2013 IEEE/RSJ International Conference on
Intelligent Robots and Systems (IROS)
September 14-18, 2014. Chicago, IL, USA,

Unsupervised object individuation from RGB-D image sequences

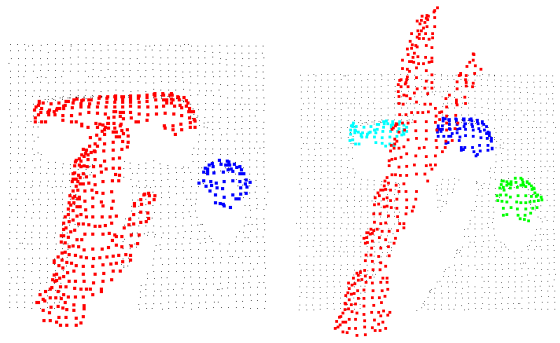
Seongyong Koo[1], Dongheui Lee[1] and Dong-Soo Kwon[2]

Abstract— **In this paper, we propose a novel unified framework for unsupervised object individuation from RGB-D image sequences. The proposed framework integrates existing location-based and feature-based object segmentation methods to achieve both computational efficiency and robustness in unstructured and dynamic situations. Based on the infant's object indexing theory, the newly proposed ambiguity graph plays as a key component of the framework to detect falsely segmented objects and rectify them by using both location and feature information. In order to evaluate the proposed method, three table-top multiple object manipulation scenarios were performed: stacking, unstacking, and occluding tasks. The results showed that the proposed method is more robust than the location-only method and more efficient than the feature-only method.**

I. INTRODUCTION

How can a robot distinguish individual objects from the visual sensor data? This is an important question for a robot to manipulate unknown objects in cluttered environments and understand human activities with various objects in indoor environments. Object individuation is a cognitive process of identifying each object as distinguished from others. In order to implement the process from visual sensor data, previous research broadly falls into two categories: individuation-by-location and individuation-by-feature. Individuation-by-location identifies each object by referring to their locations in one image [16], [10] or in sequential images [1], [20], while individuation-by-feature utilizes general features such as edge or color differences to characterize each object [3], [15] or distinguish each by using specific feature information defined in the object database [8], [2].

However, there are practical limitations to the existing approaches for a robot operating in *unstructured* and *dynamic* environment where unknown target objects constantly change their positions and shapes. First, the *unstructured* condition prohibits supervised methods in that a robot cannot obtain ground-truth information of unknown target objects in advance. For example, the approach needs prior knowledge of target objects as specified features [3], pre-defined models [1], [20], [15] and off-line learning [2]. On the other hand, unsupervised methods suffer from the *dynamic* condition in robustness against dynamic interactions of multiple objects. As an example, Fig. 1 shows two failed cases of an unsupervised clustering method [16] when a moving object becomes adjacent (contacted) to another or when some parts of the object are undetected due to occlusion or detection error.



(a) Contact (b) Occlusion

Fig. 1. Point-cloud data from typical close-point clustering methods in the presence of contact and occlusion. Three objects are represented by segmented point sets, each of which is represented in the same color.

There have been recent noteworthy works that tackle the robustness problem in unsupervised object individuation-by-location against *unstructured* and *dynamic* situations. In order to rectify segmentation errors in one point-cloud image, [13], [14] proposed detecting partial surfaces of occluded objects and grouping them using a SVM algorithm with trained object relations. Later, [19] improved real-time segmentation performance in a model-free manner. The method uses general features such as edge and surface normals to represent object hypotheses by using a graph-cut algorithm, and the occlusion problem is handled by a coplanarity check and curvature matching. Robustness problems in dynamic situations have been investigated by combining unsupervised segmentation and multiple object tracking (MOT) in point-cloud image sequences. [11] combined video object segmentation (VOS) and particle filtering based tracking of supervoxels, and their approach considers spatial-temporal coherent of the segments. [6] proposed a hierarchical spatiotemporal data association framework to unify the unsupervised clustering and multiple target tracking processes.

On the other hand, the robustness problem has been tackled by applying individuation-by-feature, which is normally a robust and supervised method, into an *unstructured* environment [7], [5]. Without prior knowledge of the target objects, multiple objects that move independently are represented as a adaptive Gaussian Mixture Models (GMM), respectively, and each model is updated from the feedback point-cloud data of simultaneous multiple object segmentation and tracking.

[1] Department of Electrical Engineering and Information Technology, Technical University of Munich, 80290 Munich, Germany. koosy@lsr.ei.tum.de, dhlee@tum.de

[2] Human-Robot Interaction Research Center, Department of Mechanical Engineering, KAIST, Deajeon, Republic of Korea. kwonds@kaist.ac.kr

The adaptive object model represents not only location but also feature information, and the model updated in the previous time step can be a reference for individuation-by-feature. Unsupervised individuation-by-feature method, however, requires expensive computation time to learn each object model at every time step.

Both individuation-by-location and individuation-by-feature have trade-off relations between efficiency and robustness, which in turn necessitates proper choice of the method depending on the given conditions such as the number of objects, shapes, and their spatial relations, e.g, stacking, occlusion, and adjacency. In *unstructured* and *dynamic* situations, in particular, the situation changes constantly, and one specific individuation method cannot achieve optimal performance in terms of robustness and computational efficiency in general. In order to achieve both aims, in this paper, we propose a unified framework for unsupervised object individuation where location and feature information of unknown objects are selected to be used depending on the situation changes. This framework is theoretically based on the human infants' object indexing mechanism, which has been proved in the area of cognitive science on the basis of the observation that an infant deploys two types of information for object individuation by cases [9]. The proposed framework provides an efficient information flow by adopting various existing methods as components of individuation-by-location, individuation-by-feature, and multiple object indexing. In addition, an ambiguity graph is proposed as a core bridge of the components, and it determines an individuation strategy depending on the object situation.

