How Computer Vision Can Help in Outdoor Positioning

Ulrich Steinhoff[1], Dušan Omerčević[2], and Roland Perko[2],
Bernt Schiele[1], and Aleš Leonardis[2]

[1] TU Darmstadt, Germany
[2] University of Ljubljana, Slovenia

Abstract. Localization technologies have been an important focus in ubiquitous computing. This paper explores an underrepresented area, namely computer vision technology, for outdoor positioning. More specifically we explore two modes of positioning in a challenging real world scenario: single snapshot based positioning, improved by a novel high-dimensional feature matching method, and continuous positioning enabled by combination of snapshot and incremental positioning. Quite interestingly, vision enables localization accuracies comparable to GPS. Furthermore the paper also analyzes and compares possibilities offered by the combination of different subsets of positioning technologies such as WiFi, GPS and dead reckoning in the same real world scenario as for vision based positioning.

Keywords: computer vision based positioning, local invariant features, sensor fusion for outdoor localization.

1 Introduction

Today, there exist a wide variety of location systems both commercial and research with different levels of accuracy, coverage, cost of installation and maintenance, and frequency of location updates [1,2]. Not surprisingly however any system has its respective strengths and weaknesses. For example standard GPS-based systems can achieve good levels of accuracy but the coverage is often limited in practice. Positioning based on WiFi (802.11 access points) is an interesting alternative but the achievable accuracy strongly depends on the density of mapped beacons [3]. Also GSM-based positioning has been advocated to achieve meaningful levels of accuracy, usable for various applications, even though not (yet?) in the range of GPS-based positioning [4].

Another far less explored alternative is computer vision-based positioning. There are several reasons why computer vision-based positioning presents an interesting alternative to more traditional positioning systems. First, the ubiquity of camera-equipped mobile devices and proliferation of high-speed wireless data connections increases the chances of acceptance of such a technology. Second, recent advances in computer vision technology suggest that vision-based positioning may enable highly accurate positioning based on single camera snapshots

B. Schiele et al. (Eds.): AmI 2007, LNCS 4794, pp. 124–141, 2007.

[5]. And third, digital map providers such as Tele Atlas, as well as Google and Microsoft, already possess and continue to record large image databases of urban areas and other environments. All these reasons, individually and taken together motivate us to explore and analyze vision-based technology for positioning, alone and in combination with other outdoor positioning methods.

This paper aims to address the following questions related to vision-based positioning: what is a realistic level of accuracy which can be obtained with current vision technology using different cameras including today's cell-phone cameras; how can snapshot based positioning be augmented to provide continuous location estimates; how does vision-based positioning compare to other technologies such as GPS, WiFi and Dead Reckoning and how may these technologies complement each other. In order to better understand the applicability of this technology, we will also discuss requirements and constrains of today's vision technology.

In this context we develop, describe and analyze two basic modes of vision-based positioning. The *first mode* uses single snapshots (called query images in the following) taken e.g. with the camera of a standard cell-phone, while the *second mode* employs additionally pedestrian dead reckoning. The first mode of operation positions the single query image w.r.t. a database of geo-referenced images and therefore allows to estimate the user's location as well as the user's orientation in world coordinates. As can be seen from the experiments presented in this paper the achievable median accuracies are comparable to or even better than standard GPS-based positioning. This first mode can be used to position oneself at the beginning of a trip e.g. after exiting a building or a subway station and to obtain directions and local map-information for further navigation. Due to the high levels of accuracies one could also use a single snapshot to correct a position-estimate obtained by another, less accurate technology such as GSM (e.g. close to the destination of a trip). This mode can be used also to support augmented reality applications where additional information is provided e.g. for a tourist site or a shopping mall. The second mode of vision-based positioning enables continuous positioning of a walking person by combining single snapshots for accurate positioning and an inertial sensor to enable incremental update of the current position estimate between single snapshots. To allow continuous position estimates we deploy a dead reckoning (DR) device. As the incremental update will obviously degrade over time we use a Kalman filter formulation to integrate multiple position estimates based on different snapshots and the DR estimate. The applicability of this mode of vision-based positioning is similar to that of other continuous positioning technologies such as GPS and WiFi. Therefore, the accuracies obtained by this mode of operation will be directly compared to these technologies.

