Indoor Positioning and Navigation Using Time-Of-Flight Cameras

Chapter · March 2013

DOI: 10.1007/978-3-642-27523-4_8

4 authors, including:

Tobias K Kohoutek
Digimapas Chile Aerofotogrametria Ltda.
22 PUBLICATIONS **96** CITATIONS

SEE PROFILE

David Droeschel
University of Bonn
61 PUBLICATIONS **1,496** CITATIONS

SEE PROFILE

Sven Behnke
University of Bonn
508 PUBLICATIONS **7,991** CITATIONS

SEE PROFILE

Indoor Positioning and Navigation Using Time-Of-Flight Cameras

Tobias K. Kohoutek, David Droeschel, Rainer Mautz and Sven Behnke

1 Introduction

The development of indoor positioning techniques is booming. There is a significant demand for systems that have the capability to determine the 3D location of objects in indoor environments for automation, warehousing and logistics. Tracking of people in indoor environments has become vital during firefighting operations, in hospitals and in homes for vulnerable people and particularly for vision impaired or elderly people [1]. Along with the implementation of innovative methods to increase the capabilities in indoor positioning, the number of application areas is growing significantly. The search for alternative indoor positioning methods is driven by the poor performance of Global Navigation Satellite Systems (GNSS) within buildings. Geodetic methods such as total stations or rotational lasers can reach millimeter level of accuracy, but are not economical for most applications. In recent years, network based methods which obtain range or time of flight measurements between network nodes have become a significant alternative for applications at decimeter level accuracy. The measured distances can be used to determine the 3D position of a device by spatial resection or multilateration. Wireless devices enjoy widespread use in numerous diverse applications including sensor networks, which can consist of countless embedded devices, equipped with sensing capabilities, deployed in all environments and organizing themselves in an ad-hoc fashion [2]. However, knowing the correct positions of network nodes and their deployment is an essential precondition. There are a large number of alternative positioning technologies

T. K. Kohoutek (✉) · R. Mautz
ETH Zurich—Institute for Geodesy and Photogrammetry,
Wolfgang-Pauli-Str. 15 8093 Zurich, Switzerland
e-mail: kohoutek@geod.baug.ethz.ch

D. Droeschel · S. Behnke
Rheinische Friedrich-Wilhelms-Universität Bonn—Institute for Informatics VI,
Friedrich-Ebert-Allee 144 53113 Bonn, Germany
e-mail: droeschel@ais.uni-bonn.de

F. Remondino and D. Stoppa (eds.), *TOF Range-Imaging Cameras*,
DOI: 10.1007/978-3-642-27523-4_8, © Springer-Verlag Berlin Heidelberg 2013

(Fig. 1) that cannot be detailed within the scope of this paper. An exhaustive overview of current indoor position technology is given in [3]. Further focus will be on optical methods.

Optical indoor positioning systems can be categorized into static sensors that locate moving objects in the images and ego-motion systems whose main purpose is the position determination of a mobile sensor (i.e. the camera) [4]. Some optical system architectures do not require the deployment of any physical reference infrastructure inside buildings, which can be a requirement for a widespread implementation.

This article investigates the use of Time-Of-Flight (TOF) cameras for ego-motion determination in indoor environments. TOF cameras are suitable sensors for simultaneous localization and mapping (SLAM), e.g. onboard of autonomous Unmanned Vehicle Systems (UVS), or the detection and localization of objects in indoor environments. They are an attractive type of sensor for indoor mapping applications owing to their high acquisition rate collecting three-dimensional (3D) data. TOF cameras consist of compact, solid-state sensors that provide depth and reflectance measurements at high frame rates of up to 50 Hz independent from surrounding light.

