3D-Reconstruction of Indoor Environments from Human Activity

Barbara Frank   Michael Ruhnke   Maxim Tatarchenko   Wolfram Burgard

Abstract— **Observing human activities can reveal a lot about the structure of the environment, the objects contained therein and also their functionality. This knowledge, in turn, can be useful for robots interacting with humans or for robots performing mobile manipulation tasks. In this paper, we present an approach to infer the geometric and functional structure of the environment and the position of certain relevant objects in it from human activity. We observe this activity using a full-body motion capture system consisting of a set of inertial measurement units. This is a hard problem since our data suit provides odometry estimates only, which severely drift over time. Therefore, we regard the objects inferred from the activities as landmarks in a graph-based simultaneous localization and mapping problem, which we optimize to obtain accurate estimates about the poses of the objects and the trajectory of the human. In extensive experiments, we demonstrate the effectiveness of the proposed method for the reconstruction of 3D representations. The resulting models not only contain a geometric but also a functional description of the environment and naturally provide a segmentation into individual objects.**

I. INTRODUCTION

Tracking people and observing their activities is an important prerequisite for today's robots to provide better services for their users, to learn how to carry out tasks and how to behave in a socially compliant way. Since humans naturally interact with the real world, tracking their activities also provides interesting information about the environment and the functionality of objects in the environment. When learning accurate models of the environment, this knowledge can be used to better understand the structure and meaning of the environment and the manipulated objects. In the context of mobile robots, 3D models are typically built from the sensor data of either 3D laser range finders or RGB-D cameras. Such models provide an accurate metric representation of the entire environment, but do not contain information on individual objects or even their function. Thus, in this paper, we propose a method to build 3D environment models from tracked human postures by observing specific activities. In addition to geometric information about objects, this allows us to learn about the function of objects and possibilities for interaction with them. A chair, for instance, is typically used to sit down, a table allows for placing objects on it. Such models can also be useful for mobile robots, for instance, to carry out cooperative tasks together with humans or to plan tasks that involve interactions with objects and require knowledge about their function. Furthermore, a mobile robot could directly use the geometric information for localization or planning of actions.
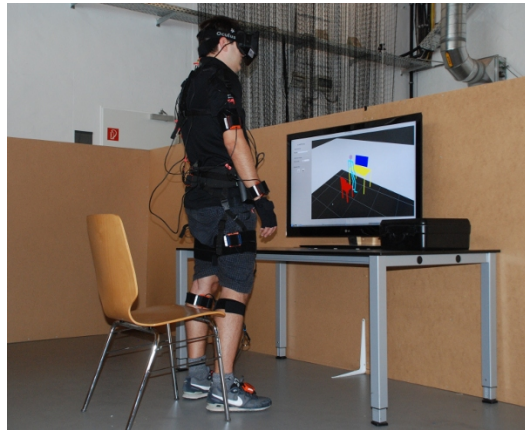
Fig. 1.   Environment and corresponding 3D model reconstructed from human activity using our approach.

In our system, the user is wearing a motion capture suit [14], which provides an estimate of the full body posture from 17 inertial measurement units (IMUs) at a frequency of 120 Hz. With the data suit, the human posture is fully observable, independent of the view point. In contrast, a mobile robot equipped with an RGB-D camera and a skeleton tracker would have to change its positions and might fail in situations, in which a human itself occludes parts of his body. Our approach shares some ideas with the work of Grzonka et al. [7, 8]. They estimate the 3D trajectory of a human based on door-handling and stair-climbing activities and also consider these activities as landmarks in a graph-based simultaneous localization and mapping (SLAM) problem. This formulation allows them to correct for odometry errors that accumulate due to the inherent drift of the suit. Based on the detected doors, they are able to build topological 2.5D maps of environments containing multiple levels. Our method also uses specific activities as landmarks to construct a graph-based SLAM problem and it reduces the rapidly accumulating drift in the odometry. In contrast to their work, however, our method aims at building geometric 3D models of the basic structure of an environment including certain relevant objects by inferring walls, tables, screens, and chairs. With our method, it is possible to quickly build accurate 3D reconstructions of indoor environments and to simultaneously reduce the drift of the motion capture suit.

