2013 IEEE/RSJ International Conference on
Intelligent Robots and Systems (IROS)
November 3-7, 2013. Tokyo, Japan

Multiple Object Tracking Using an RGB-D Camera by Hierarchical Spatiotemporal Data Association

Seongyong Koo, Dongheui Lee and Dong-Soo Kwon

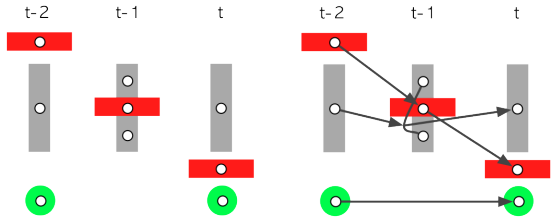
***Abstract*— In this paper, we propose a novel multiple object tracking method from RGB-D point set data by introducing the hierarchical spatiotemporal data association method (HSTA) in order to robustly track multiple objects without prior knowledge. HSTA is able to construct not only temporal associations between multiple objects, but also component-level spatiotemporal associations that allow the correction of falsely detected objects in the presence of various types of interaction among multiple objects. The proposed method was evaluated using the four representative interaction cases such as *split*, *complete occlusion*, *partial occlusion*, and *multiple contacts*. As a result, HSTA showed significantly more robust performance than did other temporal data association methods in the experiments.**

I. Introduction

The emergence of RGB-D cameras as one of several standard visual sensor systems for intelligent robots has promoted the development of point cloud processing technology [19] and its many applications such as environment reconstruction [18], object pose and shape estimation [4], [13], and tracking of human behavior [15], [16]. Tracking multiple objects from the visual information is one of several important and necessary abilities for a robot, allowing it to observe and perform complex tasks. This system produces each object's movement history, which is useful for learning complex actions by human demonstration, such as locating tasks [1] and assembly tasks [5].

Multiple object tracking from a set of point clouds involves many complex problems. The first issue is the representation and tracking of a single object. Once a target object can be specified, the object model is designed prior to the tracking process that is performed based on the pre-defined model [16], [14]. On the other hand, there have been two approaches for tracking unknown targets. One approach is model-based tracking, which uses the supervised learning technique to construct a model of an unknown object [4], [13]. Another more recent approach is model-free tracking, which uses particle filtering [21], interaction with a human [7], and incremental construction of an arbitrary model using the Gaussian Mixture Model (GMM) [10].

The second issue is robust tracking of multiple targets in the presence of interaction between objects such as occlusions, contact, and split. In addition, there are typical issues pertaining to multi-target tracking, such as different numbers of tracks, temporally missing targets, and mismatched temporal associations [22]. These problems can be formulated by constructing true associations (tracks) between nodes in a spatiotemporal graph in which each node represents each detected object, as shown in Fig. 1. When each single target is identified perfectly, temporal data associations, depicted as directed edges in Fig. 1, can be achieved using a multi-hypothesis tracker (MHT) in a stochastic manner [17] or a multi-frame tracker (MFT) in a deterministic way [20]. For example, [15] presented multiple people tracking from RGB-D data using a Combo-HOD person detector and MHT for temporal data association. Without previously constructed target models, however, distortion of each detected object can arise from interactions among multiple moving objects. At $t-1$ in Fig. 1, the gray object is occluded by the red object and separated into two different entities. In this case, the spatial association process is needed to combine the two falsely segmented parts, which are represented as the undirected edge in Fig. 1.



(a) Before association (b) After association

Fig. 1. Example of a spatiotemporal graph of multiple objects. Colored entities are used to represent each object. Directed edges show temporal associations; an undirected edge shows spatial association of two partitions of an identical object (gray), which are separated by the occlusion of another object (red).

In the area of 2D image processing, there have been several recent works that have attempted to construct spatiotemporal data associations between multiple targets wherein the above two problematic issues were tackled synthetically [2], [23]. However, in the filed of 3D RGB-D point cloud data processing, these issues have not been investigated sufficiently despite their importance in many applications, especially in the field of intelligent robots. In order to robustly track multiple objects without prior knowledge, we have proposed the model-free multiple object tracking framework in [11]. This framework has the merits of both flexibility in incremental learning and robustness in tracking multiple

S. Koo is with Mechanical Engineering Department and HRI Research Center, KAIST, Deajeon, Republic of Korea. koosy@robot.kaist.ac.kr