The rest of this paper is organized as follows. The next chapter introduces the theoretical background of the infants' object indexing theory. In chapter III, details of the proposed method are presented, and it is evaluated through experimental results in chapter IV. Finally, we conclude with discussions of further work in chapter V.

II. OBJECT INDIVIDUATION FRAMEWORK

In this chapter, we propose a framework for an unsupervised object individuation process from a point-cloud image sequence. The framework unifies individuation-by-location and individuation-by-feature methods based on the infants' object indexing theory, which is described in the following section.

A. Infants' object indexing theory

In the cognitive science area, the mechanism of the human object individuation process has been unveiled through several theories and observations regarding infant behavior. One finding is that an infant individuates objects by their locations not by features. For example, when two individual toys with different colors and shapes have been attached to each other and just one object moves independently, an infant shows very surprised action because of this unexpected situation [22]. However, there has been other evidence showing that an infant can distinguish independent moving objects by their features while excluding the location-difference effects. In the experiment in [21], a screen is introduced and two different shaped toys are brought out from the screen and returned in series, so that just one object is shown to an infant at a time. The infant shows expectation of two objects behind the screen [21]. These observations are evidence of infants' developmental steps for object individuation ability by using objects' locations and feature information in cases.

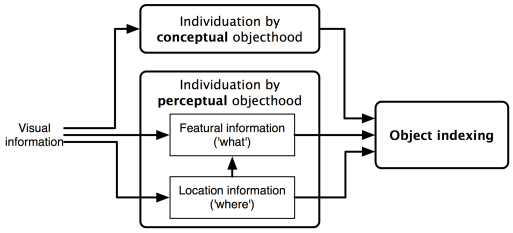


Fig. 2. A framework of the mature object indexing system by [9]

Based on the observations of the infant' object individuation process, [9] proposed an integrated mechanism by combining an adult object indexing theory, called *the mature object indexing system*. They suggested a combination of feature and location information to describe object individuation and object indexing processes as shown in Fig. 2. Basically 'where' information takes a primary role in individuating objects located in different positions. When object location is ambiguous because of occlusion or contact, the feature information that gives 'what' information of objects can play a significant role for individuation. However, in the cases of dynamically deformable and articulated objects where the feature information changes over time, the location history, which can be constructed by indexing each object in a time series, still plays a primary role in estimating existences and locations of individual objects. FINST (Fingers on INSTansiation), a theory of tracking multiple targets by introducing a 'finger pointing' metaphor to indicate on each object, is used for assigning an index on each object by only using the object's location information [12], [18] . In this manner, the mutual complementary relation between 'what' and 'where' information and object indexing allows robust and efficient individuation of dynamic and unknown objects.

B. Overview of the proposed system

Based on the infant's object indexing theory, an unsupervised object individuation framework that uses both location and feature information is proposed, as shown in Fig. 3. The observed point-cloud sensor data, $\mathbf{P}_k = \{\mathbf{p}_1, ..., \mathbf{p}_n\}$, wherein each point, $\mathbf{p}_i \in \mathbb{R}^6$, contains RGB color and 3-d position, is first individuated by only using location data; this entails a process of segmenting the initial point set data into several candidates of objects, $\mathscr{O}_k$, using the Euclidean clustering algorithm [17], [16]. It produces a set of pairs of each point and corresponding ID, $\{\mathbf{p}_i, o_i\}$. An individual object

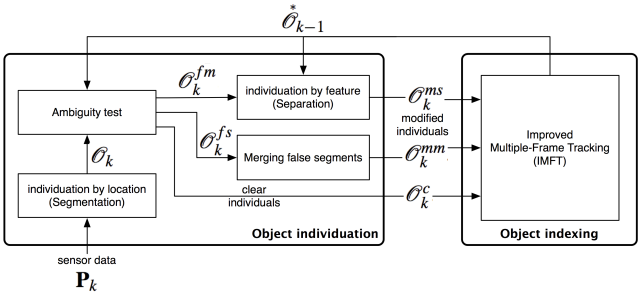


Fig. 3. A framework of object individuation process

candidate, $\mathbf{O}_j$, consists of points with the same IDs.

$$\begin{aligned} \mathscr{O}_k &= \{\mathbf{O}_1, ..., \mathbf{O}_{m_k}\}, \\ \mathbf{O}_j &= \{\mathbf{p}_i | o_i = j, j \in \{1, ..., m_k\}\}_{i=1}^{n} \end{aligned} \tag{1}$$

The initially constructed individuals are then modified in the rectification process which detects the false individuals and correct them by estimating true IDs of points involved in the false individual objects. Without any prior information of the objects from outside, the past objects in the previous time step $\overset{*}{\mathscr{O}}_{k-1}$ can be good references to estimate points involved in the false individual objects by comparing their features in a time sequence as long as the following assumption holds: an object does not change substantially in terms of its shape and position. In a real-time tracking task, this assumption can be thought of as valid, even for moving objects with a fast sampling rate. The ambiguity graph classifies the input object individuals $\mathscr{O}_k$ into the following three categories: clear individuals $\mathscr{O}^c_k$, falsely separated individuals $\mathscr{O}^{fs}_k$, and falsely merged individuals $\mathscr{O}^{fm}_k$. They are modified to be a set of true individuals by separation, $\mathscr{O}^{ms}_k$ and mergence, $\mathscr{O}^{mm}_k$. The rectification process will be described in the next chapter in detail.

