2016 International Conference on Advanced Robotics and Mechatronics (ICARM)

Robust Object Localization Based on Error Patterns Learning for Dexterous Mobile Manipulation

Boyang Gao, Fei Chen, *Member, IEEE*,
Franceso Trapani, Mario Selvaggio, and Darwin Caldwell

***Abstract*— In this article we describe an approach for object detection and pose estimation from stereo RGB frames for robot manipulation in manufacturing scenarios. This solution was developed in the framework of the second challenge of the EuRoC project, and meets the need of a registration method invariant to the view perspective and robust to the structural symmetries and ambiguities of the target objects. Our contribution consists of automatic correction of sub-optimal results of registration algorithms. As most registration algorithms only converge on local optima, a tool for recognizing and correcting wrong alignments is highly desirable. Our insight is that, for a given target point cloud, it is important to study the alignment space offline and identify sub-optimal solutions before the registration. The convergence of the algorithm leads to the error pattern knowledge that can be used to discard the wrong solutions, and recover the correct alignment. Experiments on synthesized and real data show that exploiting the known information about the spatial properties of the objects, together with appropriate pre-processing and refining of the data, we can have a substantial improvement in discarding wrong hypothesis for geometrically ambiguous items.**

I. INTRODUCTION

During the last decade we witnessed a substantial growth of the use of robotics for manufacturing applications, and an increased interest toward safe interaction in industrial environments between workers and machines. The current industrial realities demands for a broader mobile manipulator usage, not only in production lines, but also for logistics, and generic manipulation tasks. This led to the development and deployment of more and more dexterous mobile manipulators who are largely guided by vision systems. With the continuous increase of the working environments complexity, precise, reliable and fast robotics vision system becomes a key requirement for mobile manipulators to prosper.

To boost the application and innovation of advanced technologies on industrial robots, the European Robotics Challenge (EuRoC) for European manufacturing industry was initiated in 2014. Among the challenges proposed, Category 2: *Shop Floor Logistics and Manipulation* described in [17] focuses on the use of mobile manipulators in manufacturing environments to accomplish logistics and manipulation tasks. The work presented in this paper has been developed for the framework of the EuRoC project in category 2 challenge.

The work leading to this publication has received funding from the European Union's Seventh Framework Programme under grant agreement no. 608849 EuRoC.

Boyang Gao, Franceso Trapani, Fei Chen, Mario Selvaggio, Gennaro Notomista and Darwin Caldwell are with the Department of Advanced Robotics, Istituto Italiano di Tecnologia, Via Morego 30, 16163, Genova, Italy. E-mail: {boyang.gao, francesco.trapani, fei.chen, mario.selvaggio, gennaro.notomista, darwin.caldwell}@iit.it.

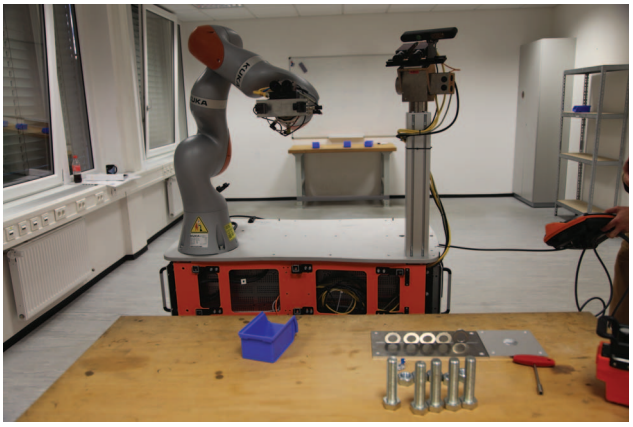

Fig. 1. The Miiwa robot at the EuRoC working environment. On the table in the foreground are visible some of the target objects: a plastic box, some nuts and washers.

