Indoor Positioning Using Smartphone Camera

Martin Werner* and Moritz Kessel* and Chadly Marouane**
*Ludwig-Maximilians-Universität/Mobile and Distributed Systems Group, Munich, Germany. Email: firstname.lastname@ifi.lmu.de
**Ludwig-Maximilians-Universität, Munich, Germany. Email: mchedly@googlemail.com

***Abstract*—With the increasing computational power of mobile devices and the increase in the usage of ordinary location-based services, the area of indoor location-based services is of growing interest. Nowadays indoor location-based services are used mainly for personalized information retrieval of maps and points of interest. Advanced location-based functionality often suffers from imprecise positioning methods. In this paper we present a simple, yet powerful positioning method inside buildings which allows for a fine-grained detection of the position and orientation of a user while being easy to deploy and optimize. The main contribution of this paper consists of the combination of an image recognition system with a distance estimation algorithm to gain a high-quality positioning service independent from any infrastructure using the camera of a mobile device. Moreover this type of positioning can be operated in a user-contributed way and is less susceptible to small changes in the environment as compared to popular WLAN-based systems. As an extension, we propose the usage of very coarse WLAN positioning to reduce the size of the candidate set of image recognition and hence speed up the system.**

***Index Terms*—Indoor Navigation; Image Processing; Location-based Services**

I. Introduction

In recent years, location-based services (LBS) began to form an increasingly important factor in industry and research. The growing spread and computational power of mobile phones and the rising number of applications result in app stores full of different location-based apps such as restaurant finders, tourist guides and navigation systems.

One of the key enablers of location-based services was the adoption of the easy-to-use and accurate GPS positioning technology in mobile phones. Unfortunately, GPS is not able to track people in indoor environments with acceptable accuracy. Signals might get lost due to attenuation effects of roofs and walls or lead to position fixes of very low accuracy due to multipath propagation.

Even worse, indoor location-based services require much higher precision guarantees than outdoor services. Errors should not exceed a few meter to allow for a differentiation between several floors or rooms. Otherwise, the service could provide information for places which are quite far away from the actual position of the target.

Existing indoor positioning techniques can be grouped by their level of precision and the expenses for additional infrastructure. Dedicated indoor positioning systems such as ultra wide band or ultrasonic systems consist of several components with the sole purpose of determining the positions of possibly multiple targets in indoor environments. The precision is often high, but an expensive infrastructure is needed and hence the navigable space is usually limited to a small area, where higher accuracy compensates the high cost. Another class of systems is build on existing infrastructure such as WLAN, Bluetooth or inertial sensors for positioning. The precision of such systems is limited, but the system can be deployed with little additional expenses.