A set of clear individuals and modified individuals then undergo the object indexing process. There are commonly occurring issues pertaining to assigning track indexes on multiple objects, such as different numbers of tracks, temporally missing points, and mismatched temporal associations. These issues arise mainly due to occlusion, objects moving in/out of a scene, and objects situated in densely populated environments [23]. In order to assign a track index to each object at every time step, the improved multi-frame tracking (IMFT) method is used to handle the typical tracking problems of generating a new track, deleting an old track, and correcting false matches due to noise and occlusions [6], [4].

The similarity measure between two objects must be defined to determine multi-object spatial associations in the ambiguity graph and temporal associations in IMFT. Here, each object individual consists of its point-cloud, $\mathbf{O}_j = \{\mathbf{p}_{j,i}\}_{i=1}^{n_j}$, and the distribution of the point-cloud in 3-d space can be approximately represented by a 3-d normal distribution, $\mathbf{p}_{j,i}|\mathbf{O}_j \sim \phi(\mathbf{x}|\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)$, where

$$\phi(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{\sqrt{(2\pi)^d |\boldsymbol{\Sigma}|}} \exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right). \tag{2}$$

The 3-dimensional mean $\boldsymbol{\mu}_j \in \mathbb{R}^3$ and covariance matrix $\boldsymbol{\Sigma}_j \in \mathbb{R}^{3\times 3}$ are calculated from the point-cloud data in the object $\mathbf{O}_j$. The similarity between two probability density functions can be calculated by using KL divergence and L2 distance representatively. For a single Gaussian case, both distances can be expressed by a closed-form solution as follows.

$$\begin{aligned} d_{L2}(\mathbf{O}_1, \mathbf{O}_2) &= \int (p(\mathbf{x}|\mathbf{O}_1) - p_(\mathbf{x}|\mathbf{O}_2))^2 d\mathbf{x} \\ &= 2 - 2\phi(\mathbf{0}|\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2, \boldsymbol{\Sigma}_1 + \boldsymbol{\Sigma}_2), \end{aligned} \tag{3}$$

$$\begin{aligned} d_{KL}(\mathbf{O}_1||\mathbf{O}_2) =& \frac{1}{2}(tr(\boldsymbol{\Sigma}_2^{-1}\boldsymbol{\Sigma}_1) + (\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1)^T \boldsymbol{\Sigma}_2^{-1}(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1) \\ & - \ln\left(\frac{|\boldsymbol{\Sigma}_1|}{|\boldsymbol{\Sigma}_2|}\right) - d) \end{aligned} \tag{4}$$

The weight function between the objects in IMFT is characterized by the symmetric KL distance between two objects.

$$d_{KL}(\mathbf{O}_1, \mathbf{O}_2) = d_{KL}(\mathbf{O}_1||\mathbf{O}_2) + d_{KL}(\mathbf{O}_2||\mathbf{O}_1) \tag{5}$$

Because the KL distance presents a smaller number with greater closeness of the two objects, the weight function is defined by (6), which has a value between 0 and 1 by introducing the maximum value of the distances in the all possible associations among the objects.

$$weight(\mathbf{O}_1, \mathbf{O}_2) = 1 - \frac{d_{KL}(\mathbf{O}_1, \mathbf{O}_2)}{\max_{i,j}(d_{KL}(\mathbf{O}_1, \mathbf{O}_2))} \tag{6}$$

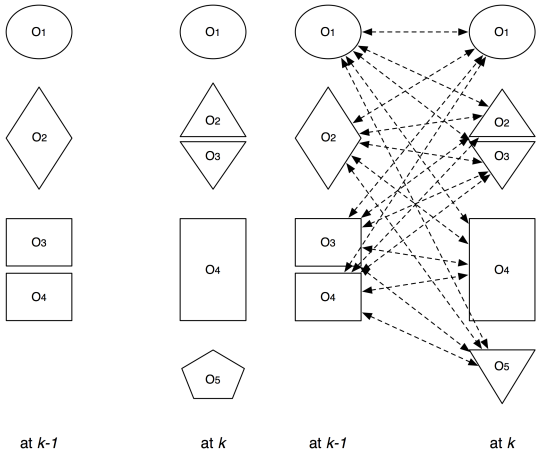
III. RECTIFICATION OF FALSE INDIVIDUALS

This chapter explains the main part of the proposed system: constructing an ambiguity graph to detect false object individuals by using graph theory, and rectifying them by using individuation-by-feature.

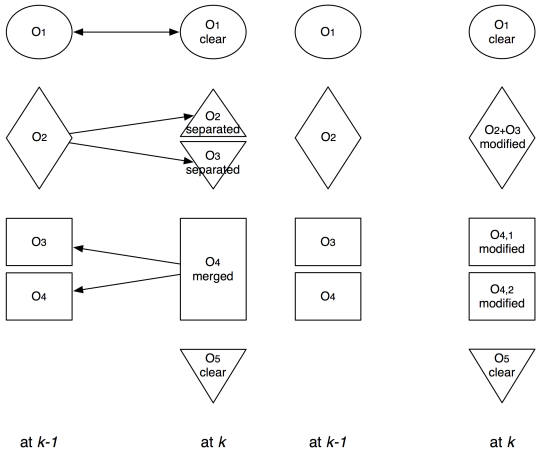
A. Ambiguity graph construction

Ambiguity graph is a directed graph whose nodes are grouped by true individual objects in $\overset{*}{\mathscr{O}}_{k-1}$ and object candidates in $\mathscr{O}_k$. When there exist falsely separated or merged objects in the two time frames, the graph shows the relations between an merged object and their segments by constructing directed edges. Here, we represent the *parent-child* relation as an edge heading from an original object (*parent*) to its segment (*child*).

