Human Workflow Analysis using 3D Occupancy Grid Hand Tracking in a Human-Robot Collaboration Scenario

Claus Lenz, Alice Sotzek, Thorsten Röder,
Helmuth Radrich, Alois Knoll
Robotics and Embedded Systems
Technische Universität München
{lenz,sotzek,roeder,radrich,knoll}@in.tum.de

Markus Huber, Stefan Glasauer
Center for Sensorimotor Research
Department of Neurology
Ludwig-Maximilians-Universität München
{mhuber,sglasauer}@nefo.med.uni-muenchen.de

***Abstract*— In this work, we present a Hidden Markov Model (HMM) based workflow analysis of an assembly task jointly performed by a human and an assistive robotic system. In an experiment subjects had to assemble a tower by combining six cubes with several bolts for their own without the influence of a robot or any other technical device. To estimate the current action of the human, we have trained composite HMMs. After the successful evaluation on disjunct experimental data sets, the models are transferred to the assistive robotic system *JAHIR*, where the same assembly tasks was executed. A new 3D occupancy grid approach was used to determine the hand positions of the worker. The positions were then used to compute the inputs of the analysis HMMs. The workflow of the right hand could be recognized with an accuracy of $92.26\,\%$ which is nearly as good as the recognition rate of reference experiments.**

I. Introduction

The combination of human flexibility and machine efficiency can essentially reduce the amount of fixed production costs in relation to variable costs [1]. Keeping the human *in the loop* of technical system advances the overall system due to the cognitive and senso-motoric advantages of the human for highly flexible assembly. In this way, flexibility and adaptivity are important for future developments in the automation of production processes. As current trends aim to combine the advantages of both fully automated and fully manual production steps (*hybrid assembly* [2]), possible future automation lie in the area of value creation by humans supported by robotic co-workers. To enable the needed flexibility and adaptability, such assistive robotic systems need to make use of information provided by multiple sensors, that are embedded in the "real world" to perceive, reason, learn and plan in a context-aware manner [3].

Context-aware technical system can pro-actively assist the human and are successfully employed in the domain of modeling and monitoring of standard surgeries including laparoscopic cholecystectomy [5], [6], [7]. This allows context-aware operating rooms assisting the surgeon by context-sensitive user interfaces [6]. But also assembly tasks can be pro-actively supported, if the current state of a task can be identified. To achieve that, [8] uses body worn accelerometers and microphones to estimate the progress of an assembly task.

Several applications in the area of gesture recognition show that a recognition of actions or special movements can also be achieved just by means of visual tracking. [9] uses a visual-based system to recognize both isolated and continuously spoken sentences in Greek Sign Language. Another visual approach to recognize sentences in American Sign Language is presented in [10]. Instead of recognizing words of a sign language, [11] concentrate on single characters and numbers. All data used in their feature vectors is derived from motion tracking of a single hand.

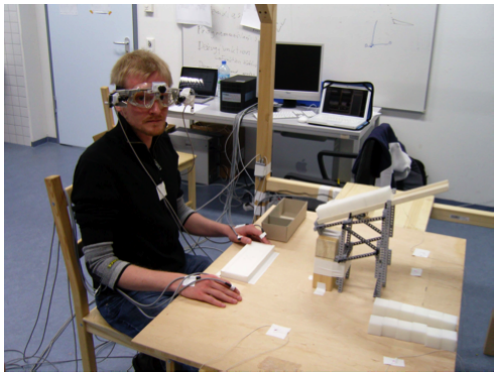


Fig. 1: **Baja experimental set-up** - Subjects assemble a tower by combining six cubes provided by a cube vendor in front of them with several bolts taken from a box on their left. During the experiment sensor record the position of the thumbs, the forefingers, the back of both hands, the head, the torso, and the gaze [4]

In the works presented above, Hidden Markov Models (HMMs) has been successfully employed for workflow analysis in different settings varying from the recognition of executed actions including gestures to the detection of different phases of an executed task. As a consequence, we present in this paper a HMM-based workflow analysis of collaborative assembly tasks between human and robot. Composite HMMs were trained with several variations in the presented sensor variety (see Section IV) in a "human-only" reference experiment with 25 subjects (see Section II-A) and then transferred to the assistive robotic system *JAHIR* (see Section II-B). To allow a natural collaboration without invasive sensors, we introduce a new approach to do 3 dimensional multiple hand tracking using 3D occupancy grids (see Section III). The results of both experiments are presented in Section V.

