Towards Bringing Robots into Homes

Markus Vincze, Walter Wohlkinger, Sven Olufs, Peter Einramhof, Robert Schwarz

Vienna University of Technology, Automation and Control Institute
Gusshausstrasse 27-29/E376, 1040 Vienna, Austria

Abstract

Home robots need to navigate in a partially structure environment and detect objects on the floor and on tables. Navigation is considered solved using laser sensors, however object recognition requires other sensor modalities. The objective of this work is the merging cognitive robot tasks using one sensing modality: stereo vision. Obstacle avoidance and navigation use the detection of free space - the traversable space in front of the robot. Localisation in a home environment is solved by matching the free space to the map, which gives results similar to laser sensors but takes objects at all heights into account. Finally, the approach naturally leads to object detection: attention is placed around the free space region and structure recognition methods can be applied to their advantage. We will show results for all tasks in lab and home settings.

1 Introduction

Commercial service robots navigate in hospitals, museums, shops, or office corridors. However, today there is no mobile platform that can learn the room layout in a home and then successfully navigate from the living room to the kitchen. Present technology enables mobile robots to move in open terrain, in-door environments with dominantly vertical structures, or in between crowds. However it is interesting to note that the home environment imposes another set of challenges yet to be re-solved. Diverse floor covers render odometry inherently unreliable, protruding surfaces such as tables are invisible to the typically used laser or sonar sensors, and in-door GPS (Global Positioning System) is too expensive to in-stall everywhere. Market deployment requires that the robots are able to learn a new environment within a short time without manual map-making in order to reduce the installation costs. Finally, in order to be able to offer services to humans, the robot's perception and representation of the environment and the objects contained within it has to be compatible with that of humans. Hence, it would be great if the robot would detect main objects in an environment and is able to communicate in human symbolic representation.

To close this gap, we set out to build a mobile platform that moves freely in homes. From the above requirements it is clear that one of the challenges is to cope with the 3D structure of a home setting as opposed to the more structured environment of corridors or offices. Options are to use depth sensors such as a tilting laser ranger to obtain a 3D scan of the environment, e.g., [4], time-of-flight sensors, e.g., [5], or stereo cameras, e.g., [2]. While it is debatable which of these sensor modalities will supersede in the long run, the intention of this contribution is to show that stereo vision is useful to achieve both navigation (detection of drivable free surface) and the detection of relevant items to give a user well-known references to places in her home. **Figure 1** presents the basic concept: first the ground plane is extracted that is needed to assure save navigation and obstacle avoidance. We then show that the ground floor is perfect to localise the robot and performs better than laser information given the smaller viewing angle (Section 2). In the next step we can use the remaining image information to find regions of attention and detect relevant object classes with much higher likelihood due to the constraints given by the ground plane, see Section 3.

Figure 1: Free space detected with stereo for the scene in the image to the right.

2 Navigation in Homes using Stereo Vision

To move from the classical laser based navigation and localisation [8] to the use of stereo we need to cope with a smaller view angle and less accurate measurements while

gaining the advantage of seeing over the full height range of the robot. To this end we propose a new area-based observation model that tracks the ground area inside a the "free space´´ (that is, the not occupied cells) of a known map. We will give more details of the method below. Figure 2 gives an example of the environment and the laser respectively the stereo detection of free space that is used for localisation.

Figure 2: Free space detected with stereo for the scene in the image to the right.

Stereo processing first uses the CENSUS transform [**?**] to obtain disparity images. Free space from stereo is the computed from the v-disparity and the known geometric set-up robot and camera. The advantage of stereo is that it takes data at all height levels into account, which is otherwise only possible with panning laser scanners at much lower rate of acquisition. All data points are then projected into the ground floor. Figure 3 gives an examples.

Figure 3: Free space detected with stereo. Green area detected on the floor. Red are indicated protruding edges.

Obviously the ground plane information gives traversable space in front of robot. It detects steps downwards as well as obstacle necessary for safe navigation.
The added advantage is that this free space is now also used for localisation. It contains more information than edge information from laser scans, because it gives the complete free space in front of the robot and hence takes viewing constraints into account that are not modelled in standard approaches [8]. Furthermore we will see that this approach can also be used for limited sensor range.

For localisation it is ideally required to match the irregular shape of free space against the map. This is computationally expensive. To make the free space calculation suitable for any-time execution we propose an approximation with integral images and image decomposition into squares. This provides a significant speed improvement with an adequate additional error in the range of a few percent and adjustable to demands or processing power given [6]. Fig. 4 gives three examples of approximations that achieve an estimate of the ideal fit. In the experiments we use 64 iterations, which can be executed at a rate of 20 Hz on today's PCs.

Figure 4: Approximation of 92.2, 97.2 and 98.5 % of the free space using integral images with 16, 64 and 256 iterations, respectively.