Fig. 4 shows an example of the ambiguity graph and the rectification process for the object individuals at time k, $\mathscr{O}_k = \{\mathbf{O}^k_1, \mathbf{O}^k_2, \mathbf{O}^k_3, \mathbf{O}^k_4, \mathbf{O}^k_5\}$, with the previous objects at time $k-1$, $\overset{*}{\mathscr{O}}_{k-1} = \{\mathbf{O}^{k-1}_1, \mathbf{O}^{k-1}_2, \mathbf{O}^{k-1}_3, \mathbf{O}^{k-1}_4, \}$. The objects $\mathbf{O}^k_1$ and $\mathbf{O}^k_5$ are clear individuals because there exists a clearly similar object of $\mathbf{O}^k_1$ ($\mathbf{O}^{k-1}_1$) and no similar object of $\mathbf{O}^k_5$ at time $k-1$. The objects $\mathbf{O}^k_2$ and $\mathbf{O}^k_3$ are separated individuals from $\mathbf{O}^{k-1}_2$ by a certain occlusion or sensor noise, while the object



(a) Object individuals at time t and $t-1$ (b) Calculating weight values between the individuals



(c) Classified object individuals (d) Modified object individuals at time t

Fig. 4. A simplified example of the ambiguity test

$\mathbf{O}_4^k$ is composed of two merged objects $\mathbf{O}_3^{k-1}$ and $\mathbf{O}_4^{k-1}$ by their contact. As a result of the ambiguity test in Fig. 4(b), the three categories can be constructed as follows.

$$\begin{aligned} \mathscr{O}_k^c &= \{\mathbf{O}_1^k, \mathbf{O}_5^k\}, \\ \mathscr{O}_k^{fs} &= \{\{\mathbf{O}_2^{k-1}, (\mathbf{O}_2^k, \mathbf{O}_3^k)\}\}, \\ \mathscr{O}_k^{fm} &= \{\{\mathbf{O}_4^k, (\mathbf{O}_3^{k-1}, \mathbf{O}_4^{k-1})\}\} \end{aligned} \tag{7}$$

In the falsely segmented case, a *parent* object is $\mathbf{O}_2^{k-1}$, and $\mathbf{O}_2^k, \mathbf{O}_3^k$ are *child* objects. In the falsely merged case, similarly, $\mathbf{O}_4^k$ is a *parent* object, and $\mathbf{O}_3^{k-1}, \mathbf{O}_4^{k-1}$ are *child* objects.

Constructing the ambiguity graph can be illustrated as a combinatorial optimization problem to obtain a directed graph in Fig. 4(c), with which the three categories can be easily distinguished. In order to construct the directed graph, initially, all objects between $\overset{*}{\mathscr{O}}_{k-1}$ and $\mathscr{O}_k$ are fully connected with certain weight values on all arcs as in Fig. 4(b). Each weight value describes how similar two object individuals are, and the optimization problem is maximizing the sum of the constructed weights with following constraints.

- *constraint 1*: An object does not change substantially in terms of its shape and position.
- *constraint 2*: A segment of an object is not generated from more than two objects. (A *child* object only has one *parent* object.)
- *constraint 3*: A segment of an object cannot generate other object segments. (A *child* object cannot be a *parent* object, or vice versa.)

The first constraint results in defining the weight function between two objects, and the second and third conditions elicit the proposed optimization algorithm to generate the modified object individuals as shown in Fig. 4(d).

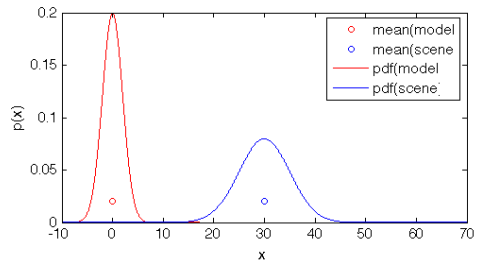
Because the tail and the head of an edge represent a *parent* object, $\mathbf{O}_p$, and a *child* object, $\mathbf{O}_c$, respectively, the value of an edge should reflect the relation of a *parent* and a *child* properly. In order to examine the effects of two distances in Eq. (3) and Eq. (5) for the *parent-child* relation, Fig. 5 shows an example of the distances between a fixed scene Gaussian distribution and a moving model Gaussian distribution. Let's assume that a scene distribution with a wide variance ($\sigma = 5$) is a *parent* object and a model distribution ($\sigma = 1$) is a *child* object. Figs. 5(b) and 5(c) show the change of each distance value when a model is moving from 0 to 60 in x direction. The similarity value of the ordered pair of two *parent* and *child* objects should represent the degree of belongingness of the *child* object into the *parent* object. The L2 distance in Fig. 5(b) remarkably indicates the similarity of two objects when their distributions are overlapped enough. This is a good measure for comparing closely-located two distributions locally, but cannot show the relative differences with other distributions globally. On the other hand, the unsymmetric two KL distances, $d_{KL}(\mathbf{O}_m||\mathbf{O}_s)$ and $d_{KL}(\mathbf{O}_s||\mathbf{O}_m)$, show the monotonic distance change in global area. In addition, the unsymmetric two distances show the relation of $d_{KL}(\mathbf{O}_m||\mathbf{O}_s) < d_{KL}(\mathbf{O}_s||\mathbf{O}_m)$, which is useful to represent the *parent-child* relation. Because the similar relation of two objects shows the smaller value of the KL distance, $d_{KL}(\mathbf{O}_c||\mathbf{O}_p)$ represents the value of the directed arc from *parent* node to *child* node, a_{pc}.

