2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)
Madrid, Spain, October 1-5, 2018

3D Deep Object Recognition and Semantic Understanding for Visually-Guided Robotic Service*

Sukhan Lee*, *Fellow, IEEE,* Ahmed M. Naguib *Member, IEEE,* and Naeem Ul Islam

Abstract— **For the success of visually-guided robotic errand service, it is critical to ensure dependability under various ill-conditioned visual environments. To this end, we have developed Adaptive Bayesian Recognition Framework in which in-situ selection of multiple sets of optimal features or evidences as well as proactive collection of sufficient evidences are proposed to implement the principle of dependability. The framework has shown excellent performance with a limited number of objects in a scene. However, there arises a need to extend the framework for handling a larger number of objects without performance degradation, while avoiding difficulty in feature engineering. To this end, a novel deep learning architecture, referred to here as FER-CNN, is introduced and integrated into the Adaptive Bayesian Recognition Framework. FER-CNN has capability of not only extracting but also reconstructing a hierarchy of features with the layer-wise independent feedback connections that can be trained. Reconstructed features representing parts of 3D objects then allow them to be semantically linked to ontology for exploring object categories and properties. Experiments are conducted in a home environment with real 3D daily-life objects as well as with the standard ModelNet dataset. In particular, it is shown that FER-CNN allows the number of objects and their categories to be extended by 10 and 5 times, respectively, while registering the recognition rate for ModelNet10 and ModelNet40 by 97% and 89.5%, respectively.**

I. Introduction

The advancement of robotic technologies in the last decades allows robots to serve human in our daily lives with ever increasing capability of visually guided task autonomy. The key to the success of such robotic service lies in ensuring dependability in visually guided task autonomy under various ill-conditioned 3D visual environments, including severe occlusions, poor illuminations, varying camera perspectives, etc. Many of the invariant photometric and geometric features developed in the past in 2D and 3D represent an attempt of achieving such dependability as described above [1-5]. Although successful in their own sake, they are too limited to serve as a solution for ensuring dependability of service robots with visually guided task autonomy in 3D cluttered environments. For one thing, biological systems including human seem to achieve perceptual dependability based more on proactive collection of as much sufficient evidences as possible with a proper execution of attentions. This indicates that a dependable vision system may require a spatial-temporal integration of not only effective visual information processing but also proactive control of attention and evidence collection.

In spite of great progress in 3D visual information processing, conventional approaches have shown some fundamental limitations of their own: 1) Scalability implying the rapid performance degradation in terms of the increase of the number of objects to be handled, 2) Difficulty in feature engineering implying laborious manual implementation of features optimal for given tasks. Recent advancement in deep learning may provide a potential for overcoming fundamental limitations of conventional approaches as described above. This is because deep learning is capable of being trained for recognition with a large number of objects while a hierarchy of layer-wise features are automatically constructed during training.

As far as the recent advances in deep learning are concerned, the most impressive results have been associated with 2D vision [30, 31]. Compare to the advancement in 2D, deep learning applications to 3D vision [12-21] are yet to be successful to show their potential, partly because of the issues involved in 3D representation, computational complexity and database availability. In a sense, applications to visually guided robotic service in 3D environments provide 3D deep learning with an opportunity as its test-bed for further advancement.

This paper first presents Adaptive Bayesian Recognition Framework as a framework of visual dependability. The framework is configured with the in-situ selection of multiple sets of optimal features and proactive collection of sufficient evidences as a means of implementing the principle of dependability adopted by biological systems. To extend the framework for handling a larger number of objects without performance degradation while avoiding difficulty in feature engineering, as well as for endowing a capability of semantic understanding of 3D objects and scenes, this paper extends the Adaptive Bayesian Recognition Framework by incorporating a novel deep learning architecture, referred to here as FER-CNN. FER-CNN not only extracts but also reconstructs a hierarchy of 3D features with the layer-wise independent feedback connections that can be trained. Reconstructed features representing part of 3D objects are then semantically linked to ontology for obtaining object categories and properties, thus elevating the framework potentially to the level of visual understanding. The main contribution of this paper is threefold: 1) Adaptive Bayesian Framework for the dependability in object recognition, 2) FER-CNN integrated into Adaptive Bayesian Framework for handling a large number of objects with automatically generated features and 3) the capability of FER-CNN for

* This research was supported, in part, by the "3D Recognition Project" of Korea Evaluation Institute of Industrial Technology (KEIT) (10060160) and, in part, by the "Robot Industry Fusion Core Technology Development Project" of KEIT (10048320), sponsored by the Korea Ministry of Trade, Industry and Energy (MOTIE).

Authors are with Intelligent Systems Research Institute, School of Information and Communication Engineering Sungkyunkwan University, Suwon, Korea (Corresponding author is Sukhan Lee. lsh1@skku.edu).

reconstructing features that opens a possible way of linking data-driven recognition to symbolic ontology.

II. Related Work

Convolutional Neural Networks (CNN) has shown significant performance in 3D object classification recently. This rapid advance in 3D object classification has been possible by the abundance of computational power, the availability of open source large-scale annotated 3D datasets and open source libraries. Despite the availability of all these resources, the task of classification of 3D objects is difficult because of the variations, sparsity and size of 3D data representations.

A. Deep Learning Based 3D Object Classification and Reconstruction

It is well known that, for deep learning based 3D object recognition and reconstruction, representation matters.

