2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)
Congress Center Hamburg
Sept 28 - Oct 2, 2015. Hamburg, Germany

Probabilistic graph based spatial assembly relation inference for programming of assembly task by demonstration

Yue Wang, Jie Cai, Yabiao Wang, Youzhong Hu, Rong Xiong, Yong Liu, Jiafan Zhang, Liwei Qi

Abstract— **In robot programming by demonstration (PBD) for assembly tasks, one of the important topics is to inference the poses and spatial relations of parts during the demonstration. In this paper, we propose a world model called assembly graph (AG) to achieve this task. The model is able to represent the poses of all parts, the relations, observations provided by vision techniques and prior knowledge in a unified probabilistic graph. Then the problem is stated as likelihood maximization estimation of pose parameters with the relations being the latent variables. Classification expectation maximization algorithm (CEM) is employed to solve the model. Besides, the contradiction between relations is incorporated as prior knowledge to better shape the posterior, thus guiding the algorithm find a more accurate solution. In experiments, both simulated and real world datasets are applied to evaluate the performance of our proposed method. The experimental results show that the AG gives better accuracy than the relations as deterministic variables (RDV) employed in some previous works due to the robustness and global consistency. Finally, the solution is implemented into a PBD system with ABB industrial robotic arm simulator as the execution stage, succeeding in real world captured assembly tasks.**

I. Introduction

Robot programming by demonstration (PBD) enables the robot to execute tasks demonstrated by the human users. It gives a new way to transfer the commands from human to robot rather than the traditionally machine programming, thus making the robot accessible to the non-programmer users. In the new generation of industry, the personalized demands are emphasized, requiring the assembly solution to be highly flexible and fast adaptive to various tasks. Therefore, application in industrial assembly of robot PBD system is directive and promising considering the significant lower cost for programming. A key problem in PBD based assembly is how to obtain the accurate poses of parts from human demonstration. Conventional techniques estimate the pose of each part independently from the image captured by a camera. In this way, the estimated poses may be inaccurate due to the image noises and occlusion, which might fail the assembly task as the tolerance of error is very limited in connection of two parts. For example in Fig. 1 left, the tracking and segmentation based methods give a noisy result with the axes of connections not aligned exactly, resulting in failure of assembly during execution.

In this paper, we propose a spatial assembly relation inference system to indirectly estimate the poses of all parts at the same time. The core idea is a probabilistic assembly graph (AG) that can represent all parts and their spatial assembly relations in an assembly task, but also the visual observations and contradictions between relations in a unified model. By inferring this model, the pose estimation can be expected to satisfy the PBD requirement. An example of result is shown in Fig. 1 right, in which the axes are aligned exactly. The framework of the spatial inference system is shown in Fig. 2. In our approach, the vision based techniques recognize the parts and the initial poses. Then these information are assigned to the observations of the AG. An expectation maximization (EM) is derived from AG to find the optimal configuration of relations. Finally a refinement is implemented to polish the final poses. The contributions of the paper are presented in four steps: 1) AG is presented that can represent poses of all parts, assembly relations, observations and prior knowledge, modeling all uncertainty generated during the demonstration of assembly task; 2) The contradiction between relations are modeled in AG as prior knowledge to improve the modeling of uncertainty; 3) An algorithm is proposed to simultaneously recognize the relations and estimate the pose, avoiding the divergence caused by the incorrect inference of relations; 4) The integration of model in the spatial inference system pipeline is introduced with a final refinement of results.

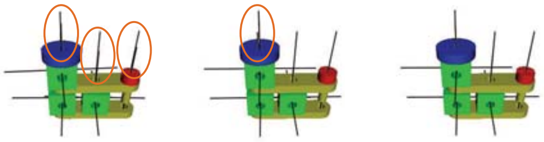


Fig. 1. The poses of parts using vision based techniques (left), RDV (middle) and AG (right). The orange circles indicate the axes are not aligned exactly.

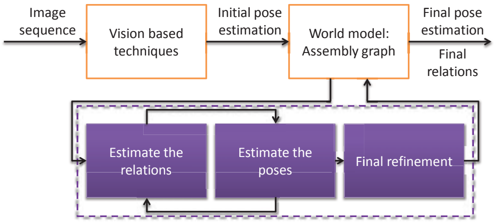


Fig. 2. The framework of spatial inference system to extract the assembly knowledge from the sequence of the images.