The contributions of this paper are the following. First, we analyze the state-of-the-art in computer vision for snapshot based positioning using different types of cameras under realistic conditions. Second, we improve upon the existing vision-based position systems (e.g., [5]) by employing a novel high-dimensional feature matching method of [6]. Third, to enable continuous positioning we complement vision-based positioning with an incremental movement estimate

obtained from DR. In particular we will discuss the relation of average distance between snapshots and accuracy of position estimate. Fourth, we record realistic data of pedestrians tracks in the city of Ljubljana including GPS, WiFi and DR sensor data as well as snapshots. Fifth, we analyze and compare the different technologies individually and in combination in terms of accuracy and in terms of requirements.

In the following we present related work (Sec. 2), then we give an overview and evaluation of the vision based positioning sytem (Sec. 3). Secondly (Sec. 4) we employ this system in combination with pedestrian dead reckoning to enable continuous positioning. Thirdly (Sec. 5) we evaluate the positioning by meaningful combinations of GPS, WiFi and DR in our scenario for a comparison to vision based position methods, and close with a summary and outlook (Sec. 6).

2 Related Work

Vision-based positioning recently attracted a lot of attention, especially because of the proliferation of digital cameras and camera phones. The authors of [7] and [8] were among the first to present a camera phone based prototype for personal navigation in urban environments. They first recognize the building in the query image and then estimate the geometric relation between the facade in the query image and the images or models in the database. However, they rely on the presence of dominant planes and estimation of vertical vanishing direction for alignment. In [9], the authors also use buildings as landmarks for localization in a two-stage recognition process but rely only on geometric regularities of man-made structures, such as parallel and orthogonal lines. A different approach [10] aims to recognize user location from images captured with a camera-equipped mobile device by first performing image-based search over a database that indexes a small fraction of the Web utilizing hybrid color histograms. Then, relevant keywords on these web pages are automatically identified and submitted to a text-based search engine (e.g., Google). The resulting image set is filtered to retain images close to the original query image thus enabling approximate user localization if the matching images have the location attributed or the web pages include some location information (e.g., street address).

All methods described up to now provide only very approximate positioning that is inferior compared to the accuracy of the GPS. Recent advances in computer vision algorithms and hardware performance have however enabled approaches that can provide in near real time not only similar but in many circumstances even better positioning accuracy than provided by the GPS. For example, given a database of reference images of city streets tagged by geographical position, the system of [5] computes the geographical position of a novel query image within a few meters accuracy. In their approach local visual features are used first to select the closest reference images in the database and then to estimate the camera motions between the query image and the closest reference images. The estimated camera motions are then used to position the query image by means of triangulation.

Among the systems enabling the continuous positioning, the positioning systems using WiFi access points either use "fingerprinting" [11,12] or simply the estimated positions of receivable access points from public or company owned databases [13,14,15]. While fingerprinting can obtain high levels of accuracies (close to 1m) it also requires a tedious calibration effort which seems unfeasible for large areas with dynamic changes in the access point infrastructure (often the case in outdoor scenarios). Other approaches like PlaceLab have been designed with a low barrier-of-entry and a wide availability in mind which is why we build on this framework in the paper. GSM-based positioning is not as well developed today even though network-providers and researchers are working towards better levels of accuracies [4].

