Indoor Positioning Using Smartphone Camera

Martin Werner and Moritz Kessel
Mobile and Distributed Systems Group
Ludwig-Maximilians-University
Munich, Germany
Email: firstname.lastname@ifi.lmu.de

Chadly Marouane
Ludwig-Maximilians-University
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.

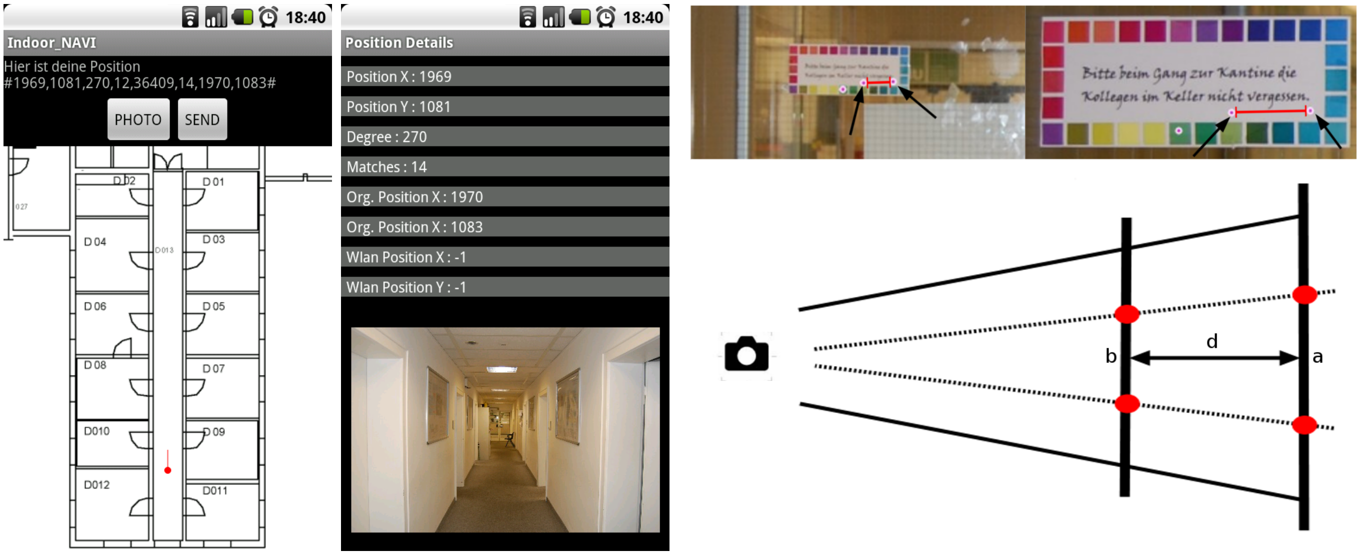
Existing indoor positioning techniques can be grouped by their level of precision and the expenses for additional infrastructure. Dedicated indoor positioning systems such as ultra wide band or ultrasonic systems consist of several components with the sole purpose of determining the positions of possibly 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 an approach for cheap and easy indoor positioning is presented 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 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, is used. For the moment the database is created in an offline phase, but purely user-generated databases or self-calibrating systems are also possible.

The system supports three modes for the position estimation. The first mode is based on a picture taken by the user which is analyzed by the system (photo mode). Then the corresponding database image is detected and finally the actual position is computed by a comparison of the object scale in both images. The second and the third mode are based on a video-stream from the mobile device. The stream consists of low quality images which are used for a continuous position estimation, but require a more sophisticated processing due to the low resolution and motion blur. While the second mode (averaging) utilizes an averaging of positions derived by matching all frames within a sliding window of the video stream, the third mode (voting) utilizes a voting algorithm for the position

978-1-4577-1804-5/11/$26.00 © 2011 IEEE



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 X : 1083

Wlan Position X : -1

Wlan Position Y : -1

18:40
PHOTO
SEND
D02
D01
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

retrieval.

