Urban Positioning Using Smartphone-Based Imaging

Deyvid Kochanov[1], Stephan Jonas[1], Nader Hamadeh[1], Ercan Yalvac[1], Hans Slijp[2], Thomas M. Deserno[1]

[1] Dept. of Medical Informatics, Uniklinik RWTH Aachen, Aachen, Germany
[2] I-Cane Social Technologies, Sittard, Netherlands
sjonas@mi.rwth-aachen.de

Abstract. Orientation and navigation in a world dominated by visual signs is still a major problem for blind and visually impaired people. The Global Positioning System is of limited use due to its inaccuracy particularly in urban environments. Therefore, we propose a novel approach of precise localization on predefined routes with the help of smartphones and image processing techniques. From an initial acquisition of a given route, a three-dimensional reconstruction is created. A query image is submitted to the database and the location and direction of the camera are calculated. Here, we demonstrate our approach on a evaluation-dataset with a mean positioning error of 5.51 meters.

1 Introduction

According to global estimates, 15.5 million people in Europe are visually impaired [1]. The same study concludes that further growth is expected due to the increase of diabetes and its effect on elder people. This emphasises the need for reducing the impact or burden of sight loss by delivering improved support through technology and innovation. Foremost, independency and mobility of the blind and visually impaired has to be improved supporting free navigation through unknown terrain. However, only very few technical advances towards mobility, navigation and independence have been made since the introduction of the white cane in the early 20th century.

New technological developments like the Global Positioning System (GPS) [2], accelerometers and digital compassess provided novel opportunities to assist visually impaired blind people in the interpretation of environment and navigation tasks. Moreover, satellite-based GPS positioning is not sufficiently accurate to guide pedestrians through traffic, especially in urban environments. In particular, larger cities have tall buildings, which block or reflect satellite signals reducing the accuracy to 34 meters [3].

Another approach uses the number of steps detected by an accelerometer, reference-points and a mobile compass for navigation assistance. Fallah et al. [4] presented a successful example of this method also with combining probabilistic algorithms with natural capabilities of visual impaired persons to detect landmarks like corners with touch. However, this system is designed for indoor

environments, where maps are precisely defined by landmarks like corners and doors.

There is also supporting research that exploits stereo vision cameras or depth cameras like the Microsoft Kinect to navigate around obstacles [5]. However, the range of these devices is only a few meters and they are not fit for general navigation tasks that require localization.

Technologies that can be used in navigation systems today either have poor reliability in different conditions because of inaccuracies in measurement devices or are too expensive or too large to be integrated into mobile devices. Our system aims at using previously calculated 3D reconstruction and real time image processing on mobile phones to solve the accuracy problem while keeping the system affordable and easy to carry.

2 Materials and methods

Current three-dimensional (3D) reconstruction methods like structure from motion [6] can be used to generate city-scale models of real world scenes using unordered sets of 2D images. These reconstructions provide compact 3D models in the form of sparse point clouds (Fig. 1). The construction of these models is computationally expensive but it can be done off-line and incrementally.

Modern mobile devices like smartphones have sophisticated cameras and enough computational capacity to perform image processing tasks like feature extraction and detection. Free and open source libraries with implementations of the algorithms performing these tasks are readily available, for example MATLAB (The Mathworks[1]) and OpenCV[2].

[1] www.mathworks.com
[2] www.opencv.org

(a) Sample image (b) Sample 3D point cloud

Fig. 1. Example 3D reconstruction of Aachen central market. The 3D reconstruction was created from a large dataset acquired for visualization reasons, not the evaluation dataset.

2.1 Reference acquistion

The images for the reference model are acquired while walking a route with a specialized smartphone application. The application records about one image per second along with other sensory data like GPS, accelerometer, magnetometer and gyroscope. The images and embedded metadata are buffered on the device and transferred to the server via the Internet connected by third (3G) or fourth (4G) generation mobile communication standards.

