Archives of Photogrammetry, Cartography and Remote Sensing, Vol. 22, 2011, pp. 159-172
ISSN 2083-2214

PEDESTRIAN MOBILE MAPPING SYSTEM FOR INDOOR ENVIRONMENTS BASED ON MEMS IMU AND RANGE CAMERA

Norbert Haala, [1] Dieter Fritsch [2], Michael Peter [3], Ali M. Khosravani [4]

[1, 2, 3,4] Institute for Photogrammetry (ifp), University of Stuttgart,
Geschwister-Scholl-Str. 24D, D-70174 Stuttgart, Germany
(Norbert.Haala, Dieter.Fritsch, Michael.Peter, Ali.Khosravani)@ifp.uni-stuttgart.de

KEY WORDS: Urban, Reconstruction, IMU, Navigation, Building, Modeling, Architecture

ABSTRACT This paper describes an approach for the modeling of building interiors based on a mobile device, which integrates modules for pedestrian navigation and low-cost 3D data collection. Personal navigation is realized by a foot mounted low cost MEMS IMU, while 3D data capture for subsequent indoor modeling uses a low cost range camera, which was originally developed for gaming applications. Both steps, navigation and modeling, are supported by additional information as provided from the automatic interpretation of evacuation plans. Such emergency plans are compulsory for public buildings in a number of countries. They consist of an approximate floor plan, the current position and escape routes. Additionally, semantic information like stairs, elevators or the floor number is available. After the user has captured an image of such a floor plan, this information is made explicit again by an automatic raster-to-vector-conversion. The resulting coarse indoor model then provides constraints at stairs or building walls, which restrict the potential movement of the user. This information is then used to support pedestrian navigation by eliminating drift effects of the used low-cost sensor system. The approximate indoor building model additionally provides a priori information during subsequent indoor modeling. Within this process, the low cost range camera Kinect is used for the collection of multiple 3D point clouds, which are aligned by a suitable matching step and then further analyzed to refine the coarse building model.

1. INTRODUCTION

Mobile Mapping Systems combine high performance navigation components usually based on GNSS/inertial measurement with mapping sensors like multiple CCD cameras and/or laser scanners. By these means georeferenced 3D point clouds or video streams are collected, which frequently cover street scenes of complete city areas. These data are then further evaluated i.e. to provide large scale 3D reconstructions of urban areas for planning or navigation purposes. While such high-performance systems are mainly used in the context of commercial geo-data collection, projects like OpenStreetMap established so-called Volunteered Geographic Information (VGI) as a powerful alternative. There, volunteers use low-cost systems to provide geo-data at considerable accuracy and amount of detail. Frequently, GNSS tracks are applied to georeference the objects of interest as captured by the user based on semi-automatic geo-data collection. In this sense, the success of VGI is closely coupled to GNSS as prevalent and inexpensive sensor system for the navigation of the user. Thus, VGI has become very popular in outdoor areas, while the lack of a suitable positioning sensor prevents corresponding developments in indoor

environments. However, low-cost inertial measurement systems can provide user tracks at sufficient quality, if the positional accuracy during pedestrian navigation is improved using ZUPTs (zero velocity updates). This can be realized by using a foot-mounted MEMS as a pedometer.
As it will be discussed in section 2 of the paper, this approach can considerably reduce drift errors within the captured position. However, especially for measurement of longer periods further improvement is still required. For this reason, MEMS IMU indoor navigation is further supported by alignment of the user tracks. This assumes that most indoor routes will be parallel or perpendicular to one of the principal directions of the building. These principal directions usually correspond to the building outlines. Information on the building contour is frequently available from existing maps or is given by 3D outer building shells.
In our system, information on the building structure is provided automatically from evacuation plans, which are compulsory for public buildings in a number of countries. Such plans usually consist of a generalized floor plan and escape routes. As it will be demonstrated in section 3, the ground plan information can be made explicit again by a suitable raster-to-vector-conversion. Thus, we assume, that our "pedestrian mobile mapping platform" as realized by the foot mounted MEMS IMU is complemented by a simple camera. Using this simple sensor platform, the user captures an image of the evacuation plan, which is then interpreted automatically by suitable software tools. In addition to location, size and shape of different rooms, such a floor plan can provide semantic information like stairs, elevators or the level number. This information is then used to further support and improve pedestrian navigation similar to map matching in outdoor environments. Finally, this process provides a track of the user related to the coarse indoor model as given from the evacuation plan.
In principle, our pedestrian mobile mapping platform can be used by the respective volunteer to manually model indoor environments in an OpenStreetMap like process (OpenStreetMap Wiki, 2011). As an example, the collected tracks of the user can be used to enhance the coarse model as provided from the interpretation of the evacuation plan by modeling missing features and indoor facilities like corridors, rooms, doors and stairways. However, efficient modeling is especially feasible, if 3D point clouds have been measured. For this purpose, our mobile mapping system is equipped with a low-cost range camera. As presented in section 4 of the paper, we realize 3D point cloud collection by integration of the Kinect stereo camera to our sensor system. Originally, the Kinect was developed as a user interface for the Xbox 360 gaming platform. For our purpose of data collection in the context of indoor modeling it is ideally suited since it combines a RGB camera with a depth sensor. This sensor consists of an infrared projector combined with a monochrome CMOS camera for stereo measurement. As discussed in section 4, this allows for the collection of coregistered image and point cloud data in indoor environments to be used for the refined modeling of the building interior.