Fig. 1 illustrates the application of our system. It shows the user wearing the data suit in an example environment along with some considered objects, namely walls, tables, screens, and chairs. Furthermore, on the screen, we can see the 3D model of the environment reconstructed with our approach.
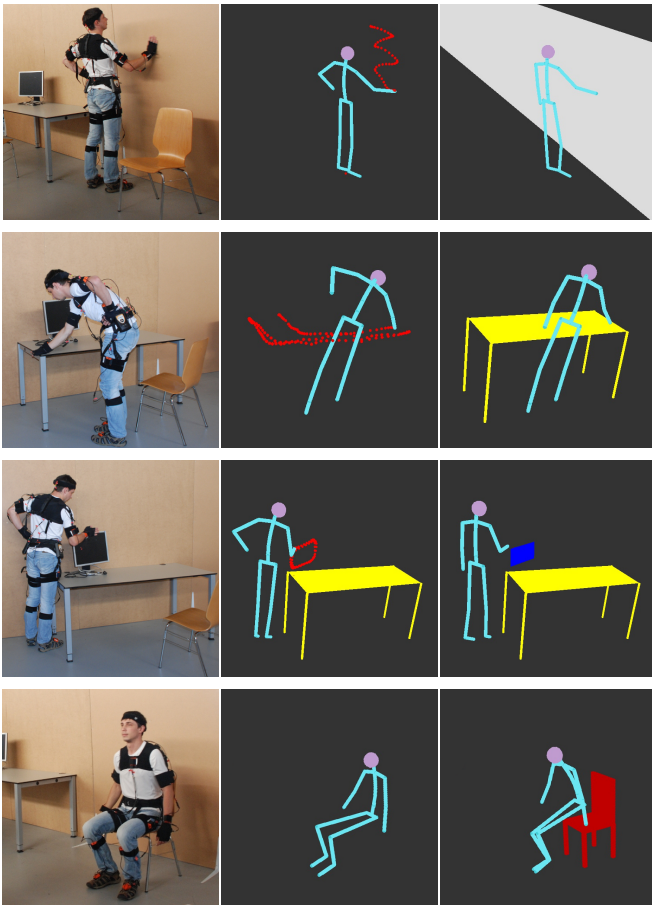
Fig. 2. Activities, collected data points and corresponding reconstructed objects.

II. RELATED WORK

Localizing and tracking humans in indoor environments is an active area of research and has received considerable attention in the past. Applications range from search and rescue missions [15, 4] over reconstruction of environment models [7, 8] to tracking of the user in a virtual environment [3]. Fischer and Gellersen [4] present an overview over the advantages and disadvantages of different technologies for tracking emergency responders. Using IMU data to track persons is a challenging task due to the drift inherent in the measurement principle.

Lee and Mase [13] use wearable accelerometers and dead reckoning to recognize different activities such as walking, standing and climbing stairs as well as changes in the user's location in a known environment. Foxlin [6] tracks pedestrians using shoe-mounted inertial measurement units and handles drift using zero-velocity updates in an extended Kalman filter. Ojeda and Borenstein [15] pursue similar ideas, and also include heuristics to deal with gyroscope errors [1]. Ziegler et al. [17] presented an approach to deal with the drift by tracking the human with a mobile robot using a particle filter.

While the approaches discussed so far deal with the localization of users in known environments, there exist different approaches to simultaneously generate a map from activities of the user in a new environment. Schindler et al. [16], for instance, use a wearable gesture interface equipped with an infrared proximity sensor and a dual axis accelerometer to build a topological map of the environment. The accelerometer detects footsteps, while the proximity sensor identifies doorways. To distinguish different doorways and to uniquely label doors, the user performs finger pointing gestures. Cinaz et al. [2] use a laser scanner and an inertial measurement unit mounted on a helmet to build 2D maps of the environment and to localize the user wearing the helmet within the map. Grzonka et al [7, 8] use door-handling and stair-climbing activities to build approximate 2.5D maps of environments. A multi-hypothesis tracker handles different possible data associations and a graph-based optimization corrects for the drift of the suit. Our system shares ideas with this approach. In contrast to their approximate topological-metric maps, the outcome of our system is a functional 3D model that contains typical objects humans are dealing with.