1) 3D Networks with Multiview 2D representation of 3D objects

One of the earliest 3D object recognition based on Multiview 2D representation of the 3D objects was presented by Su et al [6]. They used view-based CNN architecture for each view and finally combine the information from each CNN into a single complete shape descriptor, where they show improvement in the performance of recognition using single view image CNN architecture. DeepPano [7] projected 3D shapes to panoramic views using cylindrical projection around their principle axis and then used those views as input to CNN for learning the representation of the object. Row-wise max pooling is used between fully connected and convolution layer to make the learned representation rotation invariant. In [8] the authors proposed RotationNet, which takes input as multi-view images and estimate object category along with pose information. For pose information the viewpoint variables are treated as latent variable and are optimized during training in an unsupervised manner using an unaligned dataset. In [9], the authors proposed 3D deep dense shape descriptor. They used 2D multilayer dense representation of 3D volumetric data for feature extraction using design of the network which jointly train a set of convolution neural network, recurrent neural network and an adversarial discriminator. LonchaNet [10] dilate projected 2D slices of the input 3D point cloud and use independent GoogLeNets to extract discriminative representations from each slice. They concatenate extracted features and use decision layers for classification.

2) 3D Networks with Point Cloud based 3D Representation

Recently, Charles et al proposed PointNet in [20], where they directly used point clouds, taking advantage of the permutation invariance of points in the input. They showed improvement in recognition performance at the expense of computational cost.

3) 3D Networks with Voxel based 3D representation

Voxel based representation of 3D objects plays an important role in computer graphics community. It provides a simple, uniform and robust description to the objects and found the basis of volume graphics [11]. The first seminal work was carried out in 2015 by 3D ShapeNets [12] on ModelNet dataset, where they performed 3D recognition along with shape completion. They build generative network using Convolutional Deep Belief Network [13, 14] by learning the probability distribution over class labels and voxel representation of the data. Although the recognition rate was low (77% accuracy), it was a precursor for true volumetric CNNs. Maturana and Scherer proposed VoxNet [15] where they integrated volumetric occupancy Grid representation with 3D convolutional neural network and showed that the recognition rate of 3D shapes increased. In [16] the author used the combination of 3D convolutional neural network along with the Generative Adversarial Network (GAN) to capture 3D shape descriptor. To generate 3D shapes, they trained 3D GAN for each object category separately. To evaluate the performance in term of classification they used the learned features from different layers of discriminator architecture, concatenated them and then applied linear SVM for classification. In [17] the authors proposed unsupervised approach for object recognition where they used full convolutional volumetric auto encoder that learns volumetric representation from the noisy data. Brock et al. [18] explored voxel based variational autoencoder for unsupervised feature learning and Voxception-ResNet architecture for 3D object recognition. The latent space of variational autoencoder is used to interpolate between classes, whereas, the Voxception-ResNet architecture is used for classification. Inspire by the work of 3D ShapeNets and VoxNet Garcia et al. propose PointNet in [19]. They used density occupancy grids representations for the input data and integrating them into a supervised Convolutional Neural Network architecture.

4) 3D Networks with Octree based 3D Representation

To further reduce the burden of computational cost and memory requirement of Voxel based representation, OctNet [21] is proposed to exploit the sparsity in the input data by hierarchically partitioning the space using a set of unbalanced octrees, where each leaf node stores a pooled feature representation. This allows to focus memory allocation and computation to the relevant dense cells and enables deeper networks without compromising resolution.

B. Semantic Part Detection

Although pixel-wise semantic segmentation has been successful using 2D CNNs, Volumetric CNNs for 3D part segmentation is a relatively new subject. Charles et al, [20] reformulate 3D part segmentation as a fine-grained classification task. They trained their PointNet architecture on annotated ShapeNet part dataset and achieved 83.7% mIoU 3D part segmentation.

Large-Scale 3D Shape Reconstruction and Segmentation from ShapeNet Core55 benchmark [22] has been recently announced in 2017 by Universities of Stanford, Princeton, Massachusetts–Amherst, Oxford, Korea Advanced Institute of Science and Technology (KAIST), UC Berkeley, Texas, Facebook AI Research … etc. One track of this competition is 3D object semantic parts segmentation. The winner for the competition this year is [23] by Oxford university and Facebook AI Research. They defined a new Submanifold Sparse Convolution (SSC) operator and used it to construct two architectures: Fully Convolutional Network (FCN), and U-Net (Autoencoder with skip connections). FCN SSCN was able to achieve 85.98% mIoU on the test set of 3D object

semantic parts segmentation competition on ShapeNet Core55 which outperformed other methods.

III. Dependability in Object Recognition for Visually Guided Robotic Errand Service

The success of visually guided robotic errand service relies very much on the dependability of a robot in object recognition and pose estimation for finding, picking up and delivering the ordered object to the user in an unstructured and cluttered environment. However, in practice, it is quite challenging to ensure such dependability in object recognition and pose estimation for errand service due to various adverse conditions such as occlusion, overlapping, distance, orientation, illumination etc. Here, we propose a fundamental framework for building a dependable recognition system for robotic errand service. The proposed framework is based on "the principle of perceptual dependability" that we consider how biological systems such as human achieve dependability in perception. Specifically, we conjecture that human perception is dependable as, upon perceptual stimuli, human self-defines the perceptual mission associated with the stimuli and, subsequently, mobilizes its resources into a collective cognitive process and behavior to have the mission accomplished. An example of such a collective process and behavior may be the behavior of perceptual attention to and proactive search for salient evidences as a means of collecting sufficient evidences for a decision of acceptable confidence. To implement the aforementioned principle of perceptual dependability, we design Adaptive Bayesian Recognition Framework, as illustrated in Fig. 1, as a simple manifestation of a collective cognitive process