2. INDOOR NAVIGATION USING MEMS IMU

In order to provide the path of a user moving through a building, a wide range of experimental and commercial positioning systems are available. However, they either require infrastructure like WLAN-Networks or RFID beacons, or are based on high quality

indoor models. Therefore, they do not allow for an inexpensive and generally available indoor navigation at sufficient accuracy.
In contrast, inexpensive and generally available indoor navigation is feasible by low-cost MEMS IMUs (in our case an XSens MTi-G), which in principle allow for position determination by integration of inertial measurements. However, since such naive integration suffers from large drifts after short time periods, positional accuracy from such MEMS IMU measurements can be improved for pedestrian navigation by the well-known zero velocity updates (ZUPTs). Thus, the IMU is attached to the user's foot and is in principle used as a pedometer (Godha & Lachapelle, 2008). For this purpose, the swing and stance phases of the user's foot may be found within the gyro measurements during walking. If the norm of the gyro measurement vector drops below 1 rad/s, the beginning of a stance phase is detected. After integrating the accelerometer measurements once, the resulting velocities are supposed to be zero during this phase. However, the measured values will differ from zero due to sensor drift. The computed difference may then be used to correct all values since the last stance phase and the final position values relative to the starting point can be computed by a second integration step.

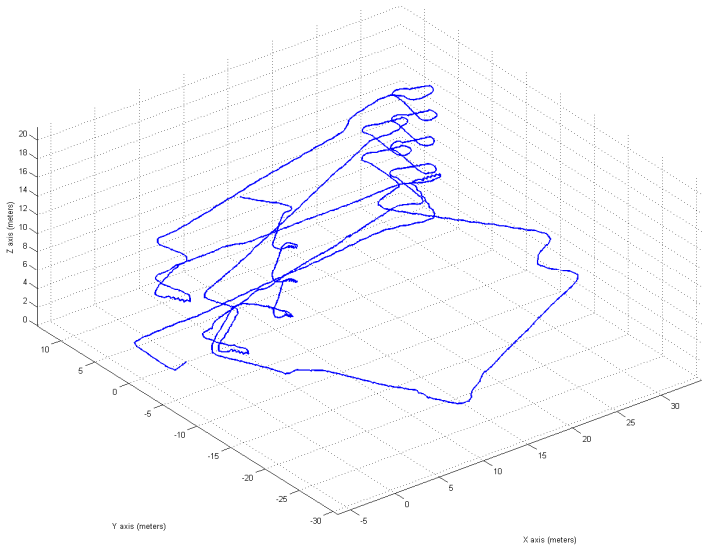

Fig. 1: Indoor track from foot mounted MEMS IMU

Fig. 1 exemplarily depicts the result of MEMS IMU indoor navigation for a user moving through the building of our institute. This building has seven floors and a size of approximately 60x25m. Even though the use of ZUPTs significantly reduced the drift errors from IMU measurements, considerable errors are still available for such a long track. For this reason, the accuracy of the measured trajectory is further improved by assuming that most parts of such indoor routes will be parallel or perpendicular to the main direction of the respective building (due to hallways etc.). Thus, the IMU track is aligned to this principal direction. For this purpose, straight lines are detected in the track by searching for at least six consecutive steps within the positions derived using ZUPTs which only feature changes in moving direction below a selected threshold. The angular difference between the

direction of this straight trajectory and the principal building direction is then eliminated by a rotation around the z-axis. Remaining effects in the vertical component of the trajectory are eliminated by strictly limiting vertical movement to stairs and elevators. The actual detection of stair steps is straightforward: all steps with height differences greater than 15cm are annotated as stair candidates. This corresponds to the common step height for public buildings (Neufert et al., 2002). The resulting indoor track can be seen in Fig. **2**.