Virtual and augmented reality applications are becoming more and more popular. Tracking the activities of users in such virtual environments seems desirable to allow for inter-activity. Damian et al. [3], for instance, track the movements of agents with a data suit similar to ours. Since they consider only local movements and no walking activity, they do not account for drift in the global pose estimate.

Knowledge about human activity can also be useful in the context of scene recognition, inference of scene geometry and object arrangement. Higuchi et al. [9] introduced a method for scene recognition and segmentation that employs conceptual knowledge on relationships between human actions and objects. In their setup, stereo cameras are used to track human motions and interactions with objects. The resulting 3D environmental map contains objects labeled according to their function. Fouhey et al. [5] use human actions for inferring constraints on the 3D structure of the environment from single-view time-lapse images. Observing and recognizing actions such as sitting, walking, or reaching allows them to identify functional regions in the environments corresponding to floors, free space, space for placing objects or sitting, and to reason on occupied voxels in the scene, which finally is useful for scene understanding and recognition. The work of Jiang and Saxena [10, 11] deals with arranging objects in the environment and inferring their best placement. They do not directly observe humans but reason about interactions between humans and objects by considering the likelihood of different possible human postures in different places and in relation with different objects. As they demonstrate, a robot can use the resulting models for semantic scene labeling and object arrangement.

In contrast to these approaches, our contributions are two-fold: first of all, we are able to reconstruct 3D models containing objects relevant to humans, such as walls, chairs, and tables, together with their functionality. Furthermore, since we consider recurring encounters of the user with these reconstructed objects, we are able to correct for the drift in the user's pose estimate.

III. System Overview

In our approach, we observe the activities and gestures of a person using a data suit that provides us with the full-body pose estimate at high frequencies. We extract different types of 3D objects from predefined activities and gestures. To obtain a consistent model of the environment, we use a graph-based solver for the simultaneous localization and mapping (SLAM) problem. In such a graph, the nodes correspond to the poses along the trajectory and the edges correspond to the relative movements between these poses as measured with the motion capture suit. Additionally, we consider the individual reconstructed objects as landmark observations connected to the user poses. Whenever the user revisits an object that is already contained in the model, we insert a loop closure constraint between the current pose and the observed landmark into the graph. In our current system, we make data associations using a nearest-neighbor approach. Whenever a loop is closed, we perform a least-squares optimization of the pose graph to correct for accumulated errors and to obtain a globally optimal solution. The resulting graph contains the pose estimates of landmarks as well as pose estimates of the person. Finally, we apply a post-processing step to convert the planes for the walls into rectangles. In the remainder of this paper, we first describe our landmark extraction scheme, followed by a detailed description of the graph construction and the objective function we solve with a state-of-the-art least-squares optimization framework [12].

IV. Activities and Landmark Extraction

Since the data suit only provides an odometry estimate, we extract landmarks from gestures and activities of the user. In our current system, we consider different kinds of activities that correspond to different types of objects: chairs, walls, tables and general rectangular structures (e.g., screens and TV sets). We represent chairs by their position and orientation in 3D space, and specify all other objects as planes in hessian normal form (additionally we store the centers of masses for the planes). We extract planar objects from the motion of the user's right hand. We defined a specific gesture indicating that points should belong to the model: left hand on hip. If we detect this gesture, that is if the angle between the upper arm and the forearm is below a threshold, we use the corresponding positions of the right hand for modeling.

Chairs are extracted and inserted into the model whenever the user is sitting. We detect this type of activity by measuring the distance between the sensor attached to the pelvis bone and the sensor attached to the upper leg of the user; if this is below a certain threshold, we derive the position of the landmark from the hip position (with some constant offset) and the orientation from the user's upper body.

Walls (as well as other planar objects) are modeled by fitting a plane to a set of data points. We classify a plane as a wall if its orientation vector is approximately parallel to the floor. While collecting points, the user can move his right hand arbitrarily over the surface of the wall, the only constraint is that the first and the last point should
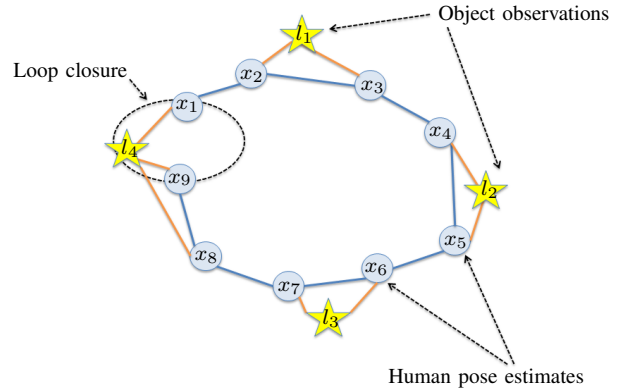