The working environment proposed by EuRoC consists of a realistic manufacturing set-up, furnished with tables and shelving units (generically referred to as working surfaces). Target items such as bolts, nuts, washers, or plastic boxes (SLC) which occasionally containing sets of the other objects are randomly arranged on the working surfaces in a room. The challenge is organized as series of tasks: the robot is required to autonomously find and collect five plastic boxes, place them in goal locations and assemble together bolts, nuts and washers. The robotic platform provided by the project, MIIWA, is composed of an omnidirectional mobile platform and a light weight 7DOF compliant robot, LWR iiwa, equipped with a gripper. Stereo RGB cameras (Mako and Manta of Allied Vision) are mounted on both the end-effector and the top of the platform.

Mobile manipulators like MIIWA are designed to autonomously interact with the physical environment. To this end, spatial awareness is a fundamental requirement. Current technology provides several options for robot and object localization. Time of flight cameras, and more sophisticated LIDAR systems, offer a direct and accurate depth measurement with opened fields of view independently of lighting conditions. However devices empowered by this technology are still quite expensive [9], [7]. Stereo cameras, on the other side, provide at once full views of the scene with adequate resolution, low power consumption, and affordable expense [10]. Because stereo vision requires to compute disparities from two or more rectified images, the quality of depth

978-1-5090-3364-5/16/$31.00 ©2016 IEEE

estimation can be affected by computational resources and adverse lighting conditions [3], [8], [18]. On MIIWA 3D vision relies on a stereo camera setup, and is implemented according to the algorithm proposed by Hirschmller in [6], where disparity matrices are computed through pixels-wise semi-global matching basing on mutual information. Hence, for each stereo RGB camera, the provided output consists of streams of RGB-Depth frames, in turn used by our system to generate the scene point clouds to describe 3D structure of the environment observed.

Based on point cloud description of 3D objects, point cloud registration can be performed to detect and localize 3D object and estimate its pose. Among point cloud registration algorithms, the Iterative Closest Point algorithm (ICP) constitutes a dominant solution [1], [12]. ICP's principle is to iteratively find correspondences between two point clouds, and refine the alignment by minimizing objective function related to matching distance. This approach is effectively used for the recognition and localization of known objects by aligning point cloud templates to the scene clusters.

The high modularity and conceptual simplicity of the algorithm promoted the development of several variations and extensions [13], making ICP an highly customizable algorithm. Many different error metrics have been proposed, based on quaternions, orthonormal matrices, normals, point to plane distances [15], color [11], [13]. In the same way, several methods for the correspondences computation and selections have been developed, considering simple Euclidean distance, plane orientations, RGB information. Approaches based on k-d tree closest point searches are widely used to speed up the matching procedures [13], since the metric based closest point searching dominates the correspondence construction. According to the dataset peculiarities, uniform sampling or random sampling can be performed for preselection cloud processing. Other approaches take advantage of known properties of the observational data to improve accuracy and speed, for example rejecting point correspondence that do not satisfy some given constraints. An example of a noticeable expansion of ICP is provided by Segal *et al.* in [15] with the Generalized-ICP, where the the point to plane error metrics is improved by introducing a probabilistic model for point matching, with the misalignment minimization based on MLE. Another notable ICP variant is Normal Iterative Closest Point (N-ICP)[19]. N-ICP considers each point together with the local features of the surface (i.e. normal and curvature) and it takes advantage of the 3D structure around the points for the determination of the data association between two point clouds.

Despite of its general effectiveness, standard ICP has shown to be particularly vulnerable to the imprecision of real world data. Sensor noise, occlusion of targets, sparse discretization of scene and models are all elements which can lead ICP to failure [11]. ICP relies on point to point correlations, however inevitable distortion in real data prevents it to find perfect correspondences, and hence can make the algorithm converge on a sub-optimal solution, i.e. a wrong registration. Good initial alignments are often helpful, but in many cases they still cannot guarantee the optimal registration. As a consequence, in most of real scenarios, ad hoc approaches are needed to validate the algorithms results and redirect it toward the correct solution.