D. Lee is with Faculty of Department of Electrical Engineering and Information Technology, Technical University of Munich, 80290 Munich, Germany dhlee@tum.de

D. Kwon is with Faculty of Mechanical Engineering Department and the Director of HRI Research Center, KAIST, Deajeon, Republic of Korea kwonds@kaist.ac.kr

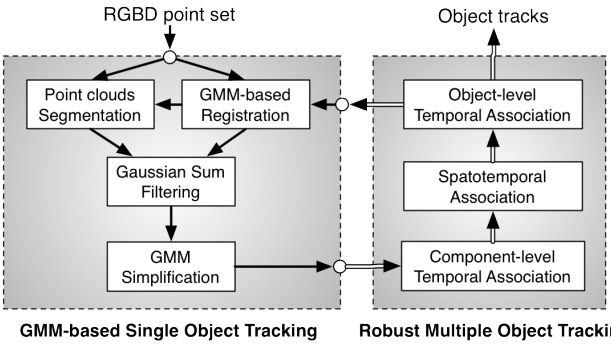


Fig. 2. Overview of the model-free multiple object tracking method

unknown objects because of the feedback connection of the two components: GMM-based single object tracking and robust multiple object tracking, as in Fig. 2. The details and performance of the single object tracking part can be found in [10], which will be briefly summarized in the next chapter. In this paper, we propose a novel hierarchical spatiotemporal data association (HSTA) method for robust multiple object tracking.[1] HSTA is able to construct not only temporal associations between multiple objects, but also component-level spatiotemporal associations that allow the correction of falsely detected objects in the presence of various types of interaction among multiple objects, as in Fig. 1. The proposed method was evaluated using the four representative interaction cases such as *split*, *complete occlusion*, *partial occlusion*, and *multiple contacts*. As a result, HSTA showed significantly more robust performance than did other temporal data association methods in the experiments.

II. GMM-BASED SINGLE OBJECT TRACKING

In order to represent each object stochastically, a GMM is constructed from the RGB-D point set data involved in the object $O = \{p_1, ..., p_n\}$, each of which consists of the 3D position and the RGB color information of a point, $p_i \in \mathbb{R}^6$.

$$p(\mathbf{x}) = \sum_{i=1}^{k} w_i \phi(\mathbf{x}|\mu_i, \Sigma_i), \quad \sum_{i=1}^{k} w_i = 1 \tag{1}$$

$$\phi(\mathbf{x}|\mu, \Sigma) = \frac{1}{\sqrt{(2\pi)^6|\Sigma|}} \exp\left(-\frac{1}{2}(\mathbf{x}-\mu)^T \Sigma^{-1}(\mathbf{x}-\mu)\right) \tag{2}$$

Equation (1) represents the probability density of a 6-dimensional point $\mathbf{x}$ belonging to the object. At each time step, the captured RGB-D point set data is segmented into each object by comparing the maximum likelihood estimations of each point to the updated objects' GMMs from the previous multiple object tracking results. The segmented point set of each object is then down-sampled with a constant sampling distance using the VoxelGrid filter in [19] in order to construct the initial GMM by evenly weighted n Gaussians centered at each point with the same spherical covariance matrix. The initial GMM with n components is simplified into k components using the hierarchical clustering method [6] for computational efficiency. The two control parameters, the sampling distance and the ratio between n and k, determine the representation capacity of the object models, which, in turn, affect on the trade-off between the computation time and the tracking accuracy.[2]

Each constructed single object as a GMM (measurement distribution) is then filtered to estimate the current probability distribution (filtering distribution) of an object in the form of GMM. Once a dynamic model of a moving object can be obtained, the filtering distribution can be estimated using Gaussian Sum Filtering (GSF) by performing the time update and measurement update steps[3] recursively [12]. The time update step in GSF follows the Extended Kalman Filtering (EKF) method with a linearized one-step prediction model to produce the predictive distribution from the prior filtering distribution. In this study, the unknown dynamics of an object is approximated as a piece-wise linear function, and the prediction model is estimated using the GMM-based robust 3D registration method [8] at each time step. In the measurement update step, the target distribution can be obtained from the measurement distribution and the predictive distribution according to the Bayes' theorem and the Markov property. This contains the recursive products of two GMMs that result in the exponential growing number of Gaussians. The GMM simplification process is needed to limit the size of the Gaussians to the given number. In this study, the HC method [6] with L2 distance is used for the simplification process; the number of Gaussians is determined proportional to the size of the point set constructing an object.