The paper is structured as follows: It begins with a short description of the algorithms for image comparison used for the matching with the database. In Section II-B related vision-based indoor positioning systems are shortly reviewed which is followed by a description of our system. In Section III the results achieved with a prototypical implementation in a university building are presented. Section IV concludes the paper and describes further improvements and future work.

II. Markerless Indoor Positioning 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 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. In the following paragraphs two algorithms are described:

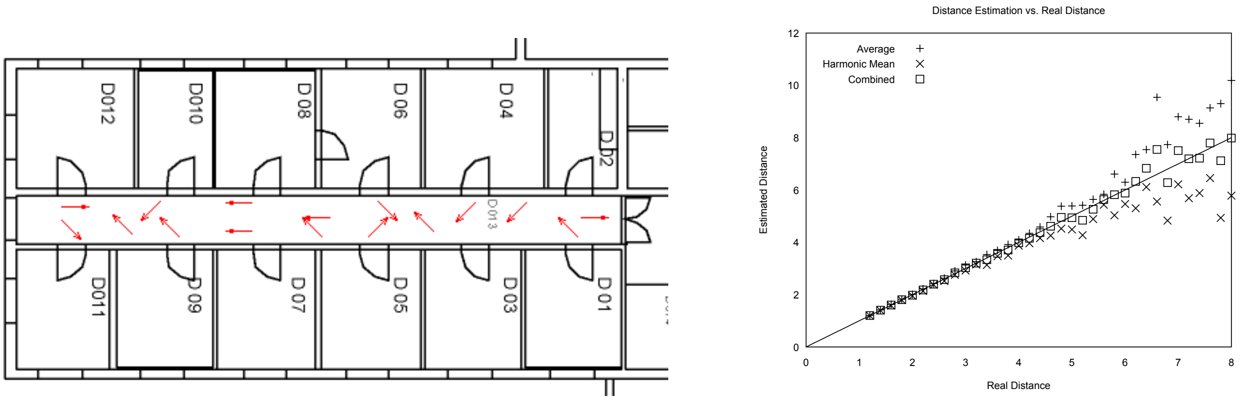
The Scale-Invariant Feature Transform (SIFT, [6]) algorithm uses a Gaussian blur along with a scale-invariant matching of local extrema to find a list of interest points. For each interest point a local and rotation-invariant descriptor is calculated which resembles some illumination-invariant properties of the surroundings 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 Positioning and 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



(a) The locations of the test images

(b) Performance of Distance Estimation

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

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 of wireless indoor positioning methods is given in [8] while several indoor navigation systems are presented in [9]. We focus on computer vision methods for accurate indoor positioning without the need for any additional infrastructure. Those methods can be broadly categorized into two classes. The first class is applied in an unknown environment where simultaneous localization and mapping (SLAM) is carried out. The VSLAM system builds a map from camera images and simultaneously extracts features from the images used for localization [10]. The second class of systems utilizes computer vision methods to estimate the location with the help of images in well-known environments. In this specific field of indoor positioning, Hile et. al [11] use edge detection for counting doors and derive the position from a map matching step. However, they report problems with the concealed doorsteps and larger rooms with furniture. Kawaji et. al [12] use omni-directional indoor images along with a SIFT variant called PCA-SIFT [13] to find the position. The algorithm works well inside of large rooms, but the accuracy depends on the density of panoramic images since only image recognition and no position correction scheme is applied.

In this paper another method of indoor positioning is proposed using a smartphone camera which is simpler to calibrate as panoramic images are not needed and which allow for position correction using a novel flexible distance estimation scheme. By using a database of images taken with mobile phones, the 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 is 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 or record a video (i.e., a continuous sequence of low resolution images) of the surroundings. Each image is uploaded to a server component which compares the image with a database of correctly located and oriented images. These were taken from the surroundings in an initial calibration step. The mobile application has been implemented as an Android application and is capable of performing the SURF transformation or uploading images 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.

