Supporting a Human-Aware World Model through Sensor Fusion

Dominik Riedelbauch, Tobias Werner, and Dominik Henrich

Lehrstuhl für Robotik und Eingebettete Systeme
Universität Bayreuth, D-95440 Bayreuth, Germany
dominik.riedelbauch@uni-bayreuth.de,
http://robotics.uni-bayreuth.de

Abstract. Recent research in robotics aims at combining the abilities of humans and robots through human-robot collaboration. Robots must overcome additional challenges to handle dynamic environments within shared workspaces. They especially must perceive objects and the working progress to synchronize with humans in shared tasks. Due to unpredictable human interaction, local information about objects detected by eye-in-hand cameras and stored within a world model falls in value as soon as respective objects get out of sight. Our contribution is an approach to making world models aware of human influences and thus allowing robots to decide, whether information is still valid. To this end, we annotate pieces of information with certainty values encoding how trustworthy they are. Certainty is adapted over time according to additional knowledge about human presence within the workspace, provided by a global sensor. Thus, we achieve human-awareness through fusion of local and global sensor data. Our concept is validated through a prototype implementation and experiments that regard certainty of objects in different scenarios of human presence.

Keywords: world model, data aging, sensor fusion

1 Introduction

Recent research in robotics aims at systems combining the abilities of humans and robots through human-robot collaboration. Those systems provide the flexibility needed to apply robots in small businesses, the service sector, and domestic homes. However, robots must be able to perceive, construe and react to unpredictably changing environments outside industrial work cells.

Application scenarios for human-robot collaboration often incorporate the manipulation of objects that are small in relation to humans and robot manipulators moving within the workspace (Fig. 1, left). A high level of occlusion may be expected when using cameras in fixed positions, rendering their usage impractical. Consequently, eye-in-hand cameras are more suitable for maintaining a world model of objects used within shared tasks. However, humans may arbitrarily change the state of objects previously stored within the world model.

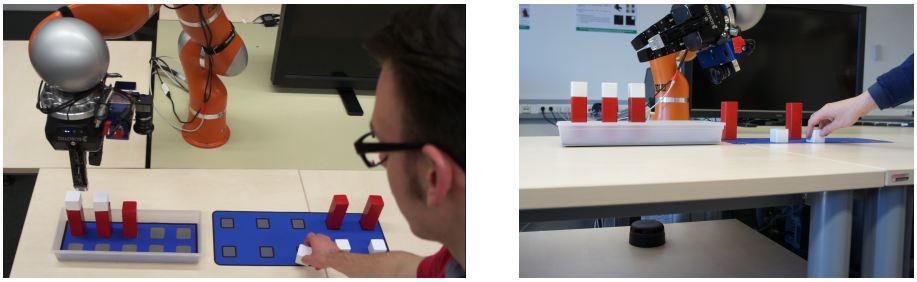

Fig. 1. Human-robot collaboration scenarios often incorporate the manipulation of objects that are small in relation to humans and robots (left). Data gathered with eye-in-hand cameras falls in value, as humans may change the environment while the robot looks away (right). Additional global sensors, e.g. a 2D LIDAR scanner (right, below table) can be used to raise human-awareness in world models constructed from local sensor data.

This makes local sensor data fall in value as soon as respective objects get out of sight (Fig. 1, right). Although global sensors are not suitable for task-related perception, they can be used to make world models aware of human influences. Such human-awareness helps robots to decide whether data is outdated or can be reused, e.g. for task planning. We contribute an approach to making world models human-aware through sensor fusion. Each object detected by local sensors is annotated with an individual certainty value. This value encodes the confidence in the correctness of object properties currently stored in the world model. Certainty is abated over time based on information about human presence within the workspace, gathered through a global sensor.

We review related work in Section 2. Details concerning our approach are reported in Section 3. The results of experiments with a robot-mounted, calibrated pair of color and depth camera as local sensors and a 2D LIDAR scanner as a source of global information are presented in Section 4.