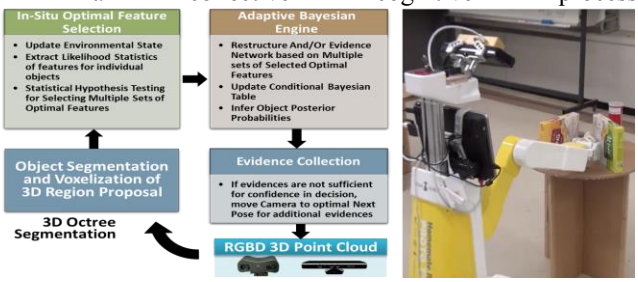

Figure 1. Flowchart of Adaptive Bayesian Recognition Framework used by HomeMate Service Robot.

and behavior. We adopt the following three key processes as fundamental building blocks of our framework: 1) in-situ selection of multiple sets of optimal features or evidences, 2) proactive search for additional evidences in case the collected evidences are insufficient for a decision of acceptable confidence and 3) adaptive Bayesian evidence reasoning for incorporating varying sets of evidences, continuously updated and accumulated, into a decision. Since the available evidences and their statistical uncertainties are varying in time, the structure of Bayesian evidence net and its inference table are to be updated accordingly. In what follows, we present more details of the above three key processes.

A. In-Situ Selection of Multiple Sets of Optimal Features

The in-situ selection uses a predefined pool of features for selecting multiple sets of optimal features. Then, the conditional probability distribution that a particular feature is measured for a given object under a given environmental condition is obtained by experimentation or by physics simulation. This allows to compute the t-test table, $T(f_k; o_i, o_j)$, that measures the effectiveness of f_k to discriminate a given target object, o_i, from other object, o_j, for all i, j, k, as shown in Table I. Note that $T(f_k; o_i, o_j)$ is assumed to be conditioned on environmental parameters too, although not explicitly expressed so. Then, multiple sets of optimal features are selected from the t-test table in such a way as to provide sufficient confidence in recognizing the target object, where an optimal set Ψ_s consists of the minimum number of features satisfying the following constraints:

For any o_i, there exist $f_k \in \Psi_s$ such that

$T(f_k \in \Psi_s: o_i, o_j) > \Omega$ for any $j \neq i$

Ω represents the t-value of the minimum acceptable two-tailed confidence interval.

TABLE I. Conditional Features T-Test Table

	O_1	O_2	…	O_N
f_1	$T(f_1: o_i, o_1)$	$T(f_1: o_i, o_2)$	…	$T(f_1: o_i, o_N)$
f_2	$T(f_2: o_i, o_1)$	$T(f_2: o_i, o_2)$	…	$T(f_2: o_i, o_N)$
…	…	…	…	…
f_n	$T(f_n: o_i, o_1)$	$T(f_n: o_i, o_2)$	…	$T(f_n: o_i, o_N)$

$$T(f_k: o_i, o_j) = \frac{\bar{x}_i - \bar{x}_j}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

$\bar{x}_i = E(f_k|o_i)$
$\sigma_i = STDEV(f_k|o_i)$
When population statistics is known, $n_1 = n_2 = 1$

B. Adaptive Bayesian Recognition Engine

The multiple sets of optimal features selected serve as multiple sets of evidences for Bayesian recognition, as shown in Fig. 2. Note that the features listed in a set serve as the evidences under Logical AND, while multiple sets are under Logical OR, as illustrated by an And/Or Graph(AOG) in Fig. 2. The in-situ selection of multiple sets of optimal features makes Adaptive Bayesian Recognition Engine reconstruct AOG and update its conditional probability tables accordingly. Then, the reconstructed AOG and its conditional probability tables are instantiated with the measurement values of individual features involved, such that the posterior probability of the target object is inferred along with the decision confidence [24]. Note that the use of multiple sets of optimal features improves robustness in decision with multiple independent reasoning paths, especially when measurements fall in low populated regions of feature space.

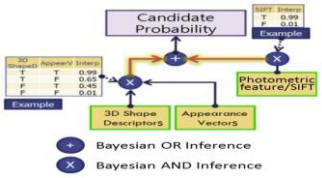

Figure 2. Figure 2. Adaptive Bayesian Recognition Engine where a Bayesian And/Or Graph (AOG) is reconstructed and its conditional probability tables are updated in order to adapt to the in-situ selection of multiple sets of optimal features.

C. Proactive Evidence Collection

In case no viable set of optimal features is found or the confidence level of a decision is insufficient, the robot is to move to the next best pose to collect further evidences. This is done by searching the following two spaces: the object space

where the probability that the target object exists is computed and the action space where the best next camera pose is obtained out of the free navigation space that supports the recognition of the target object with a maximum confidence. To this end, we first tessellate both spaces into grid cells on a global map, as shown in Fig. 3. Then, we assign the probability that the target object exists to individual cell of the object space based on the posterior probability distribution resulted from the previous recognition cycle and the occlusions currently remaining in the object space. This is followed by determining the next best camera pose by maximizing the utility function representing how effective a camera pose is for recognizing the target object [25]. Specifically, the utility function takes into consideration the probability distribution over the object space, the measurement uncertainty associated with a particular camera pose, as well as and the observability due to occlusion and distance to the object space, etc.

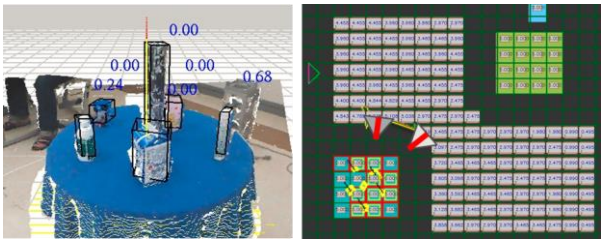

Figure 3. Left: A camera view of the objects on the table defined as the object space (green cells). The objects on the table are represented by the bounding boxes and Bayesian posterior probabilities. Right: The search for the best next camera pose in the action space (white cells) by maximizing the utility function with observing the occluded regions.