$$weight(a_{ij}) = weight(\mathbf{O}_i, \mathbf{O}_j) = d_{KL}(\mathbf{O}_j||\mathbf{O}_i) \tag{8}$$

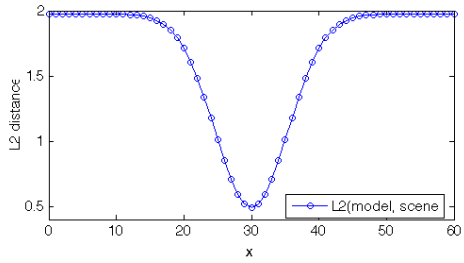
With the *constraint 1*: an object does not change substantially in terms of its shape and position; the fully connected directed graph in Fig. 4(b) is simplified by cutting weak edges, each of which has a smaller weight value than a certain threshold value. The threshold can be defined by using L2 distance in Fig. 5(b) because of its local effect and the bounded maximum value, 2.

$$d_{L2}(\mathbf{O}_i, \mathbf{O}_j)/2 > \alpha \Longrightarrow erase \;\; a_{ij}, \;\; 0 < \alpha < 1 \tag{9}$$

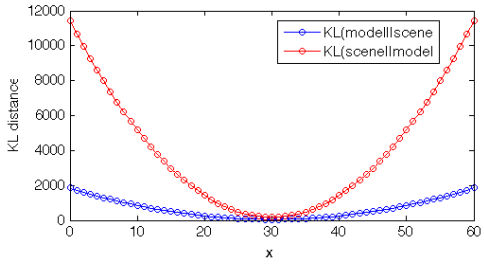
The *constraint 2* and *constrain 3* describing the relation of a *child* and a *parent* object are applied to derive a second graph cut algorithm. The process inspects the result graph of the first graph cut, and finds and deletes all false edges breaking the two conditions, as shown in Fig. 6(a), which results in the true edges remaining, as shown in Fig. 6(b).



(a) A model Gaussian distribution with a small variance and a scene Gaussian distribution with a wide variance.



(b) L2 distance between the fixed scene Gaussian distribution and the moving model Gaussian distribution from 0 to 60 in x.

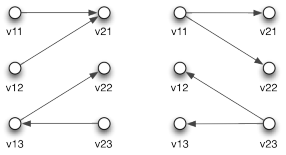


(c) Two KL distances between the fixed scene Gaussian distribution and the moving model Gaussian distribution from 0 to 60 in x.

Fig. 5. An example of the L2 distance and KL distance between a fixed wide scene Gaussian distribution and a moving small model Gaussian distribution.

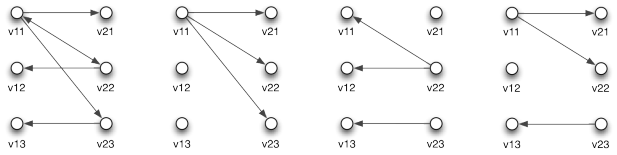
The second graph cut has multiple solutions from the results of the first graph cut as input. For example, Figs. 7(b) to 7(d) show possible solutions by deleting false edges of the graph in Fig. 7(a). In order to find the optimal solution, the summation of all weight values on the optimal true edge set should be minimal among all possible solutions. The greedy algorithm is proposed according to the following steps as illustrated in Fig. 8.

- Step 1: Finding all problematic nodes (red circle) and the related problematic arcs.
- Step 2: If there is no problematic node, the ambiguity graph is constructed.
- Step 3: If not, selecting a minimum valued problematic arc (black bold dotted arrow), and other problematic arcs that cause the minimum valued arc be problematic (red dotted arrow).
- Step 4: Deleting the selected problematic arcs and remaining the selected minimum valued arc to be true (black bold arrow).



(a) False edges. (b) True edges.

Fig. 6. Examples of the false and true edges according to the constraint 2 and 3 of the *parent* and *child* relations.



(a) Input graph (b) Solution 1 (c) Solution 2 (d) Solution 3

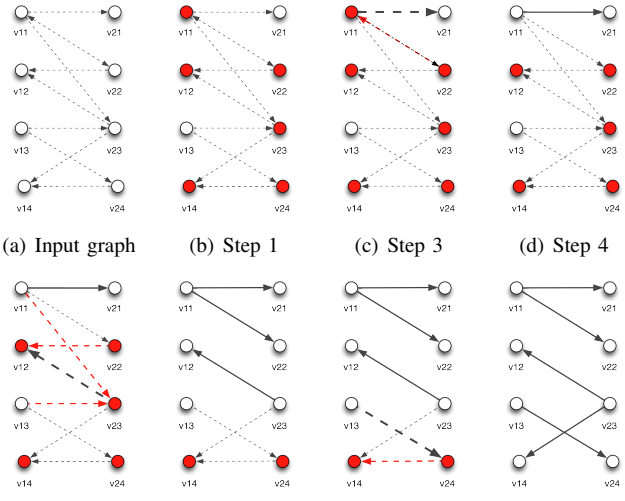
Fig. 7. Examples of possible multiple solutions of the second graph cut.

- Step 5: Go to step 1.

B. Rectification

Among the results of the ambiguity test, the separated individuals, $\mathcal{O}_k^{fs}$, and the merged individuals, $\mathcal{O}_k^{fm}$, must be manipulated to generate the modified individuals, $\mathcal{O}_k^{mm}$ and $\mathcal{O}_k^{ms}$, respectively. For the cases of separated individuals, $\mathcal{O}_k^{fs}$, the modified individuals can be constructed by a merging process which integrates all points in the false segments at time $k-1$.