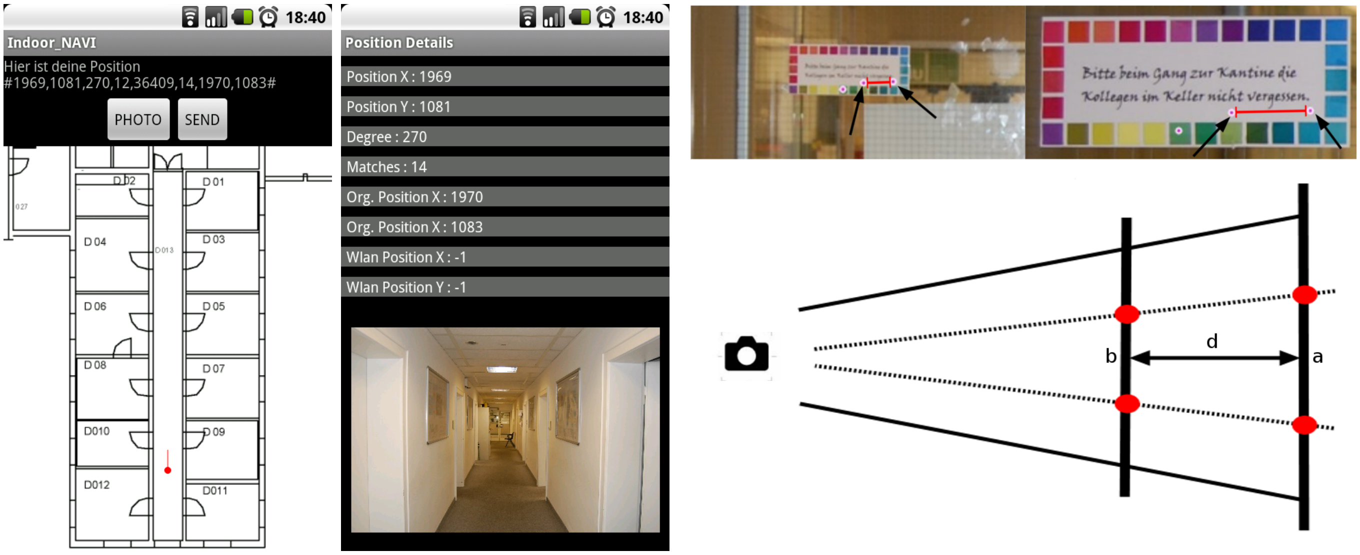
In this paper we present an approach for cheap and easy indoor positioning with no need for infrastructure components, in the sense that the positioning can be carried out with a mobile device using its camera. The system achieves a high precision of a few meters, detects the viewing direction of the user with high accuracy, is easy to setup and optimize and is very sensitive for semantic differences in navigation space. While a position error of one meter can lead to a wrong room assignment, these rooms are visually different and our system has generally a lower risk of assigning wrong semantic positions as opposed to purely geometric positioning systems. The approach is furthermore well-suited for indoor navigation, as the image used for positioning can easily be augmented with navigation instructions.

Similar to WLAN fingerprinting we use a database of images with the additional information of the corresponding position, the viewing direction, and a scale- and rotation-invariant description of the image, generated by the well-known SURF [1] algorithm. For the moment the database is created in an offline phase, but purely user-generated databases or self-calibrating systems are also possible. For position retrieval, a picture taken by the user is analyzed and the corresponding database image detected and the actual position is computed by a comparison of the object scale in both images.

The paper is structured as follows: We begin with a short description of the algorithms for image comparison we use. In section II-B we shortly review related vision-based indoor positioning systems which is followed by a description of our system. In section III we present the results we achieve with a prototypical implementation in our university building and discuss further improvements and future work.

II. Markerless Indoor Navigation Using Smartphone Camera

In the past few years, a wide variety of image analysis algorithms have been developed in the field of computer vision. On the one hand, there are image transformations which augment several visual properties of the image (e.g., edges [2] and corners [3]). The transformations process an image and create a simpler version of the same image that



Indoor_NAVI

Hier ist deine Position
#1969,1081,270,12,36409,14,1970,1083#

Position Details

Position X : 1969

Position Y : 1081

Degree : 270

Matches : 14

Org. Position X : 1970

Org. Position Y : 1083

Wlan Position X : -1

Wlan Position Y : -1

18:40
PHOTO
SEND
D02
D01
627
D04
D03
D013
D06
D05
D08
D07
D010
D09
D012
D011
Bitte beim Gang zur Kantine die Kollegen im Keller nicht vergessen.
b
d
a

(a) Screenshots of the mobile client

(b) Distance estimation algorithm

Fig. 1. Screenshots from the mobile client showing the position and orientation estimate and the reference image and position information from the database and the distance estimation algorithm using the ratio of matched pixel distance as a measure for viewpoint-to-image distance

can be used for further analysis. On the other hand, algorithms for the extraction of local, highly recognizable image features provide for more stable, rotation- and scale-invariant image processing. Both kinds of algorithms have been applied widely to the field of image hashing [4] and object recognition [5]. The first class of image analysis algorithms suffers from registration problems and are very sensitive to small changes in the environment. The second class of algorithms is very stable with respect to these problems, but suffers from more calculational overhead and problems of local similarities (e.g., corners of doors that look similar throughout a complete building). We decided to use feature transformation algorithms as the main ingredient of localization and describe them in more detail below.