The approximate 3D position accuracy for objects seen by the used MESA® TOF camera SwissRanger SR-4000 (in terms of a 1-σ standard deviation) is 1 cm

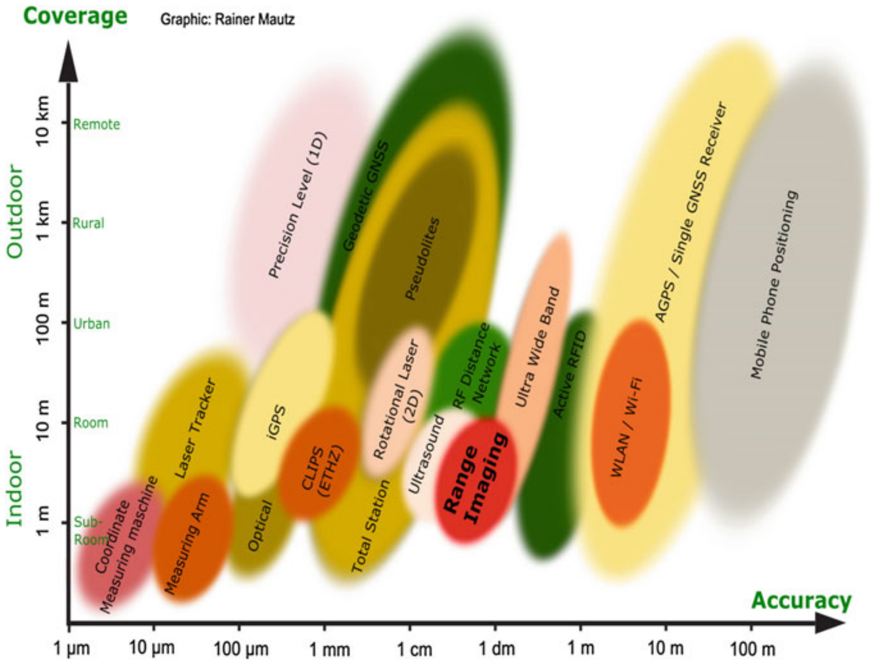

Fig. 1 Today's positioning systems in dependence to accuracy and coverage [1]

for distances of up to 5 m and 1 dm for distances up to 15 m. Such a level of accuracy is sufficient for some indoor applications, e.g. collision avoidance. Currently, ranges larger than 15 m and accuracies better than 1 cm are not applicable to TOF cameras. In these cases 3D laser scanners or stereo/multiple camera systems need to be used instead. As a drawback of two-dimensional (2D) cameras, the prerequisite for multiple views induces a high computational load since point correspondences between at least two images from different perspectives have to be determined. In addition, distances to structureless surfaces cannot be measured, because the correspondence problem [5] cannot be solved. Furthermore, passive 2D vision suffers from shadowing effects and sensitivity to changes in illumination. The use of 3D laser range finders [6] that actively illuminate the scene can avoid these issues but needs mechanical moving parts and have high power consumption as well as a low frame rate due to sequential point acquisition.

Our procedure is as follows. Image features, e.g. edges, corners or flat surfaces are detected based on reflectance data for object recognition in the indoor environment. In Sect. 2 we will show how the indoor positioning with the TOF camera can be realized. As a novelty, the proposed method combines absolute and relative orientation of a TOF camera without the need for dedicated markers or any other locally deployed infrastructure. This can be achieved, because in comparison to other methods range imaging directly provides 3D point clouds that are compared with a spatio-semantic 3D geoinformation model offered by the City Geographic Markup Language (CityGML) that supports any coordinate system and enables the missing link between the indoor and outdoor space. As higher the level of semantic information as more accurate is the geometrical integration. The entrance door of a building for example is always connected to a walkable surface. The camera motion is estimated based on depth data and will be explained within the mapping process in Sect. 3. Collision avoidance becomes important if the navigation path is unknown. Section 4 will show that TOF cameras are ideally suited for that task. A welcome side effect of our approach is the generation of 3D building models from the observed point cloud.