MoVIPS supports three modes: The first mode or photo mode is used for precise positioning. A high resolution image is taken and sent to the server where the SURF feature descriptors are extracted. These are compared to the features of 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: 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 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, a geometric position correction scheme was implemented. 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, no two points are known that definitely match correctly. Hence, the system relies on the calculation of the respective ratio for each pair of matching points. In Figure 2b it can be seen that the average of these values tends to overestimate the distance while the harmonic mean has the same tendency to underestimate the result. Hence, the average of both values is used 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 used to push back the position of the image along its stored orientation to get the real position.

In the second mode or video mode (averaging) and in the third mode or video mode (voting) a low resolution video is continuously taken and the frames are sent to the server. Since the video mode is much more error prone due to the lower resolution and motion blur, an enhanced positioning method is utilized. The features of each incoming frame are extracted, matched and the position is corrected in the same way as in the photo mode. To reduce the impact of matching with a wrong image an additional correction scheme is applied. Instead of returning a position for every frame, the video mode (averaging) returns the average of the estimated positions of all frames within a sliding window as the position of the target. The video mode (voting) utilizes also a sliding window, but returns the position of the database image which has the highest number of matches within the sliding window of frames.

Finally the corrected position estimation and the matched image from the database are downloaded to the mobile phone and the location and orientation of the phone are displayed on a map. Figure 1a shows screenshots of the prototype. Additionally a WLAN positioning system was implemented to reduce the number of images to be considered (the candidate set) from the database. However, the description of this system is outside the scope of the paper.

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.

The SURF algorithm allows influencing 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 with respect to performance and precision.

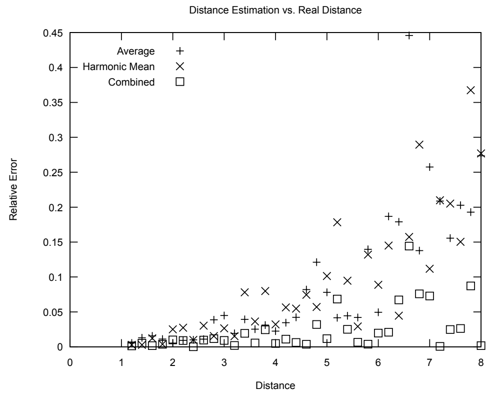


Fig. 3. Relative Error of Distance Estimation

For the correction of the position using the distance estimation, the camera field of view is calibrated. 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 the exact distance between two points is not known due to possible wrong matches and therefore one has to rely on the mean values of the ratios between pairs of matches in both images.

Furthermore, the system is evaluated 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 (with a resolution of 2560x1920) in a regular pattern and storing the position and orientation along with them. The initial calibration took approximately 10 minutes including taking the pictures and storing them. Another 5 minutes were needed to compute the database of image features. The calibration process is time consuming, but only needs to be repeated locally in the case of serious changes in the visual environment.

A floor plan and an image of this corridor is depicted in Figure 1a. The positions and orientations of the evaluation images are depicted in Figure 2a.

The storage cost of the images and the database on the server depends on the resolutions of the camera and the threshold value of the SURF transformation. With a resolution of 2560x1920 for the images and a threshold of 0.0004 for the SURF algorithm the database of feature points was 32.2MB in size and all images added to 134.3MB.