2 Related Work

The ability to detect humans is a prerequisite for realizing any human-aware application. With application scenarios ranging from human-robot collaboration over autonomous cars to smart devices and ubiquitous computing, research has already proposed a wide variety of algorithms specifically designed to detect humans in incoming sensor data.

Algorithms for human detection differ mainly in their input (e.g. color images [17], depth images [21], range scans [2]) and output (e.g. estimated location [5], pose [13], or occupied volume [9]). In terms of output data, algorithms additionally fall into one of two variants: Binary variants (e.g. [13]) yield a single hypothesis per human, while probabilistic variants (e.g. [8]) offer one or more hypotheses, each with an estimated certainty. Internally, both binary and probabilistic variants for human detection work with an assortment of deterministic

and probabilistic strategies, such as codebooks (e.g. [3]), neural networks (e.g. [16]), particle filters (e.g. [10]), or Kalman filters (e.g. [15]). Additional features include support for human tracking over time (e.g. [15]), intrinsic support for sensor fusion (e.g. [9]), or gesture recognition (e.g [18]).

Most algorithms for human detection depend on a specific type of sensor. In this context, two algorithms for human detection are particularly relevant to our contribution: Both [2] and [14] use a 2D laser scanner at knee-height to detect humans. The former approach derives a strong classifier for human legs by applying AdaBoost on a variety of weak classifiers for 2D point sets. The latter approach combines object tracking with 2D point clustering and heuristics to estimate probabilities for human presence in individual 2D clusters. Opposed to both approaches, our contribution desires to exploit the results of human detection to derive a human-aware certainty for objects within the robot workspace.

Once humans have been detected within sensor data, the remaining application must consider detection results through an appropriate system reaction. Reactions depend strongly on the application scenario. In the scenario of human-robot collaboration, current research studies a variety of application scenarios, including risk-minimized path planning (e.g. [7]), gesture-based robot programming (e.g. [18]), and human-aware task planning (e.g. [1]). Our contribution does not depend on a specific application scenario, but instead we attach certainty to objects within a world model for use by arbitrary applications.

To intuitively integrate the results of human detection and the final system reaction, a structured approach proposes benefits over ad-hoc solutions. Popular structured approaches to software engineering for robot systems — robot system architectures (e.g. ROS [11]), knowledge databases (e.g. RoboBrain [12]) or geometric world models (e.g Octomap [20]) — have individual flaws in efficiency, extensibility, or simplicity. Our contribution avoids these flaws by relying on the alternative ENACT (ENtity-ACTor) world model [19], which has been designed specifically to be efficient, extensible, and intuitive. In ENACT, a set of entities $E = \{e_1, ..., e_{|E|}\}$ models each relevant object of the physical world (e.g. the robot, humans, workpieces) as an individual entity e_i. Aspects a_j from a set of aspects $A = \{a_1, ..., a_{|A|}\}$ govern entity attribute classes (e.g. pose, color, weight, certainty). One or more world contexts W bind data to pairs of entity and aspect (e_i, a_j). Finally, threaded actors $\mathfrak{a}_\mathfrak{k}$ from a set of actors $\mathfrak{A} = \{\mathfrak{a}_1, ..., \mathfrak{a}_{|\mathfrak{A}|}\}$ perpetually update data within world contexts through a sophisticated locking and synchronization mechanism. For instance, actors realize perception, sensor fusion, path and task planning, or robot control.

3 Our Approach

The structure of our approach to fusing local and global sensor data for achieving a human-aware ENACT world model is depicted in Fig. 2. Local information about objects is generated through an *object recognition actor* $\mathfrak{a}_{\text{recognition}}$. This actor processes images provided by the camera actor $\mathfrak{a}_{\text{camera}}$ and updates values of aspects within the world context W, e.g. the object pose a_{pose}. The *human de-*