The difficulty that must be mastered is that the standard observation model is not suited for a relatively small field of view as given by stereo (typically 60-100 degrees). Hence, we adapt the model to fit free space [6]. In practical experiments carried out with a real robot this approach shows good results in partially mapped environments where the traditional Monte Carlo Localisation [8] fails or shows poor performance. Tests have been performed in a laboratory environment given in Fig. 5

Figure 5: The laboratory environment where the method was tested against other sensor data for localisation with and area of 80 m^2.

The results show that the model is able to cope with stereo data and a limited field of view. Accuracy of localisation is better than five centimetres. Also objects not fixed in the environment are coped with. The comparison in Fig. 6 presents the angular average error and Fig. 7 the translational error. The angular error is larger than for laser (however at a much larger field of view of 180 degrees) and bet-

ter if laser is restricted to the same field of view, because the full line of view is considered in the adapted observation model. Translational error does not depend as much on viewing angle, hence stereo performs nearly as good as laser at nearly a third of the viewing angle.

Figure 6: Average Error of the approximated model compared to the reference implementation: Angular error.

This indicates that using stereo images to obtain the free space in front of the robot is well suited to avoid obstacles and to provide sufficiently accurate localisation of a robot in home settings in real time.

Figure 7: Average Error of the approximated model compared to the reference implementation: Translational error. Legend as in Fig. 6.

3 Furniture Detection in Homes using Stereo Vision

Seen from the perspective of a user, locations in a home are linked to rooms and certain items of furniture, e.g., the sofa in the living room or the table in the kitchen. The first section showed how stereo is used to localise the robot and detect free space for robot safe navigation. This is used for learning places in the rooms and linking these places to the data from the sensors, for example using map topology as in [3]. However, a user will naturally link locations to semantically meaningful items in a room. She will name places with terms intuitively clear to the user, for example by using names of furniture or rooms. Hence, the objective is to classify the main items of furniture in homes. This Section will present how this is approached.

Object classes of interest are chair, table, couch, cupboard and door. In human vision, shape is one of the most powerful cues for classifying objects. However, if shape detection is performed on the whole input image without any restriction, computation times become prohibitively long for the use on a robot. Inspired by findings from attentional processes in human vision [7] a two-step process is proposed. An attention mechanism exploits the result of free space detection (see above) and identifies the closest object in front of the robot. The robot approaches the object of interest so that it is dominant within the camera image. This reduces the negative effect of background clutter. We can then use probabilistic shape matching to classify the attended to object. Furthermore, the mobility of the robot is useful to provide the shape detector with various views of the object of interest.

In the next section we will outline the attention procedure. Section 3.2 will then show how object detection is performed.

3.1 Attention to Furniture

Attention builds on the results of free space detection. The ground plane is removed, which greatly simplifies the attention process to immediately reduce the image area of interest to all object above ground. The next processing step is to cluster the point data into convex bounding boxes exploiting the constraint of expected object size, which can rather conservatively be applied for typical items of furniture in homes. Finally, a verification step can use shape-based or 2D approaches. The advantage of this approach is that the strength of stereo processing are exploited and that this focusation on a region of attention simplifies the detection of object classes such as chair, couch, table and door. Figure 8 shows examples from the home setting in our lab, the free space region on the floor region and the centrally segmented chairs. Fig. 9 shows the resulting bounding boxes for the examples that are then used for object classification.

Figure 8: Two examples of segmenting the region of attention in the center of the robot view.

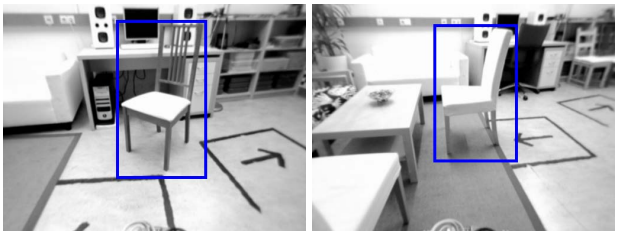

Figure 9: The bounding box from attention that is used for object classification for the examples in Fig. 8.

3.2 Classification of Furniture

In this Section we outline the approach to classify furniture. Using a region of attention resulting from the processing above aids to improve results. The method is however applicable to large image regions as well.

Object class recognition, especially in indoor environments, is an important task for mobile robotics applications as it paves the way for robots to operate successfully in home environments. Human machine interaction, robot object interaction, robot navigation and localization and mapping can greatly benefit from a system which is able to categorize objects in a home environment.