The modification of the merged individuals, $\mathcal{O}_l^{fm}$, is processed by the individuation-by-feature process. Individuation-by-feature refers to the process of separating an falsely merged object individual, $\mathbf{O}^k$, by using shape and/or color information of their true object parts at time $k-1$, $\{\mathbf{O}_i^{k-1}\}_{i=1}^n$, where n is the number of object parts of $\mathbf{O}^k$. Because a point cloud measured from an object



(a) Input graph (b) Step 1 (c) Step 3 (d) Step 4

(e) Step 1 and 3 (f) Step 4 (g) Step 1 and 3 (h) Step 4

Fig. 8. Examples of greedy algorithm for the second graph cut process

reflects its shape and color information, the individuation-by-feature process performs point-level likelihood evaluation of all points in the merged object, $\mathbf{p}_j^k \in \mathbf{O}^k$, into the estimation of true object parts at time k, $\{\widehat{\mathbf{O}}_i^k\}_{i=1}^n$, thus producing new ID of each point to be modified, o_j^k.

$$o_j^k = \arg\max_i L(\widehat{\mathbf{O}}_i^k | \mathbf{p}_j^k) \tag{10}$$

In dynamic situations, the estimation of $\mathbf{O}_i^k$ can be predicted by using filtering-based tracking or robust 3D registration methods such as [1], [20], [11], [15], [3]. In this research, GMM-based robust 3-d registration with Gaussian Sum Filtering (GSF) method is used, which is proposed in the authors' previous work in [5], with considering many outliers which are points belonging to another object as in the *contact* case of Fig. 1(a).

IV. EXPERIMENTS AND RESULTS

The proposed object individuation framework aims to achieve both computational efficiency and robustness. In this chapter, the performance of the framework in these two aspects has been investigated with following three cases involving manipulating objects on a table-top. These cases frequently happen and cause difficulties in object individuation to understand human demonstrations.

1) Stacking objects: A human hand approach to and grasps one object and stack it up on another object.

2) Unstacking objects: After all objects are piled up, they are unstacked in series by the hand.

3) Human hand occluding objects: A human hand moves over other objects to partially and completely occlude them from the camera view.

The proposed method was evaluated based on the point-cloud data sequence in the tasks. A RGB-D camera (ASUS Xtion) established at a height of 90cm on a table captured the point-cloud data sequence at 30Hz. In order to reduce the data size, a workspace was defined as a half sphere with 50cm radius on the table. The captured point-cloud was down-sampled with 10mm sampling distance by using VoxelGrid filter, and the surface of the table was excluded by using plane extraction in [17]. All experiments were performed using an Intel i7-3770 3.4GHz CPU, and the software was implemented based on ROS (Robot Operating System) platform[1].

A. Qualitative evaluation

Fig. 9 shows the point-cloud image sequences of the object individuation results for the given three tasks. The task involves the problems of one object coming into contact with others and moving together in contact. In particular, multiple contacts and partial occlusions arise between objects at the same time in the stacking and unstacking tasks. As shown in the second rows in Figs. 9(a) and 9(b), the individuation-by-location process results in falsely merged segments which

[1]The proposed algorithm was implemented based on ROS, and the open source code written in C++ and dataset can be found in http://www.hri.ei.tum.de

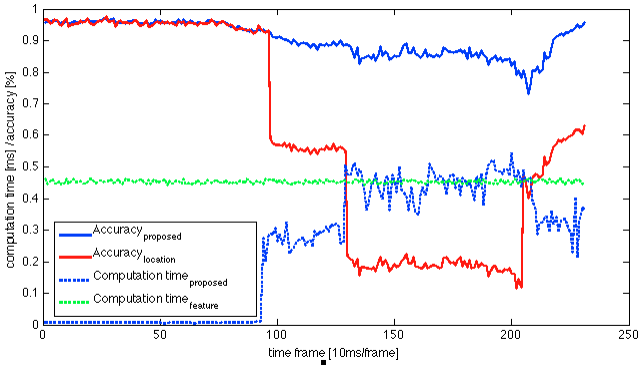


Fig. 10. Performance comparison in the two-object stacking case.

are represented with same colored point-clouds. The results of the proposed method, bottom rows in each figure, show the robust object individuation and assigned track IDs in the multiple-contact cases and even dynamic changes of the objects' shapes, positions and orientations. In the occlusion cases, the partially and completely occluded object segments in the second rows in Fig. 9(c) can be recovered as the result of the proposed method, as shown in the bottom rows.

B. Quantatitive evaluation

In order to evaluate the performance of the proposed method, the individuation accuracy and the computation time were measured and compared with the individuation-by-location [16] and individuation-by-feature [7] methods. The accuracy is computed as PMOTA (Point-level Multiple Object Tracking Accuracy) in [6]. In order to obtain the ground-truth data, the stacking and unstacking task were performed again with distinct colored two objects (white and black). The test algorithms used only 3d position data.

$$PMOTA = 1 - \frac{\sum_t \sum_i^{\mathcal{O}_t} m_i^t + fp_i^t + mme_i^t}{\sum_t \sum_i^{\mathcal{O}_t} n_i^t} \tag{11}$$