2.2 Model reconstruction

Image acquisition and processing for the 3D model construction is performed by first extracting image features with the scale invariant feature transform (SIFT) using MATLAB and OpenCV. Then, a 3D point cloud model is constructed using VisualSFM [7, 8]. This structure from motion methods also estimates the locations of the cameras in the initial set, which then can be used to construct navigation routes. The 3D model can be stored in a database of possible routes and used for localization of users identified by query images. Existing models can also be augmented by acquiring a route multiple times.

2.3 Query image localization

To estimate the location of the user, SIFT features are extracted from the image taken at the user's current position. Correspondences to the features of the points in the 3D point cloud are determined by calculating the distances between all features of the image against the point cloud and applying a threshold. The obtained matches are used to estimate the camera pose of the initial image with the following process. First, the camera matrix P is reconstructed. The camera matrix P maps the 3D world points X to their 2D image coordinates x within an unknown scaling factor λ

$$PX = \lambda x \tag{1}$$

In (1) P can be estimated from a set of corrspondences using the pseudoinverse.

$$PXX^T = \lambda xX^T \tag{2}$$

$$P = \lambda xX^T \left(XX^T\right)^{-1} \tag{3}$$

A more sophisticated and numerically stable method to estimate the matrix exists [9]. At least six correspondences are needed for the estimation of the camera matrix P. Howerver, noise caused by errors in the reconstruction and outlier matches can lead to errors in the estimation of the projection matrix.

To make the process more robust, random sample concensus (RANSAC) [9, 10] is used to discard outliers. RANSAC chooses a random sample from our set of correspondences, reconstruct the matrix and checks for consensus with the rest of the matches. Once a large enough set of inliers is found, the final camera matrix is computed. If no concensus is reached, the best model so far

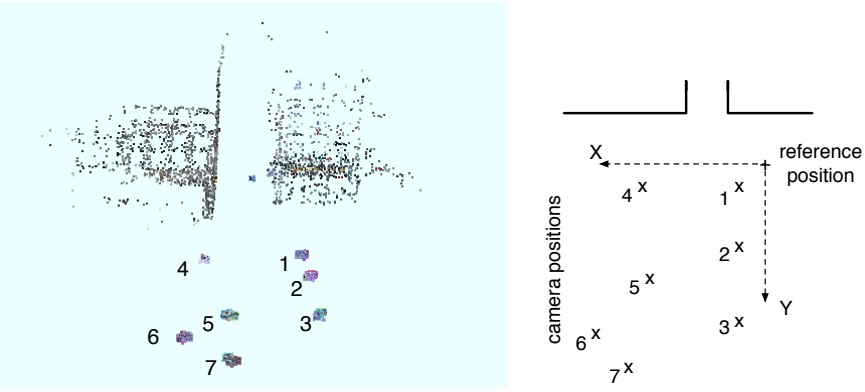
is returned. The camera position and orientation can be extracted from the camera projection matrix

$$P = K[R|t] \tag{4}$$

which is a composition of the projection matrix K and the camera motion matrix $[R|t]$ with R being a rotation matrix and t a translation vector. Because K is upper triangular matrix and R is orthogonal, QR decomposition can be used to compute the K and R from P. Thereby, the camera intrinsic parameters K and the relative camera position and orientation (R and t) are obtained [9]. After estimating a relative coordinates in the model, a real world position is computed by using real world locations of the cameras or keypoints in the initial dataset (e.g., GPS coordinates from a large number of cameras).

2.4 Evaluation

For ground truth data acquisition, an imaging protocol was set up. The protocol differs from the regular reference acquisitions by using a tripod to account for changes in height and making accurate measures of each camera location towards a reference point. This allows for easy reproduction of the set under differenct environmental conditions. The central marketplace of the city of Aachen was chosen as test location as it features both high buildings as well as pedestrians and other moving objects partly occluding the visual landmarks surrounding the buildings. Images at seven different locations and with three different angels were acquired with our acquisition app resulting in a total of 21 images (Fig. 2). In addition, a spreadsheets with the exact location of each camera towards a reference-point was created.



(a) 3D Reconstruction (b) Reference positions

Fig. 2. Evaluation set 3D reconstruction and acquisition protocol of Aachen central market based on the complete evaluation set of 21 images. In the ground truth data, each camera position contains three shots in different direction (angulation of 30 degree).