II. Experimental set-ups

A. Set-up used to gain base-line data

To be able to analyze and model the workflow of a natural collaborative task, we use as basis an experiment in which the subjects are not influenced by a robot or any other technical device. Subjects had to assemble a tower by combining six cubes with several bolts. Each cube features one to five holes on two opposing sides. With the number of bolts needed to stack two cubes, the complexity of the assembly step increases.

As depicted in Fig. 1, subjects were sitting on a desk and had to build towers upon a board. A box containing the bolts was positioned to the left. Cubes were available to the human from a slide placed in a way that the foremost cube lied at roughly the handover-position of an imaginary cooperation partner [12]. The sequence of the cubes on the slide with respect to the number of holes was varied among the persons. However, the number of holes of two subsequent cubes always matched each other. Furthermore, the board in front of the subjects initially contained the correct number of bolts for the first cube. That way, the assembly task was reduced to taking six cubes and connecting them with bolts five times.

25 persons participated in the experiment and were asked to build six towers in a row. A video of the tower building experiment can be accessed online[1]. Their movements were recorded by a Polhemus Liberty tracking device which measures with $240\,\mathrm{Hz}$ the position and quaternion of multiple sensors. The sensors were attached to the thumbs, the forefingers, the back of both hands, the head, and the torso. Moreover, the gaze of the persons on the table was recorded with the eye-tracking device EyeSeeCam [13]. A detailed description on the experimental set-up can be found in[4], where the same experiment was used and evaluated with the focus on optimal assistive timing.

B. Application scenario: the collaborative robotic system JAHIR

The results and models for a workflow analysis of the human-robot collaboration gained in the "human" experiments as described above should then be transferred to the robotic demonstration platform *JAHIR (Joint-Action for Humans and Industrial Robots)* [14], [15]. *JAHIR* is a hybrid assembly system [1], [2] created and embedded in a *Cognitive Factory* scenario [16] in order to bring human and robotic co-worker closely together in a common workspace for collaborative applications.

Human and robot jointly use a workbench, which is divided into several workspaces as shown in Fig. 2. The human is sitting on the right side of the table and can only act in a limited space (Fig. 2(b)). The slides on the left (Fig. 2(d)) and right side (Fig. 2(c)) of the table and the storage of presorted parts (Fig. 2(g)) are not within reach of the human. The robot can act in its own workspace covering the storage spaces and an interaction area overlapping with the human workspace. A standard position controlled industrial robot with six degrees of freedom, a maximum payload of six kilograms, and a manipulation sphere of $0.902\,\mathrm{m}$ radius is used as collaborative robot. The tool center point is extended with a force/torque sensor that triggers the gripper during the hand-over.

[1]http://www.youtube.com/watch?v=tfW4L7Idpqk

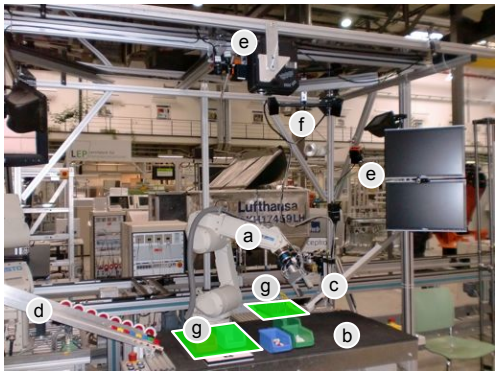


Fig. 2: **The hybrid assembly demonstration platform *JAHIR*** - (a) industrial robot, (b) shared workbench, (c) slide for tower parts, (d) slides other parts, (e) sensor devices for in- and output, (f) *ARTrack* system, (g) storage space for the robot, that is unreachable for the human

Several sensing devices (input) and a projection unit (output) are mounted on a scaffolding around the shared workspace (see Fig. 2) to survey, inform, and interact with the human worker (Fig. 2(e)). Microphones capture utterances of the human to allow a natural way of interaction with the system [17].