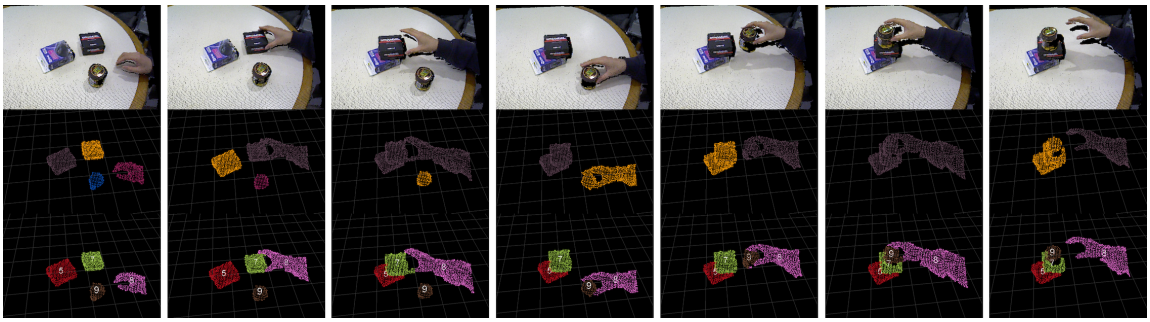
Table I and Fig. 10 show the results. The proposed method performes better than individuation-by-location in terms of accuracy in two-object contacts (frame 97 to 129, frame 205 to 231) and three-object contacts (frame 130 to 205), and efficiently computes more than individuation-by-feature in no contact and two-object contacts.

TABLE I
PMOTA OF THREE METHODS FOR THE STACKING CASE

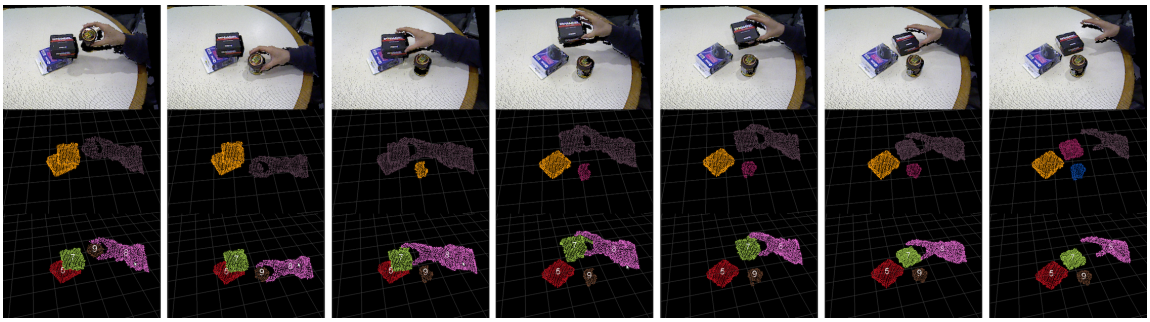
Method	Accuracy		Computation time	
	mean	std	mean	std
Proposed method	0.9034	0.0526	0.2346	0.1989
Individuation-by-location	0.5995	0.3305	0.0062	0.0005
Individuation-by-feature	0.9042	0.0474	0.4532	0.0056

V. CONCLUSION

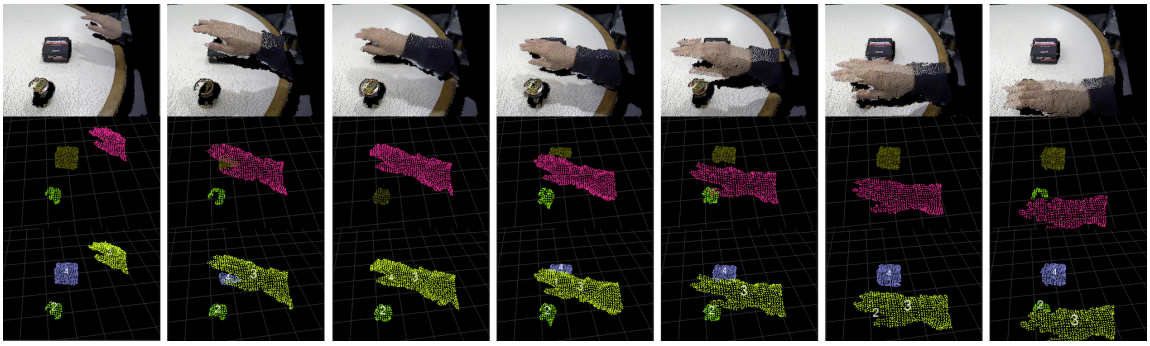
In this paper, we proposed a novel framework for unsupervised object individuation with an RGB-D camera. The method contributes to providing a theoretical background



(a) Stacking three objects



(b) Unstacking three objects



(c) Occluding objects

Fig. 9. Object individuation results of the three table-top manipulation tasks. The first rows show original captured RGB-D image sequence, the second rows show the point-cloud image sequence of the individuation-by-location results, and the bottom rows show the results of the proposed method. Each color shows each object individual and the numbers on the objects represent the track ID of the objects.

from infant's cognitive developmental theory and integrating diverse algorithms to achieve two trade-off objectives: efficiency and robustness. The newly proposed ambiguity graph plays a key component as a bridge between two different individuation methods by evaluating the given multiple object situations.

Although the result showed the feasibility and efficiency of the method, the performance of robustness is limited due to the two-frame ambiguity test and the *constraint 1*. When a object moves fast enough to break the *constraint 1* or the ambiguity graph is falsely constructed due to the noise at a certain frame, the graph cannot have a chance to be recovered after that frame. In further work, this problem will be investigated and the ambiguity graph will be improved in multiple time frames. In addition, the method can be expanded to learning human demonstrations in multiple object manipulation tasks by using dynamic spatial relations of the multiple objects represented in the ambiguity graph.

ACKNOWLEDGMENT

This work is supported partially by Technical University Munich - Institute for Advanced Study, funded by the German Excellence Initiative, and partially by the Industrial Strategic Technology Development Program (10044009, Development of a self-improving bidirectional sustainable HRI technology), funded by the Ministry of Knowledge Economy(MKE), Korea.