Our insight is that for applications such as object recognition, where an estimate of the input cloud is available, the alignment space can be explored offline, and important information on wrong solution patterns can be obtained to direct the registration process. In particular, both optimal and sub-optimal alignments can be identified, and mapped respect to each others on the alignment space. Our experiments also show that a substantial improvement of ICP accuracy can be achieved from a prior investigation of the alignment space. To our knowledge, approaches based on this principle are so far not present in literature.

The rest of paper is organized as follows. Our front-end object pose estimation method and modified ICP algorithm is presented in section II. Experiments and results analysis is summarized in section III. We draw the final conclusion in section IV.

II. METHODOLOGY

In this section we discuss the method that we developed for the pose estimation and grasping of target objects. The system has been designed to compute from each single input RGB-D image the grasping points of each target object in the scene, expressed in the scene reference frame. The core of the algorithm is a 2D-constrained modification of ICP, which estimates the pose of the target clusters by registering it to a point cloud template. Target clusters are segmented from the RGB image and projected into the 3D space. Each cluster is then pre-processed to filter noise and uniform their density. A learning procedure is carried on offline to acquire information about the relative pose of each solution (optimal and sub-optimal) in the alignment space. After the registration, the erroneous alignment pose is corrected by identifying the solution obtained among the learned ones, and using it as a reference to retrieve the pose of the correct alignment.

A. Template Construction

To compute the point cloud template of the target items, we opt for a mesh based approach. First, we use an open

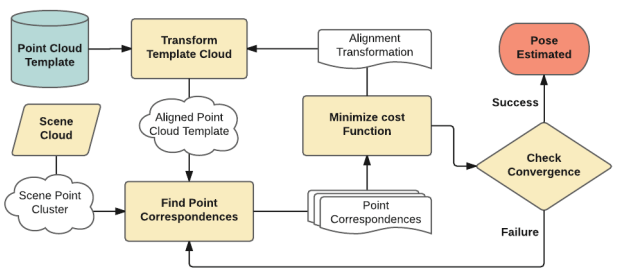

Fig. 2. Standard ICP implementation for object recognition.

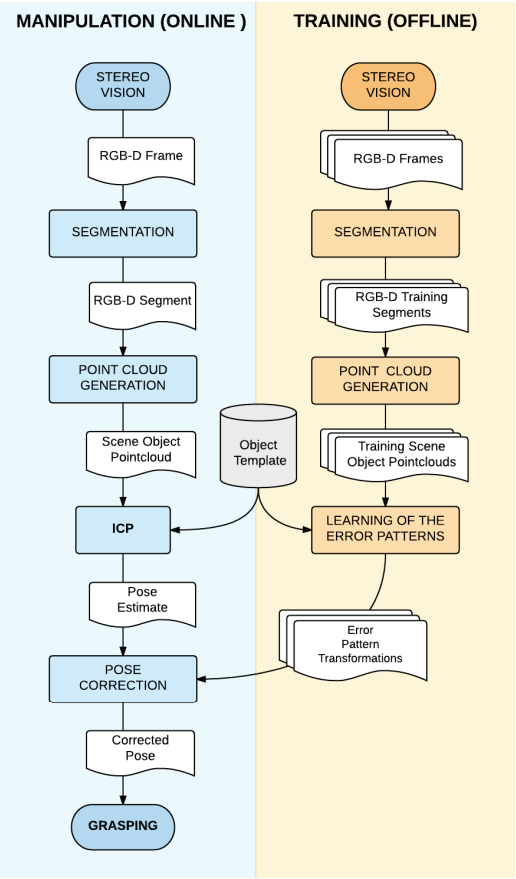

Fig. 3. Flow chart representing the key steps of the proposed method. Parts in blue represent the offline processes, performed during training time. Parts in orange represent the online processes

source graphics software, Blender, to create a 3D mesh representing each object. Each of the meshes have then been uniformly discretized into point cloud.