III. 3D Occupancy Grid Tracking

The determination of the hand position(s) and further the continuous estimation of the hand motion are crucial steps in the workflow analysis. Occupancy maps are a well known technique used in mobile robotics to solve path planning and localization problems [18], [19], [20], [21], [22]. Recently, the application of such grid maps becomes more and more popular to be employed in tracking tasks. [23] uses discretized areas on the ground plane and fits GMMs to estimate the likelihood of persons standing on a specific location. The motion of humans is modeled by a Kalman filter. A combination of probabilistic occupancy maps with models of color and motions is presented in [24]. To follow and distinguish multiple persons in the synchronized camera streams, the Viterbi algorithm and a greedy approach is used. [25] uses hierarchical likelihood grids based on intensity edges followed by a global nearest neighbor data association approach to perform the tracking of multiple persons in a multiple camera set-up. All these approaches use generative and generic models to compare "ideal" measurements with the real—and probably noisy—sensory data on discretized locations on one layer.

The discretization of the problem space allows a precomputation of expected measurements for all possible (discrete) locations. This reduces the computational complexity

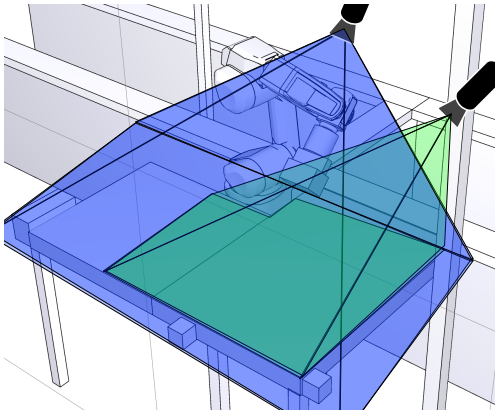


Fig. 3: **Camera set-up** - Two cameras are mounted on the *JAHIR* set-up and calibrated intrinsically and extrinsically to each other. One camera is facing the human from the front and the other is facing towards the workspace from the side

during run-time and makes such approaches scalable to multiple cameras and real-time capable. With the extension of the occupancy grid to three dimensions, a reliable, fast, and robust hand tracking in world coordinates becomes possible. Two cameras are mounted on the *JAHIR* set-up and calibrated intrinsically and extrinsically to each other. One camera is facing the human from the front and the other is facing towards the workspace from the side as depicted in Fig. 3.

We define the volume in which the hands of the human worker are likely positioned and are to be tracked. This volume is set in the world coordinate frame, that is located in the *JAHIR* set-up at the left corner of the desk. The volume of interest starts at $x = 0.3\,\mathrm{m}$, $y = -0.1\,\mathrm{m}$, $z = 0\,\mathrm{m}$ and has the width $w = 1.1\,\mathrm{m}$, the depth $d = 0.5\,\mathrm{m}$, and the height $h = 0.3\,\mathrm{m}$. This results with a discretization step of $0.05\,\mathrm{m}$ in 1694 locations (22 in x; 11 in y; 7 in z direction) as shown in Fig. 4 (a) and (b).

The hand of a human is approximated by a cube with a side length of $0.05\,\mathrm{m}$ which is roughly the dimension of the palm. This model is projected to all 1694 locations in each camera view leading to rectangular areas (*screen rectangles*) that approximate the expectation of a hand being at a specific location. Hence, the screen rectangles can be interpreted as expected measurement of a hand being at a location.

If a projected model is not visible in one camera, the corresponding screen rectangle is marked as invisible and will not be evaluated in this camera. Partly visible screen rectangles are truncated to fit the camera screen. The projection to the two camera views used here is depicted in Fig. 4 (c) and (d). The chosen camera arrangement offers the advantage that the views are aligned with two world axes which leads to axis aligned screen rectangles. All of these steps are computed off-line.

