixels size or less. Increasing A_{min} threshold will result in ignoring such small visible sub-parts. Hence, to keep A_{min} as small as possible, we propose a mechanism to refine the classification results based on the classification results produced by CNN-RNN classification model. In details, we proposed a top-down approach to correct classification errors of the tree nodes.

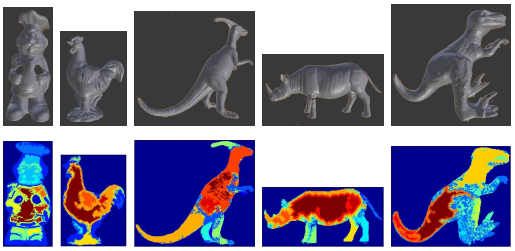
Let us consider one node N_{ind} of level l having a j children nodes N_i that is $i \in [1, j]$. Thus, we generate a temporary set S_{temp} consists of N_{ind} node with its children. S_{temp} has $j+1$ number of nodes, each of which has the size of A_i and classified as class cl_i with c_i confidence. Predicted class cl_{pred} of all nodes in S_{temp} is given by equation 2, where $class_q$ is a set of all S_{temp} elements classified as class q. We

implement this procedure on all tree nodes in a top-down scheme.

$$cl_{pred} = argmax\{\forall\ class_q |\ cl_{score} = \frac{\sum A_i.c_i}{\sum |class_q|}\} \qquad (2)$$

Experimental results presented in Section 4 shows that our proposed classification refinement significantly enhanced the recognition performance compared with pure CNN-RNN classifier.

4 EXPERIMENTS



(a) (b) (c) (d) (e)

Figure 2. Sample views for the 3D models of the training dataset (first row) and their corresponding sub-surfaces segmentation images (second row) where each color represents a single node of the segmentation tree.

We compared the performance of our proposed approach with the original CNN-RNN approach and two other recognition approaches using a 3D objects models acquired with a laser beam scanner by Mian et. al., [29]. The dataset includes five training 3D objects captured with a resolution up to 640×480 (c.f., Figure 2). We generate a 36 depth frames for each object, each of which is generated from a unique viewpoint distributed evenly on a horizontal circle around the targeted object. The test data includes 50 scenes generated randomly by placing four or five of the objects together in a scene. Finally, objects in each scene were scanned from a single viewpoint and converted to a depth test frame. The implementation of our proposed hierarchical segmentation on train/test depth frames with $A_{min} = 500$ results in total 1924, 1804 training and testing key-subsurface respectively. In other words, in the case of $A_{min} = 500$, a dataset contains 3728 key-subsurfacewas exploited for our experiments.

During experiments, we realized that the minimum sub-surface size threshold A_{min} plays an important rule in recognition performance. Choosing a small threshold will create a large number of sub-surfaces which lack of discriminating features, hence it is easier to be misclassified by recognition methods. In Figure 3, nodes recognition rate represents the rate of correctly classified tree nodes to the total number of nodes. On the other hand, area-based recognition rate represents the rate of correctly classified area size to the total area size of all nodes. A bigger difference between those two measures illustrates that a higher percentage of smaller nodes are misclassified. Since increasing A_{min} allows to generate a fewer number of small subsurface, it helps to enhance the final recognition rate.

In order to choose classifier parameters configuration such as number of CNN filters n_F and number of RNNs n_R, we tested our approach with $n_F \in \{32, 64, ..., 256\}$, $n_R \in \{16, 32, ..., 128\}$ (c.f., Figure 4). For the results presented in

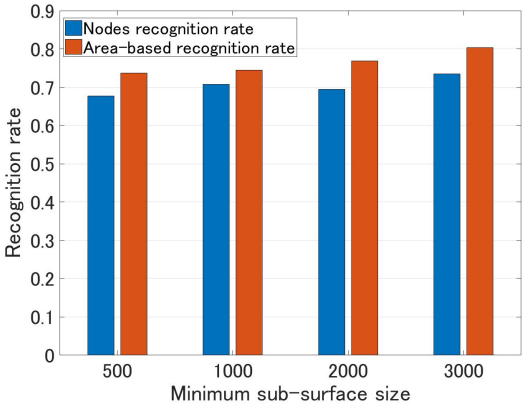


Figure 3. Recognition rate for our proposed approach versus minimum sub-surface size. Increasing the threshold will allow fewer number of small sub-surface to be generated, that explains the enhancement of the recognition rate.

rest of this paper we adopt the following experimental parameters: sub-surface minimum $A_{min} = 500$, number of CNN filters $n_F = 128$, number of RNNs $n_R = 32$ and, number of initial patches $n_P = 2000$.