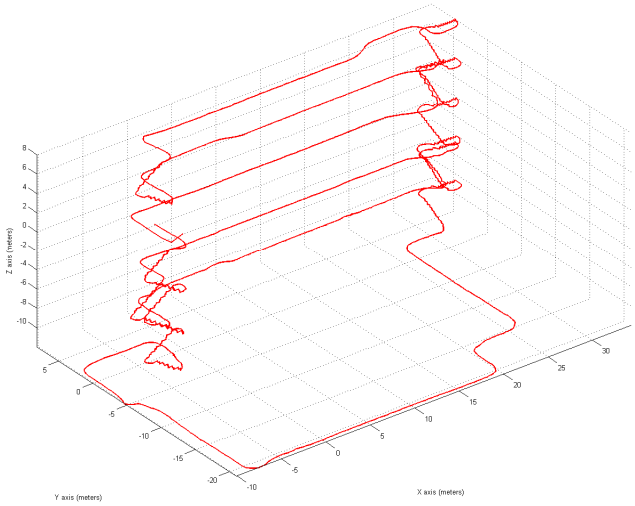

Fig. 2: Improved indoor track by alignment to principal building directions and height correction.

Despite the considerable quality of this trajectory, it still provides only coordinates relative to an initial position. Thus, final geocoding requires the availability of the starting point in the reference coordinate frame. One option is to use the entrance into a building, to which the user has been guided by GNSS. However, consumer-grade single-frequency GNSS receiver accuracies of 5-10 meters or worse in urban canyons prevent the measurement of an initial position with sufficient accuracy for the navigation in building interiors. Alternatively, the building entrance can be provided from a detailed semantic façade model as e.g. generated by (Becker & Haala, 2009).
In our application the initial position is deduced from a photographed evacuation plan. This interpretation, which is realized by a suitable raster-to-vector conversion, additionally provides ground plan information and thus a coarse map of the respective indoor environment.

3. INTERPRETATION OF EVACUATION PLANS

Indoor positioning and navigation is mainly required for large public buildings like shopping centers and administrative office buildings. In many countries, evacuation plans are compulsory for such buildings. An example of such a plan is given in fig. **3**. As it is visible, in addition to evacuation routes, these plans contain a large amount of useful information for the modeling of building interiors like inner walls, stairs, elevators and some of them even doors and windows. This information can be made explicit again by a suitable interpretation of the collected raster image, following the steps depicted in fig. 4.

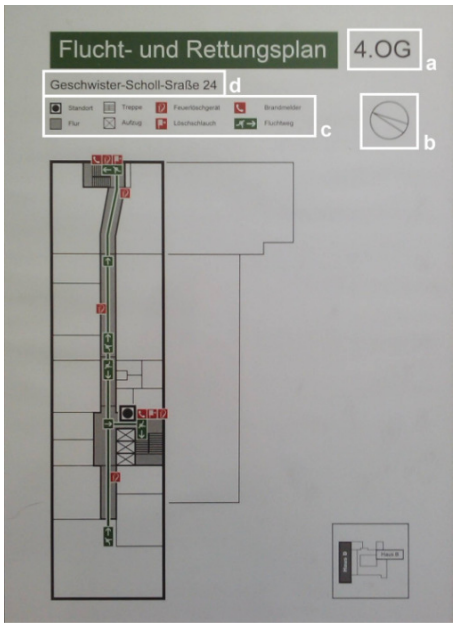

Fig. 3: Photographed evacuation plan with available information (left)

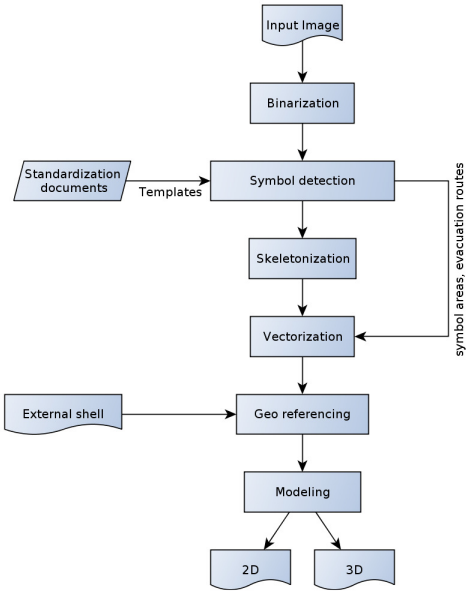

Fig. 4: Processing steps of the proposed reconstruction approach

Within our approach described in more detail in (Peter et al., 2010) the image of the evacuation plan is binarized as a first step. Since evacuation plans are optimized to be legible also in emergency situations, their content is limited to the most necessary information like ground plan and symbols, which have to be clearly distinguishable from a white background. Thus, a simple adaptive thresholding can be used. As it is also visible in fig. **3**, specific parts of the plan provide additional information on the floor (a) north direction (b), the legend (c) and the address (d).
The binarized image is then segmented to distinct regions by a boundary tracing algorithm. In order to detect the building boundary, which is symbolized by a bold line, binary erosion is used. This step eliminates thin lines representing the inner walls not required in this step. A contour finding algorithm (Suzuki et al., 1985) is then applied on the remaining bold lines to detect the corner points of the building boundary. Usually, one can assume a rectangular shaped boundary polygon from the ground plan. This information is used to rectify the ground plan image to generalize and thus simplify further processing.
Within the next step, evacuation symbols are detected. Thus, image parts occluded by these symbols can be cleaned before further digitization. In order to identify candidates for symbol regions a connected components analysis is realized. Symbol templates, which can be extracted from the plan's legend, are then detected using cross correlation. Since such symbols are standardized, high quality templates taken from standardization documents can be used for this step, alternatively.