A. Image Comparison Using Feature Points

Feature Points are points inside an image which allow for a local description that makes them highly recognizable. We describe two algorithms:

The Scale-Invariant Feature Transform (SIFT, [6]) algorithm uses a Gaussian blur along with a scale-invariant matching of local extrema to find a list of interest points. For each interest point a local and rotation-invariant descriptor is calculated which resembles some illumination-invariant properties of the surrounding of the point. The Euclidean distance between descriptors can be used for feature recognition as well as for object and image recognition.

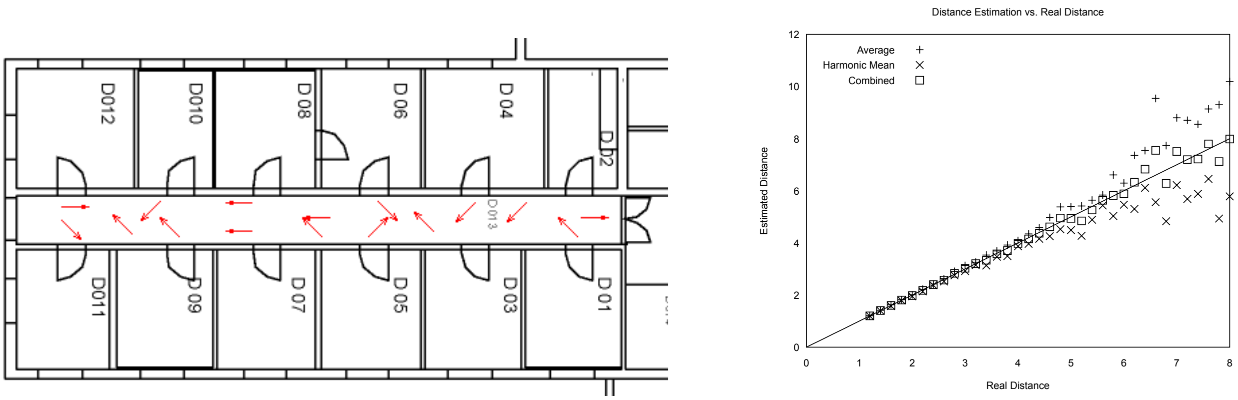
A comparable algorithm called Speeded Up Robust Features (SURF,[1]) applies less accurate but faster approximations for finding extrema and hence provides a faster and more memory-efficient extraction and description of local image features which is even possible to conduct on mobile devices.

For the recognition of images or objects, feature points can be matched using the Euclidean distance between their descriptors. Difficulties arise from interest points in the first image which can be matched to multiple points in the second image. The resulting problems are empirically solved in [6] using an ad-hoc matching process. Empirical results from this paper state that more than three feature points suffice to recognize dominant objects in the focus (e.g., a main motif) of the image and that more than ten feature points suffice to recognize more uniform images (e.g., natural scenery, buildings).

B. Visual Indoor Navigation Systems

Indoor Navigation is an important emerging field in the area of pervasive computing which tries to provide navigation services in buildings that are comparable to the navigation in the outside world. The main problems in buildings are the absence of accurate and cheap positioning systems and the unavailability of floor plans and maps with acceptable quality. Finding the position of a mobile asset inside a building is a difficult task. Several techniques have been developed, some based on existing infrastructure (mostly based on WLAN [7]) and some on additional active infrastructure (radio, audio and IR beacons) or on additional passive infrastructure (RFID, 2D barcode, etc.). A good overview is presented in [8]. We focus on computer vision methods for accurate indoor positioning without the need for any additional infrastructure. In this specific field of indoor positioning, Hile et. al [9] tried to use edge detection for counting doors and derive the position from a map matching step. Kawaji et. al [10] use omni-directional indoor images along with a SIFT variant called PCA-SIFT [11] to find the position.

In this paper we propose another method of indoor positioning using a smartphone camera which is simpler to calibrate



(a) The locations of the images in the evaluation set

(b) Performance of Distance Estimation

Fig. 2. The evaluation set and results on the distance estimation algorithm

as we do not need panoramic images and which allow for position correction using a flexible distance estimation scheme. By using a database of images taken with mobile phones, our system is easily extendable to user-contributed calibration (e.g., by including the images taken by users into the database). Moreover such a growing database easily adapts to smaller changes in the environment. Furthermore, the feedback of the actual database image which we found during matching allows for high confidence in the quality of the service.