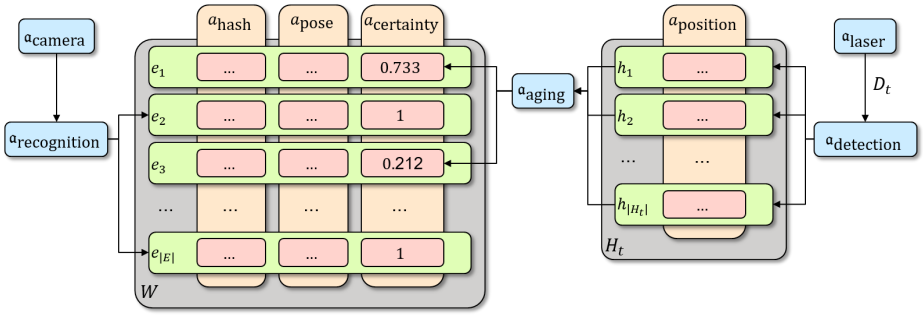

Fig. 2. Our human-aware world model is realized using the ENACT framework. Actors provide camera images ($\mathfrak{a}_{\text{camera}}$) and 2D LIDAR scans ($\mathfrak{a}_{\text{laser}}$). Object recognition ($\mathfrak{a}_{\text{recognition}}$) is used to update aspect values within a world context W. The detection actor ($\mathfrak{a}_{\text{detection}}$) stores information about the position of humans in another context H_t. H_t is used by the aging actor $\mathfrak{a}_{\text{aging}}$ to derive the current certainty for all entities.

tection actor $\mathfrak{a}_{\text{detection}}$ receives global sensor information and stores the position of humans within the workspace in another world context H_t. The actual sensor fusion is performed by the *aging actor* $\mathfrak{a}_{\text{aging}}$. This actor determines a certainty value $C_t(e)$ for each entity e within W at time t and updates the respective value of the *certainty aspect* $a_{\text{certainty}}$. Certainty is decreased with time depending on how comfortably an entity can be accessed by the humans currently present at the workbench (Sections 3.1 and 3.2). Thus certainty encodes confidence in the current values of the other aspects of e, similar to a probability of their correctness. A robot system that uses the human-aware world model can decide to reuse previously extracted information with a high certainty value, and may purge aspect values of entities with low certainty from the world context.

We favor a 2D LIDAR range scanner placed below the workbench that humans and robots work at as global sensor. This type of sensor is better suited for our small scale collaboration scenarios, as it is less invasive and needs a less complex setup routine than e.g. a multi camera system for human tracking. Moreover, LIDAR data encodes information about the position of humans more compactly than camera images. This way, the computing effort of $\mathfrak{a}_{\text{detection}}$ is reduced from human tracking in large images to detection of leg silhouettes in significantly smaller data sets. However, our approach can easily be adapted to other global sensor systems by exchanging the actors $\mathfrak{a}_{\text{laser}}$ and $\mathfrak{a}_{\text{detection}}$.

3.1 Formal Definition of Human-Aware Certainty

In formal terms, we are looking for some *certainty function* $C_t : E \to [0, 1]$ that maps an entity to a certainty value at time t to realize $\mathfrak{a}_{\text{aging}}$. C_t may use two types of sensor data as input: The local data is given as a subset $V_t \subseteq E$ of those entities currently within the field of view of the eye-in-hand camera, output by $\mathfrak{a}_{\text{recognition}}$. The detection actor determines the *human presence map*

$H_t = \{h_1,\ h_2,\ ..., h_{|H_t|}\}$ of points $h_i \in \mathbb{R}^2$ in the sensor plane that belong to silhouettes of humans. Based on this input, the certainty $C_t(e)$ of entity e at time t is calculated incrementally from $C_{t-1}(e)$ by evaluating

$$C_t(e) = \begin{cases} 1 & \text{if } e \in V_t, \\ \max\left(0,\ C_{t-1}(e) - \lambda \cdot A_{\text{acc}}(e, H_t)\right) & \text{if } e \notin V_t \wedge t > 0, \\ 0 & \text{else} \end{cases}.$$