Yue Wang, Jie Cai, Yabiao Wang, Youzhong Hu, Rong Xiong and Yong Liu are with State Key Laboratory of Industrial Control Technology, Zhejiang University, Hangzhou, P.R. China. Jiafan Zhang and Liwei Qi are with ABB Corporate Research, P.R. China. Yong Liu is the corresponding author. yongliu@iipc.zju.edu.cn

978-1-4799-9994-1/15/$31.00 ©2015 IEEE

The remainder of the paper is organized as follows: In Section II, the related works are reviewed. In Section III, we introduce the statement of the problem as an AG inference and the problem solution. In Section IV, the implementation of spatial inference system is presented by integrating the AG. The experiments on both simulation and real world collected data are carried out to verify the correctness and effectiveness of the AG in Section V. Finally, a conclusion and future work is given in Section VI.

II. Related works

The pose estimation of parts were usually proposed in PBD systems. In [1], [2], [3], a system simulating the cognition, knowledge acquiring and transferring was proposed. A world model called EgoSphere was used to maintain the spatial information based on the tracking. This method was feasible when the part was visible in most time. When the occlusion occurred, the method would fail. In [4], [5], the hand tracker was combined to better localize the parts. In [6], supervoxel was used to track the trajectory of parts and hands with the Kinect sensor using the supervoxel. As these methods considered each part independently, the error in the connections of parts may fail the execution of the assembly task when the parts are relatively small and the task is complex. In [7], the spatial assembly relations were utilized to increase the localization of parts. In [8], the face-to-face (co-planar) test was used to estimate the pose of a new part. In [9], the co-planar was abstracted to relations to inference the manipulation. These methods utilized the relations between parts to improve the accuracy in a local scale. However, at each step, the relations and poses inferred before were regarded as determined variables. In this way, the incorrect recognition of relations would never be revised. In addition, the correctness of the recognition sometimes is unknown when the subsequent parts are not assembled. An illustrative experiment will be shown in Section V. In [10], a graph model was proposed to encode a human manipulation task. Then the semantics of object-action relations were learned from the graph model. As the target was not assembly, only simple relations were supported and the accuracy of the object pose was also not highlighted. The graph model was also applied in [11] as a world model of the workspace for robot manipulation. Their methods stated the problem as a sequential maximum-a-posteriori (MAP) problem and regarded object temporally add/remove as probabilistic dynamics. Therefore, the graph model based method is shown to be a possible effective representation for modeling PBD tasks.

Our prior work proposed in [12] applied a probabilistic graph to link the all parts with all relations, thus developing a global measure of the compatibility between relations. In this paper, we furthermore model the relations with uncertainty, so that the relations can be inferred from the graph model with contradictions between relations being taken into the consideration.

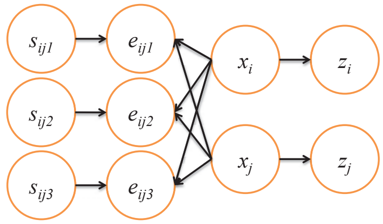


Fig. 3. A factor of AG graph model defined on two parts with 3 relations as example. Each virtual observation, e_{ijk}, is determined by x_i, x_j and the switch variable s_{ijk}. When $s_{ijk} = 1$, meaning that the relation exists, the virtual observation is a gaussian centering on 0, indicating that the poses of two parts satisfy the relation exactly. The visual observation z_i and z_j are independent when the poses are known.

III. Assembly graph representation, solution and refinement

The input of the AG is a series of parts $\{P_i\}$ with each one having two attributes: unique identity id_i and the observation of pose z_i. Denote $Z = \{z_i\}$, a set of all observations of poses. Both attributes are acquired from the precedent vision module. The output of the AG are $S = \{s_{ij}\}$ and $X = \{x_i\}$. The relation between P_i and P_j is encoded by s_{ij} with each element s_{ijk} as a binary variable indicating that whether the kth relation exist between P_i and P_j. The pose of P_i is x_i. As the relation s_{ijk} between the two parts assigns a constraint to the poses of two parts, we represent this constraint as a function $e_{ijk} = f_k(x_i, x_j)$. If the constraint is exactly satisfied, e_{ijk} is 0, which can be regarded as a virtual observation. Furthermore, e_{ij} is a concatenation of e_{ijk} and $E = \{e_{ij}\}$. The overall definition is represented in Fig. 3.