2 Positioning Inside the Room Based on a CityGML Model

The standard CityGML [7] defines a data model and an XML data format for 3D city and topography models. CityGML defines several Levels of Detail (LoD) with the highest LoD 4 having the capability for modeling the interior of buildings. In particular for the purpose of indoor modeling, the semantic model provides an object class ‘Room’ that can capture semantic data [8], including attributes for the intended and current use of the room such as ‘Living Room’ or ‘Office’. An object of the class ‘Room’ can be associated with its geometry in two different ways. In one way the outer shell of a room can be defined by establishing a link to a geometric object of type *Solid* or *MultiSurface* (both types are defined by the GML 3.1.1 specification [9]). Alternatively, the outer shell can be decomposed into semantic objects of the types *InteriorWallSurface*, *CeilingSurface* and *FloorSurface*, which are referred to geometric objects of type *MultiSurface*. Openings in the outer shell of a room can be modeled with the object classes 'Window' and 'Door' that can belong to one or two *InteriorWallSurfaces*. This data structure can be used to express topological relationships between rooms.

The semantic object class *IntBuildingInstallation* can be used to model permanent fixed objects belonging to a room e.g. radiators, columns and beams. In order to model the mobile components of a room such as desks and chairs, the object class *BuildingFurniture* can be used. *IntBuildingInstallation* and *BuildingFurniture* provide the attribute class for semantic description of the objects (Fig. 2). The geometry of these fixed installed objects can be defined by the standard GML 3.1.1. So-called implicit geometries are used to model simplified shapes of the movable objects in a room. Hereby the shape of an object is stored only once in the library even if multiple objects of the same shape are present (e.g. pieces of furniture). The shapes could be obtained directly from the 3D CAD drawings of pieces of furniture in the manufacturer's catalog. For each occurrence of such an object, only the local coordinates of an insertion point and the object's orientation are stored. The orientation parameters are linked to the geometry that has become an object of CityGML.

Nowadays, Building Information Models (BIMs) are created within the planning and construction phase of a building [10]. The acquisition of BIMs for already existing buildings requires manual measurements using total stations, terrestrial laser scanners or photogrammetric techniques. Figure 3 illustrates semantic classification of CityGML exemplified with an indoor model of a room that has been obtained by total station survey.

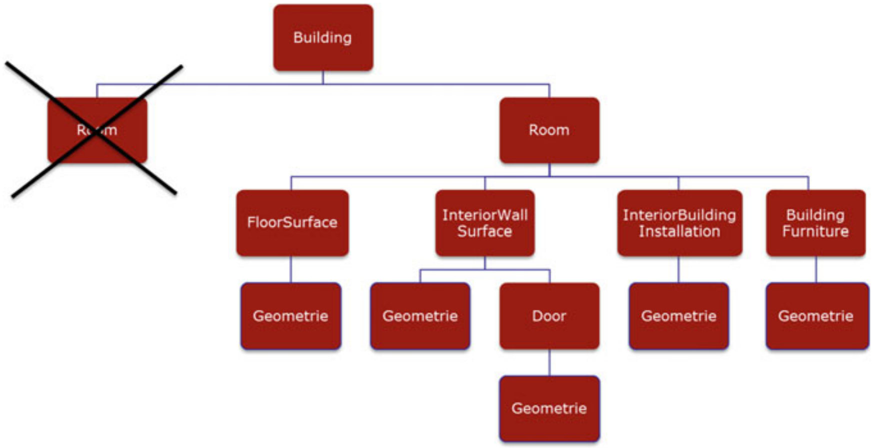

Fig. 2 Decision tree for room identification

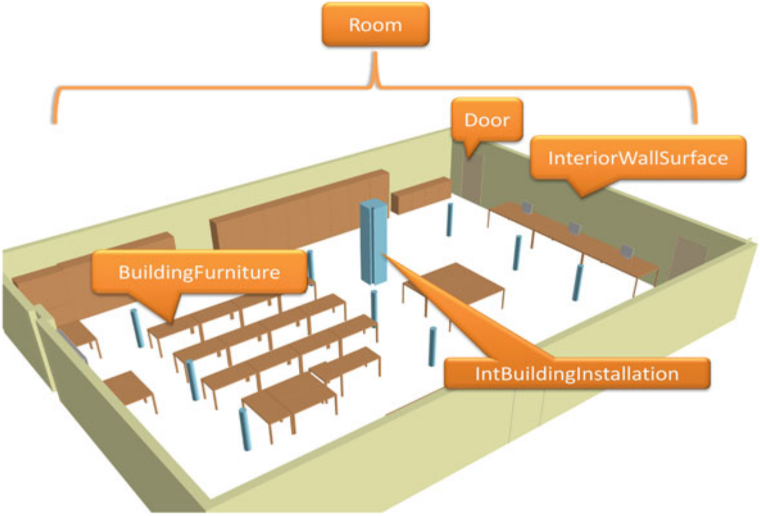

Fig. 3 ETH Zurich lecture room modeled in CityGML [11]