In pedestrian navigation by dead reckoning (DR), the authors of [16] were one of the first to use inertial sensors, compass and fusion with GPS, which was the basis for the Pointresearch DRM Module, commercially available in the 5th generation [17]. There is also an extensive body of academic research on the subject. The work of [18] was among the first to introduce the use of additional gyroscope sensors and leading to commercially available devices [19], furthermore a range of other approaches exist [20,21,22,23,24]. Dead reckoning enables a relative or incremental position estimation due to the user's orientation and walking speed, typically using the dependency of step frequency and step length [25]. To realize an absolute position estimate this information is in most approaches fused with GPS data only.

3 Vision-Based Positioning

In the following we first give an overview of our approach to vision-based positioning. Then we present the employed algorithms in detail. Finally, we describe the setup of experiments and present the results of evaluation of our approach in a real world scenario.

Method overview. Fig. 1 illustrates our approach to vision-based positioning. In our approach, similarly as in [5], the user takes a query image (blue frame) at his current position and this image is matched to a database of reference images (green frames) by matching local visual features. Matched local visual features are further used to estimate geometric relations between the query image and the matching reference images. The estimated geometric relations are then employed to position and orient the query image (and consequently the user) w.r.t. the reference images. An important requirement for such a vision-based positioning system is that the reference images must adequately cover the area where vision-based positioning shall be performed. It is important that the query image contains sufficient overlap with at least one, but preferably more reference images. The number of reference images thus required depends on the type of the scene and can be substantial (i.e., several thousand reference images per square kilometer). While this is a significant number of images, companies such as Tele Atlas, Google and Microsoft work on such databases already today for many cities around the world. An additional requirement regarding the

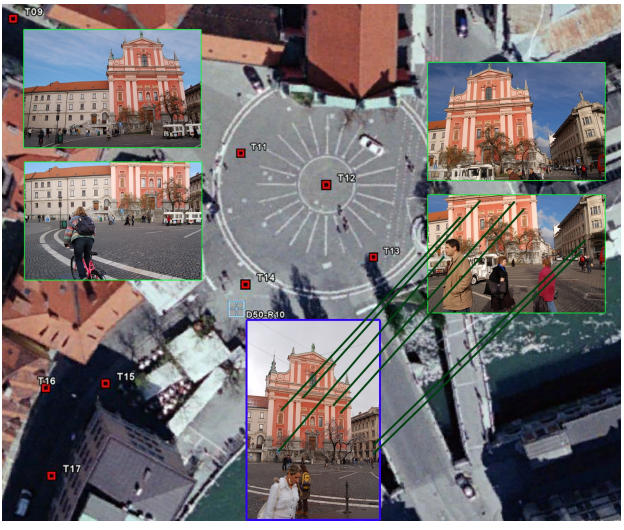

Fig. 1. A query image D50-R10 (thick, blue frame) and some reference images (thin, green frames) used to position and orient the query image and consequently the user. The two reference images on the left were shot from the standpoint T14, while the two reference images on the right were shot from the standpoint T13. Some geometric relations relating the query image with one of the reference images are indicated by the dark green lines.

reference images is that the geographic positions from where the reference images were taken must be measured accurately, e.g. by GPS or by more traditional surveying techniques (e.g. theodolite), while the absolute camera orientations of the reference images can be estimated in the preprocessing stage by estimating geometric relations between reference images.

Matching image regions. The local visual features are detected independently in each image using some subregion saliency measure (e.g., specific area around corners, area of similar color, area bounded by curves) that is invariant to common image transformations such as viewpoint or photometric changes. The image subregions selected as the local visual features are then described in a compact form (typically as a feature vector) retaining only information on image structures that is independent of common image transformations. Based on recent performance evaluations [26] and [27] we have chosen Maximally Stable Extremal Regions (MSER) as local feature detector [28] and Scale Invariant Feature Transform (SIFT) as feature descriptor [29]. The SIFT descriptor describes the image subregion by a 128-dimensional vector representing a histogram of gradient orientations.