In a series of experiments we estimated the accuracy of

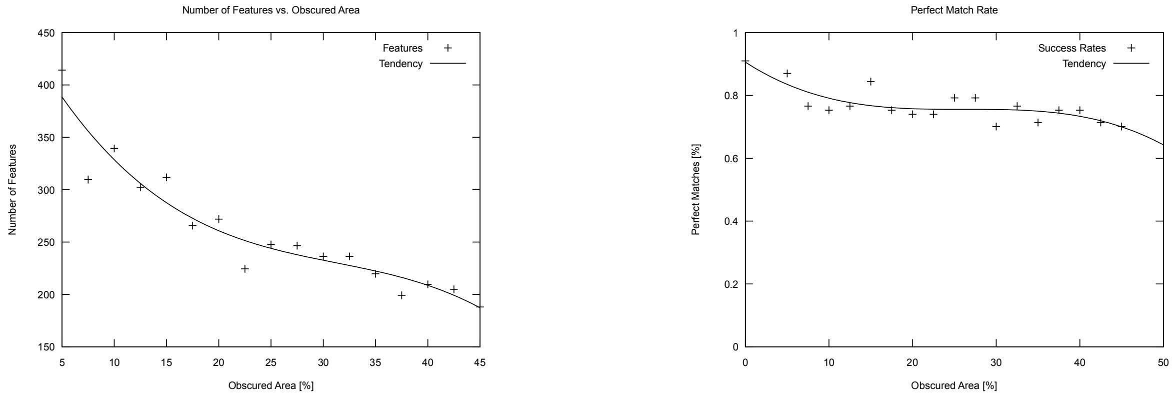


Fig. 4. The influence of overlaying the images with black boxes of the specified fraction of the image

TABLE I
THREE PERCENTILES OF THE CUMULATIVE DISTRIBUTION FUNCTION (CDF) OF THE POSITION ERROR

System	25% [m]	50% [m]	75% [m]
Photo Mode (Stationary)	0.28	0.68	1.25
Video Mode (Averaging)	2.39	3.88	4.4
Video Mode (Voting)	1.0	2.85	4.4
RADAR	1.92	2.94	4.69

our positioning system in three modes: In the first mode, stationary pictures were taken at 17 reference positions once with a smartphone camera with a resolution of 2560x1920. In the second and third mode, images were extracted from a live video-stream with a resolution of 640x480. In the second mode, the average of the positions induced from the frames of a sliding window of approximately 15 frames (i.e., in 500 ms intervalls) was selected and used for positioning with the stationary algorithm. In the third mode, the position of the image which is induced most often from the individual frames is taken. Table I shows the position error of three percentiles for those three modes.

As is to be expected, the positioning quality of both video modes is worse than that of the stationary mode due to lower resolution, encoding artifacts and motion blur.

Previous research on indoor localization using images uses the image recognition performance as the quality measure. As this does not give any insight in the usability of the system with respect to the accuracy, we decided to compare our results with the accuracy of RADAR [7], which is given in meters. They report results with a median position error of 2.94m and a worst-case distance around 23m. Our testbed installation using the 17 test images resulted in a median position error of 0.68m in the photo mode (see Table I). However, as the position does not continuously depend on the image, the worst-case position error can be arbitrary high. The reason for this is the fact that two images which are very far from each other could 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 the 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 the 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.

Investigating the database in more detail, we performed a one-against-all cross-validation in the following manner: To test the internal ambiguity of the dataset, we estimated the position of each calibration photo with our algorithms. This was done twice: Once against the complete database including the specific image and once using only the other images as the database.

Using the algorithm with the calibration images showed good performance as expected. Almost all (97%) images were recognized correctly. The maximal position error due to matching with the wrong image was 10m. The influence of selecting the wrong image on the total error of positioning is 73%. The maximal positing error due to a wrong distance estimation is 2.8m. Nevertheless the average total error of 0.28m can be neglected. However, this experiment indicates that some quality analysis of the content of the image database and a preselection of possible images using a coarse WLAN positioning can be of great help. Though the maximal errors of WLAN positioning and our approach are comparable it is unlikely that they will occur in parallel. Hence, a combination of both approaches could gain a lot.

Omitting the test image from the database showed, that 70% of the images matched with another image and hence led to a position estimate. The other 30% did not match with any other image and hence revealed no position. The position error for the cases, where a position has been returned has an average of 4.71m. This average is considerably better than the average distance between images, which is 5.38m. Hence, the algorithm selected neighboring images in many cases.