III. HIERARCHICAL SPATIOTEMPORAL DATA ASSOCIATION

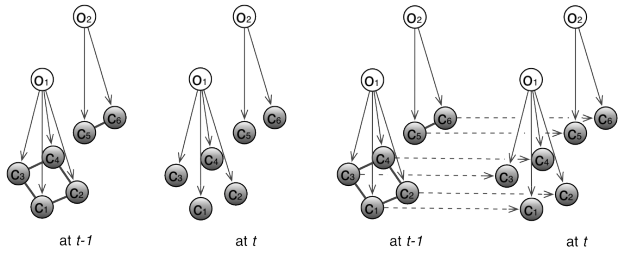
This chapter illustrates the proposed hierarchical spatiotemporal data association method (HSTA) in order to solve the second issue of multiple object tracking, addressed in the chapter I. An object O_i, represented as GMM (1), consists of k_i-components, $O_i = \{c_1^i, c_2^i, ..., c_{k_i}^i\}$, each of which represents a cluster of points. Once multiple objects are detected at time t, $\mathcal{O}_t = \{O_1^t, O_2^t, ..., O_{n_t}^t\}$, tracking is the process of constructing temporal associations between objects in $\mathcal{O}_{1:t}$. The proposed hierarchical spatiotemporal association not only constructs a temporal association on the object-level but also establishes temporal and spatial associations between components of each object. This component-level associations allow the correct generation of new objects with the spatiotemporal relations between components, which results in a modifying of the set of detected objects $\mathcal{O}_t$ into the correct objects $\mathcal{O}_t^\star$.

Fig. 3 shows an example of hierarchical spatiotemporal association, which corrects the detected objects $\mathcal{O}_t = \{O_1, O_2\}$ to the modified objects $\mathcal{O}_t^\star = \{O_1, O_2, O_3\}$, as in Fig. 3(a) to 3(c), and then constructs the temporal associations for

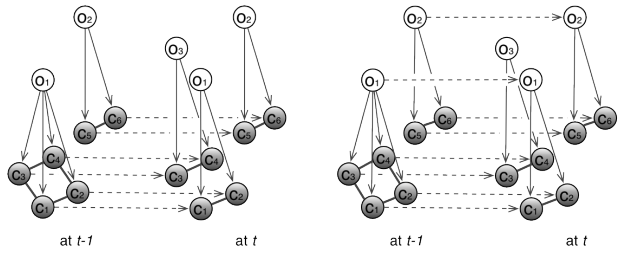
[1]The components of the method were originally proposed in the authors' previous paper [11]. In this paper, they are unified in the proposed HSTA and their performances are elaborately investigated in comparison with other approaches.

[2]An empirical study of this effect was performed and analyzed in [10].

[3]The details of the time update and measurement update processes are explained in [10].



(a) Detected clusters (b) Low-level temporal association



(c) association (d) High-level temporal association

Fig. 3. Example of hierarchical spatiotemporal data association. White nodes stand for objects and gray nodes represent detected components. Dashed directed edges are temporal associations in each level; undirected edges are spatial associations.

multiple objects, as in Fig. 3(d). The newly detected components at t in Fig. 3(a) make their temporal associations with existing components in the same object at $t-1$, as in Fig. 3(b). The temporal associations allow the estimation of the movement differences between the components, and with the position differences, produce spatiotemporal relations between components, which disconnect the relations of $c_1 - c_3$ and $c_2 - c_4$ and split $O_1^{t-1} = \{c_1, c_2, c_3, c_4\}$ into $O_1^t = \{c_1, c_2\}$ and $O_3^t = \{c_3, c_4\}$, as in Fig. 3(c).

The temporal data association process is required to solve typical multiple object tracking problems, i.e., generating and removing an object, correcting false matching, and the robustness against occlusion. Multi-Hypothesis Tracking (MHT) [17] is one of the representative probabilistic methods which can give a globally optimal solution for the problems. Multi-Frame Tracking (MFT) [20], which find the optimal pairs of associations to maximize the sum of weight values on the matched correspondences using the greedy method, is more robust for a large variety of motions and more computationally efficient for a large number of targets than the MHT method. In [9], improved MFT (IMFT) has been proposed to enhance the computational efficiency in the long-term complete occlusion case. The IMFT method is used as a preliminary tool for the robust and efficient temporal association in the HSTA.