During the on-line tracking, every incoming image is first transformed into a scale space and then segmented using a histogram back projection in the H-S color space resulting in skin-colored regions. Every screen rectangle $S_{1\ldots R}^{1\ldots C}$ is tested on the binary image z^c of camera c and the likelihood of a hand being positioned in the rectangle r in camera view c is evaluated by

$$P(S_r^c|z^c) = \frac{F(S_r^c, z^c)}{A(S_r^c)} \tag{1}$$

where $F(S_r^c, z^c)$ estimates the number of skin-colored pixels in the screen rectangle with an integral image approach [26] and $A(S_r^c)$ is the area covered by the screen rectangle. The overall likelihood for a hand in a specific rectangle is then given by

$$P(S_r|z) = \prod_{c=1}^{C} P(S_r^c|z^c). \tag{2}$$

Given this three-dimensional likelihood distribution, we use all rectangle candidates that are above a chosen global prior value and compute the weighted average. This average is used to divide the data set to left hand and right hand candidates. This assumption is only valid because we assume to have exact two hands in the volume of interest. For the hand candidates we apply again a weighted average and get the three-dimensional position of both hands.

These positions are then used as input values for two Kalman filters. Since we are working directly in the three-dimensional space, we comply with the linearity and Gaussian requirements of a Kalman filter [27]. The motion of the hands is modeled by a constant velocity motion model with white noise acceleration (WNA).

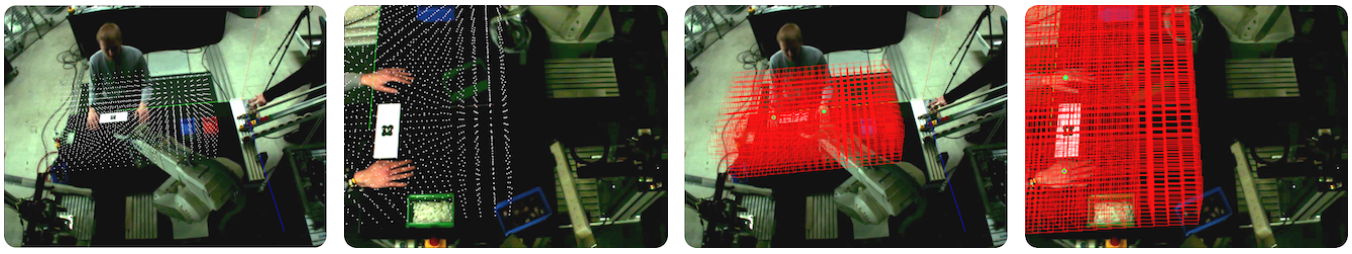
The results of the tracking as depicted in Fig. 7 show that the presented tracking approach delivers a good estimation of the hand position compared to ground truth data. To gain the ground truth, the hand positions were labeled in every frame for every camera and then the 3D position was reconstructed using the *direct linear transform (DLT)* algorithm. The standard deviation of the position error is given with $0.0207\,\mathrm{m}$ in x, $0.0179\,\mathrm{m}$ in y, and $0.026\,\mathrm{m}$ in z direction. The approach works in real-time with over 20 fps on a standard machine.

IV. HMM Models

The tower assembly task used in the "human" experiment—described in Section II-A—consists of roughly three different actions: taking cubes, taking bolts and assembling both. The taking actions of cubes and bolts can further be divided to the sub-actions reaching out, grasping and moving the object (bolt or cube) to the tower for assembly (retraction). Hence, seven different actions need to be distinguished.

Separately for both hands, we train a left-to-right continuous HMM [28] for each of the seven actions using the Baum-Welch algorithm. For training, we label the data with the different action categories based on several conditions. In particular, the position of the right or left hand, its velocity and the knowledge about past/future positions are relevant.

In contrast, the position is not included in the feature vector used as input to the individual HMMs for training. Instead, the three-dimensional velocity, acceleration and jerk are taken which can be derived from the position.



(a) (b) (c) (d)

Fig. 4: **Discretized locations and screen Rectangles** - The discretized volume of interest starting at $x = 0.3\,\mathrm{m}$, $y = -0.1\,\mathrm{m}$, $z = 0\,\mathrm{m}$ results with a discretization step of $0.05\,\mathrm{m}$ in 1694 locations (22 in x; 11 in y; 7 in z direction)

Furthermore, we constrain the frequency to $60\,\mathrm{Hz}$ (instead of the $240\,\mathrm{Hz}$ available from the Polhemus system). As described in Section II-A, not just one but eight different sensors were fixed to several spots on the subjects. So, every sensor provides a velocity, acceleration and jerk vector. Different permutations of used sensor data were individually investigated.