D. Performance with HomeMate Errand Service Robot

To test the performance of the proposed adaptive Bayesian recognition framework, we selected 11 objects from ISRI_DB [26] and placed them randomly on a table in a cluttered setting, as shown in Fig. 4(top). The proposed framework implemented was then applied to the recognition of all the objects, along with their poses estimated, in the scene. Using a pool of 3D shape descriptors, comprised of 15 global and local geometric shape features including height, width, top shape, mouth opening, top-middle width ratio, etc., we achieved 92% accuracy from a single view [24]. Similarly, we experimented on 10 industrial objects using not only 3D Shape Descriptors, but also 3D SIFT and Closed Loop Boundary (CLB), and achieved 97.5% accuracy from a single view [27]. The higher accuracy we obtained for the latter was due to the inclusion of additional features: 3D SIFT and CLB. Finally, the pose estimation was conducted based on the following two steps processes: the coarse pose estimation based on a geometric feature such as 3D line feature, which was followed by the fine tuning based on ICP. This resulted in the mean error of less than 1 mm in translation and 0.559 degree in orientation.

To carry out a full-fledged experiment including the proactive evidence collection, we randomly placed the 11 objects from ISRI_DB used in the above experiment on two separate tables. Then, we commanded our home service robot, "HomeMate" shown in Fig.1 [31] to fetch each object under various environmental conditions. Due to the proactive evidence collection implemented, HomeMate was able to recognize and fetch all the objects successfully, but with the execution of different number of camera poses, as illustrated in Fig. 4 (bottom).

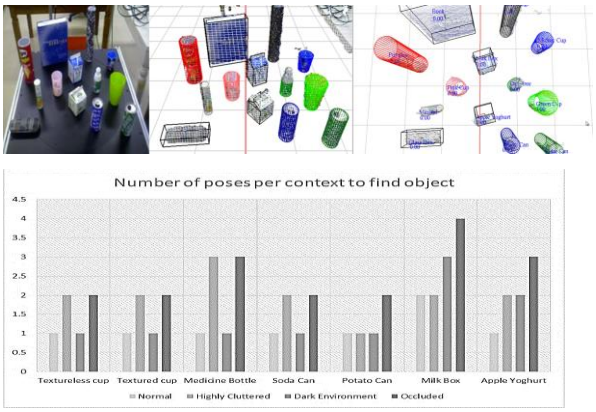

Figure 4. Top: Adaptive Bayesian Recognition Framework recognizing 11 objects from ISRI_DB using a pool of 3D Shape Descriptors. Bottom: the number of camera views used to recognize the target object under different environmental conditions.

IV. Deep Learning Extension of Adaptive Bayesian Recognition framework

The proposed adaptive Bayesian recognition framework based on the principle of dependability works well as intended. However, we encounter its limitation on the following two accounts: the first is when we scale up the number of objects it handles and the second is when we level up its scope from classification to semantic understanding of objects and scenes. The first issue stems from the fact that the framework does not scale up with the number of objects it handles, as it works well when the number of objects to be handled are modest, say up to 30. This is because 1) the hand-crafted feature engineering based on the collection of sufficient statistics becomes impractical with the rapid growth of the number of features required, 2) the computational cost of an online adaptive framework is increased quadratic with the number of objects, and 3) the recognition performance is degraded due to the feature space becoming more crowded. The second issue is more fundamental as the data-driven bottom-up classification has no effective means of being linked to the symbol-driven semantic ontology to date.

In order to tackle the issues described above, we propose the following: First, we take advantage of the capability of deep convolutional neural networks (CNN) for dealing with a large number of objects, for instance, over 150K in case of ModelNet40 dataset [12], and for building a hierarchy of features automatically, thus avoiding the limitation from scalability and the difficulty of hand-crafted feature engineering. Second, we combine the voxel based 3D segmentation of adaptive Bayesian recognition framework with a 2D region proposal network integrated with CNN to effectively handle a highly occluded and cluttered environment. Third, we devise a method to identify and reconstruct the features formed in each layer of CNN such that the identified and reconstructed features play a role as semantic part proposals that can be linked to ontology. This results in a deep learning extension of adaptive Bayesian recognition framework in which 3D region proposal, CNN based deep feature extraction and classification as well as deep

feature reconstruction for linking to ontology are integrated, as shown in Fig. 5. One thing to note is that, as a key enabler, we develop a novel convolutional neural network, referred to here as Feature Extraction and Reconstruction Convolutional Neural Network (FER-CNN), that not only extracts but also reconstructs a hierarchy of deep features based on supervised or unsupervised training, as shown in Fig. 6.

To help grasp the proposed deep learning extension more clearly, the following scenario is presented: After the robot moves to the next best pose, an RGBD camera captures both 2D image and 3D point cloud of the scene. Then, 3D octree segmentation [28] and 2D Faster R-CNN [29] are applied in combination to achieve the object segmentation with region proposals. This combination takes advantage of 3D octree segmentation to propose the regions of novel objects suppressed by Faster R-CNN, while making use of Faster R-CNN to propose regions of small, shiny, transparent and occluded objects otherwise difficult to obtain from 3D octree segmentation alone. The resulting region proposals are represented by 32x32x32 voxels using the same octree representation as before with a fixed cell size. Then, the input sample of 32x32x32 voxels is fed into FER-CNN for the extraction of a hierarchy of deep features, followed by object recognition and coarse orientation estimation through the classification layers.