For the plastic box we also try a different approach based on a multi-view reconstruction of the real object. We created a composite point cloud by merging several different clouds obtained from RGB-D frames from four different views. Even after post processing, the obtained point cloud was too noisy and inaccurate compared with the one obtained from the mesh. We then just employ mesh based template for the registration steps.

B. Data Pre-processing

For each stereo frame, potential objects are identified at image level by segmenting the RGB frames according to color ranges. As the working environment of our case provides a neutral background, color proved to be a good feature to distinguish items from the scene. Each segment is then projected in the 3D space, and a set of point clouds is obtained. In order to deal with the possible overlapping of more objects in the image, each segment is clustered based on point distances, so that each cluster depicts one and one only object. Clusters are finally filtered using a statistical outlier removal, and uniformly sampled to the same resolution of the template point clouds. For this last processing steps we relied on the statistical outlier removal and the uniform sampling functions offered by the Point Cloud Library (PCL) [14].

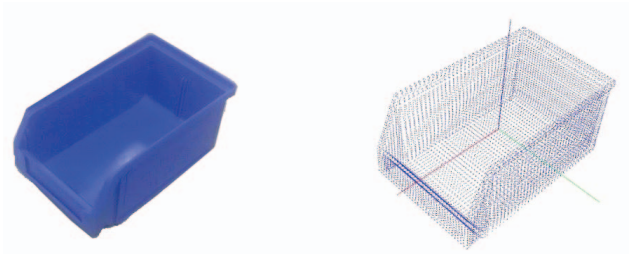

Fig. 4. The mesh based point cloud template representing a blue plastic box. On the left, a picture of the real box. On the right, a rendering of the final point cloud template used in our experiments.

C. Constrained ICP

The pose estimation of the objects identified in the scene is accomplished by registering the point cloud templates on the scene clusters. For the registration, we relied on the standard ICP variant as implemented by Pomerleau *et al.* in [11]. This solution uses the point-to-point (Besl and McKay, [2]) distance metrics combined with the trimmed-ICP outlier rejection (Chetverikov et al. [4]). In our scenario, most of the target objects have constraints on their poses. In particular, nuts, bolts and boxes always lay on the flat working surfaces leaning against their base. As a result, the registration algorithm should only consider alignment poses satisfying the constraint that the alignment space can be reduced to 4DOFs. To better represent this constraint we define α as the angle representing, for each given alignment, its horizontal orientation (i.e. its rotation around the axis vertical to the working surface). Because of our constraint, α is the only variable characterizing the orientation of the alignment pose, as the other two rotational components are always fixed. We hence implemented an ICP variation, orientation constrained, which relies on a different minimization function that updates the transform orientation by only varying α. This solution simplifies the problem, discards out unfeasible poses from the alignment space, and hence speeds up the registration process. As initial alignment, we use the transformation which aligns the centroid of the template on the cluster centroid, without altering its orientation.

D. Confidence Metric

In order to evaluate and compare the results of our corrections, we need to define a confidence metric. We use the reciprocal of average squared Euclidean distance between matching points to define the confidence of a given registration:

$$Confidence = \frac{n}{\sum_{i=0}^{n} Euclidean(p_i^m, p_i^t)^2}$$

where n is the number of correspondences, p_i^m the i-th point of the template cloud, and p_i^t the i-th point of the

target cloud. This metric allows the system to easily compare results of the registration and safely select the best correction.

E. Error Patterns Learning

As explained above, the main drawback of ICP is a blind convergence over the closest minimum of the objective function, which is not guaranteed to be the correct solution. In order to verify the correctness of an alignment, some further knowledge about the objective function is necessary.

We define as *Alignment Space* the set of admissible poses of the template cloud with respect to the scene frame. As the objective function depends on the template pose, each minima will correspond to a particular pose in the alignment space. Nevertheless, the morphology of the objective function, and hence the number and relative poses of its minima in the alignment space, is only determined by the structure of template and target clouds. In fact, different relative positions of template and target only result in different rigid transformations on the alignment space. As a result, given the template cloud and an estimate of the target, we can compute an approximation of the objective function offline, and explore it to predict the number of sub-optimal alignments and their relative poses respect to the correct one. This can be achieved by registering template to the estimated target multiple times, using different initial alignments, and comparing the obtained solutions. If the set of initial alignments provides a good coverage of the alignment space, the registration steps will generate a significant portion of all the reachable minima.

