Semantic-Aware Environment Perception for Mobile Human-Robot Interaction

1st Hempel, Thorsten  
EIT  
Otto-von-Guericke-University  
Magdeburg, Germany  
thorsten.hempel@ovgu.de

2nd Fiedler, Marc-André  
EIT  
Otto-von-Guericke-University  
Magdeburg, Germany  
marc-andre.fiedler@ovgu.de

3rd Khalifa, Aly  
EIT  
Otto-von-Guericke-University  
Magdeburg, Germany  
aly.khalifa@ovgu.de

4th Al-Hamadi, Ayoub  
EIT  
Otto-von-Guericke-University  
Magdeburg, Germany  
ayoub.al-hamadi@ovgu.de

5th Dinges, Laslo  
EIT  
Otto-von-Guericke-University  
Magdeburg, Germany  
laslo.dinges@ovgu.de

Abstract—Current technological advances open up new opportunities for bringing human-machine interaction to a new level of human-centered cooperation. In this context, a key issue is the semantic understanding of the environment in order to enable mobile robots more complex interactions and a facilitated communication with humans. Prerequisites are the vision-based registration of semantic objects and humans where the latter are further analyzed for potential interaction partners. Despite significant research achievements, the reliable and fast registration of semantic information still remains a challenging task for mobile robots in real-world scenarios. In this paper, we present a vision-based system for mobile assistive robots to enable a semantic-aware environment perception without additional a-priori knowledge. We deploy our system on a mobile humanoid robot that enables us to test our methods in real-world applications.

Index Terms—human-robot interaction, semantic objects, interaction willingness

I. INTRODUCTION

Recently, assistive robotics became an increasingly popular research field as it provides use-case applications for solutions from a wide range of disciplines. Whereas special-purposed service robots such as autonomous cleaning systems or lawn mowers already found their way into many households, more general-task directed robotic assistants still entail challenges for the safe and reliable use in unconstrained environments. In general, these systems are aimed at providing various services in the household and workspace of humans that are assigned in direct and human-oriented interactions. This requires the assistive robot to enhance the mostly geometric perception as solely needed for cleaning tasks to a more general semantic understanding of its environment. In particular, a common task for a service robot is to find and bring back a specific kind of object. In order to execute such a command the robot must be able to interact with the human by understanding its needs and subsequently having the capabilities to navigate and find the dedicated object autonomously.

In this paper we focus on the perceptual semantics for processing meaning-oriented information for objects and humans. Multiple approaches [1]–[3] have been presented targeting the detection and recognition of semantic data for mobile robots. Common methods harness model-based solution approaches where each object is identified by its shape or color using handcrafted descriptors. Given that the target object has an unique appearance, these methods can be very performant for...
recognition tasks. On the other side, low image resolution, small objects, an ordinary appearance and multiple instances of similar looking objects can impair the detection and recognition of such systems. In addition, the target models have to be taught beforehand with training images of the object showing it from multiple different perspectives.

We tackle this challenge by using neural networks and combine them with geometric constraints provided by the robot. Our system can be divided in two separate modules. The first one targets the enrichment of geometric maps with semantic information by detecting and localizing objects in a robust manner (Fig. 1). Additionally, we make sure to constantly correct and extend currently mapped objects based on updated perception data from the robot. This makes our system also suitable for localization methods that provide retrospective optimization capabilities. The procedure is carried out on the fly in real-time, when the robot is exploring the surroundings making it a well benefiting add-on for geometric-based mapping algorithms. The second module detects and analyzes potential human interaction partners where we aim to additionally predict cooperation willingness. This provides a first step of interactions with humans in a proactive manner while at the same time semantic objects can be adhoc incorporated in resulting tasks. We implemented our system on the TIArGo, a mobile humanoid robot platform. The rest of this paper is organized as follows: Section II describes our proposed system. In section III we give details of the current state of implementation. Section IV concludes and gives an insight about our future work.

II. SYSTEM COMPONENTS

Our system (Fig. 2) consists of two main pipelines for extracting and evaluating semantic information from visual data. The first one is dedicated to detect and register volumetric objects. The second one classifies and estimates the attention related behavior of humans. In the following subsections, we describe each component in detail.