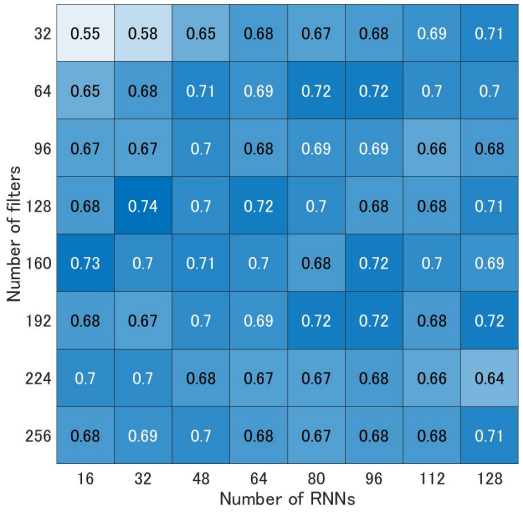


Figure 4. Area-based recognition rate of our proposed method changes depends on the chosen values of n_F, n_R parameters. These results are obtained with $n_P = 2000$ and $A_{min} = 500$.

We compare our proposed approach to other methods such as CNN-RNN [9], ELM-LRF [4] and DHONV+SVM [5]. We employ two measures to evaluate the performance of all methods that are: the nodes recognition rate and the area-based recognition rate. We perform experiments on two types of test frames with and without occlusion. Results presented in Table 1 show that our proposal outperforms all other 3D depth based recognition methods. In particular, recognition performance of our proposed method shows a higher robustness under occlusion compared with other methods.

Figure 5 shows six samples of test 3D frames in the first

Table 1. Comparison of our proposed approach to multiple 3D depth based recognition approaches. Results show that our proposal outperform other 3D depth based recognition methods. Moreover, recognition performance of our proposed method shows a hinger robustness under occlusion compared with other methods including CNN-RNN and ELM-LRF approaches.

Method	Test frames with occlusion		Test frames with no-occlusion	
	Nodes recognition rates	Area-based recognition rate	Nodes recognition rates	Area-based recognition rate
ELM-LRF [4]	45.527	46.4967	45.07	41.72
DHONV+SVM [5]	43.7585	50.94	54.44	65.8
CNN-RNN [9]	55.21	68.79	77.47	78.70
Proposed method	65.73	73.1	85.36	86.62

row and their corresponding sub-surfaces segmentation images are shown in the third row where each color represents a single node of the segmentation tree. Area-based recognition results of our proposed approach on those six depth scenes are presented in the bottom row. All objects are correctly recognized except for the rhino in (d) and (e). The second raw illustrate the recognition results obtained by CNN-RNN approach proposed by Socher et al., in [9]. Figure 5 demonstrates that our proposed approach is more capable of successfully recognize smaller sub-surfaces which will allow us to obtain a prediction with higher confidence.

5 CONCLUSION

We introduced a novel recognition approach for 3D depth occluded objects. Our model is based on a combination of hierarchical segmentation with convolutional-recursive deep learning recognition approach. A refinement of the classification results was possible based on the knowledge obtained from hierarchical clustering of the test frames. This architecture outperforms other 3D depth based recognition approaches including CNN-RNN deep learning based method.

References

[1] S. Boubou and E. Suzuki, "Classifying actions based on histogram of oriented velocity vectors," *Journal of Intelligent Information Systems*, vol. 44, no. 1, pp. 49–65, Feb 2015. [Online]. Available: https://doi.org/10.1007/s10844-014-0329-0

[2] L. Zhang, P. Shen, J. Ding, J. Song, J. Liu, and K. Yi, "An improved rgb-d slam algorithm based on kinect sensor," in *2015 IEEE International Conference on Advanced Intelligent Mechatronics (AIM)*, 2015, pp. 555–562.

[3] S. Boubou, T. Narikiyo, and M. Kawanishi, "Differential histogram of normal vectors for object recognition with depth sensors," in *2016 International Conference on Autonomous Robot Systems and Competitions (ICARSC)*, 2016, pp. 162–167.

[4] ——, "Object recognition from 3d depth data with extreme learning machine and local receptive field," in *2017 IEEE International Conference on Advanced Intelligent Mechatronics (AIM)*, 2017, pp. 394–399.