Furthermore, the extracted features are fed to a semantic part detection network so as to identify the local geometrical parts defined in the object ontology DB. Finally, both the classification probabilities and the semantic parts identified are input to a high-level Bayesian reasoning network in order to recognize, categorize and semantically understand the input object. In what follows, we present more details of the key processes described above:

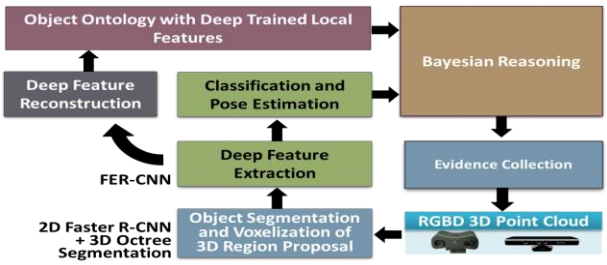

Figure 5. Extension of Adaptive Bayesian Recognition Framework by integrating Faster R-CNN for region proposal and FER-CNN for hierarchal deep feature extraction, reconstruction, classification, and semantic part detection

A. 3D Region Proposal

The individual 3D objects of a scene can be segmented by identifying their geometric separations using 3D Octree Representation. This method is effective for 3D object segmentation due to its invariance to illumination and texture. However, it fails to segment objects when the point clouds captured are noisy with outliers and the objects are geometrically non-separable. On the other hand, the detection of objects in 2D images has been quite successful based on the supervised training of feedforward convolutional neural networks [30]. However, this method may be vulnerable to the occlusion incurred by a camera perspective as well as to the variation of illumination and texture. We propose to combine the two methods into a 3D region proposal.

One thing to note is that deep learning based object detection relies highly on the availability of a large number of annotated training samples. In case of robotic applications dealing with 3D objects, objects are usually stored as a textured CAD database. This arises an issue of data dependency in training due to the discrepancy between CAD and real datasets. However, neither creating a large-scale of annotated real 3D dataset nor adding new real 3D data to an existing CAD dataset is easy to be done. Here, we pre-train Faster R-CNN [29] on simulated scenes rendered from the textured CAD models of scene objects. Then, we fine-tuned Faster R-CNN using several manually annotated real scenes. Although the classification performance of Faster R-CNN is poor due to undertraining, the resulting region proposals are sufficient for segmentation. After that, the resulting bounding boxes are projected back and intersected with the 3D point cloud to obtain 3D region proposals. Finally, we combine both Octree based segmentation and Faster R-CNN segmentation by concatenating their proposed regions. For each region proposal, we generate 32x32x32 voxels within the bounding box of an object.

B. Deep Feature Extraction and Reconstruction with FER-CNN

For the automatic extraction and reconstruction of a hierarchy of features, we propose a new CNN architecture with layer-wise inverse connections, termed as Feature Extraction and Reconstruction CNN (FER-CNN), as shown in Fig. 6. The FER-CNN implemented consists of 3 sub-networks: Encoder, Decoder and Classifier. Encoder and Decoder sub-networks are mirrored in their configurations but with independent weights assigned to their respective convolutional and deconvolutional connections. Encoder sub-network consists of five 3D convolutional layers, each followed by the batch normalization and the ReLu activation function. Each of the five 3D deconvolutional layer of Decoder sub-network accepts an input from either of the following three sources: the corresponding encoder layer in the form of a skip connection, the upper deconvolutional layer, or the externally injected latent code. Classification sub-network, consisting of 3 fully-connected layers with a softmax layer at the last, takes the concatenated dense feature output of all encoder layers as an input.

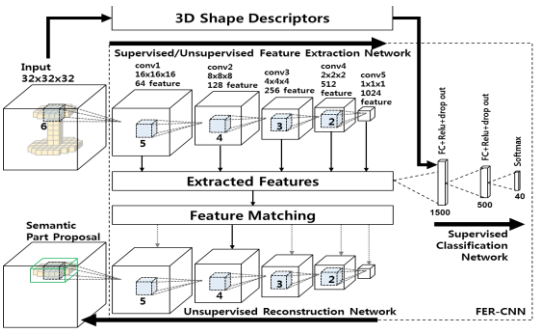

Figure 6. Architecture of FER-CNN for HomeMate Robot

The training of FER-CNN follows the following 3 phases in general:

- Phase 1: Supervised end-to-end training of the encoder and classification sub-networks using cross entropy loss.

- Phase 2: Unsupervised training of the decoder weights of each layer from layer 1 to 5 using the following loss function with the weights of Encoder sub-network either fixed or tunable:

$$\min_{E_i,D_i}\left(\left(x - D_{i\sim1}\left(E_{1\sim i}\left(x\right)\right)\right)^2 + \left(E_{1\sim i}\left(x\right) - E_{1\sim i}\left(D_{i\sim1}\left(E_{1\sim i}\left(x\right)\right)\right)\right)^2\right)$$

- Phases 3: Unsupervised training of the weights joint decoder layers from 1-2 to 1-5 with the following loss function with the weights of Encoder sub-network either fixed or tunable:

$$\min_{E_{1\sim i},D_{1\sim i}}\left(\left(x - D_{i\sim1}\left(E_{1\sim i}\left(x\right)\right)\right)^2 + \left(E_{1\sim i}\left(x\right) - E_{1\sim i}\left(D_{i\sim1}\left(E_{1\sim i}\left(x\right)\right)\right)\right)^2\right)$$