A. Object registration

The first step of the object registration is the classification of image regions in the image stream provided by the camera in the head of the robot. For this matter deep neural networks have been proven to reliably detect wide ranges of different types of objects. Popular architectures are the RCNN model family [4], [5] and the YOLO model family [6], [7]. Due to its outstanding speed we chose YOLOv4 [8] for our framework. The network is pretrained on the COCO database [9] and able to predict up to 80 different classes. However, when testing it outside of the database we experienced a quite decreased accuracy rate and repeating occurrences of false positive detections. Accordingly, the unfiltered processing of the provided predictions will result with false labeled map parts in later stages. To overcome this problem, we apply an adapted Intersection over Union (IOU) tracker [10] that associates similarly located and equally labeled bounding boxes from consecutive images. Associated detections form a so called track, where the confidence of a true positive object classification increases with the tracks length. After reaching a specific track length threshold the 2D detection is projected into the 3D space using the intrinsic camera parameters and the extrinsics provided by the robots localization algorithm. For this purpose the depth estimation from the camera active stereo module is used together with the RGB images to generate metric scaled point clouds. Depending on the bounding box sizes we then cut out object cuboids that are afterwards projected into the robots 3D point cloud map. While this method for registering of semantic objects can be performed in a very efficient way, it is combined with supportive solutions to overcome the following challenges:

1) Object Recognition: When the robot is moving around it may detect objects in the image stream that already has been registered earlier to the map. Processing it a second time will lead to multiple mapped instances of the same physical objects and must be therefore prevented. Consequently, the object mapping must be able to recognize objects that has been seen and registered before. In general, 3D object recognition has been extensively researched in the recent time, but it commonly requires huge computational efforts. In addition, objects from the same class can resemble each other in a way that taking unique visual fingerprints fails. We approach the issue by comparing the projected localization of a new object candidate with the localization of earlier registered objects of the same class. Is the candidate significantly overlaying with its counterpart it is likely that both belong to the same origin. This approach is implemented using a nearest neighbor association process, where we compare the average distance between the entire point clouds instead of their centroids. This way, we incorporate the objects size in the association decision as point clouds of large objects can contain centroids that are wide apart, while the clouds themselves heavily overlap.

2) Object Optimization: The probabilistic localization of the robot involves uncertainties and will only be a convergence of its real pose. This effect is further enhanced when navigating in an unknown environment and, thus, a simultaneous mapping process is required. For localization methods with optimization capabilities the detection of distinctive or familiar map areas is exploited to re-evaluate the past trajectory and, where appropriate, to correct drifts. Concurrently, we use this mechanism to recalculate the poses of our semantic objects derived from the corrected robot trajectory. At the same time, we analyze if this leads to any strong overlaps between their updated point clouds. This provides us the opportunity to even correct mapping errors induced from the erroneous trajectory. Objects with salient overlap are assumed to belong to the same physical instance and are therefore merged whereby the points clouds are concatenated and the localization and object dimensions are recalculated accordingly. This way, we can not only maintain and correct the perception of semantic objects around us, but with the aid of point cloud merging we are able to build up more complete point cloud appearance models of
Fig. 2: Overview of the proposed method. We use RGB-D images to feed our process lines for object registration and human condition estimation. The object registration consist of 2D object detections that are mapped and associated to 3D objects. The human condition estimation starts with multi task detection process for prediction persons and faces that are in a subsequent step further analyzed for interaction willingness.

B. Human Behavior Estimation

To create optimal preconditions for a successful human-robot interaction we apply a dedicated process pipeline to search for and analyze human interaction partners. Instead of using one network for person detection and one for face detection, we utilize a custom neural network that is able to predict person and face bounding boxes simultaneously.

1) Face-Person Detection: The network architecture is based on the SSD approach [11] that provides a higher inference rate than two-stages detectors. Moreover, both detectors for persons and faces are combined using multi-task learning which enables them to share their first layers of feature extraction. This equally boost the efficiency and the generalization capabilities.