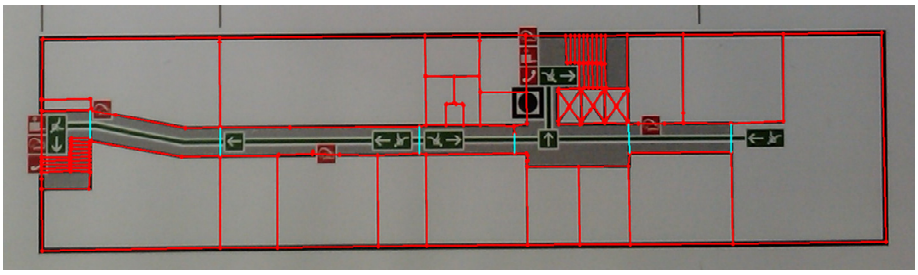

Fig. 5: Result of vectorization with extracted room outline.

After cleaning the binary image a skeleton is computed similar to the approach described by (Zhang & Suen, 1984). Fig. 5 shows the result of this process with the implemented cleaning for symbol regions. To further use this ground plan skeleton during navigation and modeling, a transformation to a geocoded reference coordinate system is required. For this purpose, again the contour of the building outline as represented by the bold line in fig. **3** is used. The comparison of the automatically derived model edges to a paper plan (scale 1:100) yields maximum differences of up to 0.1m. The photographed plan's scale used here was approximately 1:500, however, this may vary greatly with different plans and accuracy is further influenced by the process of photographing.

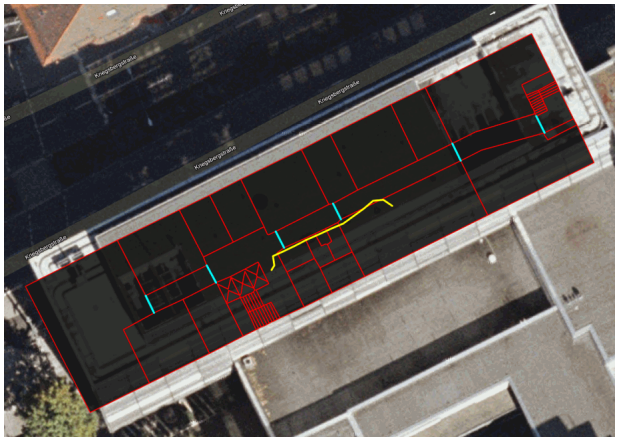

Fig. 6: Reconstructed indoor model using evacuation plan in Google Maps (red); IMU track using ZUPTs, alignment and height correction (yellow)

The building contour extracted from the evacuation plan is also represented within the corresponding 3D building model. In our test scenario this 3D building model is freely available from the virtual city model of Stuttgart (Bauer & Mohl, 2005). Corresponding nodes between the building outline in the evacuation plan and the virtual city model can then be used to define the transformation between image coordinates of the interpreted evacuation plan and the global reference system of the city model. Afterwards, the vectorized ground plan in fig. 5 is available in real world coordinates. This information can e.g. be used to compute dimensional properties like width and area of different features and thus allows distinguishing between rooms using e.g. a minimum area threshold and stairs e.g. using a maximum width threshold. Furthermore, this georeferencing step is also used to provide world coordinates for the "you are here" symbol, which was detected previously in the evacuation plan. This information then can be used as initial position for the IMU navigation, thus, the measured trajectory of the user is also available in the reference coordinate system of the virtual city model. This result is demonstrated in fig. **6**. It visualizes the final indoor model from the interpretation of the evacuation plan within Google Maps with an overlaid user track, which was measured and processed by our system.

4. POINT CLOUD COLLECTION FROM LOW-COST RANGE CAMERA

In order to refine or update such a coarse indoor model e.g. by extracting features like doors or to detect missing walls, 3D point clouds are especially suitable. Within Mobile Mapping Systems this is frequently realized by laser scanning, however, these systems are usually too large and expensive for pedestrian applications. One suitable alternative are range cameras also known as Time-of-Flight (TOF) cameras using a Photonic Mixer Device (PMD) for point cloud measurement at video rate. An example is the CamCube operating at a 204×204pixel resolution with a 40° viewing angle and a maximum distance (ambiguity interval) of 7.5m. Within our low-cost system the Kinect range camera, offered by Microsoft as a user interface for the Xbox 360 video game platform is used.