2.1 Room Identification Through Object Detection

Object detection is the key challenge for the correct identification of the room where the sensor is located. The detection of objects can be achieved by exploiting the amplitude image. In order to identify objects such as chairs, tables, etc., the known or "learned" primitives, features and image templates that have previously stored in the database are matched with the current image. The detected object properties such as the size, geometry or quantity of a certain object are the main criteria for the comparison with the database. This way, the unknown camera position can be limited to a small number of possible rooms in the building. The room can be identified uniquely by detecting its distinct properties, e.g. position of installations. After a successful identification additional semantic and geographic information can be extracted from the 3D geo database.

2.2 Accurate Positioning Using Distance Measurements

This step compares and transforms the in real time acquired Cartesian 3D coordinates of the objects into the reference coordinate system of the database. All room and object models in the CityGML database are saved as Virtual Reality Modeling Language (VRML) files. Suitable reference points for the transformation (with 6 degrees of freedom) are the corners of the room, vertices of doors, windows and other fixed installations. The accuracy of the objects in CityGML should be at centimeter level and should lead to position determination of the camera with centimeter-accuracy using a least squares adjustment with a redundant number of reference points to determine the 3D camera position. One requirement for the camera is that its interior orientation has been determined previously. The exterior camera orientation (3 translations and 3 rotations) is determined by a Cartesian 3D coordinate transformation with 3 shift and 3 rotational parameters. There is no need to estimate a scale parameter, since calibrated TOF cameras measure the absolute distance.

3 Mapping and Ego-Motion Estimation

Dense depth measurements from TOF cameras enable the generation of 3D maps of the camera's environment. However, the accuracy of measurements in unknown scenes varies considerably, due to error effects inherent to their functional principle. Therefore, a set of preprocessing steps to discard and correct noisy and erroneous measurements need to be applied in order to achieve accuracy according to the specification.

3.1 Sensor Data Processing

First, mixed pixels at so-called jump edges are filtered out. Mixed pixels are a result of false measurements that occur when the signal from the TOF camera hits an edge of an object. Then, the signal is partially reflected at the foreground, but also at the background. Both signal parts arrive at the same CCD element. The true distance changes suddenly at the object border, but the values of the mixed pixels consist of an average between the foreground and background distance. In the point cloud, these pixels appear as single unconnected points that seem to float in the air and that do not belong to any object. This is also a common problem in terrestrial laser scanning. Jump edges are filtered by local neighborhood relations comparing the opposing angles of a point $\mathrm{p_i}$ and its eight neighbors $\mathrm{p_{i,\,n,}}$ [12]. From a set of 3D points $\mathrm{P} = \left\{ \mathrm{p_i} \in R^3 | i = 1, \ldots, N_p \right\}$, jump edges are detected by comparing opposing angles $\theta_{\mathrm{i,n}}$ of the triangle spanned by the focal point $\mathrm{f} = 0$ and its eight neighbors $\mathrm{P_n}=\left\{ \mathrm{p_{i,n}} \in R^3 | i = 1, \ldots N_p : n = 1, \ldots, 8 \right\}$ and filtered with a threshold θ_{th}:

$$\theta_i = \max \arcsin\left(\frac{\|p_{i,n}\|}{\|p_{i,n} - p_i\|} \sin\varphi \right), \tag{1}$$

$$\mathrm{J} \ = \{p_i | \theta_i > \theta_{th}\}, \tag{2}$$

where φ is the apex angle between two neighboring pixels. Since the jump edge filter is sensitive to noise, a median filter is applied to the distance image beforehand. Besides mixed pixels, measurements with low amplitude are neglected since the accuracy of distance measurements is dependent on the amount of light returning to the sensor.