As far as reconstruction of a specific feature is concerned, it is not trivial as a receptive field of the input space is not only a function of the corresponding feature of a particular layer but also of its neighboring features of that layer, referred to here as a response field, that affect the receptive field, as shown in Fig. 7. This happens due to the overlapping convolution windows dictated by the choice of layer configurations with window size and stride. Taking the above response field into consideration, we present the following, so called, contribution-based trained reconstruction algorithm:

Input: feature "f" at location "p" in layer "L"

Output: reconstructed receptive field in input space "x"

1. Identify "Rx" receptive field of "p" in input space

2. Identify "Rf" response field in layer "L" by adding the response field of each cell in "Rx"

3.Copy "Rf" from a training sample in which it has the closest response to "f" at location "p" in layer "L"

4.Overwrite the value of location "p" in "Rf" with "f"

5.. Reconstruct "x*" in input space from layer "L" using "Rf" padded with zeros

6.Crop "x*" using receptive field "Rx" to obtain "x"

Note that the unique nature of the proposed feature reconstruction based on a trained deconvolution sub-network and on a response field, compared to conventional approaches such as simple cropping of a receptive field, layered relevance propagation [32] and iterative input optimization [33], as it provides a means of filter code-based reconstruction without iterative optimization.

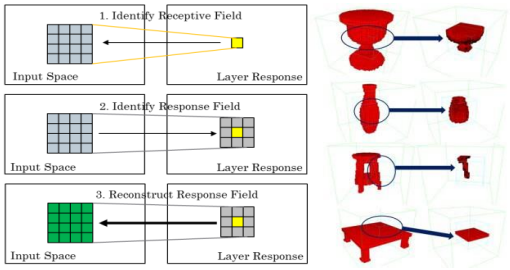

Figure 7. Left: Procedure for Contribution-based Trained Reconstruction algorithm. Right: a few examples of the reconstructed deep features.

C. *Semantic Part Proposals*

As a means of leveling up the proposed framework from classification to semantic understanding of objects, we propose a method of connecting the features extracted by FER-CNN with an ontology database. The proposed method consists of three steps: 1) identification of dominant features, 2) connection to ontology part classes or sub-classes and 3) semantic feature matching.

1. Identification of Dominant Features: DB-SCAN clustering algorithm is applied to clustering the feature space of each layer across all training data. Clusters with a higher density are approximated using multivariate Gaussian distributions.
2. Connection to Ontology Part Classes: Connection to ontology part classes or sub-classes is performed by a simple classification network that learns to map a distribution of features to a specific semantic label. In case semantic part labels are not readily available, we can rely on feature reconstruction to generate region proposals. We may reconstruct both the cluster mean and a few training samples of the detected region that belongs to the said cluster. This visualization assists a human operator to label semantic parts and to update ontology database with the reconstructed features.
3. Semantic Feature Matching: Those features extracted from the higher layers of FER-CNN match the global geometric shapes while those features from the lower layers, say, the first two layers, match the local geometric shapes of objects of different categories. We train a classification network with the same architecture as the classification sub-network used for object classification based on the semantic part proposals from feature reconstruction. As shown in Fig. 8, this allows to match given features with the corresponding clusters of FER-CNN as described in the previous section. Note that the proposed reconstruction provides more information than simple cropping of a receptive field as it encodes the importance of local regions in terms of either a strong evidence with high intensity or a novelty with low but non-zero intensity as illustrated by Fig. 8.

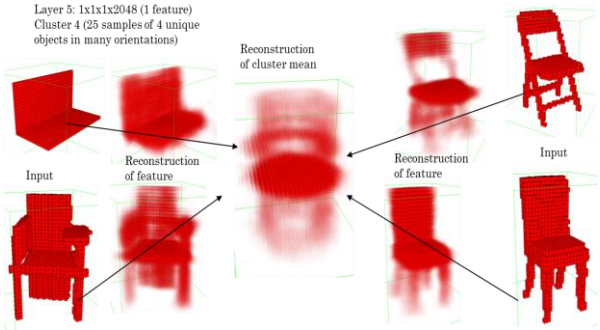

Figure 8. Example of a semantic cluster. Closest 4 objects to cluster center are shown at the edges along with their feature reconstruction. Reconstruction of the mean of the cluster is shown at the center of the figure. High intensity in reconstruction corresponds to strength of the feature. Low (non-zero) intensity in reconstruction encodes novel features for specific objects.

V. Experimental Results

We evaluate the proposed framework extended with FER-CNN based on two different datasets: ISRI_DB [26] real dataset for testing under a real visually guided robotic service environment and ModelNet40 CAD dataset [12] for quantitatively evaluating recognition performance with a large-scale of 3D objects.

A. 3D Object Recognition using Real ISRI_DB

One of the target applications of our proposed framework is visually guided robotic errand service based on HomeMate [31] developed for services to elderly and disabled. HomeMate is to be operated in an indoor environment under visually adverse conditions due to clutter, occlusion, poor illumination, etc.

To evaluate our extended framework, we trained FER-CNN on ISRI_DB consisting of RGB and 3D point cloud data of 247 household objects in 33 categories. The 3D point cloud of each object is a full-blown model generated by registering a number of data from different viewpoints. Based on the 3D point cloud models stored in the DB, we generated 1,000 random poses per object as training samples. For testing, we first placed about 10 real objects defined in ISRI_DB in a cluttered setting on a table. And, then, all the objects on the table were labeled automatically from a single view of RGB-D data for scene understanding. Fig. 9 (top) illustrates an example of 3D region proposals based on the combination of 3D octree segmentation and 2D region proposal of Faster RCNN. We observed that this combination was necessary to outcome a complete list of 3D region proposal. Fig. 9 (bottom) illustrates examples of object classification and categorization based on the voxel representation of the segments from the 3D region proposal. We repeated the experiments 10 times, each with roughly 15 3D objects per scene. The average precision of FER-CNN for classification is about 91%. Then, the object pose was estimated first roughly by regressing a fully connected layer with the features of FER-CNN and then refined by ICP, resulting in about 10 degrees of orientation error.