To allow a general sense of the position, we divide the table in three zones: one around the cube vendor, another one around the box of bolts and the last one around the tower. A zone is activated as soon as the sensor on the back of a hand is within its specified dimensions. For each hand separately, the activation of the zones is taken as further dimensions of the feature vector. Since the left hand should never be near the cube vendor, we neglect this zone for the left hand resulting in five dimensions for the table zones.

Experimental results revealed that six states with skipping transitions per model are appropriate to be utilized for all individual action HMMs. These HMMs are connected by means of a grammar and form the workflow analysis composite HMM for the left and the right hand separately. As the left and the right hand act mostly parallel and therefore a large number of movement combinations are possible, an inclusion of both hands in one model would be quite unhandy. The grammar allows taking cubes (for the right hand), taking bolts and assembling to follow in an arbitrary manner. However, it restricts the three minor movements of taking a cube to succeed in the correct order. The same holds with respect to taking a bolt. Since the subjects were asked to only use the right hand to take cubes, the movement vocabulary of the left hand does not encompass the taking of a cube. The structure of the composite HMM is depicted in Fig. 5 where the grey HMMs are only available for the right hand model.

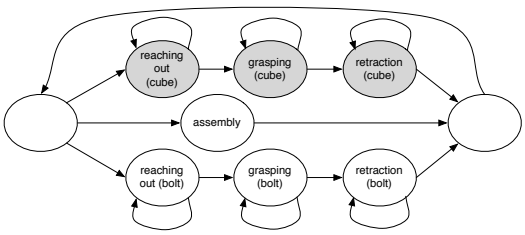


Fig. 5: **Composite HMM** - Individual continuous HMMs are trained and connected in a composite HMM for each hand. Grey action HMMs are only available for the right hand model

V. Evaluation

The "human" experiment was evaluated using a 11-fold cross-validation. As three out of 25 subjects did not follow the experminental instructions and performed the assembly task in a different way compared to the others—i.e. they placed the bolts not in the existing tower, but in the cube that should be mounted—we exclude them from the data set. Hence, the HMMs are trained on 20 persons and tested on the remaining two ones.

For each person, all appropriate sequences of building a tower are taken which can be up to six ones. Unfortunately, some subjects dropped bolts near the cube vendor or wrongly decided to take a cube and therefore stopped the execution of the movement in the middle. For this reason, we exclude some sequences but at least three ones per person always remain.

Using all available data that is the different table zones, the gaze data and the velocity, acceleration and jerk of the sensors of the head and both thumbs, forefingers and back of hands, the feature vector consists of 70 dimensions. Like that, we achieve an average accuracy of $(95.67 \pm 5.07)\,\%$ for the right hand and $(87.68 \pm 5.32)\,\%$ for the left hand. Accuracy means the percentage of labels which correspond to the true ones. Correspondence for us also includes being in the same general movement. So for example, if a sample in the recognition sequence is labeled as reaching the hand to the cube vendor and the true label stands for taking a cube, no error will be registered.

However, by reducing the data in the feature vector to just the table zones and the velocity, acceleration and jerk of the back of both hands, we still get an average accuracy of $(95.11 \pm 5.20)\,\%$ for the right hand and $(83.48 \pm 7.39)\,\%$ for the left hand. Compared to the above results, these values are not remarkably lower. That means, it is sufficient to focus on the hands. The recognition results are summarized in Table I.

In an industrial setting, equipping workers with sensors on their hands is impossible, though. Besides perturbing the

natural flow of actions, the cables connecting the sensors to a computer will not be accepted by most of the workers. Hence, it is desirable to extract the same data by different means. A good possibility is the use of the occupancy grid approach presented in Section III due to its robust and fast results.

Having the three-dimensional position of both hands, all necessary data for the feature vector can be derived. That is straight forward with respect to the velocity, acceleration and jerk of both hands. Regarding the activation of the table zones, their dimensions need to be adjusted to the geometry of the recorded workspace. Thereupon, the activation can be determined from the position. An alternative approach would be to use extra cameras for each table zone which get activated as soon as a hand is recognized in the area of the associated camera.