TOF cameras gain depth information by measuring the phase shift between emitted and reflected light, which is proportional to the object's distance modulo the wavelength of the modulation frequency. As a consequence, a distance ambiguity arises: measurements beyond the sensor's wavelength are wrapped back causing artifacts and spurious distance measurements. Wrapped distance measurements can be corrected by identifying a number of so-called phase jumps in the distance image, i.e., the relative wrappings between every pair of neighboring measurements. Droeschel et al. proposed attempt a probabilistic approach that detects discontinuities in the depth image to infer phase jumps using a graphical model [13]. Every node in the graphical model is connected to adjacent image pixels and represents the probability of a phase jump between them. Belief propagation is used to detect the locations of the phase jumps which are integrated into the depth image by carrying out the respective projections, thereby correcting the erroneously wrapped distance measurements. The application of phase unwrapping for an indoor scene is shown in Fig. 4.

3.2 Mapping and Ego-Motion Estimation

To estimate the camera's motion between two consecutive frames, image features in the reflectance image of the TOF camera are extracted to determine point correspondences between the frames. To detect image features, the Scale Invariant

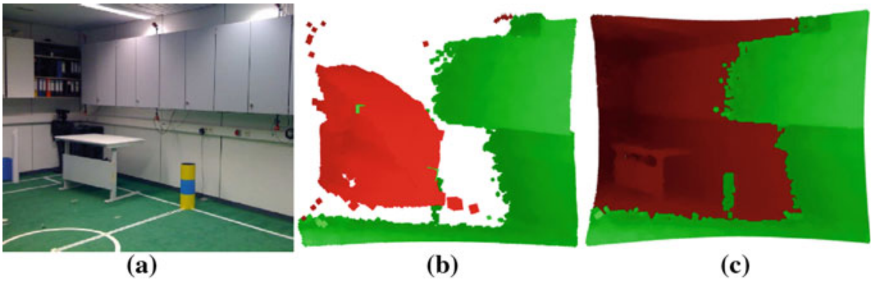

Fig. 4 Phase unwrapping of an indoor scene. **a** Image of the scene. **b** and **c** 3D point clouds that have been generated based on the camera's depth image. *Color* of the points indicates the result of the algorithm; wrapped measurements are shown in *red*. *Brightness* encodes distance to the camera center. **b** Point cloud without unwrapping. Measured distances beyond the sensor's non-ambiguity range are wrapped into it, which results in artifacts between distances of 0 and 3 meters. **c** Unwrapped depth image

Feature Transform (SIFT) [14] is used. SIFT features are invariant in rotation and scale and are robust against noise and illumination changes.

In order to estimate the camera motion between two frames, the features of one frame are matched against the features of the other frame. The best match is the nearest neighbor in a 128-dimensional keypoint descriptor space. To determine the nearest neighbor, the Euclidean distance is used. In order to measure the quality of a match, a distance ratio between the nearest neighbor and the second-nearest neighbor is considered. If both are too similar, the match is rejected. Hence, only features that are unambiguous in the descriptor space are considered as matches.

Figure 5a and b show the reflectance image of two consecutive frames with detected features. Figure 5c shows the matching result of the two images. Each match constitutes a point correspondence between two frames. By knowing the depth of every pixel, a point correspondence in 3D is known.

The set of points from the current frame is called the data set, and the set of corresponding points in the previous frame is called the model set. The scene is translated and rotated by the sensor's ego motion. Thus, the sensor's ego motion can be deduced by finding the best transformation that maps the data set to the model set. A common approach for estimating a rigid transformation uses a closed form solution for estimating the 3 × 3 rotation matrix R and the translation vector t, which is based on singular value decomposition (SVD) [15]. The distances between corresponding points, after applying the estimated transformation are used to compute the root mean square error (RMSE) which is often used in range registration to evaluate the scene-to-model consistency. It can be seen as a measure for the quality of the match: if the RMSE is significantly high, the scene-to-model registration cannot be consistent. On the other hand, a low RMSE does not imply a consistent scene-to-model registration, since it also depends on the number and distribution of the point correspondences.