Once the optimal solution is identified, the transformations between each local minima and the global one can be calculated. We consider this procedure as a *learning* process of the error patterns present in the alignment space. We introduce the *Error Pattern Transformation* (EPT). For each local minimum, the EPT is defined as the transformation matrix transforming the wrong alignment into the correct one. EPTs represent the relative poses of the minima in the alignment space, and are hence dependent only on template and target structures. This means that once computed, they maintain their validity for real scene data.

F. Error Patterns based Correction

When ICP registers a scene online, in the first pass it outputs an initial alignment. The alignment can be either the correct solution, or stuck in a local minima. If the matching confidence is not higher than the threshold, all EPTs are applied to correct potential error by multiplying the transformation matrices to the initial alignment. After the EPT corrections the second pass constrained ICP is performed to generate corrected pose estimations with updated matching confidences. The pose with the largest matching confidence is preserved as final result.

III. EXPERIMENTS AND RESULTS

In this section we report the results on standard, our 2D constrained ICP, and our 2D constrained ICP with EPT based correction. The algorithms have been evaluated on the task of estimating the pose of the blue plastic box used in the assembling scenario. To this end, we created two different datasets, one is based on the observational data recorded at the assembling working location, and the other one is based on samples synthesized from the mesh templates. We tested the different systems on the two datasets and measured their performances.

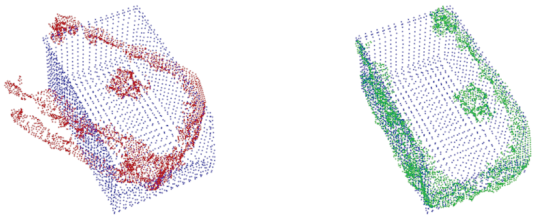

Fig. 5. Example of an ICP registration on the observational dataset. On the left, the template cloud (in white) and the target cloud (red) before the registration. On the right, the target cloud (now in green) is aligned on the template.

A. Datasets

The synthesized dataset was generated by applying Gaussian noise to the mesh based point cloud template. As suggested in [11], we defined three different noise magnitudes for position misplacement, respectively with a standard deviation of 0.1, 0.4 and 0.7 mm. For each noise level, we generated 2 different point cloud samples for each noisy condition which adds up to 6 clouds. The final data was a set of 8 point clouds representing different portions of the box and featuring different levels of noise. To build the dataset from the real data, we used actual stereo frames captured at the assembling work location through the robot stereo vision system. 8 different RGB-D frames depicting the plastic box from different angles and distances were selected, and preprocessed as stated in section III.B.

For each dataset, we used 50% of the samples as a training set to generate possible the target object cloud in alignment space and to construct the Error Pattern Transformations. We tested different algorithms on the remaining 50% testing samples.

B. Error Pattern Transformations Generation

To compute the Error Pattern Transformations, we used each of the training set samples as a distinct target cloud. We then obtained the set of minima by registering our template to each of such target cloud. In order to get an effective coverage of the alignment space, we registered each target from different initial poses. Because of the 2D constraint, the pose orientation can only variate by rotating around the axis vertical to the working surface, that is of different values of α.