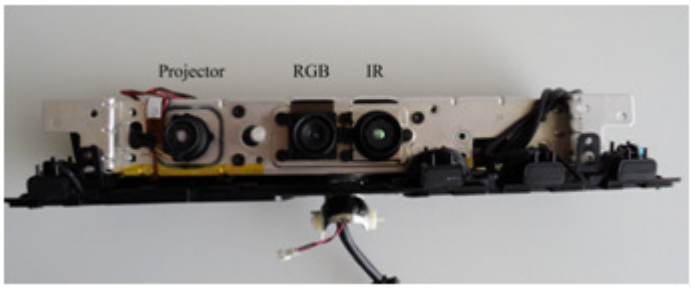

Fig. 7: Disassembled Kinect system

This system, depicted disassembled in Fig. 7, consists of an infrared (IR) laser projector and a monochrome IR CMOS sensor. Stereo measurement is realized by projecting a fixed pattern of light and dark speckles on the respective object surfaces. The IR camera then collects a video stream of the continuously-projected infrared structured light in 30 Hz at VGA resolution (640×480pixels) with 11-bit radiometric depth. By analysis of the IR speckle pattern automatic stereo measurement is realized in order to compute a range image from spatial intersection in the following step. When used with the Xbox software, the Kinect has a practical ranging limit of 1.2-3.5m, the angular field of view is 57° horizontally and 43° vertically. The area required to play Kinect is roughly 6m², although the sensor can maintain tracking through an extended range of approximately 0.7-6m.

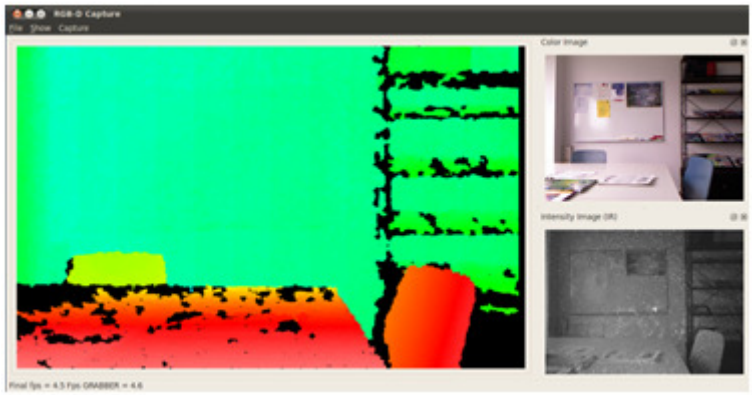

Fig. 8: User interface of Kinect RGB Demo

Open source software libraries and drivers like libfreenect (OpenKinect, 2011) and ROS OpenNI (ROS OpenNI Kinect, 2011) provide suitable interfaces to use the Kinect as a low-cost range camera. Fig. **8** provides a snapshot of the open source software "Kinect RGBDemo" (Burrus, 2010), which was used during our measurements. The snapshot shows the range image from stereo matching on the left. On the right the synchronized video streams from the RGB and the IR camera are available on top and bottom, respectively. As it is visible in Fig. 7, the RGB camera is mounted between the IR camera and projector within the Kinect system.

4.1 System Calibration

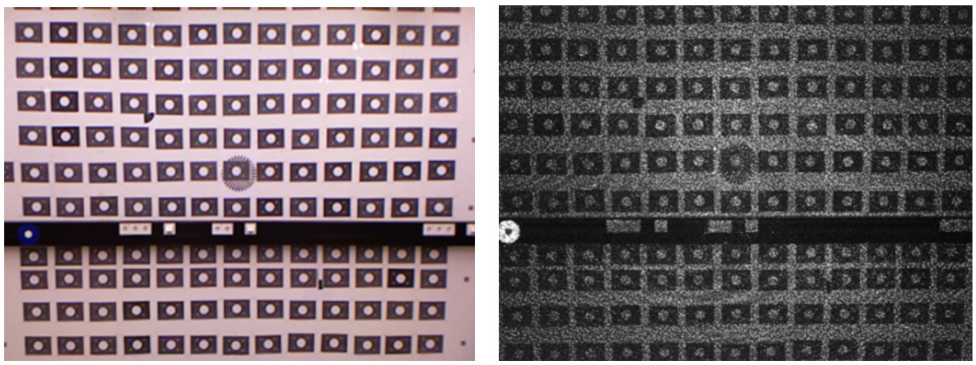

Fig. 9: RGB (left) and IR (right) calibration images.

In order to allow for the generation of textured point clouds, the Kinect camera system was calibrated at a suitable calibration field. Image blocks from the RGB and the IR camera were collected and used for stereo camera calibration in a photogrammetric bundle block adjustment using the Australis software system. By these means lens distortions could be eliminated based on the Brown parameter set. The test field is depicted in Figure 9, which shows an example of a RGB image on the left and an IR image on the right. Within the IR image the projected speckle is visible in addition to the photogrammetric target points.