This paper addresses the problem of vision-based categorization of objects on a robot platform in home environments. We are interested in categorizing types of furniture as well as doors found in home environments. This is a challenging problem due to poorly-textured scenes with many wiry objects like chairs and piecewise planar surfaces. Traditional approaches like dense stereo fail in such uniform textured scenes as the correspondence problem is not solvable and the resulting data is too sparse or too inaccurate. Laser scanners, on the other hand, have insufficient resolution and power to detect narrow objects such as the legs of chairs or the black metal hat stand found in our test environment. Furthermore, the single scan-line of the laser at a specific height makes its use for detecting furniture challenging, as many legs are displaced towards the center of the object and horizontal layers may therefore function as invisible obstacles. The use of multiple lasers or rotating lasers is also not applicable in our set-up for cost and security reasons.

To overcome the disadvantages of the environment and target categories, we use a pure-vision based system for the task. We employed straight line segments detected in a multi-camera set-up as our elementary features since man-made environments are full of straight lines and these serve well as a complement to dense-stereo.

For matching of line segments over multiple views, a minimum of four views is necessary to achieve robust matching as stated in [9]. Therefore, we use a calibrated four camera set-up. Two cameras out of the four are arranged without vertical displacement, due to the additional use of these cameras as a dense stereo sensor, providing the pose of the camera rig with respect to the ground plane. The cameras are mounted on a mobile robot at a height of about 130 cm on a pan-tilt unit, mimicking a small person, able to look down at a desk or chair. The cameras are four synchronized wide-angle cameras, which have been calibrated in a previous off-line calibration stage using a modified planar- target-based calibration and subsequent bundle-adjustment.

Processing then proceeds as follows. Section 3.2.1 discusses the object class representation. The 3D line calculation and refinement by non-linear minimization together with a additional constraint minimization for accuracy increasing is presented in Section 3.2.2. Section 3.2.3 the decision tree for the retrieval of the objects is described. Finally, the experimental results demonstrate our approach on a large dataset in Section 3.2.4.

3.2.1 Object Class Representation

Our system has access to a large amount of domain information such as the fact that the features are extracted by a mobile robot of a given size and that typical furniture has a consistent set of geometric properties. We have used this information to model desired object classes by the geometric and spatial relationships between their features. This representation is tightly coupled with our chosen features, but, since we have a metric reconstruction of the environment, search for objects is simple. That is, multi-scale search is not required. Our use of priors for object geometry clearly excludes toy-sized furniture or huge doors from being detected, but as the target platform is a robot for the home-environment, this restriction does not affect us. Fig. 10 gives a qualitative summary of the class models for door, chair and table.

Figure 10: Visual representation of our concept classes door, chair and table: red depicts minimum, green maximum, ground plane in grey. Best viewed in color.

3.2.2 3D Line Extraction

The 3D lines are created with a non-linear optimization step by minimizing the re-projection error in all four images. To increase the stability of the 3D lines, the lines segments corresponding to the vertical vanishing point are optimized to go through this point in three space. For an additional enhancement, the intersections of line pairs in

2D are calculated and verified with the epipolar constraint against the four views. These intersection points are then also used as 3D points for an optimization like the vertical vanishing point.
As a intersection of two lines in three space defines a plane and we are searching for planar structures in man-made environments, we use these planes as our primary features. Together with the 3D line segments, these planes are fed into the detection stage. Figure 11 shows the matched lines and the resulting 3D reconstruction.

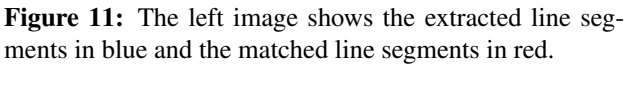

Figure 11: The left image shows the extracted line segments in blue and the matched line segments in red.

3.2.3 Object Detection

The concept models are used for detection by forming a decision tree. In this detection stage, each plane hypotheses is fed into a decision tree which evaluates the plane attributes 'orientation', 'size' and 'distance to ground floor' and assigns the plane to one of the four classes consisting of 'chair', 'table', 'door' and 'no detection'. For the classes table and door this decision tree is sufficient, but for the chair class an additional search algorithm for the detection of the back support has to be called. At each node of the decision tree, a probability is assigned to the plane hypotheses. This enables us to keep track of planes which maybe belong to the categories but do not fulfil all requirements. This partial affiliation has its source at the feature extraction stage, where line segments may be too short or broken.

3.2.4 Experimental Results

This section shows the results, both positive and negative, for object class detection of chairs, tables and doors. The dataset of chairs consists of 39 chairs with approximately five views of each chair. On our chair database, 33 out of the 39 chair models were detected correctly. We counted a correct detection if the chair was detected correctly at least in one of the views of the chair. The dataset of doors consists of 13 distinct doors in about 70 images. We could detect 8 out of the 13 doors correctly, the remaining doors failed due to too narrow viewing angle, too far distance or the inability to extract long non-broken vertical line segments. In Figure 12 a representative sample of correct object class detection of class chair.

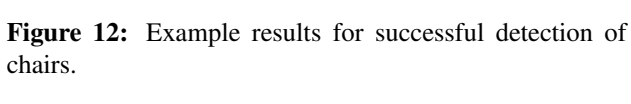

Figure 12: Example results for successful detection of chairs.