C. MoVIPS - Mobile Visual Indoor Positioning System

Our proposed system, the Mobile Visual Indoor Positioning System (MoVIPS), is based on a distributed architecture. A mobile application is used to take images of the surroundings and upload them to a server component which compares this image with a database of correctly located and oriented images taken from the surroundings in an initial calibration step.

The mobile application has been implemented as an Android application and is capable of taking images, performing SURF transformation or uploading the image to a server. For easier evaluation and estimation of configuration parameters, our testbed usually does not analyze the image on the phone, but transfers the complete image to the server instead. Once an image has been taken and uploaded to the server, the results of the image comparison and the resulting position estimation are downloaded by the phone. The application then shows the location and orientation on a map. Figure 1a shows screenshots of the prototype. Additionally, we implemented a WLAN positioning system to reduce the number of images to be considered (the candidate set) from the database.

The server component extracts SURF feature descriptors from the incoming image. These are compared to every image in the candidate set from the database using the method described in [1] applied in both directions. The criteria for choosing the best image out of the database is then as follows: We take the two images which result in the best and second best rating with respect to the total count of matches in both directions. From these two choices we select the one with the smallest difference in the number of matches in the respective directions. The reason for this is that images with few features tend to have a small difference in the number of features while the number of feature matches is a general measure for the probability of a correct recognition.

At this point, the server component has found the most probable image along with its interest points and the reference image from the database. As the position, where the reference image has been taken, usually differs from the position, where the actual image has been taken, we implement a geometric position correction scheme. As depicted in figure 1b, the distance d between two images of the same object taken from different distances to the object is proportional to the ratio of the distance between those points in the image.

$$d = \alpha \frac{a}{b}, \quad \alpha \text{ constant describing camera field of view}$$

The constant parameter α describes the field of view of the camera and can be calculated from the field of view or simply calibrated from two images with known distance to each other. While comparing images, we do not know two points that definitely match correctly. Hence we rely on the calculation of the respective ratio for each pair of matching points. In figure 2b you can see that the average of these values tends to overestimate the distance while the harmonic mean has the same tendency to underestimate the result. Hence we use the average of both values as the ratio value for distance estimation. Relative errors for the three cases (harmonic mean, average and the combination of harmonic mean and average) are depicted in figure 3. This results in a distance estimation, which is then used to push back the position of the image along its stored orientation to get the real position.

III. Results

The results from our prototype implementation of MoVIPS include a detailed evaluation of the position corrections schema as well as empirical results concerning the position accuracy.

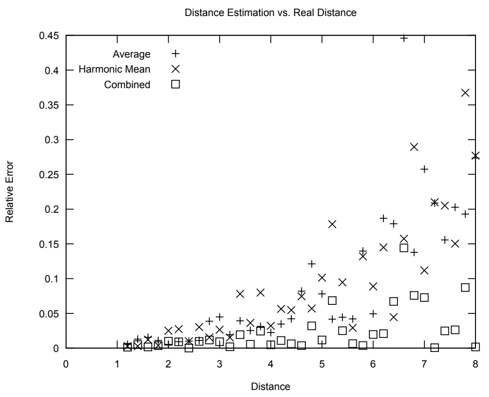


Fig. 3. Relative Error of Distance Estimation

The SURF algorithm allows to influence the number of features by setting a threshold value indicating how distinguishable a feature point inside an image has to be. For lower thresholds, more points are reported as interest points, but the number of wrong matches will increase. High threshold values could miss important features resulting in no recognition at all. In our setting, the mobile application shall perform the SURF transformation locally and hence the number of features is proportional to the communication cost. Moreover the complexity of the matching procedure is quadratic in the number of features. From our experiments with a smartphone camera, a threshold value of 0.0004 for the incoming and the reference images led to optimal results.

For the correction of the position using the distance estimation, we first calibrated the camera field of view. The quality of the resulting distance estimation is evaluated against a synthetic set of images to reduce the impact of noise and quality on the evaluation. The results are depicted in figure 2b and 3. The main source of errors in this case is the fact, that we do not know the exact distance between two points due to possible wrong matches and have to rely on the mean values of the ratios between pairs of matches in both images.