Fig. 3. Graph-based formulation of the simultaneous localization and mapping problem.

be sufficiently far away from each other (to not confuse walls with general rectangular structures). For each wall, we initially assume that it extends infinitely, when we add more walls to the model, however, we cut them at intersections.

Rectangular structures (like screens, posters etc.) are modeled by fitting a plane to a set of data points collected when the user moves his hand along the edges of the object. We extract this type of landmark when the first and the last point of the model are close to each other. To find the extents of the object after fitting a plane, we detect its corners in the data. Let $\{x_i\}$ be an ordered set of data points. For each point x_i, we consider points from its local neighborhood in positive and negative direction $\{x_i^+\}$ and $\{x_i^-\}$. We fit a line to each of those two subsets and check the angle θ between the lines. At the corner points, there are local maxima of θ.

Tables are extracted from the motion of the user's hand along the two adjacent edges. We detect the corner of the intersection for these two lines in the same way as for general rectangular structures. To obtain two more corners, we then find the two points in the point set that are maximally far away from each other. The line segment between these two points is the diagonal of the table. We obtain the last corner – we assume rectangular tables – by mirroring the first one over the diagonal. We classify an object as a table if the normal vector of the plane is approximately perpendicular to the floor plane.

Fig. 2 illustrates how we extract these different types of objects and shows the gestures performed by the user, the extracted data points and the corresponding 3D models of the objects.

V. Least-Squares Optimization

So far, we have a set of extracted objects or landmarks and the six degree-of-freedom pose estimates for every segment of the skeleton generated by the data suit. To construct a SLAM pose graph, we use the pelvis bone sensor as reference on the body of the user. Since the data suit provides us with pose estimates at 120 Hz, we down-sample the poses to segments of 20 cm, that is we insert a new odometry node whenever the user has moved for more than 20 cm

with respect to the last odometry node, in order to keep the number of nodes in the graph at a manageable level. For all resulting poses of the reference sensor, we insert nodes in the graph and link such nodes with edges that correspond to the relative data suit pose estimates as measurements. More formally, a measurement $\mathbf{z}_{ij}$ describes the relative transformation between nodes poses $\mathbf{x}_i$ and $\mathbf{x}_j$. Based on such a constraint, we can calculate the error

$$\mathbf{e}(\mathbf{x}_i, \mathbf{x}_j, \mathbf{z}_{ij}) = \mathbf{z}_{ij} \ominus (\mathbf{x}_i \ominus \mathbf{x}_j). \tag{1}$$

The weighted error for odometry constraints is then

$$\mathbf{e}^o_{ij} = \mathbf{e}(\mathbf{x}_i, \mathbf{x}_j, \mathbf{z}_{ij})^\mathsf{T} \Omega^o_{ij} \mathbf{e}(\mathbf{x}_i, \mathbf{x}_j, \mathbf{z}_{ij}), \tag{2}$$

where Ω^o_{ij} denotes the information matrix used to weight the error for a constraint $\mathbf{z}_{ij}$ according to the uncertainty of a measurement. In our implementation, we chose an uncertainty of 1 cm for translation per 20 cm segment.

Since the user touches all landmarks except the chairs with the right hand, we also obtain relative measurements for each landmark by relating the poses of the right hand to the reference joint on the body. The relative inner body pose estimates are not subject to drift since an underlying biomechanical model constrains them. Furthermore, the relative orientations reported by the IMUs are typically accurate since the gravity vector is globally consistent. Therefore, we add landmark edges between the pose $\mathbf{x}_r$ and the landmark $\mathbf{l}_k$ with a measurement $\mathbf{z}_{rk}$ to the graph similar to the odometry edges, but with a different information matrix Ω^l_{rk}, which assumes a smaller uncertainty. The measurement $\mathbf{z}_{rk}$ corresponds to the relative position of the landmarks center of mass. We estimate the orientation of the landmark from the collected data points as described in the previous section. The weighted error for landmark constraints is then

$$\mathbf{e}^l_{rk} = \mathbf{e}(\mathbf{x}_r, \mathbf{l}_k, \mathbf{z}_{rk})^\mathsf{T} \Omega^l_{rk} \mathbf{e}(\mathbf{x}_r, \mathbf{l}_k, \mathbf{z}_{rk}). \tag{3}$$