As long as an entity is seen by the camera ($e \in V_t$), certainty is 1. Entities outside the field of view are aged by decrementing the value of the previous time step. The decrement mainly depends on the *accessibility* $A_{\text{acc}} : E \times 2^{\mathbb{R}^2} \to [0, 1]$. A_{acc} can be understood as a measure for probability of some entity to be modified under the current human presence. Entities that are within the sphere of influence of many humans are highly accessible and thus need to age quickly. On the other hand, an object that is far away from any human presence has low accessibility, as there is no need to lose trust in its current state within the world model. We additionally use a constant rate λ to control the impact of accessibility on certainty. After some time, $C_t(e)$ will drop to 0.

3.2 Realization and Implementation

Our formula to determine a human-aware certainty is based on the set of currently seen entities V_t, the current human presence H_t, and the accessibility A_{acc} of an entity under influence of H_t. Our prototype implementation realizes these parameters as follows: Visible objects and corresponding aspect values are extracted from point clouds by $\mathfrak{a}_{\text{recognition}}$. However, the world context would not be consistent if all recognized objects were stored immediately — if two subsequently captured point clouds show the same object instance, two entities would exist in W for this instance. We achieve consistency by compressing the set of aspect values belonging to an entity using a hash function f_H. The hash quantizes and then concatenates all aspect values to form a character string unique to the current state of the entity. The recognition actor finally compares the hash of newly detected objects with precomputed hashes of already known entities in W, stored as an aspect a_{hash}. If an entity of matching hash exists, the aspect values of that entity are updated. Otherwise, a new entity is introduced.

We are using background subtraction to determine current human presence H_t. Let $D_t(\theta)$ be the LIDAR scan at time t, mapping each angle to the measured distance. A reference scan $R(\theta)$ is acquired at $t = 0$. Based on the assumption that no humans are present during the reference scanning process, samples belonging to human leg silhouettes are characterized through $|D_t(\theta) - R(\theta)| > \epsilon_{\text{subtraction}}$. The identified samples are converted into points $H'_t \in \mathbb{R}^2$. We apply Euclidean clustering with threshold $\epsilon_{\text{cluster}}$ to H'_t for further reduction of computation efforts in subsequent steps. The human presence map H_t then consists of the centroids of the identified clusters. More sophisticated approaches to human tracking in 2D LIDAR data can be integrated by replacing the actor $\mathfrak{a}_{\text{detection}}$ (e.g. with [2] or [14], see Section 2).

Finally, we need a specific definition of A_{acc}. Let $A_{\text{ind}} : \mathbb{R}^2 \times \mathbb{R}^2 \to [0,1]$ be the probability that a human at position $h_i \in \mathbb{R}^2$ accesses some point p within the workspace. For multiple humans as encoded in H_t, the probability $A_{\text{point}} : \mathbb{R}^2 \times 2^{\mathbb{R}^2}$ that any of them will access p is given as

$$A_{\text{point}}(p, H_t) = 1 - \prod_{i=1}^{|H_t|} (1 - A_{\text{ind}}(p, h_i)).$$

This expression is intuitively derived by applying rules for joint probabilities and complementary events from probability theory, under the additional assumption that all humans act independently from one another. A_{point} can be interpreted as a scalar field that maps each point of the LIDAR scanning plane to a value describing how likely this point will be accessed for the current human presence. The *accumulated accessibility* A_{acc} averages the accessibilities of all points p_e within an entity. A_{acc} is given as

$$A_{\text{acc}}(e, H_t) = \frac{\int A_{\text{point}}(p_e, H_t)\ dp_e}{\int 1\ dp_e}.$$

In our current implementation, we reduce entities to their centroids, and thus avoid the calculation effort of integrating over entity volumes.