Since we want to simulate the application of the model in a real environment, we do not perform a 11-fold cross-validation. Instead, we train our HMMs on all 22 persons of the "human" experiment. Thus, testing on the camera tracking data corresponds to using a pre-trained model in a real setting.

As result, we achieve an accuracy of $92.26\,\%$ for the right hand and $40.11\,\%$ for the left hand. Compared to the previous results in this section (see Table I), the estimation for the right hand is just marginally lower. In fact, a direct comparison between the true and the recognized label sequence shows that all grasps of cubes and bolts are correctly identified. As shown in Fig. 6, only the boundaries of the movements are not exactly recognized.

In contrast, the results of the left hand seem to be poor at first glance. A closer examination of the three-dimensional position sequence gained through tracking reveals the real cause. In reality, the subject does not move its left hand at all. However, the position sequence shows little movements. Those occur especially whenever the right hand takes a bolt and are probably caused by the combination of both hands in the Kalman filter. Since the left hand is located near the box of bolts and therefore the corresponding table zone is activated, these small perturbations can be easily mistaken as taking a bolt.

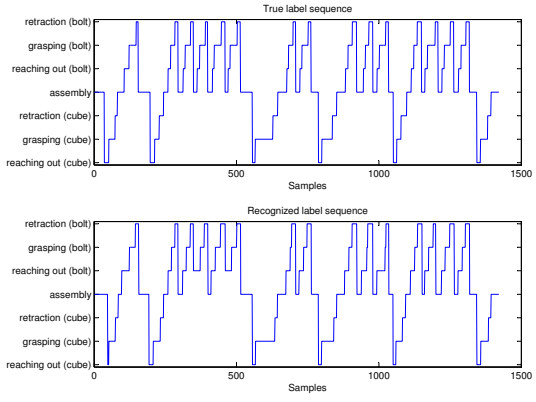


Fig. 6: **Recognized Workflow** - Ground truth label sequence (upper diagram) and recognized one (lower diagram) of the right hand using the camera tracking dataset

Set-up	**Data source of feature vector**	**Accuracy (%) left hand**	**Accuracy (%) right hand**
BAJA	All sensors + gaze	87.68 ± 5.32	95.67 ± 5.07
BAJA	Hand sensors	83.48 ± 7.39	$\mathbf{95.11 \pm 5.20}$
JAHIR	Camera tracking data	40.11	**92.26**

TABLE I: **Workflow recognition results for the left and the right hand** - The accuracy is the percentage of labels which correspond to the true ones. Correspondence also includes being in the same general movement. The results of the recognition rates for the right hand show that a transfer of the trained models to another set-up is possible.

All in all, our approach works quite well, even on the camera tracking data. Further improvement could presumably be achieved by lessening the interlinking of the position estimation of both hands.

VI. Conclusion and Future Work

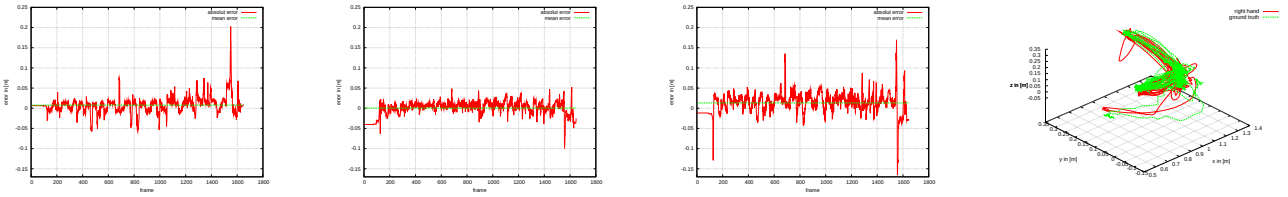
In this work, we have used composite HMMs to analyze the workflow of a joint assembly task between a human and a robot. The task consisted of building a tower by combining six cubes with a varying number of bolts. The composite HMMs used for workflow analysis were trained on a dataset collected in a reference experiment with 22 subjects and without any influence of a robot. Evaluations show that it is sufficient to use only the hand positions with the derived parameters (velocity, acceleration, jerk, and table zones) as input values. The label sequence of the right hand was recognized with an accuracy of $(95.11 \pm 5.20)\,\%$.