Furthermore the effect of partially concealed feature points, if parts of an image are hidden behind persons or other nonstationary objects, is analyzed. To investigate the effect of objects overlaying parts of the image, we created a test-set by

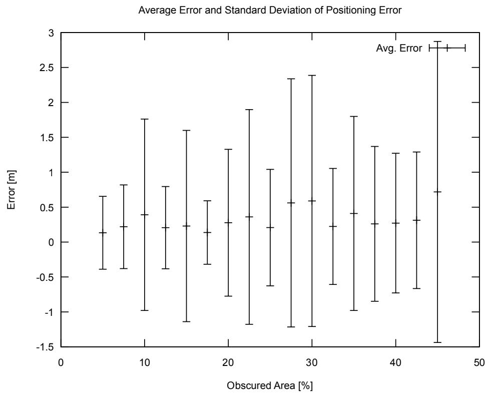


Fig. 5. Average error and standard deviation of overlaid images

blacking out a random rectangular area inside each calibration image covering a specified fraction of the total area. These images were then used to infer positions and led to the results depicted in Figure 4 and Figure 5.

Figure 5 shows the average error and the standard deviation in meters using images with an overlay. As can be seen, the influence can be neglected. The average error is always below one meter and the standard deviation is high. This is due to the fact that again most of the images were correctly matched and returned a nearly correct position whereas some images were mapped to completely different places. The left diagram in Figure 4 shows the number of feature matches between the obscured image and the selected database image. The number drops with increasing overlay area, but the influence is not really strong (see the right diagram in Figure 4). The success rate of the positioning system that is the fraction of images which were detected correctly keeps above 70%, though it shows a tendency to drop with increasing overlay area. This behavior was to be expected, because the number of features in the images with overlay is still sufficiently high.

IV. Conclusion and Future Work

In this paper we presented MoVIPS, an indoor positioning system based on a camera phone which is independent from infrastructure. MoVIPS utilizes the SURF algorithm to extract feature points from images taken with a mobile phone's camera and matches those feature points to a database of feature points from images of the surrounding environment. The system offers three modes for location estimation: In the first mode, the system is able to match a single picture with the images of the database and returns a high accuracy position estimate and the estimated orientation of the user. In the second and third mode a low resolution video-stream is analyzed and utilized for the location estimation for continuous positioning and tracking. However, the accuracy drops due to the low resolution and motion blur.

MoVIPS offers a high degree of accuracy and precision. The median position error in the first mode was 0.68m in a test environment inside a corridor of a university building. However, the worst-case position error is unpredictable, because two images which are very far from each other could match. We propose to use MoVIPS together with a coarse WLAN positioning to introduce an upper bound for the positioning error and reduce the candidate set of database images. In the third mode, where a voting algorithm is applied on a sliding window of frames from the video-stream for position estimation, the median position error was 2.85m.

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 we will 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). We are currently working on mechanisms for automatically reporting problematic areas for this indoor positioning technology to the system operator by cross-checking database images for similarity. This feature could be used together with a user-generated self-calibrating database of images.

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 *International Conference on Computer Vision*. 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 *IEEE INFOCOM 2000. Conference on Computer Communications.*, vol. 2. IEEE, 2000, pp. 775–784.

[8] H. Liu, H. Darabi, P. Banerjee, and J. Liu, "Survey of Wireless Indoor Positioning Techniques and Systems," *IEEE Transactions on Systems, Man, and Cybernetics–Part C: Applications and Reviews*, vol. 37, no. 6, pp. 1067–1080, 2007.

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

[10] N. Karlsson, E. D. Bernardo, J. Ostrowski, L. Goncalves, P. Pirjanian, and M. E. Munich, "The VSLAM Algorithm for Robust Localization and Mapping," in *International Conference on Robotics and Automation, ICRA*. IEEE, 2005, pp. 24–29.

[11] H. Hile and G. Borriello, "Information Overlay for Camera Phones in Indoor Environments," in *Location-and Context-Awareness: Third International Symposium, LoCA*. Springer, 2007, p. 68.

[12] H. Kawaji, K. Hatada, T. Yamasaki, and K. Aizawa, "Image-Based Indoor Positioning System: Fast Image Matching Using Omnidirectional Panoramic Images," in *1st ACM International Workshop on Multimodal Pervasive Video Analysis*. ACM, 2010, pp. 1–4.

[13] Y. Ke and R. Sukthankar, "PCA-SIFT: A More Distinctive Representation for Local Image Descriptors," in *IEEE Conference on Computer Vision and Pattern Recognition*. IEEE Computer Society, 2004.