[5] S. Boubou, H. Jabbari Asl, T. Narikiyo, and M. Kawanishi, "Real-time recognition and pursuit in robots based on 3d depth data," *Journal of Intelligent & Robotic Systems*, Jan 2018. [Online]. Available: https://doi.org/10.1007/s10846-017-0769-1

[6] S. Boubou, T. Narikiyo, and M. Kawanishi, "Adaptive filter for denoising 3d data captured by depth sensors," in *2017 3DTV Conference: The True Vision-Capture, Transmission and Display of 3D Video (3DTV-CON)*, 2017, pp. 1–4.

[7] B. Chen, J. Yang, B. Jeon, and X. Zhang, "Kernel quaternion principal component analysis and its application in rgb-d object recognition," *Neurocomputing*, vol. 266, pp. 293 – 303, 2017.

[8] K. Lai, L. Bo, X. Ren, and D. Fox, "A large-scale hierarchical multi-view rgb-d object dataset," in *2011 IEEE International Conference on Robotics and Automation*, 2011, pp. 1817–1824.

[9] R. Socher, B. Huval, B. Bath, C. D. Manning, and A. Y. Ng, "Convolutional-recursive deep learning for 3d object classification," in *Advances in Neural Information Processing Systems*, 2012, pp. 656–664.

[10] M. Blum, J. T. Springenberg, J. Wülfing, and M. Riedmiller, "A learned feature descriptor for object recognition in rgb-d data," in *2012 IEEE International Conference on Robotics and Automation*, 2012, pp. 1298–1303.

[11] L. Bo, X. Ren, and D. Fox, "Unsupervised feature learning for rgb-d based object recognition," in *Experimental Robotics*, 2013, pp. 387–402.

[12] S. Salti, F. Tombari, and L. Di Stefano, "Shot: Unique signatures of histograms for surface and texture description," *Computer Vision and Image Understanding*, vol. 125, pp. 251–264, 2014.

[13] Y. Guo, F. Sohel, M. Bennamoun, J. Wan, and M. Lu, "A novel local surface feature for 3d object recognition under clutter and occlusion," *Information Sciences*, vol. 293, pp. 196–213, 2015.

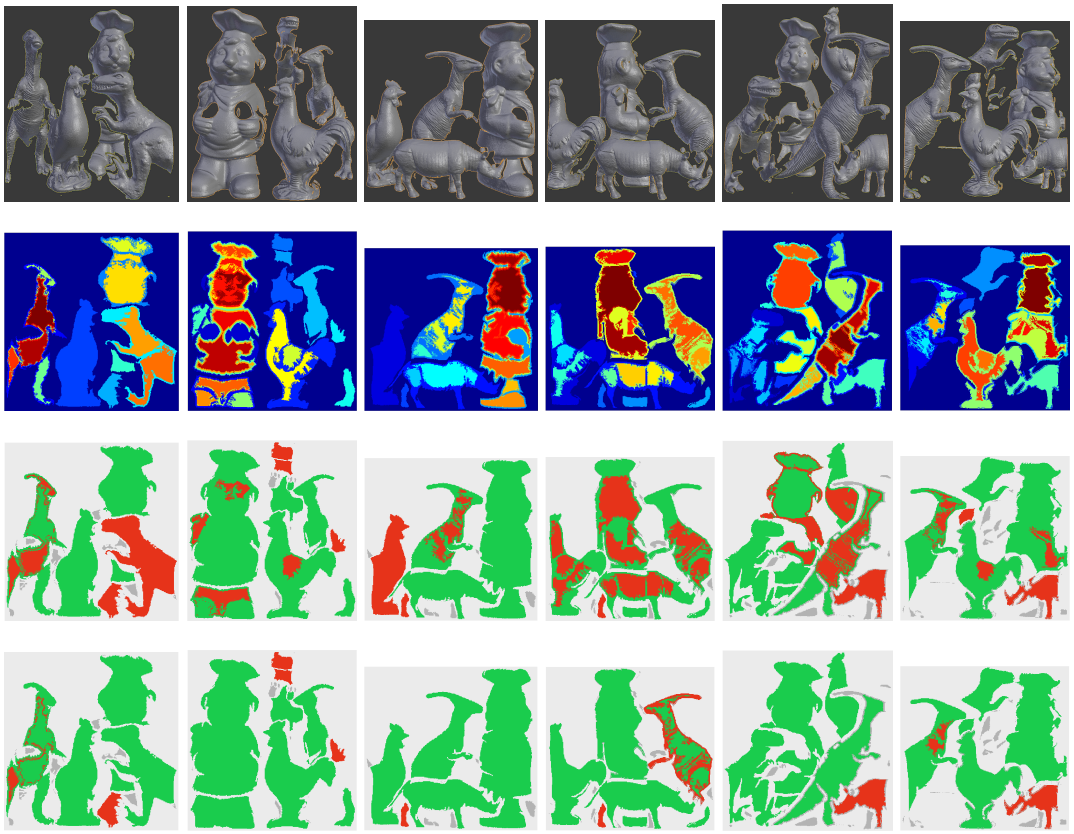
[14] Z. Li and J. Chen, "Superpixel segmentation using linear spectral clustering," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 1356–1363.