Matching images. The first step of the presented vision-based positioning system is to identify which reference images depict the same scene as the query image. For this we use a novel image matching method [6] based on the concept

of meaningful nearest neighbors. A local visual feature detected in a reference image is considered a meaningful nearest neighbor if it is sufficiently close to a query feature such that it is an outlier to a background feature distribution of irrelevant match candidates. The background feature distribution is estimated from the extended neighborhood of a query feature given by its k nearest neighbors. When applied to image matching, meaningful nearest neighbors are independently weighted for each query feature. The sum of weights then defines the similarity of a reference image to the query image.

To make search for k nearest neighbors more efficient, we employ a novel approximate nearest neighbors search method of [6] that is based on sparse coding with an overcomplete basis set and provides a ten-fold speed-up over an exhaustive search even for high dimensional spaces while retaining excellent approximation to an exact nearest neighbors search.

Estimation of geometric relations. The second step of the presented vision-based positioning system is to estimate geometric relations between the query image and the matching reference images. The type of geometric relation that we employ is called epipolar geometry [30]. Epipolar geometry is independent of scene structure and only depends on the cameras' internal parameters and relative pose. If the cameras' internal parameters are known, the relative pose of the two images can be extracted from the known epipolar geometry. The most important internal parameter of the camera is the focal length that we acquire from the EXIF header of the images while we assume standard values for other internal parameters of the camera [30]. If the internal parameters of the camera are known and if at least five correspondences are established by matching image regions of the two views, the epipolar geometry relating two views can be estimated using the *five-point algorithm* [31]. Existing matching algorithms cannot guarantee that all correspondences (matching image regions) are true correspondences, i.e., that they are projections of the same structure in 3D world [9], so we resort to a hypothesize-and-test approach in order to find a subset of true correspondences among all the matching image regions. In the hypothesize-and-test approach we first construct a hypothesis by calculating the epipolar geometry from a selected five-tuple of the correspondences and then we test how many tentative correspondences are consistent with the hypothesized epipolar geometry. The hypothesized epipolar geometry consistent with most of the tentative correspondences is selected as the representative one. It is infeasible to consider all possible five-tuples of correspondences so we construct hypotheses only from a few dozen tentative correspondences with the highest weight as provided by the feature matching algorithm [6].

Positioning the query image. In the final step of our approach we use the estimated geometric relations to position and orient the query image w.r.t. the reference images. Our positioning method chooses among a set of predefined positioning strategies based on the number and the type of estimated geometric

relations. The *first* strategy is used when the query image and one of the reference images are shot from locations that are very close together compared to the depicted scene. In such a case the query image is positioned at the location of the reference image, while the orientation of the query image is calculated by estimating rotation relating the query and the reference image. The *second* strategy uses triangulation to position the query image. This strategy is chosen either when we know the geometric relations of the query image with three or more reference images not shot at the same standpoint, or if we know the geometric relations of the query image with two reference images and we also know the camera orientation of at least one reference image. The *third* strategy is selected when only two geometric relations of the query and reference images are known and we do not know the camera orientations of the reference images. This strategy also uses triangulation to position the query image, but with an additional assumption that the query image was shot at the same height from the ground as the two reference images. The *fourth* strategy is chosen when only one geometric relation is known. In such a case the query image is positioned at the location of the respective reference image. If no geometric relation can be estimated, then we use the *fifth* strategy that positions the query image at the location of the most similar reference image, but only if the reference images is similar enough.

3.1 Setup and Experiments

In our experiments we used Ljubljana urban image data set[1] that consists of 612 reference images of an urban environment covering an area of 200×200 square meters (see Fig. 2). At each of the 34 standpoints 18 images were captured at 6 orientations and 3 tilt angles covering a field of view of $360° \times 100°$ (see Fig. 3). All images were taken with a Nikon D50 digital SLR camera with a field of view of $88° \times 52°$ at a resolution of 6 megapixel in November 2005 in overcast weather. The positions of the reference standpoints were measured by classical surveying techniques using a theodolite for relative measurements (with an accuracy of 10cm) and a high accurate GPS device for absolute positioning (accuracy 50cm).