With the estimated ego motion between consecutive frames, accumulating 3D points of every frame generates a point-based map. A resulting map is shown in Fig. 6.

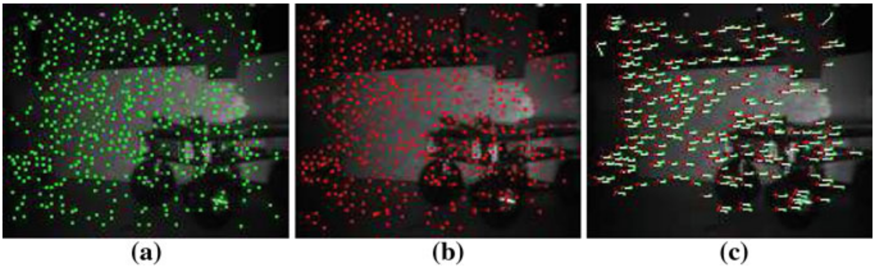

Fig. 5 SIFT feature extraction and matching applied on two consecutive camera frames on a TOF reflectance image. The numbers of detected features are 475 (**a**) and 458 (**b**). **c** Matching result: 245 features from image (**a**) are matched to features from image (**b**). *White lines* indicate feature displacement

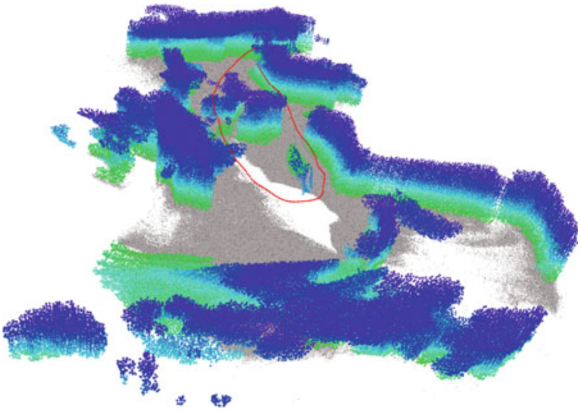

Fig. 6 The resulting 3D map based on the estimated trajectory (*red*). The *colors* of the points correspond to the distance of the point from the ground plane

4 3D Collision Avoidance

If the navigation path is unknown in dynamic environments, collision avoidance becomes important. TOF cameras are ideally suited for collision avoidance since they measure distances to surfaces at high frame rates.

A typical example of a point cloud taken in an indoor environment is shown in Fig. 7a. This point cloud can be used to build a so-called height image as shown in Fig. 7b. A point $p_{i,\ j}$ is classified as belonging to an obstacle if

$$(W_{max}\ \text{-}\ W_{min})\ > H, \tag{3}$$

where W_{max} and W_{min} are the maximum and minimum height values from a local window W, spanned by the eight-connected neighborhood around $p_{i,\ j}$. The Threshold H thereby corresponds to the minimum tolerable height of an obstacle.

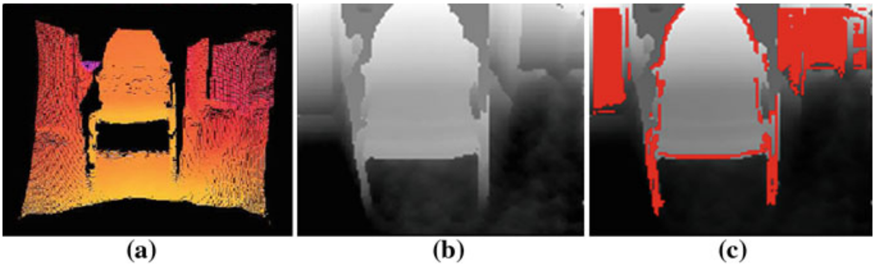

Fig. 7 **a** 3D Point cloud of an exemplary scene. The *color* of the points corresponds to the distance, *brighter color* relates to shorter distances and *darker color* to farther distances. **b** The generated height image. The grayscale value of every pixel corresponds to the z-coordinate of the respective point in the point cloud. **c** The resulting obstacle points (*red*)