Furthermore, we evaluated our system with respect to the position accuracy at a university building. An initial database modeling a long corridor with high self-similarity has been constructed by taking 68 images in a regular pattern and storing the position and orientation along with them. A floor plan and an image of this corridor is depicted in figure 1a. Additionally we took a set of 16 test images at random positions and orientations as depicted in figure 2a.

We compare our results with the accuracy of RADAR [7]. They report results with a median position error of $2.94m$ and a worst-case distance around $23m$. Our testbed installation using the 16 test images resulted in a median position error of $1.32m$. However as the position does not continuously depend on the image, the worst-case position error is unpredictable. The reason for this is the fact, that we could have two images which are very far from each other but match. A countermeasure would be to use a coarse WLAN position for the database reduction resulting in an upper bound for the error. It is worth noting that our system is able to report a high-quality orientation value which does not depend on the magnetic environment nor on the pose of the phone and hence allows for augmenting a navigation application with arrows on the screen. Moreover our system is easy to tune. The pairwise checking of all images in the database can show places with a high similarity. They can be enhanced by additional paintings or furnishing.

IV. Conclusion

In this paper we presented an indoor positioning system independent from infrastructure which showed very promising result. To create a clear picture of functionality and stability, we did not include any spatial or temporal information into the positioning method. It is obvious that the elapsed time from and the position of the last position fix can help to reduce the candidate set of database images much more efficient and completely independent from infrastructure (e.g, by including movement models). Moreover one should investigate in future work whether the image recognition technology can exploit specialties of indoor images and reduce the set of feature points by removing misleading features which do not contribute to positioning (e.g., features that match on many database images with the corresponding positions near each other) and how to report problematic areas for this indoor positioning technology to the system operator by cross-checking database images for similarity.

References

[1] H. Bay, T. Tuytelaars, and L. Van Gool, “Surf: Speeded up robust features,” *Computer Vision–ECCV 2006*, pp. 404–417, 2006.

[2] J. Canny, “A computational approach to edge detection,” *Readings in computer vision: issues, problems, principles, and paradigms*, vol. 184, pp. 87–116, 1987.

[3] C. Harris and M. Stephens, “A combined corner and edge detector,” in *Alvey vision conference*, vol. 15. Manchester, UK, 1988, p. 50.

[4] R. Venkatesan, S. M. Koon, M. H. Jakubowski, and P. Moulin, “Robust image hashing,” in *Proceedings 2000 International Conference on Image Processing*, vol. 3. IEEE, 2000, pp. 664–666.

[5] M. Martinez, A. Collet, and S. S. Srinivasa, “MOPED: A Scalable and low Latency Object Recognition and Pose Estimation System,” in *IEEE International Conference on Robotics and Automation*. IEEE, 2010.

[6] D. Lowe, “Object recognition from local scale-invariant features,” in *iccv*. Published by the IEEE Computer Society, 1999, p. 1150.

[7] P. Bahl and V. N. Padmanabhan, “RADAR: an in-building RF-based user location and tracking system,” in *Proceedings IEEE INFOCOM 2000. Conference on Computer Communications.*, vol. 2. IEEE, 2000, pp. 775–784.

[8] H. Huang and G. Gartner, “A Survey of Mobile Indoor Navigation Systems,” *Cartography in Central and Eastern Europe*, pp. 305–319, 2009.

[9] H. Hile and G. Borriello, “Information overlay for camera phones in indoor environments,” in *Location-and context-awareness: third international symposium, LoCA 2007*. Springer-Verlag New York Inc, 2007, p. 68.

[10] H. Kawaji, K. Hatada, T. Yamasaki, and K. Aizawa, “Image-based indoor positioning system: fast image matching using omnidirectional panoramic images,” in *Proceedings of the 1st ACM international workshop on Multimodal pervasive video analysis*. ACM, 2010, pp. 1–4.

[11] Y. Ke and R. Sukthankar, “PCA-SIFT: A more distinctive representation for local image descriptors,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE Computer Society, 2004.