A. EM formulation

A world model for an assembled task with multiple parts can be factorized into a product of such forms as only binary relations is considered in the paper. With the definition of graph, the likelihood of observations Z is defined as

$$p(Z|X) = \prod_i p(z_i|x_i) \tag{1}$$

where $p(z_i|x_i)$ is in $N(x_i; z_i, \xi)$. The likelihood of virtual observations is

$$p(E|X,S) = \prod_{i,j} p(e_{ij}|x_i, x_j, s_{ij}) = \prod_{i,j}\prod_k p(e_{ijk}|x_i, x_j, s_{ij}) \tag{2}$$

where $p(e_{ijk}|x_i, x_j, s_{ijk})$ is defined as

$$p(e_{ijk}|x_i, x_j, s_{ijk}) = \begin{cases} N(f_k(x_i, x_j); 0, \sigma_k) & s_{ijk} = 1 \\ \theta_k & s_{ijk} = 0 \end{cases} \tag{3}$$

with θ_k being a penalty when $s_{ijk} = 0$.

Then the log-likelihood of all observations is derived as

$$\log p(Z, E|X) = \log p(Z|X) + \log \sum_S p(E, S|X) \tag{4}$$

where S are latent variables, and marginalized out to give $p(E|X)$. A possible way to find the solution that maximize the log-likelihood is the to use expectation maximization (EM) algorithm [13]. The algorithm is guaranteed to converge to a local minimum in an iterative execution of expectation step and maximization step. In expectation step, the posterior over S is found by utilizing the current X^t. In maximization step, the lower bound of the log-likelihood is maximized given the posterior.

B. Modeling of contradiction

The task of expectation step is to find the posterior over S as

$$p(S|E,X) = \frac{p(E|X,S)p(S)}{\sum_S p(E|X,S)p(S)} \tag{5}$$

The only undefined term is $p(s_{ij})$. Its simplest definition is that each s_{ijk} is independent to others, i.e. $p(s_{ij}) = \prod_k p(s_{ijk})$, leading to a factorization as

$$p(s_{ijk}|e_{ijk},x_i,x_j) = \frac{p(e_{ijk}|x_i,x_j,s_{ijk})p(s_{ijk})}{\sum_{s_{ijk}} p(e_{ijk}|x_i,x_j,s_{ijk})p(s_{ijk})} \tag{6}$$

This prior model means no contradiction exists between these relations, which however, is very rare in practice. In most situations, several possible relations between parts can not exist at the same time. The result caused by ignoring the contradiction is equivalent to modeling an impossible thing as possible.

To model the contradiction, s_{ij} should be modeled in multivariate joint distribution. To balance the complexity and accuracy, contradictions between two parts are modeled. Given two parts, all possible relations in s_{ij} are tested: For a pair of relations s_{ijm} and s_{ijn}, a compatibility hypothesis is generated and verified, which is to tell whether $p(s_{ijm} = 1, s_{ijn} = 1) > 0$. All contradicted pairs of relations $(p(s_{ijm} = 1, s_{ijn} = 1) > 0)$ are encoded in an undirected relation graph (RG). In RG, each node denotes a relation s_{ijk} and each edge between two nodes means this pair of relations HAS contradiction $(p(s_{ijm} = 1, s_{ijn} = 1) = 0)$.

With RG integrated into the graph, the new graph model is shown in Fig. 4, we have

$$p(g_{ijc}|e_{ij},x_i,x_j) = \frac{\prod_{s_{ijk}\in g_{ijc}} p(e_{ijk}|x_i,x_j,s_{ijk})p(g_{ijc})}{\sum_{g_{ijc}} \prod_{s_{ijk}\in g_{ijc}} p(e_{ijk}|x_i,x_j,s_{ijk})p(g_{ijc})} \tag{7}$$

where g_{ijc} is a clique in RG, containing s_{ijk} belonging to it. Controlled by the prior $p(g_{ijc})$, the corresponding posterior is non-zero only when zero or one relation in the subset exists.