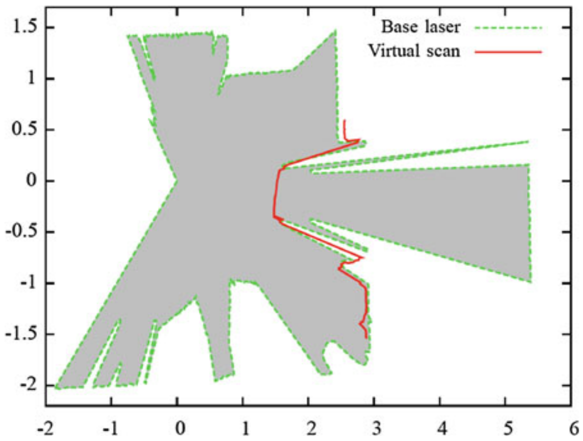

Fig. 8 The resulting virtual scan of the scene is compared with the scan from the laser range finder. The dashed *green line* illustrates the base laser scan. The *red line* illustrates the virtual laser scan. The chair shows only a few points in the base laser scan since only the legs of the chair are in the scan plane, whereas the virtual scan outlines the contour of the chair

It needs to be chosen appropriately since it should not be smaller than the sensor's measurement accuracy. Due to evaluating a point's local neighborhood, floor points are inherently not considered as obstacles. Points classified as belonging to obstacles are shown in Fig. 7c.

The resulting obstacle points are used to extract a 2D virtual scan similar to an obstacle map by (1) projecting the 3D data into the xy-plane and (2) extracting relevant information.

The number of range readings in the virtual scan as well as its apex angle and resolution correspond to the acquired 3D data. For the SR4000, the number of range readings is 176, which is the number of columns in the image array. The apex angle and the angular resolution are 43 and 0.23°, which correspond to the camera's horizontal apex angle and resolution. For every column of the TOF camera's distance image, the obstacle point with the shortest Euclidean distance to the robot is chosen. This distance constitutes the range reading in the scan. If no obstacle point is detected in a column, the scan point is marked invalid.

The resulting virtual scan is fused with a 2D laser range scan obtained at 30 cm height yielding a common obstacle map modeling the closest objects in both sensors. The obstacle map from the 2D laser range finder and the TOF camera for the aforementioned example scenario is visualized in Fig. 8. By fusing the information of both sensors, the robot possesses correct information about traversable free space (light gray) in its immediate vicinity.

5 Conclusions and Outlook

Efficient and precise position determination of a TOF camera is possible based on kinematic object acquisition in form of 3D Cartesian coordinates. The absolute position of the camera can be obtained by a transformation from the camera coordinate system into the reference coordinate system, i.e. the coordinate system of the spatio-semantic 3D model. Positions of detected objects are reported in respect to the coordinate system of the 3D model. The described mapping approach can also be used for data acquisition of such 3D building models. The advantage of such models is the use of the VRML file text format allowing data compression for the purpose of quick Internet transfer and maintenance of a small-sized database. We conclude that rooms can be identified by detection of unique objects in images or point clouds. Such method is to be implemented in further research based on a data set, which includes multiple rooms.

Due to their measurement of a volume at high frame rates, TOF cameras are well suited for applications where either the sensor or the measured objects move quickly, such as 3D obstacle avoidance, measured or gesture recognition [16].

Difficulties of the proposed method arise from TOF cameras suffering from a set of error sources that hamper the goal of infrastructure-free indoor positioning. Current range imaging sensors are able to measure distances unambiguously between 5–15 m at an accuracy level of centimeters. Until now so-called mixed pixels posed a problem in the literature. Filtering methods presented in Sect. 3 could solve this problem.

Acknowledgments The support of Andreas Donaubauer, Andreas Schmidt, Dirk Holz and Stefan May is gracefully acknowledged.