REFERENCES

[1] Changhyun Choi and Henrik I Christensen. Rgb-d object tracking: A particle filter approach on gpu. In *Intelligent Robots and Systems (IROS), 2013 IEEE/RSJ International Conference on*, pages 1084–1091. IEEE, 2013.

[2] Michael Firman, Diego Thomas, Simon Julier, and Akihiro Sugimoto. Learning to discover objects in rgb-d images using correlation clustering. In *Intelligent Robots and Systems (IROS), 2013 IEEE/RSJ International Conference on*, pages 1107–1112. IEEE, 2013.

[3] Jared Glover and Sanja Popovic. Bingham procrustean alignment for object detection in clutter. *arXiv preprint arXiv:1304.7399*, 2013.

[4] S. Koo and D.-S. Kwon. Multiple people tracking from 2d depth data by deterministic spatiotemporal data association. In *2013 IEEE International Symposium on Robot and Human Interactive Communication*, pages 656–661. IEEE, 2013.

[5] Seongyong Koo, Dongheui Lee, and Dong-Soo Kwon. Gmm-based 3d object representation and robust tracking in unconstructed dynamic environments. In *2013 IEEE International Conference on Robotics and Automation (ICRA)*, pages 1106–1113. IEEE, 2013.

[6] Seongyong Koo, Dongheui Lee, and Dong-Soo Kwon. Multiple object tracking using an rgb-d camera by hierarchical spatiotemporal data association. In *Intelligent Robots and Systems (IROS), 2013 IEEE/RSJ International Conference on*, pages 1113–1118. IEEE, 2013.

[7] Seongyong Koo, Dongheui Lee, and Dong-Soo Kwon. Incremental object learning and robust tracking of multiple objects from rgb-d point set data. *Journal of Visual Communication and Image Representation*, 25(1):108–121, 2014.

[8] Simon Kriegel, Manuel Brucker, Zoltan-Csaba Marton, Tim Bodenmuller, and Michael Suppa. Combining object modeling and recognition for active scene exploration. In *Intelligent Robots and Systems (IROS), 2013 IEEE/RSJ International Conference on*, pages 2384–2391. IEEE, 2013.

[9] Alan M Leslie, Fei Xu, Patrice D Tremoulet, and Brian J Scholl. Indexing and the object concept: developing what' and where' systems. *Trends in cognitive sciences*, 2(1):10–18, 1998.

[10] Haowei Liu, Matthai Philipose, and Ming-Ting Sun. Automatic objects segmentation with rgb-d cameras. *Journal of Visual Communication and Image Representation*, 2013.

[11] Jeremie Papon, Tomas Kulvicius, Eren Erdal Aksoy, and Florentin Worgotter. Point cloud video object segmentation using a persistent supervoxel world-model. In *Intelligent Robots and Systems (IROS), 2013 IEEE/RSJ International Conference on*, pages 3712–3718. IEEE, 2013.

[12] Z. Pylyshyn. The role of location indexes in spatial perception: A sketch of the finst spatial-index model. *Cognition*, 32(1):65–97, 1989.

[13] Andreas Richtsfeld, Thomas Morwald, Johann Prankl, Michael Zillich, and Markus Vincze. Segmentation of unknown objects in indoor environments. In *Intelligent Robots and Systems (IROS), 2012 IEEE/RSJ International Conference on*, pages 4791–4796. IEEE, 2012.

[14] Andreas Richtsfeld, Thomas Mörwald, Johann Prankl, Michael Zillich, and Markus Vincze. Learning of perceptual grouping for object segmentation on rgb-d data. *Journal of Visual Communication and Image Representation*, 2013.

[15] Christian Rink, Zoltan-Csaba Marton, Daniel Seth, Tim Bodenmuller, and Michael Suppa. Feature based particle filter registration of 3d surface models and its application in robotics. In *Intelligent Robots and Systems (IROS), 2013 IEEE/RSJ International Conference on*, pages 3187–3194. IEEE, 2013.

[16] R.B. Rusu, N. Blodow, Z.C. Marton, and M. Beetz. Close-range scene segmentation and reconstruction of 3d point cloud maps for mobile manipulation in domestic environments. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS 2009)*, pages 1–6, 2009.

[17] R.B. Rusu and S. Cousins. 3d is here: Point cloud library (pcl). In *2011 IEEE International Conference on Robotics and Automation (ICRA)*, pages 1–4. IEEE, 2011.

[18] Lana M Trick and Zenon W Pylyshyn. Why are small and large numbers enumerated differently? a limited-capacity preattentive stage in vision. *Psychological review*, 101(1):80, 1994.

[19] André Ückermann, Robert Haschke, and Helge Ritter. Realtime 3d segmentation for human-robot interaction. 2013.

[20] Manuel Wuthrich, Peter Pastor, Mrinal Kalakrishnan, Jeannette Bohg, and Stefan Schaal. Probabilistic object tracking using a range camera. In *Intelligent Robots and Systems (IROS), 2013 IEEE/RSJ International Conference on*, pages 3195–3202. IEEE, 2013.

[21] Fei Xu and Susan Carey. Infants' metaphysics: The case of numerical identity. *Cognitive psychology*, 30(2):111–153, 1996.

[22] Fei Xu, Susan Carey, and Jenny Welch. Infants' ability to use object kind information for object individuation. *Cognition*, 70(2):137–166, 1999.

[23] A. Yilmaz, O. Javed, and M. Shah. Object tracking: A survey. *Acm Computing Surveys (CSUR)*, 38(4):13, 2006.