In addition, we consider loop closure constraints whenever the user revisits previously observed landmarks. Fig. 3 illustrates the constructed graph. To identify loop closures, we make nearest-neighbor data associations. Thus, we assume that individual landmarks are sufficiently far away. Whenever two landmark observations are below a given threshold, we insert a loop closure constraint into the graph.

Loop closure constraints describe the difference in the expected and measured relative transformation between the user and the object in the map. We consider walls separately, since the user does not necessarily return to the same spot and only his distance in the direction normal to the wall is relevant when minimizing the error in the observation. Therefore, we model the weighted error between a pose $\mathbf{x}_i$ corresponding to the position of the right hand with respect to the estimated wall $\mathbf{w}_j$ as follows

$$\mathbf{e}^w_{ij} = (\mathbf{z}_{i,j} - h(\mathbf{x}_i, P_j))^\mathsf{T} \Omega^w_{ij} (\mathbf{z}_{i,j} - h(\mathbf{x}_i, P_j)), \tag{4}$$

where $h(\mathbf{x}_i, P_j)$ computes the distance of the hand to the plane, Ω^w_{ij} is the corresponding information matrix for walls and $\mathbf{z}_{ij}$ is the current observation of a plane, fit to the data points.
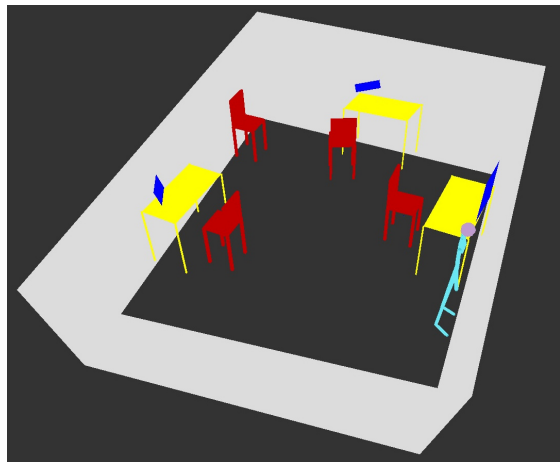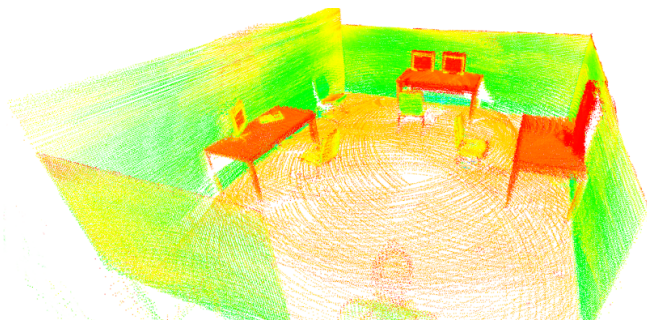

Fig. 4. 3D model computed with a traditional graph-based SLAM approach using the data of a 3D laser range finder. This example office environment consists of three tables and four chairs which are the landmarks for our SLAM method. The corresponding model reconstructed with our method is shown below.

Based on the three different types of constraints, we formulate a joint optimization problem with the following objective function:

$$\mathbf{x}^*, \mathbf{l}^*, \mathbf{w}^* = \underset{\mathbf{x}, \mathbf{l}, \mathbf{w}}{\operatorname{argmin}} \quad \sum_{i,j} \mathbf{e}^o_{ij} + \sum_{i,j} \mathbf{e}^l_{ij} + \sum_{i,j} \mathbf{e}^w_{ij}. \tag{5}$$

This function is then optimized using the g^{2}o framework of Kümmerle et al. [12]. While the user is acquiring data, we incrementally add objects to the model, check for loop closures and carry out an optimization step whenever the user revisits a landmark.