C. Maximization step

We select classification EM (CEM), a variant of the original version to optimize the likelihood, since it is has lower complexity of computation [14]. Before the maximization step, a classification step is conducted in CEM, which is to find

$$\hat{g}_{ijc} = \arg\max p(g_{ijc}|e_{ij},x_i,x_j) \tag{8}$$

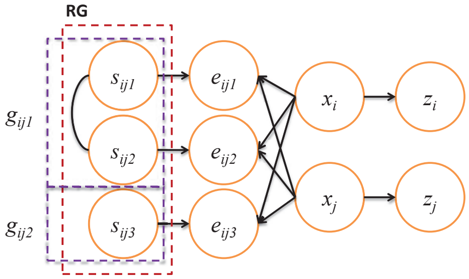


Fig. 4. The improved AG graph model definition on two parts by considering the contradiction. Now s_{ijk} are no longer independent but partitioned by RG. The undirected edge, connecting s_{ij1} and s_{ij2}, means that only one of them can exist at most because they are contradicted. s_{ij3} is still independent to relations in g_{ij1}, so g_{ij2} is defined on s_{ij3} solely.

As only one relation can exist at most in g_{ijc}, the complexity of this step is a linear with respect to the dimension of g_{ijc}. Denoting $\hat{s}_{ij} \triangleq \{\hat{g}_{ijc}\}$ then $\hat{S} = \{\hat{s}_{ij}\}$, we in maximization step optimize

$$\hat{X} = \arg\max \log p(Z|X) + \log p(E, S = \hat{S}|X) \tag{9}$$

where $\tilde{L}(Z,E|X)$ is called classification likelihood. The best X can be found by using gradient ascent algorithm easily.

D. Final refinement

At this step, the core AG model can be integrated into our spatial assembly relations inference system. The output of AG is the estimated poses of the parts $(\hat{X})$ as well as the recognized relations among them $(\hat{S})$. But the poses here may not satisfy all existed relations $\{\hat{s}_{ijk} = 1|\hat{s}_{ijk} \in \hat{S}\}$ exactly $(f_k(x_i,x_j) \neq 0)$ due to $p(Z|X)$ included in the likelihood. So in the completed system, a refinement step is conducted on the obtained relations to polish the poses of parts. The aim of the refinement is to find a solution maximizing $p(Z|X)$ among the set of X exactly satisfying the constraints of all relations, which is formally stated as

$$\hat{X} = \arg\max p(Z|X), \quad s.t. \; F(X, S = \hat{S}) = 0 \tag{10}$$

where $F(\cdot,\cdot)$ is a concatenation of all relations $f_k(x_i,x_j)$ among all possible pair i and j whose $\hat{s}_{ijk} = 1$. The AG result of pose estimation $\hat{X}$ can be employed as initial value. It means that the refined poses must satisfy the constraints of all relations. For those degrees of freedom unconstrained, the best solution we can achieve is the precedent visual observations, i.e. maximizing $p(Z|X)$. After this step, the final pose estimation of assembly tasks are finished.

IV. Implementation

In the experiments, the pose x_i of a part includes the 3D translation t_i and orientation R_i. Each part has its own coordinates defined by the user. In this local coordinates, the kth axis $\{a^o_{ik}\}$ or the kth plane $\{l^o_{ik}\}$ of this part are defined. With rigid geometry, the pose of axes and planes of this part can be computed in the world coordinates as $\{a_{ik}\}$ and $\{l_{ik}\}$. Each axis is defined by a point v and a direction u, then each

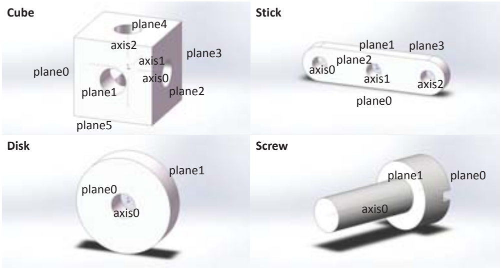


Fig. 5. The model of the parts employed in the experiment with their IDs of axis and plane.

sample as on the axis is defined as $as = v + \lambda u$ where λ is a real number. Each plane is defined by a normal vector n and a perpendicular distance d, then each sample ls on the plane is defined as $n^T ls + d = 0$. The parts employed in the experiments follow the definition shown in Fig. 5.