A. Component-level temporal association

In any object at each time frame t, each component is a new node in the t frame of IMFT. The only parts necessary to construct an IMFT for components are the definition of the weight function between two components and the size of the time frame k to make the associations in the frames. k is determined by the given situations to consider the length of the complete occlusion time. The weight function between two components that are represented as Gaussians, $c_1 = \{\mu_1, \Sigma_1\}$ and $c_2 = \{\mu_2, \Sigma_2\}$, is determined using the L2 distance. With the property of a Gaussian function, $\int \phi(\mathbf{x}|\mu_1, \Sigma_1)\phi(\mathbf{x}|\mu_2, \Sigma_2)d\mathbf{x} = \phi(\mathbf{0}|\mu_1 - \mu_2, \Sigma_1 + \Sigma_2)$, the L2 distance of c_1 and c_2 can be expressed as follows.

$$\begin{aligned} d_{L2}(c_1, c_2) &= \int (p_{c_1}(\mathbf{x}) - p_{c_2}(\mathbf{x}))^2 d\mathbf{x} \\ &= 2 - 2\phi(\mathbf{0}|\mu_1 - \mu_2, \Sigma_1 + \Sigma_2) \end{aligned} \tag{3}$$

Because the matching algorithm in MFT maximizes the sum of weight values in the matched associations, the L2 distance, that is 0 for the closeness components, is converted to a weight value between 0 and 1 by introducing the maximum value of the distance in the graph.

$$weight(c_1, c_2) = 1 - \frac{d_{L2}(c_1, c_2)}{\max_{i,j}(d_{L2}(c_i, c_j))} \tag{4}$$

B. Spatiotemporal association

In order to represent the spatial and temporal relations among the components in an object, the relations can be represented as a topological graph where each node represents each component and the undirected edge between two nodes. Each edge contains the weight value of the association. In order to reflect the spatial and temporal relations, the weight value of an edge is determined as a convex combination of the differences of position and velocity between two components, which is controlled by a parameter $0 \leq \alpha \leq 1$.

$$w(e_{i,j}) = \alpha \times w_{pos}(e_{i,j}) + (1 - \alpha) \times w_{vel}(e_{i,j}) \tag{5}$$

The position difference of two components is defined by the normalized KL distance in an object and converted into a weight of association between 0 and 1 as follows.

$$w_{pos}(e_{i,j}) = 1 - \frac{d_{KL}(c_i, c_j)}{\max_{i,j}(d_{KL}(c_i, c_j))} \tag{6}$$

Because each object has a different size, the closeness of two components in the object should be relatively determined. The normalized value regardless of the object size is useful to determine a common threshold value to cut off the association later. In the case of defining the relation between spatially distributed components, the KL distance is more appropriate than the L2 distance due to its global effect. The symmetrised KL distance (7) is used for the undirected edges.

$$\begin{aligned} &d_{KL}(c_1, c_2) = d_{KL}(c_1||c_2) + d_{KL}(c_2||c_1), \quad where \\ &d_{KL}(c_1||c_2) = \\ &\frac{1}{2}\left(tr(\Sigma_2^{-1}\Sigma_1) + (\mu_2 - \mu_1)^T\Sigma_2^{-1}(\mu_2 - \mu_1) - \ln\left(\frac{|\Sigma_1|}{|\Sigma_2|}\right) - 6\right) \end{aligned} \tag{7}$$

The weight value for referring temporal property is a normalized velocity difference between two components. Because each component already has a historical track from the component-level temporal associations, the velocity can be

calculated by the change of position vectors in a component as follows.

$$w_{vel}(e_{i,j}) = 1 - \frac{d_{vel}(c_i, c_j)}{\max_{i,j}(d_{vel}(c_i, c_j))}, \quad where$$
$$d_{vel}(c_i, c_j) = \left| (\mu_i^t - \mu_i^{t-1}) - (\mu_j^t - \mu_j^{t-1}) \right| \tag{8}$$

The fully connected topological graph with initial weight values of all edges needs to be simplified to construct a meaningful topology of the object. The weight value of each edge is tested with a threshold value, th_{edge}, to erase the edge in the graph.