The main difficulty for the combination of the two tasks face and person detection in a single neural network is the fact that publicly available databases contain only ground truths for one of the two tasks. For this purpose, we developed a custom multi-task loss function and designed an architecture consisting of a shared backbone and separate detection layers for each detection task. During training, we alternate between batches of person annotations and batches of face annotations, which are taken from different databases. The prediction of the respective class with non-existing ground truth is assumed to be correctly determined and only the gradients of the detection layers with existing ground truth information are adjusted. Thus, a completely end-to-end trainable framework could be created.

While the predicted bounding boxes for persons give us a good estimation of its localization, the face predictions are further used for behavior examination. In particular, we want to know if the human is interested in an interaction with the robot. A good indication for general interest is the gaze or head pose pointing towards the robot. As the gaze estimation is error-prone at low image resolutions, we focus on the head pose to evaluate general interaction willingness.

2) Interaction Willingness Estimation: To estimate the head pose we first determine the position of facial landmarks in the face images provided by the face predictor. These facial keypoints are distinctive spots in the face (e.g. the corners of the eyes, the sides and top of the nose) that provide us information about the current formation of the face. We estimate the landmark positions by applying an ensemble of regression trees [12] that has proven to provide very fast and reliable predictions. In the following step we project these 2D landmarks into the 3D space. This is achieved by using a default 3D model that contains the same facial landmarks and aligning it with the previously predicted 2D counterpart. As this represents a minimization problem we apply the commonly used Levenberg-Marquardt optimization to estimate a matching 3D mask. The rotation and translation for the projection that provides the smallest re-projection error implies the head pose.

The prediction is performed for every single image that contains a detected face. However, the pose information from a single frame is not sufficient to derive assumptions about an underlying general interaction willingness. Brief views in the direction of the robot can be caused by arbitrary intentions including behavior estimation (e.g. when the robot is moving) or even causal glances without deeper conscious intentions. We therefore track a humans head pose over multiple images and gain confidence about interaction intentions the longer the persons focus is directed at the robot. We assume that a viewing direction towards the robot of a duration of 3 ms represents a reliable threshold to determine interaction willingness. However, short distractions are common that result in brief interruptions of the attention towards the robot. We
therefore use a solution based on our previous idea [13] and calculate the interaction willingness in a dynamic manner. We apply a progress bar that loads faster when the attention is directed to the robot and unloads slower in case of distractions. In this way, showing interaction willingness can be resumed in natural way even though it has been abandoned for a short time.

III. IMPLEMENTATION

We implemented our system in form of multiple Robot Operating System (ROS) compatible modules for seamless intercommunication with other (ROS) components and deployed it on the TIAGo robot. For mapping and localization we use the RTAB-Map [14] as its graph-based approach suits our semantic data update and correction process. The robots IMU is used as guess to perform a laser-based ICP-SLAM that runs along with other ROS nodes for path planning, collision avoidance and motion planning on the robots onboard i7 computer. Our image processing focused methods are deployed on an external mobile system that is placed on the robots shoulders. It contains a Quadro RTX 5000 that is able to process the external mobile system that is placed on the robots shoulders.

IV. CONCLUSION

In this work, we addressed the problem of semantic meaningful perception for mobile assistive robots which constitutes a fundamental requirement for solving complex tasks.

We propose a neural network enhanced approach for successively mapping and maintaining 3D objects of the robots environment that can run alongside other geometrical mapping modules. Similarly, we process humans by utilizing a custom single-shot detector that simultaneously provides person and face predictions in the image stream. The latter are furtherly used to estimates the persons interaction willingness to enable proactive collaboration behavior on the robots side. All modules are implemented on a mobile humanoid robot platform and are compatible for intercommunication with other ROS modules such as path and motion planning. In future works we will exploit this advantage to incorporate additional text-to-speech and speech recognition modules. First, we will search we will exploit this advantage to incorporate additional text-to-speech and speech recognition modules. Future works will include evaluation on real-world scenarios.