Since it is impossible to equip workers with artificial sensors or markers in an industrial setting, we propose a method to collect the same three-dimensional spatial information about the hands. The extension of an occupancy grid approach to three dimensions using two cameras offers a reliable (error standard deviation $\sim 0.02\,\mathrm{m}$) and fast ($>20\,\mathrm{fps}$) hand tracking in world coordinates. Using the composite HMMs of the "human" experiment, the label sequence of the right hand could be recognized with an accuracy of $92.26\,\%$ which is only marginally below the recognition rate of the reference experiments.

Based on this work, future steps can include the proactive support of a human worker during a complex assembly task, the automatic generation of an assembly report for later analysis. This might include error tracing or efficiency optimization.

VII. Acknowledgements

This work was partly supported by the DFG cluster of excellence "CoTeSys" (www.cotesys.org).



(a) (b) (c) (d)

Fig. 7: **Tracking results** - The graphs (a) to (c) show the position error for x, y, and z for the tracking of the right hand compared to ground truth data. Graph (d) shows the 3 dimensional trajectory for the right hand (red) for the tower assembly task along with the ground truth trajectory (green)

REFERENCES

[1] J. Krüger, R. Bernhardt, D. Surdilovic, and G. Spur, "Intelligent assist systems for flexible assembly," *CIRP Annals - Manufacturing Technology*, vol. 55, no. 1, pp. 29 – 32, 2006.

[2] J. Krüger, T. Lien, and A. Verl, "Cooperation of human and machines in assembly lines," *CIRP Annals - Manufacturing Technology*, vol. 59, 2009.

[3] M. Buss, M. Beetz, and D. Wollherr, "Cotesys - cognition for technical systems," in *Proceedings of the 4th COE Workshop on Human Adaptive Mechatronics (HAM)*, 2007.

[4] M. Huber, A. Knoll, T. Brandt, and S. Glasauer, "When to assist? - Modelling human behaviour for hybrid assembly systems," *Proceedings-ISR/ROBOTIK 2010*, 2010.

[5] T. Blum, N. Padoy, H. Feußner, and N. Navab, "Modeling and online recognition of surgical phases using hidden markov models," in *Medical Image Computing and Computer-Assisted Intervention*, ser. Lecture Notes in Computer Science, vol. 5242. Berlin / Heidelberg: Springer-Verlag, 2008, pp. 627–635.

[6] N. Padoy, T. Blum, H. Feussner, M.-O. Berger, and N. Navab, "On-line recognition of surgical activity for monitoring in the operating room," in *Proceedings of the 20th Conference on Innovative Applications of Artificial Intelligence*, July 2008, pp. 1718–1724.

[7] N. Padoy, T. Blum, S.-A. Ahmadi, H. Feussner, M.-O. Berger, and N. Navab, "Statistical modeling and recognition of surgical workflow," *Medical Image Analysis*, 2010.

[8] P. Lukowicz, J. A. Ward, H. Junker, M. Stäger, G. Tröster, A. Atrash, and T. Starner, "Recognizing workshop activity using body worn microphones and accelerometers," in *Pervasive Computing - LNCS*, ser. Lecture Notes on Computer Science, A. Ferscha and F. Mattern, Eds. Berlin / Heidelberg: Springer-Verlag, 2004, vol. 3001, pp. 18–32.

[9] V. Pashaloudi and K. Margaritis, "Feature extraction and sign recognition for greek sign language," in *Proceedings of the 11th IEEE Mediterranean Conference on Control and Automation (MED)*, 2003.

[10] T. Starner, J. Weaver, and A. Pentland, "Real-time american sign language recognition using desk and wearable computer based video," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 20, no. 12, pp. 1371–1375, Dezember 1998.

[11] M. Elmezain, A. Al-Hamadi, S. S. Pathan, and B. Michaelis, "Spatio-temporal feature extraction-based hand gesture recognition for isolated american sign language and arabic numbers," in *Proceedings of 6th International Symposium on Image and Signal Processing and Analysis (ISPA)*, 2009, pp. 254–259.