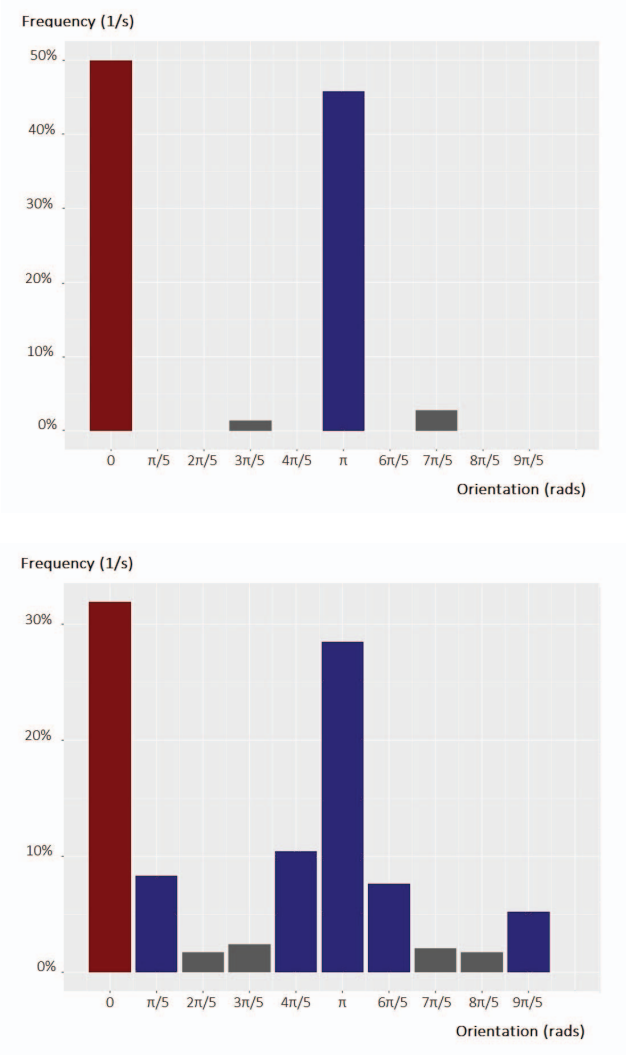

Fig. 6. Histograms representing the frequency of convergence of the ICP algorithm over solutions within the given intervals of α. Figure 3.A shows the frequencies relative to the *synthesized data*, while figure 3.B shows the results for the *observational data*. Dark red histograms indicate the correct alignment, while blue histograms represent the sub-optimal solutions selected for the computation of the Error Pattern Transformations.

We then generated the set of initial alignments by fixing the alignment position to the target centroid, and shifting α of 5 degrees over the interval $[0, 2\pi]$, for a total of 72 orientations. This proved to be a simple but effective sampling of the alignment space for two reasons. First, during the online registrations our system always uses initial alignments centered on the target cloud centroid. Hence, by considering that region, we direct the minima search over the alignment space actually explored by the online registration processing. Second, we notice that the convergence of the algorithm is mainly dependent on the initial orientation, but hardly influenced by the initial position. Hence, keeping the position fixed and only varying α largely reduces the number of initial alignments to consider, but does not induce significant loss in the computation of the minima set.

Fig. 6 describes through histograms the results of our minima search. Histograms A and B represent the distributions of minima with respect to α with bin size of $\pi/5$, for the synthesized and for the observational training sets. The distribution has been computed by merging the minima sets obtained from the registration of each sample for each set. The horizontal axes show all the possible convergence poses, identified by the respective value of α, grouped according to rotation intervals. Since the position can always be estimated by aligning two centroids, hence is not needed to characterize the pose for each α value. The vertical axes depict the frequency of convergences obtained for each given orientation interval over the whole training set. For the synthesized training set the error patterns are clearly identifiable: one only prominent sub-optimal solution is found for $\alpha = \pi$. The graph for the observational case shows similar results, but with the presence of further minima around the optimal ($\alpha = 0$) and the main sub-optimal ($\alpha = \pi$) solutions.

We considered all the poses in the same bin as belonging to the same minima. Of all the minima, we decided to consider only the ones with frequency superior to than 5%. We hence computed for each corresponding interval the mean pose, and used it to calculate the corresponding Error Pattern Transformation. For the synthesized dataset, we found 1 minimum, and hence computed only one EPTs. For the observational dataset, the EPTs set included 5 patterns.