REFERENCES

[1] M. Bajones, D. Fischinger, A. Weiss, D. Wolf, M. Vincze, P. de la Puente, T. Koertner, M. Weniger, K. Papoutsakis, D. Michel, A. Quinn, P. Pauleiters, M. Foukarakis, I. Adamu, D. Ioannidi, A. Leondis, M. Antonia, A. Argyros, P. Mayer, and S. Frennert, “Hobbit: Providing fall detection and prevention for the elderly in the real world,” Journal of Robotics, vol. 2018, pp. 1–20, 06 2018, doi: 10.1155/2018/1754657.

[2] E. Martinez and A. P. del Pobil, “Object detection and recognition for assistants,” IEEE Robotics & Automation Magazine, vol. PP, pp. 1–1, 03 2017, doi: 10.1109/MRA.2016.2615329.

[3] S. Lee, S. Lee, J. Lee, D. Moon, E. Kim, and J. Seo, “Robust recognition and pose estimation of 3d objects based on evidence fusion in a sequence of images,” in Proceedings 2007 IEEE International Conference on Robotics and Automation, 2007, doi: 10.1109/ROBOT.2007.364057 pp. 3773–3779.

[4] R. Girshick, J. Donahue, T. Darrell, and J. Malik, “Rich feature hierarchies for accurate object detection and semantic segmentation,” Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 11 2013, doi: 10.1109/CVPR.2014.81.

[5] S. Ren, K. He, R. B. Girshick, and J. Sun, “Faster r-cnn: Towards real-time object detection with region proposal networks,” in NIPS, C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, Eds., 2015, pp. 91–99.

[6] J. Redmon, S. K. Divvala, R. B. Girshick, and A. Farhadi. “You only look once: Unified, real-time object detection,” CoRR, vol. abs/1506.02640, 2015. [Online]. Available: http://arxiv.org/abs/1506.02640.

[7] J. Redmon and A. Farhadi, “Yolo9000: Better, faster, stronger,” 2016, cite arxiv:1612.08242. [Online]. Available: http://arxiv.org/abs/1612.08242.

[8] A. Bochkovskiy, C. Wang, and H. M. Liao, “Yolov4: Optimal speed and accuracy of object detection,” CoRR, vol. abs/2004.10934, 2020. [Online]. Available: https://arxiv.org/abs/2004.10934.

[9] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, “Microsoft coco: Common objects in context,” in Computer Vision – ECCV 2014, D. Fleet, T. Pajdla, B. Schiele, and T. Tuytelaars, Eds. Cham: Springer International Publishing, 2014, doi: https://doi.org/10.1007/978-3-319-10602-1_48. ISBN 978-3-319-10602-1 pp. 740–755.

[10] E. Bochinski, V. Eiselein, and T. Sikora, “High-speed tracking-by-detection without using image information [challenge winner iwor4s],” 08 2017, doi: 10.1109/AVSS.2017.8078516.

[11] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, Reed, C.-Y. Fu, and A. C. Berg, “Ssd: Single shot multibox detector,” in Computer Vision – ECCV 2016, B. Leibe, J. Matas, N. Sebe, and M. Welling, Eds. Cham: Springer International Publishing, 2016, doi: https://doi.org/10.1007/978-3-319-46448-0_2. ISBN 978-3-319-46448-0 pp. 21–37.

[12] V. Kazemi and J. Sullivan, “One millisecond face alignment with an ensemble of regression trees,” in 2014 IEEE Conference on Computer Vision and Pattern Recognition, 2014, doi: 10.1109/CVPR.2014.241. pp. 1867–1874.

[13] T. Hempel and A. Al-Hamadi, “Slam-based multistate tracking system for mobile human-robot interaction,” in Image Analysis and Recognition – 17th International Conference,ICIAR 2020, Póvoa de Varzim, Portugal, June 24-26, 2020. Proceedings, Part I, ser. Lecture Notes in Computer Science, A. Campilho, F. Karray, and Z. Wang, Eds., vol. 12131. Springer, 2020, doi: 10.1007/978-3-030-50347-5_32. pp. 368–376.

[14] M. Labbé and F. Michaud, “Rtab-map as an open-source lidar and visual simultaneous localization and mapping library for large-scale and long-term online operation: Labbé et michaud,” Journal of Field Robotics, vol. 36, 10 2018, doi: 10.1002/rob.21831.