As it is impossible to find a task-independent representation of A_{ind}, we use

$$A_{\text{ind}}(p_e,\ h_i) = \begin{cases} 1 & \text{if } d(p_e,\ h_i) \leq L_{\text{arm}}, \\ f(d(p_e, h_i)) & \text{if } L_{\text{arm}} < d(p_e, h_i) \leq L_{\text{arm}} + L_{\text{torso}}, \\ 0 & \text{else,} \end{cases}$$

as an approximation, with Euclidean distance d of two points and a monotonously falling function $f : [L_{\text{arm}}, L_{\text{arm}} + L_{\text{torso}}] \to [0,1]$, e.g. $f(x) = 1 - \frac{1}{L_{\text{torso}}} \cdot (x - L_{\text{arm}})$. This definition of A_{ind} is motivated by human body dimensions. Some point p_e within a radius of the arm length L_{arm} is highly accessible for a human at position h_i. With increasing distance, one needs to lean over the workbench, making the access to objects less comfortable. Points with a distance greater than the sum $L_{\text{arm}} + L_{\text{torso}}$ of arm and torso length L_{torso} can not be accessed at all. Values for the parameters L_{arm} and L_{torso} can be derived from industrial standards, e.g. DIN 33402-2 [4] or ISO/TR 7250-2 [6].

4 Experimental Validation

Our experimental setup is shown in Fig. 1. We use a calibrated pair of an IDS uEye UI-1220SE-C-HQ color camera and an Ensenso N10-308 depth camera to collect point clouds for object recognition while a lightweight robot is moving. A RPLIDAR A2 laser scanner is placed on knee-height below the workbench. Fig. 3 (left) partially shows a typical scan (black dots) of the sensor located at the position marked with a gray hexagon. Two humans are detected due to their leg

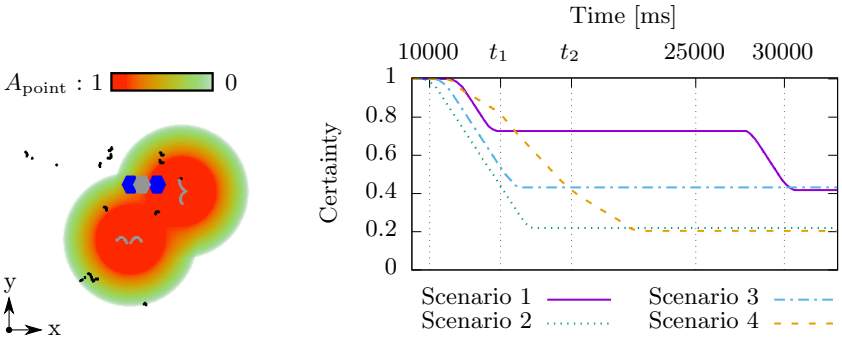

Fig. 3. Two humans (gray samples) detected by the LIDAR sensor (gray hexagon) induce a field of high (red) to low (green) accessibility. The plot shows how the certainty value of two entities (blue hexagons) develops in different scenarios of human presence.

silhouettes (gray samples). The humans induce a scalar field of high accessibility ($A_{\text{point}} = 1$, red) that drops to $A_{\text{point}} = 0$ (green) with increasing distance.

The behavior of entities under the influence of human-aware certainty is documented in Fig. 3 (right). We regard the certainty of two entities (blue in Fig. 3, left) e_1 (right hexagon) and e_2 (left hexagon) over time in different scenarios of human presence. In Scenario 1, a human bypasses the workbench in positive y direction, and returns on the same path. The certainty of e_1 drops while the human is near on the way forth and back, but remains constant while he is out of grasping range. For Scenarios 2 and 3, a human approaches the workbench along the x axis. The certainty of e_1 starts dropping earlier (Scenario 2) and reaches a lower absolute value, as e_1 remains within the human handling area for longer than e_2 (Scenario 3). Certainty in Scenarios 1 to 3 reaches a constant, identical gradient, as the distance between regarded entity and humans falls below L_{arm} in all three cases. In contrast to this, the certainty gradient of e_2 in Scenario 4 does not reach this maximum while two humans arrive and stand at the table at positions depicted in Fig. 3 (left) — e_2 is not within a distance lower than L_{arm} to either of the humans. The certainty falloff is stronger between t_1 and t_2. This is due to the fact that one human arrives late at time t_1 and leaves early at t_2.