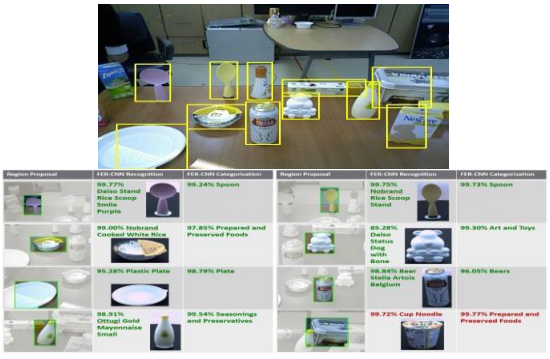

Figure 9. Top: Region proposals obtained using Faster R-CNN and Octree segmentation (projected onto the 2D image). Bottom: examples of objects classification and categorization using FER-CNN. For comparison, we illustrate FER-CNN highest probability object using its image from ISRI_DB databset. An example of failure in classification and categorization is shown and attributed to severe occlusion. FER-CNN is trained on occlusion-free ISRI_DB models.

B. 3D Object Recognition using ModelNet CAD Dataset

FER-CNN is evaluated further using a large-scale of ModelNet dataset. ModelNet has two datasets: ModelNet10 having 10 classes of 48000 3D CAD models and ModelNet40 having 40 classes of 151,128 3D CAD models. We used two fully connected layers along with one dropout layer at the output of the encoder sub-network of FER-CNN without resorting to conventional 3D shape descriptors. We trained FER-CNN on ModelNet10 and ModelNet40 datasets and achieved 97% accuracy on ModelNet10 testing dataset and 89% accuracy on ModelNet40 testing datasets, as shown in Table II.

TABLE II. Categorization accuracy on ModelNet dataset

Model	ModelNet10 Classification test accuracy	ModelNet40 Classification test accuracy
VoxNet	92%	83%
FER-CNN(**Ours**)	97%	89.5%

C. Semantic 3D Part Detection and Connection to Ontology with ISRI_DB Dataset

We show FER-CNN capability of learning meaningful features as 3D semantic parts of objects. We first detect these semantic parts from input objects as shown by their receptive fields illustrated in Fig. 10 (top). For example, a bottle has ontology structure which contains Side wall, Bottle neck and Tall pillar. Detected semantic parts are reconstructed along with the cluster means to which these features belong to and are linked to ontology structure of that specific object, as illustrated in Fig. 10 (bottom).

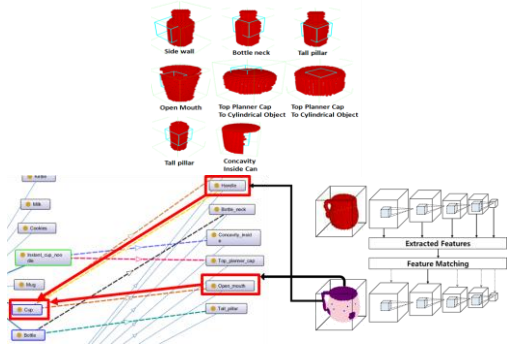

Figure 10. Top: examples of detected 3D semantic parts from ISRI_DB captured during HomeMate experimentation. Bottom: example of Semantic features reconstruction from an input 3D model of ModelNet dataset. Reconstruction of both "Open Mouth" and "Handle" features are overlayed on top of input object for visualization. Using ontology of "Cup" object, the detection of "Open Mouth" and "Handle" enhance the probability of "Cup" posterior probability.

VI. Conclusion

We first present an adaptive Bayesian recognition framework for achieving dependability in visually guided robotic service. The framework composed of in-situ optimal feature selection, proactive evidence collection and adaptive Bayesian evidence reasoning is to follow the principle of perceptual dependability seemingly adopted by biological systems. Then, in order to level up the framework for handling a large number of objects while solving the difficulty of feature engineering as well as for semantically understanding objects and scenes by linking to ontology, we devise FER-CNN and integrate it into the frame. FER-CNN extracts a hierarchy of features automatically by training that

can be used in classification while eliminating the burden of feature engineering. In particular, FER-CNN we proposed shows a unique capability of feature reconstruction through training, opening many possibilities of utilizing the extracted features when combined with feature clustering in the layer-wise filter spaces. One example shown is the generation of semantic part proposals that can be connected to ontology. Experiments demonstrate that our proposed extended framework is capable of classifying and categorizing a large number of objects in a scene as demonstrated by the experiments. Further research includes more extensive analysis and experiments on the capability of FER-CNN in terms of reconstruction, attention and interpretability associated with 3D deep learning networks. We hope that this paper spurs readers with a new direction of research for the robotic vision applied to robotic service in 3D environments.

REFERENCES

[1] Lowe, David G. "Object recognition from local scale-invariant features." Computer vision, 1999. The proceedings of the seventh IEEE international conference on. Vol. 2. Ieee, 1999.

[2] Lowe, David G. "Distinctive image features from scale-invariant keypoints." International journal of computer vision 60.2 (2004): 91-110.

[3] Liu, Run Zong, Yuan Yan Tang, and Bin Fang. "Topological coding and its application in the refinement of SIFT." IEEE transactions on cybernetics 44.11 (2014): 2155-2166.