4.2 Errors of Captured Point Clouds

After matching the projected IR patterns within the images collected by the monochrome IR camera, disparity measurements are available. These disparities or parallaxes are then used to compute 3D coordinates for the depicted object surfaces by spatial intersection. These are then represented by range images or 3D point clouds, respectively. If a suitable calibration of the stereo system as realized by the IR camera and the projector is available, the geometric quality of this point cloud is mainly influenced by the accuracy and reliability of the pattern matching as implemented within the Kinect. This pattern matching can be disturbed e.g. by ambient light in the IR spectrum, if the projected pattern is superimposed. Additional errors can occur depending on geometric and radiometric properties of the depicted object surfaces. Problems are to be expected at object boundaries and sharp edges due to occlusions. Moreover the object surface may absorb or reflect a major part of the projected IR pattern, which again prevents a reliable matching. Geometric errors also result from imperfect calibration of the stereo system, e.g. the relative orientation between IR projector and camera, as well as imperfect modeled distortions of the optical systems. Finally, simplifications to speed up the matching step by the Kinect software can potentially result in systematic errors.

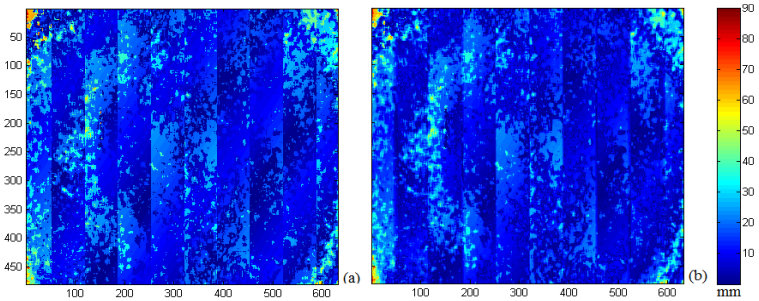

Fig. 10: Noise of the data at a distance of 2.6m for a single frame (left) and an average of 30 frames (right).

Figure 10 shows the noise of the data at a distance of 2.6m for a single frame and an average of 30 frames. The RMS of the noise is 12mm, and its maximum is 90mm in both cases. This can prove that the noise of the data mainly results from systematic errors, which can be seen in this example as vertical and radial patterns.
As the magnitude and behavior of the error pattern (e.g. the number of vertical strips) depend on the distance to the object, bulk error compensation methods like surface fitting were too complicated and impractical. A next step to get a higher accuracy from the Kinect data could be analysis of the integrated sensors and their configuration, in order to estimate effective physical models for the systematic errors.

4.3 Alignment of Point Clouds

During camera calibration, also the relative orientation between the RGB and IR cameras is determined. Thus, the pixels of the range image can be matched against the RGB images while simultaneously compensating the lens distortion effects. By these means, textured 3D point clouds can be generated as depicted in Fig. **1**1 for the scene already shown in Fig. **8**.

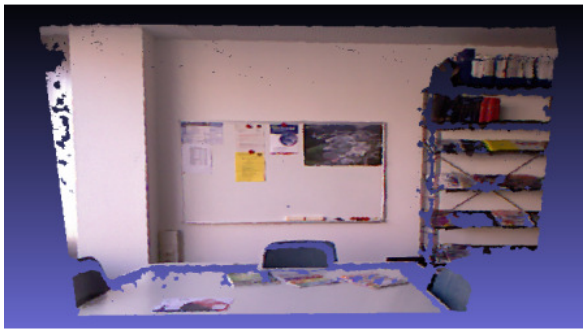

Fig. 11: A textured point cloud

The transformation between the RGB and the disparity image can be used to provide a co-registration of multiple 3D point clouds, which is even more important for our application. Such a combination of range images from multiple viewing directions and viewpoints is essential in order to cover a complete room.

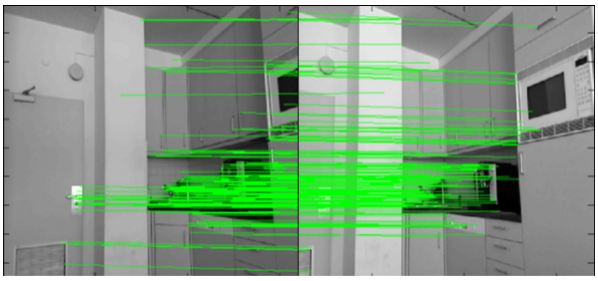

Fig. 12: Tie point matching for two RGB camera images.