5 Conclusion

In the preceding, we have contributed a novel approach that derives data certainty from human presence within the workspace. In particular, our approach allows robot systems to become human-aware through the fusion of local and global perception results. We have empirically validated our approach in an example application, where we derived data certainty by fusing local object perception results with human presence maps acquired through a global LIDAR sensor. In future work, we aspire to integrate predictive, certainty-based task planning into our example application. However, our contribution is not limited

to this application, but intuitively extends to alternative scenarios and arbitrary sensor configurations. Thus, in conclusion, our contribution carries significance for a wide array of predictive and human-aware applications from the field of human-robot collaboration.

References

1. R. Alami et al., "Toward Human-Aware Robot Task Planning", AAAI Spring Symposium, 2006.
2. K. Arras et al., "Using Boosted Features for the Detection of People in 2D Range Data", IEEE Conference on Robotics and Automation, 2007.
3. N. Dalal et al., "Human Detection Using Oriented Histograms of Flow and Appearance", European Conference on Computer Vision, 2006.
4. DIN 33402-2:2005:12, Ergonomics - Human body dimensions – Part 2: Values
5. V. Fox et al., "Bayesian Filtering for Location Estimation", IEEE Pervasive Computing 2 (3), 2003.
6. ISO/TR 7250-2:2011, Basic human body measurements for technological design, Part 2: Statistical summaries of body measurements from national populations
7. B. Lacevic, P. Rocco, "Towards a complete safe path planning for robotic manipulators", IEEE Conference on Intelligent Robots and Systems, 2010.
8. M. W. Lee et al, "Particle Filter with Analytical Inference for Human Body Tracking", IEEE Workshop on Motion and Video Computing, 2002.
9. A. Ober-Gecks et al, "Fast Multi-Camera Reconstruction and Surveillance with Human Tracking and Optimized Camera Configurations", ISR, 2014.
10. K. Okuma et al., "A Boosted Particle Filter: Multitarget Detection and Tracking", European Conference on Computer Vision, 2004.
11. M. Quigley et al., "ROS: An Open-Source Robot Operating System", ICRA Workshop on Open Source Software, 2009.
12. A. Saxena et al., "RoboBrain: Large-Scale Knowledge Engine for Robots", arXiv:1412.0691, 2014.
13. J. Shotton et al., "Real-Time Human Pose Recognition in Parts from Single Depth Images", Communications of the ACM 56 (1), 2013.
14. T. Taipalus and J. Ahtiainen, "Human Detection and Tracking with Knee-High Mobile 2D LIDAR", IEEE Conference on Robotics and Biomimetics, 2011.
15. D. V. Thombre et al., "Human Detection and Tracking using Image Segmentation and Kalman Filter", International Conference on Intelligent Agent and Multi-Agent Systems, 2009.
16. A. Toshev et al., "Deeppose: Human Pose Estimation via Deep Neural Networks", IEEE Conference on Computer Vision and Pattern Recognition, 2014.
17. O. Tuzel et al., "Human Detection via Classification on Riemannian Manifolds", IEEE Conference on Computer Vision and Pattern Recognition, 2007.
18. S. Waldherr, R. Romero and S. Thrun, "A Gesture Based Interface for Human-Robot Interaction", Autonomous Robots 9, 2000.
19. T. Werner et al., "ENACT: An Efficient and Extensible Entity-Actor Framework for Modular Robotics Software Components", 47th Symposium on Robotics, 2016.
20. K. M. Wurm et al., "OctoMap: A Probabilistic, Flexible, and Compact 3D Map Representation for Robotic Systems", ICRA Workshop, 2010.
21. L. Xia, Lu, C. Chen and J. Aggarwal, "Human Detection using Depth Information by Kinect", IEEE Computer Vision and Pattern Recognition Workshops, 2011.

The final publication is available at Springer via
https://doi.org/10.1007/978-3-319-61276-8_70