C. Experimental Results

Both ICP and constrained ICP have been tested on each testing sample over the whole set of initial poses defined in the previous section including position fixed on the centroid and 72 uniformly separated values over the interval $[0, 2\pi]$ for α. We then corrected the output of each registration of constrained ICP with the pre-computed EPTs, and evaluated the improvement of the results. For each registration, the correction with the highest confidence has been selected. Nonetheless when the initial alignment featured an higher confidence no correction was applied. The corrected alignments were then further refined by running constrained ICP again using the corrected alignment as the initialization.

The results obtained by the different algorithms are shown in Table 1. We define the accuracy of a given algorithm on a given training sample as the percentage of correct convergences over all $5\,\mathrm{deg}$ separated alignments. We consider an alignment correct if the translational error and the rotational error are inferior to 0.3 cm and 0.03 rads respectively. The Table 1 shows similar results for the synthesized and observational sets. On the synthesized data standard ICP and 2D constrained ICP gave the same results, and their performance showed to be invariant respect to the increase of noise magnitude. We still find that for half of the registration results, in particular the ones with initial alignment with $\alpha \in (\pi/2, 3\pi/2)$, the two algorithms converged on the sub-optimal solution identified in Fig. 6. On the observational data, 2D-ICP showed an higher average performance, but the correct convergences did not exceed 50% for either of the two algorithms. The EPT based algorithm, on the other

TABLE I

ALGORITHMS PERFORMANCE

Synthesized Set

		STD-ICP	2D-ICP	EPT-ICP
Samples	1	48.61	48.61	100.00
	2	48.61	48.61	100.00
	3	48.61	48.61	100.00
	Average	**48.61**	**48.61**	**100.00**

Observational Set

		STD-ICP	2D-ICP	EPT-ICP
Samples	1	48.62	48.62	100.00
	2	38.89	40.28	100.00
	3	37.50	51.39	100.00
	4	45.83	44.44	100.00
	Average	**42.71**	**46.18**	**100.00**

Table 1. Performance of the three algorithms over the proposed datasets, intended as percentage of correct convergences. The algorithms are: standard ICP (STD-ICP), 2D constrained ICP (2D-ICP), and 2D constrained ICP with EPTs based error recovery (EPT-ICP).

side, gave impressing results, as it managed to correctly recover the optimal alignment for each of the testing samples with any initial alignment. Even for the observational set, where testing and training samples featured structural differences and different morphology of the objective function, the accuracy was still 100%. This came at the cost of a reduced speed which is mainly due to the correction steps and the second ICP procedure. Still, the improved alignment accuracy justified the value of the trade off in time.

D. Experiments on the Real Robot

After testing the algorithm on the two datasets, we also applied it to the real robot MIIWA at the assembling working environment. Because of the time limitation, we did not carry out a rigorous experimentation with statistical results. But we verified the validity of our approach by observing that the MIIWA was able to correctly identify and grasp all the boxes on the working surfaces under all real circumstances. An exhaustive evaluation of the system performances in the real scenario is to be carried out in our further work.

IV. CONCLUSIONS

In this article we present an approach for object detection and pose estimation of working tools in partially structured manufacturing environments. We provide an insight of how point cloud registration, and in particular ICP, can be used to achieve satisfying results for real industrial robotics tasks.

We discuss our constrained variation of ICP, which takes into account spatial constraints specific to our case study, and exploits them to reduce the alignment space during the error minimization. We also propose an algorithm to identify offline the set of sub-optimal solutions ICP can get stuck in, and a solution to recognize and correct such solutions during online registrations.

We compared the accuracy of a traditional implementation of ICP with our 2D constrained version and with a the 2D constrained version plus error correction. All of experimental results and real robot behavior confirms the validity of our solutions, and show that the presented approaches have significantly improved the performance of ICP based registration algorithms.

Our future directions include the testing of our method on more challenging datasets. In particular, we would like to consider testing scenes with occlusion and different types of noise.