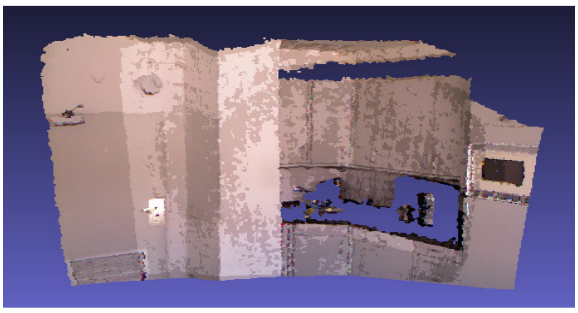

Fig.13: Alignment of two point clouds corresponding to images in Fig. 1.

Similar to the work of (Henry et al., 2010), this is efficiently realized in our application by automatic relative orientation of consecutively captured RGB images. Fig. **1**2 gives an example for two images collected by the RGB camera at two viewpoints during data collection. Within these images corresponding points are then determined by the SIFT feature extraction and matching (Lowe, 2004). These matches are represented by the green lines in Fig. **1**2. The relative orientation between the RGB and IR images was already determined during system calibration as discussed in section 4.1. Each pixel in the IR image corresponds to a pixel in the range image and thus a 3D object point coordinate. Thus, the matched image points in Fig. **1**2 can be directly used to generate correspondences between 3D points from the two consecutive views. This information allows for an approximate alignment of the two point clouds in a local coordinate frame, which is further improved using an iterative closest point (ICP) algorithm (Besl & MacKay, 1992).
Fig.13 depicts the point clouds corresponding to the images in Fig. **1**2 after alignment. This co-registration process is then performed for a complete sequence of the range and RGB images. The result of this process is given exemplarily in Fig. **1**4, which shows an aligned sequence of 40 point clouds covering a complete room. Of course, the quality of the point clouds alignment using this approach depends on the existence of enough well distributed SIFT features in each pair of RGB images, as well as on the noise of the corresponding features in the range data. For this reason weak connections have to be avoided by using a relatively large overlap between captured data sets to provide suitable alignment accuracy for the respective point clouds.

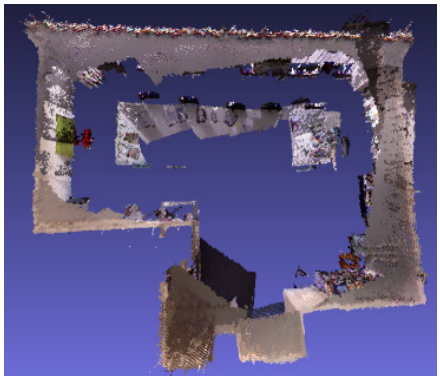

Fig. 14: Co-registration of multiple point clouds from alignment of RGB image sequence.

4.4 Matching of Point Cloud and Indoor Model

Since the result of an indoor navigation using MEMS IMU is available during acquisition of the range images, the respective user positions can be used to transform the collected point clouds to the reference coordinate frame. However, the accuracy and reliability of this georeferencing is improved considerably, if the point cloud is matched against the coarse indoor model as generated by the interpretation of the evacuation plan.
The fit of the point cloud from Fig. **1**4 to the indoor model depicted in fig. **6** is given in Fig. 1. This fitting was realized by the ICP algorithms already used for point cloud alignment. In order to use this algorithm, first a point cloud was sampled from the faces of the 3D indoor model. Then the measured Kinect point cloud was matched against this reference by the ICP approach while the required initial approximation of that transformation was derived from the user position as captured during range data measurement by the methods discussed in section 2 and 3.

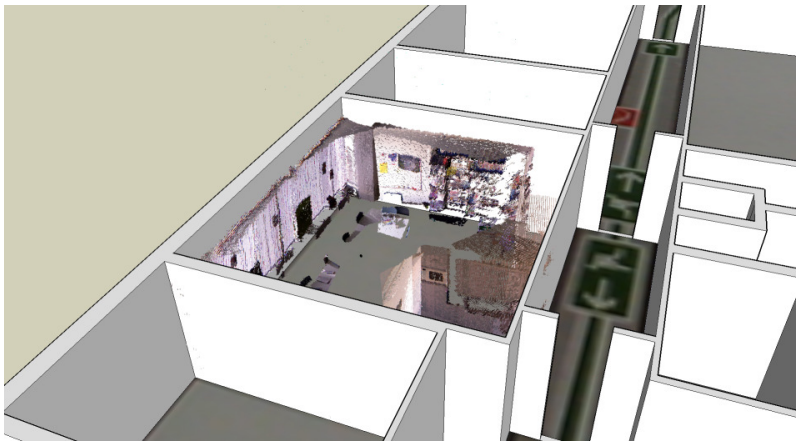

Fig. 15: Point cloud collected in a room fitted to the available indoor model.