VI. EXPERIMENTAL EVALUATION

We evaluated our approach on data sets acquired with an inertial motion capture system. The inertial measurement units (IMUs) consist of 3D linear accelerometers, 3D rate gyroscopes and 3D magnetometers, which are fused to obtain the positions and orientations of the sensors. The data suit performs a pre-processing and filtering of the raw data to prevent the individual IMUs from drifting apart. To this end, it uses a biomechanical model of the human body that constrains the movements of the IMUs. This model requires measurements of the individual body segments such as body height, arm and leg length for each user, which are provided manually. Furthermore, it requires the user to perform a calibration step in a so-called T-pose, prior

to operation. In total, we collected nine data sets in two different office environments with multiple walls, tables, chairs and screens. In each run, we collected data with different gestures and multiple loop-closures over three to six minutes. For evaluation of our system, we additionally recorded the trajectory of the person using a highly-accurate and well-calibrated optical motion capture system. For this purpose, the user wore a helmet with uniquely identifiable markers during the experiments. Due to the nature of the measurements from the IMU sensors, the open-loop pose estimates are subject to serious drift, which is already evident at the scale of small rooms. Therefore, we designed the experiments in a way to close loops as often as possible and avoided trajectories longer than 5 m to 10 m without loop closure. In this way, we confine the pose uncertainty to reasonable boundaries and are able to successfully solve the data association problem.

Interesting questions arise from the chosen sensor setup. For SLAM solutions, the accuracy of the resulting model and the pose estimate are typically the quantities of interest. To provide an intuition about the achievable accuracy of the 3D model computed by our method, we collected a set of 3D range scans and created a reference point cloud for comparison. Fig. 4 shows the resulting model computed with a traditional 3D SLAM approach and the corresponding model generated with our approach for visual comparison. We evaluated the size of the reconstructed model and compared it to the ground truth provided by the point cloud. In detail, we determined for each object the mean absolute percentage error:

$$\text{MAPE} = \frac{100\%}{n} \sum_{i=1}^{n} \frac{|\hat{y}_i - y_i|}{y_i},$$

where $\hat{y}_i$ are the edge lengths of the reconstructed objects and y_i the actual edge lengths. Tab. I summarizes the results. The error is around 10 % for walls and ranges from 10 % to 30 % for tables. In comparison to models generated with vision, laser or other sensors typically used in robotics, the models obtained with our method are less accurate. This results from the fact that we rely on the raw odometry of the suit to collect data points of individual objects, which accumulates errors and is particularly evident for larger objects such as the tables. Here, the user had to move a lot to cover all edges. Furthermore, the metal found in the table lead to disturbances and increased the error in the pose estimate of the suit.

We evaluated the accuracy of the pose estimates for the open-loop IMU trajectory as well as our optimized SLAM solution with respect to the ground truth provided by the optical motion capture system. Fig. 5 shows the resulting user trajectories for one experimental run. As this plot illustrates, after one round through the environment, starting and end pose of the open-loop estimate are more than 1 m apart. Our method, in contrast, stays close to the true trajectory of the user as reported from the optical motion capture. We observe some differences in local trajectories, for instance, in the upper left corner of the plot. These result from head movements during collection of data points from the tables,
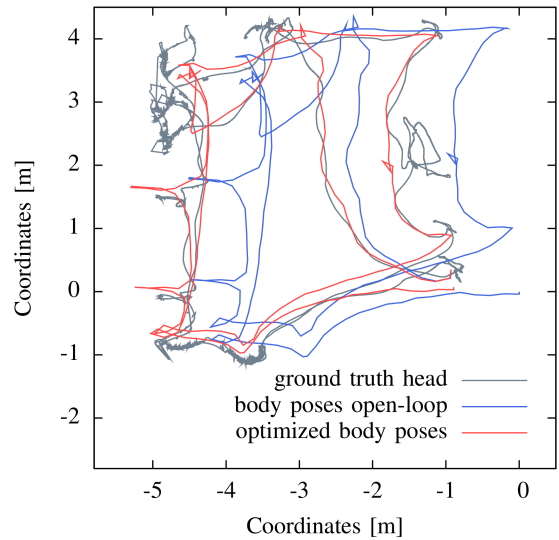