[4] Bay, Herbert, Tinne Tuytelaars, and Luc Van Gool. "Surf: Speeded up robust features." European conference on computer vision. Springer, Berlin, Heidelberg, 2006.

[5] Rublee, Ethan, et al. "ORB: An efficient alternative to SIFT or SURF." Computer Vision (ICCV), 2011 IEEE international conference on. IEEE, 2011.

[6] Su, Hang, et al. "Multi-view convolutional neural networks for 3d shape recognition." Proceedings of the IEEE international conference on computer vision. 2015.

[7] Shi, Baoguang, et al. "Deeppano: Deep panoramic representation for 3-d shape recognition." IEEE Signal Processing Letters 22.12 (2015): 2339-2343.

[8] Kanezaki, Asako. "RotationNet: Learning Object Classification Using Unsupervised Viewpoint Estimation." CoRR (2016).

[9] Ren, Mengwei, Liang Niu, and Yi Fang. "3D-A-Nets: 3D Deep Dense Descriptor for Volumetric Shapes with Adversarial Networks." arXiv preprint arXiv:1711.10108 (2017).

[10] F. Gomez-Donoso, A. Garcia-Garcia, J. Garcia-Rodriguez, S. Orts-Escolano and M. Cazorla, "LonchaNet: A sliced-based CNN architecture for real-time 3D object recognition," 2017 International Joint Conference on Neural Networks (IJCNN), Anchorage, AK, 2017, pp. 412-418.

[11] Kaufman, Arie, Daniel Cohen, and Roni Yagel. "Volume graphics." Computer 26.7 (1993): 51-64.

[12] Wu, Zhirong, et al. "3d shapenets: A deep representation for volumetric shapes." Proceedings of the IEEE conference on computer vision and pattern recognition. 2015.

[13] Hinton, Geoffrey E., Simon Osindero, and Yee-Whye Teh. "A fast learning algorithm for deep belief nets." Neural computation 18.7 (2006): 1527-1554.

[14] Lee, Honglak, et al. "Unsupervised learning of hierarchical representations with convolutional deep belief networks." Communications of the ACM 54.10 (2011): 95-103.

[15] Maturana, Daniel, and Sebastian Scherer. "Voxnet: A 3d convolutional neural network for real-time object recognition." Intelligent Robots and Systems (IROS), 2015 IEEE/RSJ International Conference on. IEEE, 2015.

[16] Wu, Jiajun, et al. "Learning a probabilistic latent space of object shapes via 3d generative-adversarial modeling." Advances in Neural Information Processing Systems. 2016.

[17] Sharma, Abhishek, Oliver Grau, and Mario Fritz. "Vconv-dae: Deep volumetric shape learning without object labels." European Conference on Computer Vision. Springer, Cham, 2016.

[18] Brock, Andrew, et al. "Generative and discriminative voxel modeling with convolutional neural networks." arXiv preprint arXiv:1608.04236 (2016).

[19] Garcia-Garcia, Alberto, et al. "Pointnet: A 3d convolutional neural network for real-time object class recognition." Neural Networks (IJCNN), 2016 International Joint Conference on. IEEE, 2016.

[20] Qi, Charles R., et al. "Pointnet: Deep learning on point sets for 3d classification and segmentation." Proc. Computer Vision and Pattern Recognition (CVPR), IEEE 1.2 (2017): 4.

[21] Riegler, Gernot, Ali Osman Ulusoy, and Andreas Geiger. "Octnet: Learning deep 3d representations at high resolutions." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Vol. 3. 2017.

[22] Li Yi et.al "Large-Scale 3D Shape Reconstruction and Segmentation from ShapeNet Core55," CVPR 2017

[23] Benjamin Graham, Martin Engelcke, Laurens van der Maaten, "3D Semantic Segmentation with Submanifold Sparse Convolutional Networks," CVPR2017

[24] Naguib, Ahmed M., and Sukhan Lee. "An adaptive evidence structure for Bayesian recognition of 3D objects." Proceedings of the 9th International Conference on Ubiquitous Information Management and Communication. ACM, 2015.

[25] Xi Chen, Sukhan Lee, "Visual Search of an Object in Cluttered Environments for Robotic Errand Service", in proceedings of IEEE International Conference on Systems, Man, and Cybernetics (SMC), pp.4060–4065, October 2013

[26] Intelligent Systems Research Institute Database of Household Objects (ISRI_DB): http://isrc.skku.ac.kr/DB3D/db.php

[27] S. Lee, L. Wei and A. M. Naguib, "Adaptive Bayesian recognition and pose estimation of 3D industrial objects with optimal feature selection," 2016 IEEE International Symposium on Assembly and Manufacturing (ISAM), Fort Worth, TX, 2016, pp. 50-55.

[28] J. Kim, D. Kim, J. Seo, S. Lee, and Y. Park, "Octree-Based Obstacle Representation and Registration for Real-Time," 2007 International Conference on Mechatronics and Information Technology (ICMIT).

[29] S. Ren, K. He, R. Girshick and J. Sun, "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 39, no. 6, pp. 1137-1149, June 1 2017.

[30] Olga Russakovsky*, Jia Deng*, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg and Li Fei-Fei. (* = equal contribution) ImageNet Large Scale Visual Recognition Challenge. IJCV, 2015.

[31] Sukhan Lee, "Cognitive recognition and the homemate robot," 2011 IEEE International Conference on Systems, Man, and Cybernetics, Anchorage, AK, 2011, pp. 1-1.

[32] Bach, Sebastian, et al. "On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation." PloS one 10.7 (2015): e0130140.

[33] Mahendran, Aravindh, and Andrea Vedaldi. "Understanding deep image representations by inverting them." (2015): 5188-5196.