5. CONCLUSION

Within the paper, we presented a pedestrian mobile mapping system, which aims at the collection of georeferenced 3D point clouds using a foot mounted MEMS IMU and a low-cost range camera. Such a measurement of area covering 3D point clouds for indoor environments from sequences of range scans or images is an important task within mobile robotics and computer vision. There, the Simultaneous Localization and Mapping (SLAM) problem aims at placing a mobile robot at an unknown location in an unknown environment. The robot then incrementally builds a consistent map of this environment while simultaneously determining its location within this map (Fietz et al., 2010). Such approaches usually align consecutive data frames, while the detection and global alignment of loop closures additionally improves the consistency of the complete data sequence. These systems generate a representation of the scene by collection of dense 3D point clouds, while the actual localization is more of a by-product. Our approach additionally integrates coarse models of the indoor environment. Thus, much more semantic information like room number, position of stairs, elevators can potentially be made available by the interpretation of ground or evacuation plans. Since the integration of this information for navigation and route planning within pedestrian indoor applications is very reasonable, these issues will be tackled in our future work.
Within the paper, a coarse 3D model for the building interior was successfully generated. However, the implemented automatic image interpretation is still highly adapted to the visual appearance of the captured evacuation plan. Similar problems have also been reported during the use of architectural drawings for the reconstruction and 3D modeling of building interiors (Yin et al., 2009). Even though a number of researchers and CAD developers aim on the automatic conversion of 2D drawings into 3D models, the lack of generality still remains the most important shortcoming. This problem is facilitated for evacuation plans since they do not contain too much and complex information. Thus, additional work to allow for an interpretation on a more abstract and thus general level is still required.

6. REFERENCES

Bauer, W. & Mohl, H.-U., 2005. Das 3D-Stadtmodell der Landeshauptstadt Stuttgart. In: 3D-Geoinformationssysteme: Grundlagen und Anwendungen, eds. Coors, V. and Zipf, A. e., Wichmann Verlag, pp. 265-278.

Becker, S. & Haala, N., 2009. Grammar supported Facade Reconstruction from Mobile LiDAR Mapping. In: ISPRS Workshop on Object Extraction for 3D City Models, Road Databases and Traffic Monitoring. Paris, France.

Besl, P. J. and MacKay, N., 1992. A method for registration of 3-D shapes. IEEE Transactions on Pattern Analysis and Machine Intelligence, 14(2), pp. 239-256.

Burrus, N., 2011. Demo software to visualize, calibrate and process Kinect cameras output: http://nicolas.burrus.name/index.php/
Research/KinectRgbDemoV5?from=Research.KinectRgbDemoV4 (Accessed 1 Apr. 2011)

Fietz, A., Jakisch, S.M., Visel, B.A., Fritsch, D., 2010. Automated 2D Measuring of Interiors Using a Mobile Platform. In: Proceedings of the 7th International Conference on Informatics in Control, Automation and Robotics (ICINCO 2010), Funchal, Madeira/Portugal.

Godha, S. & Lachapelle, G., 2008. Foot mounted inertial system for pedestrian navigation. Measurement Science and Technology, 19(7), 075202.

Henry, P., Krainin, M., Herbst, E., Renand, X., Fox, D., 2010. RGB-D Mapping: Using Depth Cameras for Dense 3D Modeling of Indoor Environments, RSS Workshop on Advanced Reasoning with Depth Cameras.

Lowe, D. G., 2004. Distinctive image features from scale-invariant keypoints. International journal of computer vision, 60(2), pp. 91-110.

Neufert, E. et al., 2002. Architects' Data 3rd ed., Wiley-Blackwell.

OpenKinect, 2011. OpenKinect open source project:
http://openkinect.org/wiki/Main_Page (Accessed 1 Apr. 2011)

OpenStreetMap Wiki, 2011. Beginners' guide - OpenStreetMap Wiki.:
http://wiki.openstreetmap.org/wiki/Beginners%27_Guide (Accessed 1 Apr. 2011)

Peter, M., Haala, N., Schenk, M. & Otto, T., 2010. Indoor Navigation and Modeling Using Photographed Evacuation Plans and MEMS IMU. IAPRS, Vol. XXXVIII, Part 4, on CD.

ROS OpenNI Kinect, 2011. ROS OpenNI open source project:
http://www.ros.org/wiki/openni_kinect (Accessed 1 Apr. 2011)

Suzuki, S. et al., 1985. Topological structural analysis of digitized binary images by border following. Computer Vision, Graphics, and Image Processing, 30(1), pp. 32-46.

Yin, X., Wonka, P. & Razdan, A., 2009. Generating 3D Building Models from Architectural Drawings: A Survey. IEEE Computer Graphics and Applications, 29(1), pp. 20-30.

Zhang, T. Y. & Suen, C. Y., 1984. A fast parallel algorithm for thinning digital patterns. Communications of the ACM, 27(3), pp. 236-239.