Based on the definition of planes and axes in the parts, we define two kinds of relations in the implemented system: the co-planar relation and the co-linear relation. In the co-planar relation, given two planes, say l_{11} and l_{21}, which are the the 1th plane of the 1th part and the 1th plane of the 2nd part in the world coordinates, we have each sample in l_{21} satisfies

$$f_1(x_1, x_2) = n_{21}^T (R_{21} ls_{11k} + t_{21}) + d_{21} = 0 \tag{11}$$

where R_{21} and t_{21} are the relative pose of plane l_{11} in coordinates of part P_2 computed using x_1 and x_2. This equation means that if these two planes are co-planar, samples on l_{11} can also be the samples in l_{21}. The samples are selected offline by the user and fixed, at least three samples are required for a plane. In the co-linear relation, given two axes, say a_{11} and a_{21}, we have each sample satisfying

$$\begin{aligned} f_2(x_1, x_2) = & v_{21} + u_{21}^T (R_{21} as_{11k} + t_{21} \\ & - v_{21}) u_{21} - (R_{21} as_{11k} + t_{21}) = 0 \end{aligned} \tag{12}$$

which indicates that samples on a_{11} can be the samples on a_{21} too. In the maximization step, these functions $f(\cdot, \cdot)$ form an objective function to be minimized (maximize the probability) while in the refinement, these functions act as equality constraints $f(\cdot, \cdot) = 0$ in (10).

The vision technique employed in the spatial inference system is off-the-shelf change detection method, which is also used in other PBD spatial inference systems. The area of change (AOC) is segmented by differentiating the image before and after a new part is assembled. The part is recognized by training a softmax classifier on the color histogram. Through investigating the overlap between AOC of the new part and the previous parts, the height of the top surface is inferred through accumulate the height of part. Finally we assign the initial pose of the new part with the pose of the minimum bounding box of the reconstructed surface.

V. Experiments

In experiments, two methods are included: The AG, which considers the relations as random variable; and the model considers the relations as deterministic variables, called RDV, meaning the relations are inferred from the result of vision algorithms and will not change anymore [9], [7], [15], [12]. We first compare AG and RDV on the simulated datasets, so that the performance of the world model can be evaluated independently without coupling the performance of prior modules, we can also add different levels of noise to test the performance thoroughly. The relation recognition is evaluated by F-measure, which is a balanced measure of precision and recall. After that the spatial assembly inference system is employed on the real world datasets for qualitative comparison. Then the system is implemented with an ABB industrial robotic arm simulator to execute demonstrated PBD tasks. The parts we used are modeled in the Section IV.

A. Simulated data

The performances of relation recognition using AG and RDV are compared using the F-measure as the evaluation metric, which is derived from precision and recall and defined as,

$$P = \frac{TP}{TP + FP}, \quad R = \frac{TP}{TP + FN} \tag{13}$$

where TP,FP and FN are true positives, false positives and false negatives respectively. Since high precision can be achieved by ignoring most relations while high recall can be achieved by keeping most relations exist, using either of the measure solely is unable to evaluate the performance. So the F-score is employed, which is computed as

$$F = \frac{2PR}{P + R} \tag{14}$$

The F-score ranges from 0 to 1 (the higher the better). As the accurate poses are known, the ground truth of this classification problem is that a relation exists if its corresponding $f(x_i, x_j) = 0$. There are 8 assembly tasks in our simulated datasets as shown in Fig. 6. The number of parts used in these datasets is from 2 to 9. We add 4 levels of noise to the real value as the observations. For each group (a noise level and a model), the test is conducted five times, of which the mean value is computed as the final result. These results are shown in Tab. I. The F-measure of both RDV and AG decreases with respect to the noise level. Besides, the AG gives better results in all noise levels.

Turn to the comparison of pose estimation, the error is evaluated by the translational error and rotational error [16] as follows.

$$err_{trans} = \frac{1}{N} \sum_i \|T_{g,i} - T_i\|^2 \tag{15}$$

$$err_{rot} = \frac{1}{N} \sum_i \|R_i^T R_{g,i} - I\|_F \tag{16}$$

where T and R are translation vector and rotation matrix derived from the pose x and g, N is the number of blocks in this case. The unit of translational error is millimeter (mm). The results are still the mean of five-time experiments. To

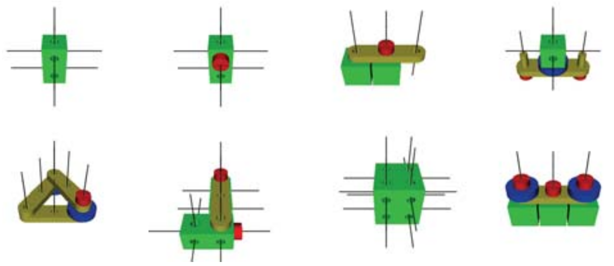


Fig. 6. The assembly charts of the 8 simulated datasets. The poses of parts are all ground truth.