Figure 13 gives examples for doors. And Figure 14 gives examples for tables. Due to the deficiencies in the line extraction stage, not all planar surfaces are detected and line segments are missed, for example for the sofa detected as table.

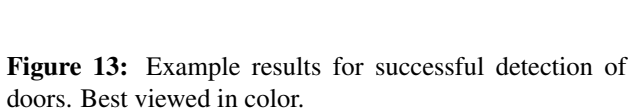

Figure 13: Example results for successful detection of doors. Best viewed in color.

The experiments showed that lines are a powerful feature for man-made environments and can be used for furniture classification. The results are good for doors and tables and satisfying for chairs. However, a main disadvantage became evident: Especially in the chair class, the use of straight lines as main features is inadequate to robustly detect the planar seat area. This is caused by a non- straight form of the seat surface or by presence of upholstery or

texture. This detection deficiency could probably solved by incorporating additional features like dense stereo. This is discussed in the next section. Another disadvantage was seen for the door class: The description of the door is too simple to disambiguate between doors, bookshelves, cupboards and other vertically aligned planar structures. This will also be targeted in the next section.

Figure 14: Example results for successful detection of tables. On the right a sofa with rectangular arm rest and back has falsely been detected as table. Best viewed in color.

We have shown a robot system which is able to categorize objects in a home environment with vision only sensors. The set-up uses a calibrated multi-camera set-up which uses lines as primary features and performs object class recognition on the classes chair, table and door. We have shown that this approach is feasible to work under real conditions and gives good results.
The future system improvements are twofold: On the one side we will port the algorithm from the existing Matlab implementation to C++ to perform real-time evaluation on a robot platform. On the other side, to overcome the deficiencies of only using straight lines we will include dense stereo and curves in the detection and triangulation stage. The planar surface detection will be extended by a plane sweep algorithm for improved generation of plane hypotheses.

4 Conclusion

We show that stereo vision can be used both for navigation and object identification. We further show that detection of free space is not only good to avoid obstacles but also is robust to obtain localisation. We discussed the influence of viewing angle and could show that the smaller viewing angle of stereo is compensated when using an improved observation model.
Furthermore, using free space is the key to find objects surrounding the floor area. They can be segmented in an attention-based mechanism and the subsequent object classification step is highly simplified.
With this combined approach the use of stereo as the main sensor modality for home robot applications could be shown. Nevertheless there are several steps to go. While classification has been shown for several types of furniture, line based approaches are not suitable for more elaborate designs. Hence, we will further exploit the features provided by stereo and combine line with area-based features given from the stereo depth image.

Acknowledgements

The research leading to these results has received funding from the European Community's Sixth Framework Programme (FP6/2003-2006) under grant agreement number FP6-2006-IST-6-043450 (robotshome).

References

[1] AmHu08 K. Ambrosch, M. Humenberger, W. Kubinger and A. Steininger: Extending two non-parametric transforms for FPGA based stereo matching using bayer filtered cameras, Fourth IEEE Workshop on Embedded Computer Vision (ECVW08), Anchorage, Alaska, June 28, 2008.

[2] Soohwan Kim, Howon Cheong, Ju-Hong Park, and Sung-Kee Park: Human augmented mapping for indoor environments using a stereo camera. In IEEE/RSJ International Conference on Intelligent Robots and Systems, 2009.

[3] M. Liu, D. Scaramuzza, C. Pradalier, R. Siegwart, Q. Chen: Scene Recognition with Omnidirectional Vision for Topological Map using Lightweight Adaptive Descriptors, IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), 2009.

[4] Zoltan Csaba Marton, Radu Bogdan Rusu, and Michael Beetz: On fast surface reconstruction methods for large and noisy point clouds. IEEE ICRA 2009, pages 3218-3223, May 2009.

[5] Stefan May, Stefan Fuchs, David Droeschel, Dirk Holz, and Andreas Nuechter. Robust 3d-mapping with time-of-?ight cameras. In IEEE/RSJ International Conference on Intelligent Robots and Systems, 2009.

[6] Sven Olufs and Markus Vincze. An efficient area-based observation model for monte-carlo robot localization. In IEEE/JRS IROS, 2009.

[7] Albert L. Rothenstein and John K. Tsotsos. Attention links sensing to recognition. Image and Vision Computing, 26:114-126, 2008.

[8] S. Thrun, W. Burgard, and D. Fox: Probabilistic Robotics. MIT Press, Cambridge MA, First edition, 2005.

[9] L. Zebedin, J. Bauer, K. Karner and H. Bischof: Fusion of Feature- and Area-Based Information for Urban Buildings Modeling from Aerial Imagery; 10th European Conference on Computer Vision, pp. 873-886, 2008.