Fig. 5. Trajectory comparison between open-loop body poses, ground truth head pose (optical motion capture) and the body poses optimized by our method.
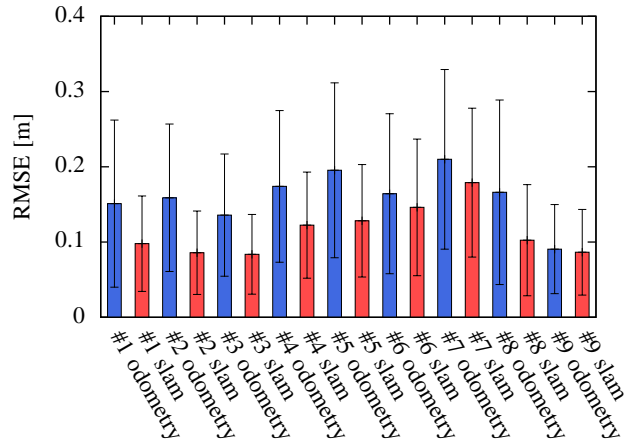

Fig. 6. Trajectory comparison of the open-loop body poses and the body poses optimized by our method to the ground truth head pose recorded with optical motion capture for nine different data sets.

while the hip position considered in our approach did not move a lot. Fig. 6 provides a quantitative evaluation of the reconstructed trajectories in terms of the root mean square error (RMSE) for all nine data sets. It shows that the RMSE is always smaller for the optimized trajectories compared to the open-loop estimate from the suit. The error for the open-loop estimates grows unbounded with the length of the trajectory. As an example, Fig. 7 shows a comparison of the generated models after three rounds through a 3.9 m × 8.7 m big environment. For this data set, we measured the size of the room manually. The reconstructed walls have a length of 8.6 × 3.8 × 8.8 × 3.5 m, thus, the average error of the walls is less than 5 % for this environment. Hence, our approach is able to globally correct the trajectory and to explain all data associations correctly. A video demonstrating our system can be found online.[1]

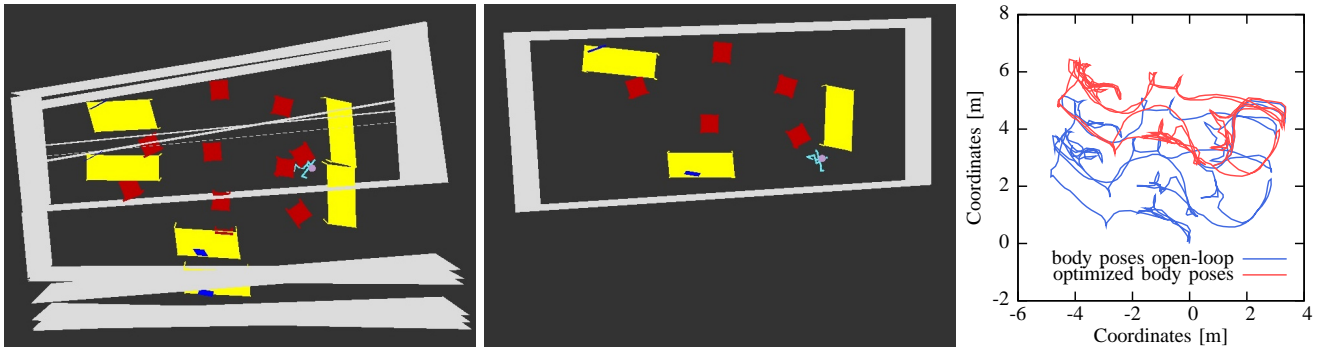[1] http://ais.informatik.uni-freiburg.de/projects/mvn/#3Drec

Fig. 7. An example environment reconstructed with our approach, shown are the models obtained with raw odometry containing several walls and tables (left) and the model obtained with our approach (middle). The plot on the right shows an overlay of the optimized trajectory and the odometry from the suit.

TABLE I

ACCURACY OF RECONSTRUCTED OBJECTS IN TERMS OF THE MEAN ABSOLUTE PERCENTAGE ERROR.