References

1. T.K. Kohoutek, R. Mautz, A. Donaubauer, Real-time indoor positioning using range imaging sensors, in *Proceedings of SPIE Photonics Europe, Real-Time Image and Video Processing*, vol. 7724 (SPIE, 2010), p. 77240K. doi:10.1117/12.853688 (CCC code: 0277-786X/10/$18)
2. R. Mautz, W.Y. Ochieng, Indoor positioning using wireless distances between motes, in *Proceedings of TimeNav'07 / IEEE International Frequency Control Symposium*, (Geneva, Switzerland, 29 May–1 June 2007), pp. 1530–1541
3. R. Mautz, Indoor positioning technologies, in *A habilitation thesis submitted to ETH Zurich for the Venia Legendi in Positioning and Engineering Geodesy*, (ETH, Zurich, 2012)
4. R. Mautz, S. Tilch, Survey of optical indoor positioning systems, in *IEEE Xplore Proceedings of the 2011 International Conference on Indoor Positioning and Indoor Navigation (IPIN)*, (Guimarães, Portugal, 21–23 Sept 2011)

5. B. Julesz, Binocular depth perception of computer-generated patterns. Bell System Tech. **39**(5), 1125–1161 (1960)
6. C. Keßler, C. Ascher, G.F. Trommer, Multi-sensor indoor navigation system with vision- and laser-based localisation and mapping capabilities. Eur. J. Navig. **9**(3), 4–11 (2011)
7. G. Gröger, T.H. Kolbe, A. Czerwinski, C. Nagel: *OpenGIS® City Geography Markup Language (CityGML) Encoding Standard Version 1.0.0*, (International OGC Standard. Open GeospatialConsortium, Doc. No. 08-007r1, 2008)
8. G. Gröger, T.H. Kolbe, A. Czerwinski: Candidate *OpenGIS® CityGML Implementation Specification (City Geography Markup Language) Version 0.4.0*, (International OGC Standard. Open Geospatial Consortium, Doc. No. 07-062, 2007)
9. S. Cox, P. Daisey, R. Lake, C. Portele and A. Whiteside: *OpenGIS® Geography Markup Language (GML) Implementation Specification Version 3.1.1*, (International OGC Standard. Open Geospatial Consortium, Doc. No. 03-105r1, 2004)
10. C. Nagel, A. Stadler, T.H. Kolbe, Conceptual requirements for the automatic reconstruction of building information models from uninterpreted 3D models, in *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, vol. 34, Part XXX (2009)
11. A. Donaubauer, T.K. Kohoutek, R. Mautz, CityGML als Grundlage für die Indoor Positionierung mittels Range Imaging, in *Proceedings of 15. Münchner Fortbildungsseminar Geoinformationssysteme*, (Munich, Germany, 8–11 March 2010), pp. 168–181
12. S. May, D. Droeschel, D. Holz, S. Fuchs, E. Malis, A. Nüchter and J. Hertzberg, Three-dimensional mapping with Time-Of-Flight cameras, J. Field Robot. **26**(11–12), 934–965 (2009). (*Special Issue on Three-dimensional Mapping, Part 2*)
13. D. Droeschel, D. Holz, S. Behnke, Probabilistic phase unwrapping for Time-Of-Flight cameras, in *Proceedings of the joint conference of the 41st International Symposium on Robotics (ISR 2010) and the 6th German Conference on Robotics (ROBOTIK 2010)*, (Munich, Germany, 2010), pp. 318–324
14. D.G. Lowe, Distinctive image features from scale invariant keypoints. Int. J. Comput. Vision **60**, 91–110 (2004)
15. K.S. Arun, T.S. Huang, S.D. Blostein, Least-squares fitting of two 3-d point sets. IEEE Trans. Pattern Anal. Mach. Intell. **9**(5), 698–700 (1987)
16. D. Droeschel, J. Stückler, S. Behnke, Learning to interpret pointing gestures with a Time-Of-Flight camera, in *Proceedings of the 6th ACM/IEEE International Conference on Human-Robot Interaction (HRI*), (Lausanne, Switzerland, March 2011), pp. 481–488