As in Fig. 3(c), the constructed spatiotemporal association needs to be separated in cases of generating new objects. Because the weight value of an edge represents the closeness of two components in terms of the spatial positions and temporal movements, two individual objects are easily disconnected when they move in different ways as in the *split* case. On the other hand, even if there is a *partial occlusion* of an object, components are not easily disconnected when they have the same movement pattern. Therefore, a connectivity test of the topological graph can decide the separation of one object only in the *split* case. Each separated graph constructs a new object with its components as shown in Fig. 3(c).

C. Object-level temporal association

Like the component-level temporal association, object-level temporal association can be constructed using the IMFT method. The modified objects $\mathscr{O}_t^\star$ after the spatiotemporal association process construct the nodes in a new frame at t; the weight function is characterized by the L2-distance of the objects' GMMs, e.g., $O_1 = \{k^1, \mathbf{w}^1, \boldsymbol{\mu}^1, \boldsymbol{\Sigma}^1\}$ and $O_2 = \{k^2, \mathbf{w}^2, \boldsymbol{\mu}^2, \boldsymbol{\Sigma}^2\}$, as follows.

$$\begin{aligned} d_{L2}(O_1, O_2) = &\sum_{i=1}^{k^1}\sum_{j=1}^{k^1} w_i^1 w_j^1 \phi(0 | \mu_i^1 - \mu_j^1, \Sigma_i^1 + \Sigma_j^1) \\ &- 2\sum_{i=1}^{k^1}\sum_{j=1}^{k^2} w_i^1 w_j^2 \phi(0 | \mu_i^1 - \mu_j^2, \Sigma_i^1 + \Sigma_j^2) \\ &+ \sum_{i=1}^{k^2}\sum_{j=1}^{k^2} w_i^2 w_j^2 \phi(0 | \mu_i^2 - \mu_j^2, \Sigma_i^2 + \Sigma_j^2) \end{aligned} \tag{9}$$

Because the L2-distance presents a smaller number with greater closeness of the two objects, the weight function is defined by (10), and takes a value between 0 and 1.

$$weight(O_1, O_2) = 1 - \frac{d_{L2}(O_1, O_2)}{\max_{i,j}(d_{L2}(O_i, O_j))} \tag{10}$$

IV. Experiments and results

The purpose of this study is to track multiple objects robustly in the presence of interaction between objects. We conducted several experiments to examine the performance of the proposed HSTA method in the following four cases.

- *Split:* One object is separated into two different objects, like a hand putting an object on a table.
- *Complete occlusion:* One object is completely occluded by another object, like a hand covering a smaller object.
- *Partial occlusion:* One object is partially occluded by another object, like a hand passing over a bigger object.
- *Multiple contacts:* Multiple objects are in contact with each other and move independently.

A. Experimental environments

The experiments involve tracking human hands and multiple objects on a table.[4] The data is captured using a RGB-D camera (ASUS Xtion) established at a height of 90cm over the table. The size of the workspace is $70cm \times 70cm \times 70cm$ and the surface of the table is not included in the space. The computation device is an Intel i7 2.8GHz CPU; RGB-D point set data, with a size of 640×480, is captured at an average of 30Hz frequency. The data is then transformed into 6-dimensional point data (x, y, z, r, g, and b) with respect to the coordinates on the table. The data in the workspace enter the proposed tracking process as shown in Fig. 2.

The four cases were evaluated according to the two experimental scenarios shown in Fig. 4. Two human hands put small objects onto the table in consecutive order. When the hands place the objects on the table, the objects are separated from the hands (*split*) as shown in Fig. 4(a). Subsequently, each hand completely covers each object for a while and uncovers it at the same time. At that moment, the small objects are completely occluded by the hands (*complete occlusion*) as shown in Fig. 4(a). In addition, the occluded objects held in the two hands change their positions. The *partial occlusion* and *multiple contacts* cases were evaluated with two white boxes as shown in Fig. 4(b). Each hand passes over the two objects one at a time and holds each object. The hands make the two objects come into contact and move them in the direction of the rotation and translation.