TABLE I

The comparison between RDV and AG on the F-measure of the relations recognition using the simulated data. N1-N4 indicate increasing noise levels.

Noise	N1		N2		N3		N4	
Method	RDV	AG	RDV	AG	RDV	AG	RDV	AG
2 blks	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
3 blks	0.97	0.99	0.98	0.97	0.97	0.97	0.98	0.97
4 blks	0.69	0.99	0.71	0.97	0.72	0.95	0.71	0.93
5 blks	0.83	0.97	0.85	0.94	0.82	0.93	0.80	0.86
6 blks	0.99	1.00	0.99	0.99	0.95	0.95	0.94	0.93
7 blks	0.53	0.95	0.55	0.91	0.53	0.91	0.54	0.88
8 blks	0.97	1.00	0.97	1.00	0.94	0.97	0.93	0.96
9 blks	0.75	0.99	0.67	0.98	0.67	0.94	0.65	0.90
Mean	0.84	**0.98**	0.84	**0.97**	0.82	**0.95**	0.82	**0.93**

TABLE II

The comparison between RDV and AG on translational error.

Noise	N1		N2		N3		N4	
Method	RDV	AG	RDV	AG	RDV	AG	RDV	AG
2 blks	0.00	0.00	0.17	0.00	0.46	0.00	0.48	0.00
3 blks	0.00	0.00	0.26	0.00	0.54	0.25	1.28	0.41
4 blks	0.35	0.00	0.64	0.00	1.32	0.00	1.69	1.23
5 blks	4.45	0.33	2.05	0.50	1.65	0.84	1.81	1.43
6 blks	0.01	0.01	0.22	0.01	0.92	0.33	2.01	0.56
7 blks	1.01	0.00	9.12	0.67	16.1	0.74	2.03	1.30
8 blks	0.00	0.00	0.00	0.00	0.12	0.00	0.44	0.00
9 blks	0.82	0.00	5.83	0.00	21.2	0.10	1.90	0.07
Mean	0.83	**0.04**	2.29	**0.15**	5.29	**0.28**	1.45	**0.62**

TABLE III

The comparison between RDV and AG on rotational error.

Noise	N1		N2		N3		N4	
Method	RDV	AG	RDV	AG	RDV	AG	RDV	AG
2 blks	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
3 blks	0.01	0.01	0.01	0.01	0.01	0.01	0.02	0.03
4 blks	0.01	0.01	0.02	0.01	0.02	0.01	0.04	0.02
5 blks	0.15	0.02	0.09	0.02	0.04	0.03	0.05	0.03
6 blks	0.01	0.01	0.02	0.02	0.03	0.03	0.04	0.04
7 blks	0.01	0.00	0.16	0.01	0.21	0.01	0.02	0.01
8 blks	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
9 blks	0.65	0.01	0.47	0.02	0.29	0.02	0.03	0.04
Mean	0.11	**0.01**	0.10	**0.01**	0.07	**0.02**	0.03	**0.02**

reduce the influence of diverged cases, we use the median to evaluate the overall performance. The results are shown in Tab. II and Tab. III. We have three observations:

- In both Tab. II and Tab. III, AG achieves better performance than RDV. This is because the relations are recognized better when AG is employed (higher F-score). With correct relations, the error in those constrained degrees of freedom is exactly zero, leading to the better pose estimation naturally. In scenario 2 and 8, both AG and RDV give high accuracy, because both of them achieve high F-scores.
- Compared to the monotonically increasing trend of error with respect to the noise level in AG, RDV is more unstable due to the lack of mechanism to remove the incorrect recognized relation. When the contradiction exist between the recognized relations (incorrect), the algorithm will diverge, leading to the unstable error. When the noise increases, the incorrect recognition may be rejected in AG, thus leading to occasionally better results.
- The error is not directly correlated to the number of parts in the assembly task. Note the task of 8 blocks, the mean error is zero even in $N4$ level when AG is employed since the relations constrain all the degrees of freedom in the pose. For 7 blocks, RDV gives large error due to the low F-score in the Tab. I.