[15] S. Gupta, P. Arbeláez, R. Girshick, and J. Malik, "Indoor scene understanding with rgb-d images: Bottom-up segmentation, object detection and semantic segmentation," *International Journal of Computer Vision*, vol. 112, no. 2, pp. 133–149, 2015.

[16] P. V. Sander, J. Snyder, S. J. Gortler, and H. Hoppe, "Texture mapping progressive meshes," in *Proceedings of the 28th annual conference on Computer graphics and interactive techniques*, 2001, pp. 409–416.

[17] Y. Lee, S. Lee, A. Shamir, D. Cohen-Or, and H.-P. Seidel, "Mesh scissoring with minima rule and part salience," *Computer Aided Geometric Design*, vol. 22, no. 5, pp. 444–465, 2005.

[18] A.-V. Vo, L. Truong-Hong, D. F. Laefer, and M. Bertolotto, "Octree-based region growing for point cloud segmentation," *ISPRS Journal of Photogrammetry and Re-*



(a) (b) (c) (d) (e) (f)

Figure 5. Samples of test 3D frame are shown in the first row. Corresponding sub-surfaces segmentation images obtained by our proposed approach are illustrated in the second row where each color represents a single node of the segmentation tree. The third row illustrates the recognition results obtained by CNN-RNN method [9]. Recognition results of our proposed approach on those six depth scenes are presented at the bottom row. All objects are correctly recognized except for the rhino in (d) and (e).

mote Sensing, vol. 104, pp. 88–100, 2015.

[19] H. Zhang, C. Li, L. Gao, S. Li, and G. Wang, "Shape segmentation by hierarchical splat clustering," *Computers & Graphics*, vol. 51, pp. 136–145, 2015.

[20] J. Papon, A. Abramov, M. Schoeler, and F. Wörgötter, "Voxel cloud connectivity segmentation-supervoxels for point clouds," in *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*, 2013, pp. 2027–2034.

[21] H. Zhang, O. van Kaick, and R. Dyer, "Spectral methods for mesh processing and analysis," in *Proceedings of Eurographics State-of-the-art Report*, vol. 122, 2007.

[22] T. Joshi, B. Vijayakumar, D. J. Kriegman, and J. Ponce, "Hot curves for modelling and recognition of smooth curved 3d objects," *Image and Vision Computing*, vol. 15, no. 7, pp. 479–498, 1997.

[23] A. E. Johnson and M. Hebert, "Using spin images for efficient object recognition in cluttered 3d scenes," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 21, no. 5, pp. 433–449, 1999.

[24] A. S. Mian, M. Bennamoun, and R. A. Owens, "A novel representation and feature matching algorithm for automatic pairwise registration of range images," *International Journal of Computer Vision*, vol. 66, no. 1, pp. 19–40, 2006.

[25] O. Carmichael, D. Huber, and M. Hebert, "Large data sets and confusing scenes in 3-d surface matching and recognition," in *Second International Conference on 3-D Digital Imaging and Modeling (Cat. No.PR00062)*, 1999, pp. 358–367.

[26] R. Socher, C. D. Manning, and A. Y. Ng, "Learning continuous phrase representations and syntactic parsing with recursive neural networks," in *Proceedings of the NIPS-2010 Deep Learning and Unsupervised Feature Learning Workshop*, 2010, pp. 1–9.

[27] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[28] R. Socher, C. C.-Y. Lin, A. Y. Ng, and C. D. Manning, "Parsing natural scenes and natural language with recursive neural networks," in *Proceedings of the 28th International Conference on International Conference on Machine Learning*, ser. ICML'11, 2011, pp. 129–136.

[29] A. S. Mian, M. Bennamoun, and R. Owens, "Three-dimensional model-based object recognition and segmentation in cluttered scenes," *IEEE transactions on pattern analysis and machine intelligence*, vol. 28, no. 10, pp. 1584–1601, 2006.