REFERENCES

[1] Bellekens, B., Spruyt, V., Weyn, R. B. M., & Berkvens, R. (2014). A survey of rigid 3d pointcloud registration algorithms. In The Fourth International Conference on Ambient Computing, Applications, Services and Technologies.

[2] Besl, P. J., & McKay, N. D. (1992, April). Method for registration of 3-D shapes. In Robotics-DL tentative (pp. 586-606). International Society for Optics and Photonics.

[3] Chen, K., Lai, Y. K., & Hu, S. M. (2015). 3D indoor scene modeling from RGB-D data: a survey. Computational Visual Media, 1(4), 267-278.

[4] Chetverikov, D., Svirko, D., Stepanov, D., & Krsek, P. (2002). The trimmed iterative closest point algorithm. In Proceedings of the 16th international conference on pattern recognition (pp. 545548).

[5] Engelhard, N., Endres, F., Hess, J., Sturm, J., & Burgard, W. (2011, April). Real-time 3D visual SLAM with a hand-held RGB-D camera. In Proc. of the RGB-D Workshop on 3D Perception in Robotics at the European Robotics Forum, Vasteras, Sweden (Vol. 180).

[6] Hirschmller, H. (2005, June). Accurate and efficient stereo processing by semi-global matching and mutual information. In Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on (Vol. 2, pp. 807-814). IEEE.

[7] Hussmann, S., Hagebeuker, B., & Ringbeck, T. (2008). A performance review of 3D TOF vision systems in comparison to stereo vision systems. INTECH Open Access Publisher.

[8] Izadi, S., Kim, D., Hilliges, O., Molyneaux, D., Newcombe, R., Kohli, P., ... & Fitzgibbon, A. (2011, October). KinectFusion: real-time 3D reconstruction and interaction using a moving depth camera. In Proceedings of the 24th annual ACM symposium on User interface software and technology (pp. 559-568). ACM.

[9] Li, L. (2014). Time-of-flight cameraan introduction. Technical White Paper, May.

[10] Murray, D., & Little, J. J. (2000). Using real-time stereo vision for mobile robot navigation. Autonomous Robots, 8(2), 161-171.

[11] Pomerleau, F., Colas, F., Siegwart, R., & Magnenat, S. (2013). Comparing ICP variants on real-world data sets. Autonomous Robots, 34(3), 133-148.

[12] Pomerleau, F., Colas, F., & Siegwart, R. (2015). A review of point cloud registration algorithms for mobile robotics. Foundations and Trends in Robotics (FnTROB), 4(1), 1-104.

[13] Rusinkiewicz, S., & Levoy, M. (2001). Efficient variants of the ICP algorithm. In 3-D Digital Imaging and Modeling, 2001. Proceedings. Third International Conference on (pp. 145-152). IEEE.

[14] Rusu, R. B., & Cousins, S. (2011, May). 3d is here: Point cloud library (pcl). In Robotics and Automation (ICRA), 2011 IEEE International Conference on (pp. 1-4). IEEE.

[15] Segal, A., Haehnel, D., & Thrun, S. (2009, June). Generalized-ICP. In Robotics: Science and Systems (Vol. 2, No. 4).

[16] Serafin, J., & Grisetti, G. (2015, September). NICP: Dense normal based point cloud registration. In Intelligent Robots and Systems (IROS), 2015 IEEE/RSJ International Conference on (pp. 742-749). IEEE.

[17] Siciliano, Bruno, et al. "EuRoC-The Challenge Initiative for European Robotics." ISR/Robotik 2014; 41st International Symposium on Robotics; Proceedings of. VDE, 2014.

[18] Tippetts, B., Lee, D. J., Lillywhite, K., & Archibald, J. (2016). Review of stereo vision algorithms and their suitability for resource-limited systems. Journal of Real-Time Image Processing, 11(1), 5-25.

[19] Serafin, J., and Grisetti, G. (2015). GICP: Dense normal based point cloud registration, Intelligent Robots and Systems (IROS), 2015 IEEE/RSJ International (pp. 742-749), IEEE.