[12] M. Huber, A. Knoll, T. Brandt, and S. Glasauer, "Handing Over a Cube," *Annals of the New York Academy of Sciences*, vol. 1164, no. Basic and Clinical Aspects of Vertigo and Dizziness, pp. 380–382, 2009.

[13] E. Schneider, T. Villgrattner, J. Vockeroth, K. Bartl, S. Kohlbecher, S. Bardins, H. Ulbrich, and T. Brandt, "Eyeseecam: An eye movement-driven head camera for the examination of natural visual exploration," *Annals of the New York Academy of Sciences*, vol. 1164, no. 1, pp. 461–467, 2009.

[14] C. Lenz, S. Nair, M. Rickert, A. Knoll, W. Rösel, J. Gast, and F. Wallhoff, "Joint-action for humans and industrial robots for assembly tasks," in *Proceedings of the IEEE International Symposium on Robot and Human Interactive Communication*, 2008, pp. 130–135.

[15] C. Lenz, M. Rickert, G. Panin, and A. Knoll, "Constraint task-based control in industrial settings," in *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems*, IEEE/RSJ. St. Louis, MO, USA: IEEE, Oct. 2009, pp. 3058–3063.

[16] A. Bannat, T. Bautze, M. Beetz, J. Blume, K. Diepold, C. Ertelt, F. Geiger, T. Gmeiner, T. Gyger, A. Knoll, C. Lau, C. Lenz, M. Ostgathe, G. Reinhart, W. Rösel, T. Rühr, A. Schuboe, K. Shea, I. S. genannt Wersborg, S. Stork, W. Tekouo, F. Wallhoff, M. Wiesbeck, and M. F. Zäh, "Artificial cognition in production systems," *IEEE Transactions on Automation Science and Engineering*, vol. PP, no. 99, pp. 1–27, July 2010.

[17] F. Wallhoff, J. Blume, A. Bannat, W. Rösel, C. Lenz, and A. Knoll, "A skill-based approach towards hybrid assembly," *Advanced Engineering Informatics*, vol. 24, no. 3, pp. 329 – 339, Aug. 2010, the Cognitive Factory.

[18] A. Elfes, "Using occupancy grids for mobile robot perception and navigation," *Computer*, vol. 22, no. 6, pp. 46–57, June 1989.

[19] B. Schiele and J. Crowley, "A comparison of position estimation techniques using occupancy grids," *Robotics and autonomous systems*, vol. 12, no. 3-4, pp. 163–171, 1994.

[20] S. Thrun, "Learning occupancy grid maps with forward sensor models," *Autonomous robots*, vol. 15, no. 2, pp. 111–127, 2003.

[21] P. Stepan, M. Kulich, and L. Preucil, "Robust data fusion with occupancy grid," *IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews*, vol. 35, no. 1, pp. 106–115, 2005.

[22] C. Fulgenzi, A. Spalanzani, and C. Laugier, "Dynamic obstacle avoidance in uncertain environment combining PVOs and occupancy grid," in *Proceedings of the IEEE International Conference on Robotics and Automation*. IEEE, 2007, pp. 1610–1616.

[23] D. Beymer, "Person counting using stereo," in *Proceedings of the Workshop on Human Motion*, 2000, pp. 127–133.

[24] F. Fleuret, J. Berclaz, R. Lengagne, and P. Fua, "Multicamera people tracking with a probabilistic occupancy map," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 30, no. 2, pp. 267–282, Feb. 2008.

[25] L. Chen, G. Panin, and A. Knoll, "Multi-camera people tracking with hierarchical likelihood grids," in *Proceedings of the 6th International Conference on Computer Vision Theory and Applications (VISAPP11)*. Algarve, Portugal: INSTICC press, Mar. 2011.

[26] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 1, 2001.

[27] R. E. Kalman, "A new approach to linear filtering and prediction problems," *Transactions of the ASME–Journal of Basic Engineering*, vol. 82, no. Series D, pp. 35–45, 1960.

[28] L. Rabiner, "A tutorial on hidden markov models and selected applications in speech recognition," *Proceedings of the IEEE*, vol. 77, no. 2, pp. 257–286, 1989.