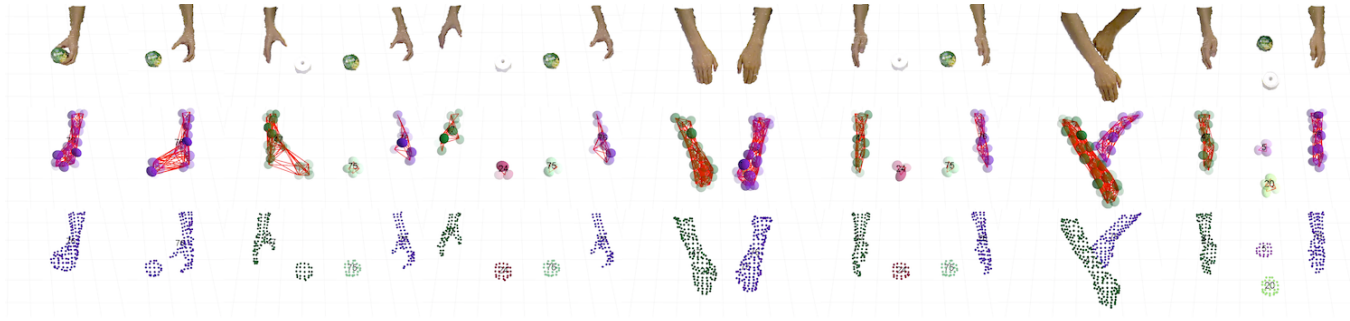
B. Evaluation results

In order to assess the performance of multiple object tracking methods, CLEAR MOT metrics were proposed in [3]. There are two metrics, multiple-object tracking precision (MOTP), which represents the ability to estimate precise object positions, and multiple-object tracking accuracy (MOTA), which is defined in order to account for all object tracking errors over all frames. In this studty, MOTA is used and modified for calculating point-level multiple object tracking accuracy (PMOTA). The three error components in MOTA, misses, false positives, and mismatches, are calculated by counting the number of error targets. Because in this study an object is represented by a set of points, PMOTA is defined by counting the number of error points in each object, as follows.

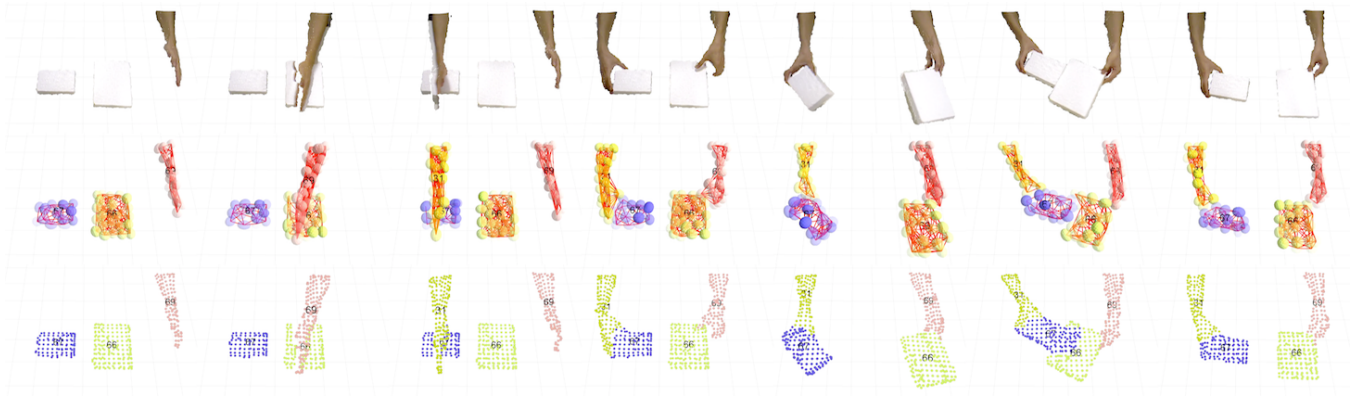
$$PMOTA = 1 - \frac{\sum_t \sum_i^{\mathscr{O}_t} m_i^t + fp_i^t + mme_i^t}{\sum_t \sum_i^{\mathscr{O}_t} n_i^t} \tag{11}$$

The number of misses of an object O_i^t, m_i^t, is defined by the entire number of points in the missed object, such as the small ball-in-hand of the first column in Fig. 4(a). The number of false positives, fp_i^t, is the number of falsely identified points in O_i^t like the blue points in the object (id:66) of the sixth column in Fig. 4(b). mme_i^t stands for the entire number of points in the object O_i^t if it is a mismatched object, like the two small balls in the last column of Fig.

[4]The movie clips of the experiments can be found at the official webpage of this work: (http://robot.kaist.ac.kr/project/pmot).