Using a position-based leaving-one-out approach, six 3D models of the scene were created, each containing only 18 images from 6 locations. The remaining three images – all from the same position – were tested against the corresponding dataset. Since the 3D reconstructions itself can already contain errors, the distances of each query image towards the closest two positions, one in X and one in Y direction are used (Fig. 2(b)). Based on the ratio of these two distances and the supposed ratio, the positioning error was computed.

3 Results

Out of the 21 images used for the evaluation, one outlier localization was removed, which was more than 10 meters off. The average localization error of our system was 5.51 m (stdev 4.39 m).

4 Discussion

This work demonstrates that image guided localization is 2-5 times more precise than GPS-based localization which has an cummulated localization error of more than 30 m when relying to GPS only and more than 10 m with map-based correction [3]. Previous research suggest that from image base localization alone we can obtain localizations with error below 3 m with probabilty 0.75 and and below 18 m with probability 0.9 [10] which is compareable to our results. An increasing number of reference-images should furhter decrease the positioning error.

Another limitation of our evaluation is the quality of the reference images, as our evaluation images are acquired with a stable tripod while reference and query images are acquired while walking. Nonetheless, our results give a good indication of the potential of image-guided localization with smartphone cameras. further limitations of this approach are, of course, possible occlusions, changes in weather and lighting.

We aplied SIFT features and the RANSAC algorithm. Another approach that is often used in robotics is simultaneous localization and mapping (SLAM) [11]. However, the mapping would require additional computational power while our localization method can run on a minimal hardware setup.

Our smart camera-based positioning can then be used to navigate blind or visually impaired people on a predefined route. Next steps will focus on making the method more robust towards these influences and faster for realtime applications and a navigation prototype. Further improvements by using additional sensory data like GPS, accelerometer, magnetometer and gyroscope are currently investigated. The position accuracy can be increased by taking into account previous positions, current walking direction, and actual speed. Other benefits of image guided navigation, like depth calculation for the detection and warning of steps, gaps and other obstacles, will also be part of our future work.

Acknowledgement. This work was co-funded by the German Federal Ministry of Education and Research (BMBF, Grant No. 16SV5846) and the European Commission's Ambient Assisted Living (AAL) Joint Programme ICT for ageing well. (EU, Grant No. 810302758160 – IMAGO).

References

1. Resnikoff S, Pascolini D, Mariotti SP, et al. Global magnitude of visual impairment caused by uncorrected refractive errors in 2004. Bull World Health Organ. 2008;86(1):63–70.
2. Loomis JM, Golledge RG, Klatzky RL, et al. Navigation system for the blind: auditory display modes and guidance. Presence. 1998;7:193–203.
3. Modsching M, Kramer R, ten Hagen K. Field trial on GPS accuracy in a medium size city: the influence of built-up. Proc WPNC. 2006; p. 209–18.
4. Fallah N, Apostolopoulos I, Bekris K, et al. The user as a sensor: navigating users with visual impairments in indoor spaces using tactile landmarks. Proc SIGCHI. 2012; p. 425–32.
5. Filipe V, Fernandes F, Fernandes H, et al. Blind navigation support system based on Microsoft Kinect. Comp Sci. 2012;14:94–101.
6. Agarwal S, Snavely N, Simon I, et al. Building Rome in a day. Proc IEEE Int Conf Comput Vis. 2009; p. 72–9.
7. Wu C. Towards linear-time incremental structure from motion. Proc 3DV. 2013; p. 127–34.
8. Wu C, Agarwal S, Curless B, et al. Multicore bundle adjustment. Proc IEEE Comput Soc Conf Comput Vis Pattern Recognit. 2011; p. 3057–64.
9. Hartley RI, Zisserman A. Multiple View Geometry in Computer Vision. 2nd ed. Cambridge University Press; 2004.
10. Sattler T, Leibe B, Kobbelt L. Fast image-based localization using direct 2D-to-3D matching. Proc ICCV. 2011; p. 667–74.
11. Karlsson N, Di Bernardo E, Ostrowski J, et al. The vSLAM algorithm for robust localization and mapping. Proc IEEE ICRA. 2005; p. 24–29.