In summary, both AG and RDV demonstrate that the utilization of relations is able to improve the pose estimation if the recognition is right since high F-score leads to low error in pose estimation. However, deterministic modeling of relations in RDV sometimes make larger error because the incorrect relations are never revised. The RDV fully 'trusts' the results from precedent modules, whose error is actually uncontrolled. In AG, the probabilistic modeling of relations enable the algorithm to 'suspect' the relations during iterations, thus reflecting the significance of AG's modeling of all uncertainties. It should be emphasized that in AG, (10) guarantees the error between recognized relations are exactly 0. Therefore it is of great value to apply AG in PBD.

B. *Real world data*

With the vision techniques mentioned in Section IV, the experiment is conducted on the real world assembly task. Furthermore, the spatial inference system is connected to ABB industrial robotic arm simulator for physical simulation of execution to realize the PBD. An assembly task is shown in Fig. 7 with 7 blocks in total. The pose estimation using vision technique solely, RDV and AG are shown in Fig. 1. One can see that the pose estimation of each part independently using vision based techniques is not accurate. For example, in the result of vision technique, the screw on the right is not aligned with the hole of the stick, which will cause failure when the robot execute this task. By using RDV, two misalignments (indicated by 2 orange circles on the right) are fixed due to recognition of co-linear relations. In the result of AG, the incorrect relations are revised during iteration, making all 3 misalignments fixed. An explanation is shown in Fig. 8 for the different configurations of relations in the final models using RDV and AG. The IDs of the axes

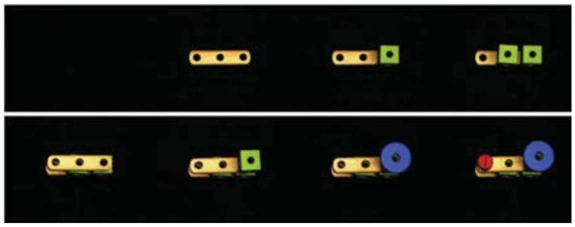
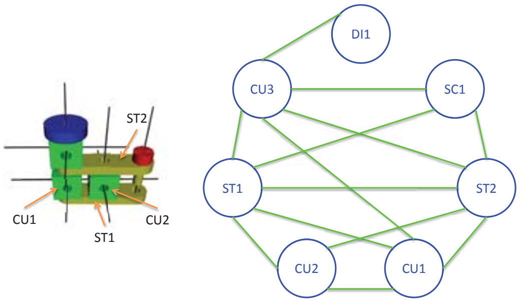


Fig. 7. The process of an wooden building blocks assembly.



	CU2	ST1
Axis	2	1
Plane	4	3
Plane	1	1
Plane	3	3

	CU1	ST1
Axis	2	0
Plane	0	0
Plane	4	3
Plane	3	3

	ST2	CU2
Axis	1	2
Plane	2	5
Plane	1	1
Plane	3	3

	ST2	CU1
Axis	0	2
Plane	0	0
Plane	2	5
Plane	3	3

Fig. 8. The assembled parts with labeled IDs (top left). The final AG model (top right). If there is an edge, it means there is at least one relation existing between the two parts. If there is no edge, no relation exists. CU, ST, DI and SC are stand for cube, stick, disk and screw respectively. Each table lists the relations with $s = 1$ between two parts (bottom). The first column gives the types of relations. The second and third columns are IDs for planes or axes in a part. The black rows are relations with $s = 1$ in the final AG models using both RDV and AG. The gray rows are relations with $s = 1$ in the model using RDV, but with $s = 0$ using AG.

and planes follows the definition in Fig. 5. RDV regards some co-planar relations as existing ($s = 1$) but AG finally find they should be with $s = 0$ during iterations, making the difference in the final results.

To show the feasibility of the proposed model in a completed PBD system, we connect the parser to the ABB industrial robotic arm simulator to execute the assembly task demonstrated by the human teacher. The simulator is in commercial level, and many projects have verified its effectiveness. So we think its result is reliable. It also avoids producing grippers for various types of blocks. The video is attached in the supplemental material. This experiment also includes LEGO cubes to show the method works even when the part is small, in which case the error in the result of vision technique is unacceptable in the execution.

VI. Conclusion

In this paper, a probabilistic graph model, AG for assembly task is proposed. The focus of the model is to probabilistically represent all information including unknown relations between parts and poses of parts, the prior knowledge of relational contradiction as well as the observations provided by vision based techniques. Then the poses and relations are estimated alternatively using CEM algorithm. With the estimated relation, a refinement is conducted to derive the final poses. This method, AG outperforms RDV due to its global modeling of all uncertainties. Finally a whole process of PBD is conducted based on our AG based spatial inference system.