(a) Test of the *split* and *complete occlusion* cases



(b) Test of the *partial occlusion* and *multiple contact* cases

Fig. 4. Snapshots of the tracking multiple objects in the sequence (from left to right) of the movements. The first row of each figure show the original captured point set data. The second row illustrate Gaussian mixture models as a set of 3D ellipsoids with tempo-spatial topological graph. The tracking results of the proposed algorithm are depicted in the figures on the third row.

4(a). In these experiments, ground truth data for all points are manually labeled by referring to the RGB-D data, as in the first rows of Fig. 4.

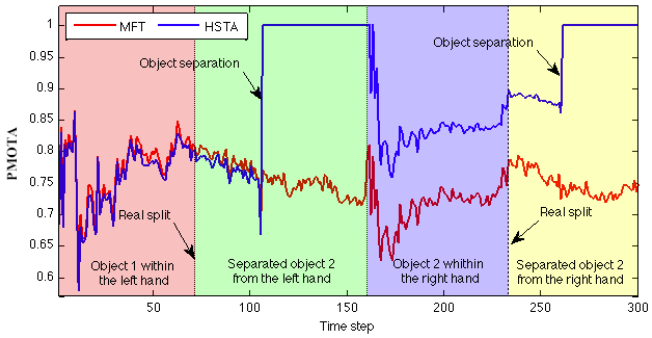
In order to analyze the proposed HSTA algorithm for multiple object tracking, two comparators were performed as the object-level temporal association methods, MFT and Nearest Neighbor (NN) wherein a new object is matched to the closest object in the previous frame within a particular region of interest. All these methods used the identical single object tracking method of [10], with a 0.02m sampling rate and 0.15 simplification ratio. The HSTA method used 0.98 of α in (5) and 0.7 of th_{edge}. Table I shows the PMOTA results of the three different methods for the four cases; Fig. 5 shows examples of the change of PMOTA according to the time step in the two meaningful cases.

TABLE I

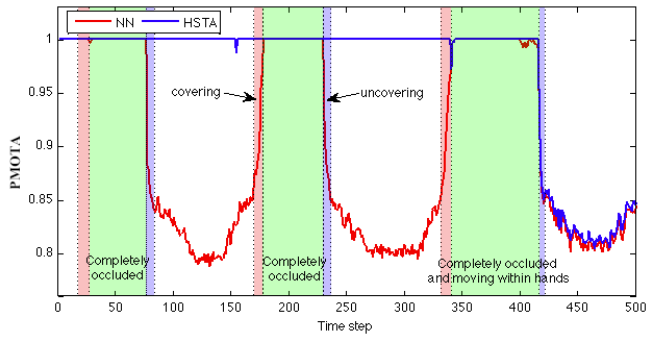
PMOTA OF THREE METHODS FOR THE FOUR CASES

Task	Split	Complete occlusion	Partial occlusion	Multiple contacts
NN	0.7326	0.8912	0.9652	0.8075
MFT	0.7346	0.9703	0.9745	0.8052
HSTA	0.8689	0.9752	0.9725	0.8095

In the *split* case, HSTA performs better than other methods because of the object separation process in the component-level spatiotemporal associations. The separation process



(a) The *split* case



(b) The *complete occlusion* case

Fig. 5. The change of PMOTA according to the time step

can be found in Fig. 4(a); PMOTA recovers to 1 after the separation as shown in Fig. 5(a). The effects of MFT allows the retaining of the tracks of a completely occluded object in a long-term period, as shown in Figs. 4(a) and 5(b). However, when occluded objects change their positions, they cannot be tracked, as is shown in the last column of Fig. 4(a) and the last part of the blue graph in Fig. 5(b). The tracking performance in the *partial occlusion* and *multiple contacts* cases is subject to the single object tracking method. In this study, all three algorithms yield similar results with the same single object tracking method. The details of the evaluation of the method can be found in [10].

V. Conclusion and further works

In this paper, we have presented a novel multiple object tracking method from RGB-D point set data by proposing the hierarchical spatiotemporal data association method (HSTA). The method uses component-level temporal association, spatiotemporal association, and object-level temporal association in order to enhance the robustness of tracking in the presence of various types of interaction among multiple objects such as *split*, *complete occlusion*, *partial occlusion*, and *multiple contacts*. In order to construct temporal associations on both levels, IMFT was applied due to its robustness and computational efficiency, and the spatiotemporal association process constructs a topological graph of an object that contributes to the generation of new objects in the *split* case.