Object	# instances	avg. reconstruction error (%)
Tables	36	26.3 ± 8.7
Big screens	9	23.1 ± 2.7
Small screens	23	19.3 ± 4.8
Walls	36	7.1 ± 3.5

VII. CONCLUSIONS AND FUTURE WORK

In this paper, we presented an approach to build object-centric 3D models environments based on the activity of a user captured with a wearable data suit. Our approach uses specific gestures of the user to infer certain objects and their positions. We employ a SLAM formulation of the overall problem and optimize the data using a least-squares method. As a result, we obtain an accurate 3D model of the environment and at the same time a corrected pose estimate for the user. Practical experiments demonstrate that the reconstructed models are topologically correct and that the reconstructed objects as well as the recovered trajectory are accurate. In addition to geometric information, our model naturally provides a segmentation into individual objects and contains information on their functionality.

In future work, we will additionally consider dynamic environments and will allow the user to move individual objects in the scene. Further directions for future work could include using information on natural movements of the user during everyday activities in his environment such as opening doors or placing objects. We also envision integrating visual information into the system and performing object class recognition on the images. Combined with the data from the suit, this could give an even better semantic understanding of the environment.

REFERENCES

[1] J. Borenstein, L. Ojeda, and S. Kwanmuang. Heuristic reduction of gyro drift for personnel tracking systems. *Journal of Navigation*, 62(1):41–58, 2009.

[2] B. Cinaz and H. Kenn. HeadSLAM - simultaneous localization and mapping with head-mounted inertial and laser range sensors. In *IEEE International Symposium on Wearable Computers*, 2008.

[3] I. Damian, F. Kistler, M. Obaid, R. Büling, M. Billinghurst, and E. André. Motion capturing empowered interaction with a virtual agent in an augmented reality environment. In *Proc. of the International Symposium on Mixed and Augmented Reality*. IEEE, 2013.

[4] C. Fischer and H. Gellersen. Location and navigation support for emergency responders: A survey. *IEEE Pervasive Computing*, 9(1):38–47, 2010.

[5] D. F. Fouhey, V. Delaitre, A. Gupta, A. A. Efros, I. Laptev, and J. Sivic. People watching: Human actions as a cue for single-view geometry. In *Proc. of the 12th European Conference on Computer Vision*, 2012.

[6] E. Foxlin. Pedestrian tracking with shoe-mounted inertial sensors. *IEEE Computer Graphics and Applications*, 25(6):38–46, 2005.

[7] S. Grzonka, F. Dijoux, A. Karwath, and W. Burgard. Mapping indoor environments based on human activity. In *Proc. of the IEEE International Conference on Robotics and Automation (ICRA)*, 2010.

[8] S. Grzonka, A. Karwath, F. Dijoux, and W. Burgard. Activity-based Estimation of Human Trajectories. *IEEE Transactions on Robotics (T-RO)*, 8(1):234–245, 2 2012.

[9] M. Higuchi, S. Aoki, A. Kojima, and K. Fukunaga. Scene recognition based on relationship between human actions and objects. In *Proc. of the 17th International Conference on Pattern Recognition (ICPR)*, 2004.

[10] Y. Jiang and A. Saxena. Hallucinating humans for learning robotic placement of objects. In *International Symposium on Experimental Robotics (ISER)*, 2012.

[11] Y. Jiang and A. Saxena. Infinite latent conditional random fields for modeling environments through humans. In *Robotics: Science and Systems (RSS)*, 2013.

[12] R. Kümmerle, G. Grisetti, H. Strasdat, K. Konolige, and W. Burgard. g2o: A general framework for graph optimization. In *Proc. of the IEEE International Conference on Robotics and Automation (ICRA)*, 2011.

[13] S.-W. Lee and K. Mase. Activity and location recognition using wearable sensors. *IEEE Pervasive Computing*, 1(3):24–32, 2002.

[14] Xsens MVN Suit. http://www.xsens.com/en/general/mvn.

[15] L. Ojeda and J. Borenstein. Non-GPS navigation for security personnel and first responders. *Journal of Navigation*, 60(3):391–407, 2007.

[16] G. Schindler, C. Metzger, and T. Starner. A wearable interface for topological mapping and localization in indoor environments. In *Lecture Notes in Computer Science*, volume 3987. Springer, 2006.

[17] J. Ziegler, H. Kretzschmar, C. Stachniss, G. Grisetti, and Wolfram Burgard. Accurate human motion capture in large areas by combining IMU- and laser-based people tracking. In *Proc. of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2011.