References

[1] S. Lallée, S. Lemaignan, A. Lenz, C. Melhuish, L. Natale, S. Skachek, T. van Der Zant, F. Warneken, and P. F. Dominey, "Towards a platform-independent cooperative human-robot interaction system: I. perception," in *Intelligent Robots and Systems (IROS), 2010 IEEE/RSJ International Conference on*. IEEE, 2010, pp. 4444–4451.

[2] S. Lallée, U. Pattacini, J.-D. Boucher, S. Lemaignan, A. Lenz, C. Melhuish, L. Natale, S. Skachek, K. Hamann, J. Steinwender, *et al.*, "Towards a platform-independent cooperative human-robot interaction system: Ii. perception, execution and imitation of goal directed actions," in *Intelligent Robots and Systems (IROS), 2011 IEEE/RSJ International Conference on*. IEEE, 2011, pp. 2895–2902.

[3] S. Lallée, U. Pattacini, S. Lemaignan, A. Lenz, C. Melhuish, L. Natale, S. Skachek, K. Hamann, J. Steinwender, E. A. Sisbot, *et al.*, "Towards a platform-independent cooperative human robot interaction system: Iii an architecture for learning and executing actions and shared plans," *Autonomous Mental Development, IEEE Transactions on*, vol. 4, no. 3, pp. 239–253, 2012.

[4] A. Levas and M. Selfridge, "A user-friendly high-level robot teaching system," in *Robotics and Automation. Proceedings. 1984 IEEE International Conference on*, vol. 1. IEEE, 1984, pp. 413–416.

[5] J. Aleotti, S. Caselli, and M. Reggiani, "Toward programming of assembly tasks by demonstration in virtual environments," in *Robot and Human Interactive Communication, 2003. Proceedings. ROMAN 2003. The 12th IEEE International Workshop on*. IEEE, 2003, pp. 309–314.

[6] J. Papon, T. Kulvicius, E. E. Aksoy, and F. Worgotter, "Point cloud video object segmentation using a persistent supervoxel world-model," in *Intelligent Robots and Systems (IROS), 2013 IEEE/RSJ International Conference on*. IEEE, 2013, pp. 3712–3718.

[7] Y. Kuniyoshi, M. Inaba, and H. Inoue, "Learning by watching: Extracting reusable task knowledge from visual observation of human performance," *Robotics and Automation, IEEE Transactions on*, vol. 10, no. 6, pp. 799–822, 1994.

[8] C.-P. Tung and A. C. Kak, "Automatic learning of assembly tasks using a dataglove system," in *Intelligent Robots and Systems 95.'Human Robot Interaction and Cooperative Robots', Proceedings. 1995 IEEE/RSJ International Conference on*, vol. 1. IEEE, 1995, pp. 1–8.

[9] K. Ikeuchi and T. Suehiro, "Toward an assembly plan from observation. i. task recognition with polyhedral objects," *Robotics and Automation, IEEE Transactions on*, vol. 10, no. 3, pp. 368–385, 1994.

[10] E. E. Aksoy, A. Abramov, J. Dörr, K. Ning, B. Dellen, and F. Wörgötter, "Learning the semantics of object–action relations by observation," *The International Journal of Robotics Research*, p. 0278364911410459, 2011.

[11] G. D. Hager and B. Wegbreit, "Scene parsing using a prior world model," *The International Journal of Robotics Research*, p. 0278364911399340, 2011.

[12] Y. Wang, R. Xiong, L. Shen, X. Sun, J. Zhang, and L. Qi, "Towards learning from demonstration system for parts assembly: a graph based representation for knowledge," in *Cyber Technology in Automation, Control and Intelligent Systems, 2014 IEEE International Conference on*. IEEE, 2014, pp. 174–179.

[13] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the em algorithm," *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 1–38, 1977.

[14] G. Celeux and G. Govaert, "A classification em algorithm for clustering and two stochastic versions," *Computational statistics & Data analysis*, vol. 14, no. 3, pp. 315–332, 1992.

[15] J. W. Tangelder and R. C. Veltkamp, "A survey of content based 3d shape retrieval methods," *Multimedia tools and applications*, vol. 39, no. 3, pp. 441–471, 2008.

[16] R. Hartley, J. Trumpf, Y. Dai, and H. Li, "Rotation averaging," *International journal of computer vision*, vol. 103, no. 3, pp. 267–305, 2013.