The proposed method was evaluated using two experimental scenarios including four types of interaction; the results show that this method successfully attains better performance than that of other temporal association methods. Although the results showed the feasibility of the algorithm, there are some areas that can be supplemented in further work. First, as shown in Fig. 5(b), occluded objects such as objects-in-hand were not correctly tracked anymore when they moved within other objects. This problem can be tackled by considering the *merge* case or by using task-level knowledge. In the future, the HSTA structure will be extended to construct object-level spatial relations. Second, the structure model of an object was constructed by introducing the topological graph, but this graph was not directly used to enhance the single object tracking performance. The automatically constructed structure model will be able to facilitate movement prediction for objects that are arbitrarily articulated and this, in turn, will improve the tracking accuracy.

Acknowledgment

This work was supported in part within the DFG excellence initiative research cluster "Cognition for Technical System-CoTeSys", and financially supported 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] E.E. Aksoy, B. Dellen, M. Tamosiunaite, and F. Worgotter. Execution of a dual-object (pushing) action with semantic event chains. In *2011 11th IEEE-RAS International Conference on Humanoid Robots (Humanoids)*, pages 576–583, 2011.

[2] C. Beleznai and D. Schreiber. Multiple object tracking by hierarchical association of spatio-temporal data. In *2010 17th IEEE International Conference on Image Processing (ICIP)*, pages 41–44, 2010.

[3] K. Bernardin and R. Stiefelhagen. Evaluating multiple object tracking performance: the clear mot metrics. *EURASIP Journal on Image and Video Processing*, pages 1–10.

[4] C. Choi and H. I. Christensen. 3d pose estimation of daily objects using an rgb-d camera. In *Intelligent Robots and Systems (IROS), 2012 IEEE/RSJ International Conference on*, pages 3342–3349, 2012.

[5] N. Dantam, I. Essa, and M. Stilman. Linguistic transfer of human assembly tasks to robots. In *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 237–242, 2012.

[6] J. Goldberger and S. Roweis. Hierarchical clustering of a mixture model. *Advances in Neural Information Processing Systems*, 17(NIPS-2004):505–512, 2005.

[7] K. Hausman, F. Balint-Benczedi, D. Pangercic, Z.-C. Marton, R. Ueda, K. Okada, and M. Beetz. Tracking-based interactive segmentation of textureless objects. In *IEEE International Conference on Robotics and Automation (ICRA)*, pages 1114–1121, 2013.

[8] B. Jian and B. Vemuri. Robust point set registration using gaussian mixture models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(8):1633–1645, 2011.

[9] 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*, 2013, accepted for publication.

[10] S. Koo, D. Lee, and D.-S 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, 2013.

[11] S. Koo, D. Lee, and D.-S. Kwon. Incremental object learning and robust tracking of multiple objects from rgb-d point set data. *Journal of Visual Communication and Image Representation*, 2013, in press.

[12] J.H. Kotecha and P.M. Djuric. Gaussian sum particle filtering. *IEEE Transactions on Signal Processing*, 51(10):2602–2612, 2003.

[13] M. Krainin, P. Henry, X. Ren, and D. Fox. Manipulator and object tracking for in-hand 3d object modeling. *The International Journal of Robotics Research*, 30(11):1311–1327, 2011.

[14] Z. Liu, D. Lee, and W. Sepp. Particle filter based monocular human tracking with a 3d cardbox model and a novel deterministic resampling strategy. In *2011 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 3626–3631, 2011.

[15] M. Luber, L. Spinello, and K.O. Arras. People tracking in rgb-d data with on-line boosted target models. In *2011 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 3844–3849, 2011.

[16] I. Oikonomidis, N. Kyriazis, and A. Argyros. Efficient model-based 3d tracking of hand articulations using kinect. *Procs. of BMVC, Dundee, UK (August 29–September 10 2011)[547]*, 2011.

[17] D. Reid. An algorithm for tracking multiple targets. *IEEE Transactions on Automatic Control*, 24(6):843–854, 1979.

[18] 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.

[19] 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, 2011.

[20] K. Shafique and M. Shah. A noniterative greedy algorithm for multiframe point correspondence. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(1):51–65, 2005.

[21] R. Ueda. Tracking 3d objects with point cloud library. pointclouds.org, 2012.

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

[23] Q. Yu and G. Medioni. Multiple-target tracking by spatiotemporal monte carlo markov chain data association. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(12):2196–2210, 2009.