The 17 query images were shot from 13 distinct positions where a user might want to position himself. The query images were captured in January 2007 in light snowfall weather. At every query location three images were shot using Nikon D50, Sony DSC-W1, and Sony Ericsson (SE) W800 cameras. Nikon D50 is a high-end digital SLR camera with 6 megapixels, Sony DSC-W1 is a standard consumer grade camera with 5 megapixels and SE W800 is a mobile phone equipped with a 2 megapixels camera. The ground truth positions were calculated by measuring the distance from each query point to two geo-referenced standpoints of the Ljubljana urban image data set. Nine query images taken by the three cameras at three different standpoints together with the most similar reference images are presented in Fig. 4.

[1] Available online at http://vicos.fri.uni-lj.si/LUIS34/

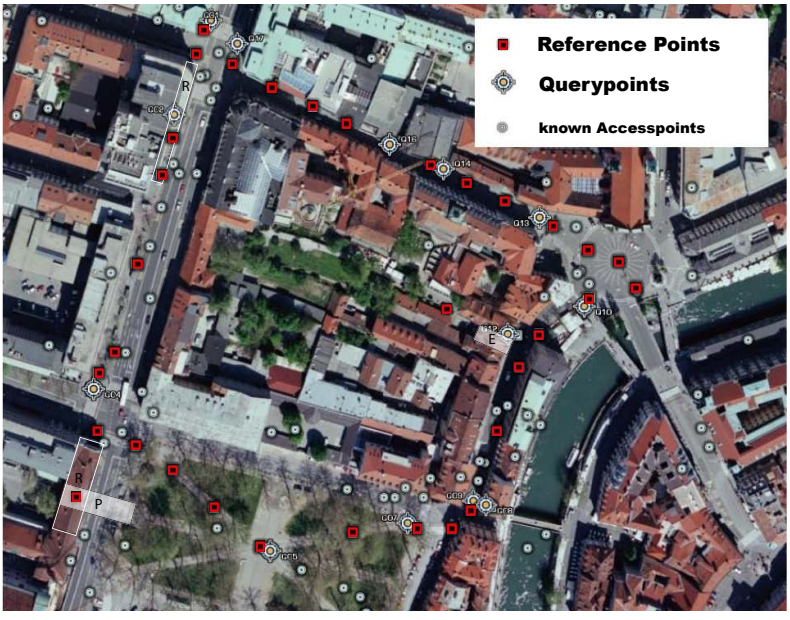

Fig. 2. Overview of the Test area. In the image the reference points, snapshot points and known access points from the wigle.net database are shown. (P) marks a pedestrian underpass, (E) a passage and (R) roofed areas.

3.2 Results

The results of the vision-based positioning system are given in Table 1. The position of the query images could be established in 14 (13 for SE W800) out of 17 cases with a median positioning error of 4.9m for Nikon D50 camera, 6.6m for Sony DSC-W1 camera and 6.2m for SE W800 camera phone. The method fails altogether if vegetation or dynamic objects compromise most of the user image.

The results of the Nikon D50 camera are discussed in more detail and these insights also hold for the other two cameras. For seven query images the first positioning strategy was selected (Q02, Q08, Q10, Q11, Q12, Q15 and Q16) so their position was set to the corresponding reference point. In the case of query image Q02 the algorithm incorrectly selected this positioning strategy, therefore the localization error is large (31.8m). For three query images (Q01, Q13 and Q14) the second positioning strategy was selected so that their positions were estimated by triangulation (an example is presented in Fig. 5). In four cases (Q04, Q07, Q09 and Q17) only one related image was found, so that their positions were set to these reference positions by choosing the fourth strategy. As seen with Q17 large errors could be produced when triangulation is not possible. The remaining three images (Q03, Q05 and Q06) were not matched at all, since any approach based on local features has problems with images holding just vegetation or if large parts of the images are occluded (examples are shown in Fig. 6).

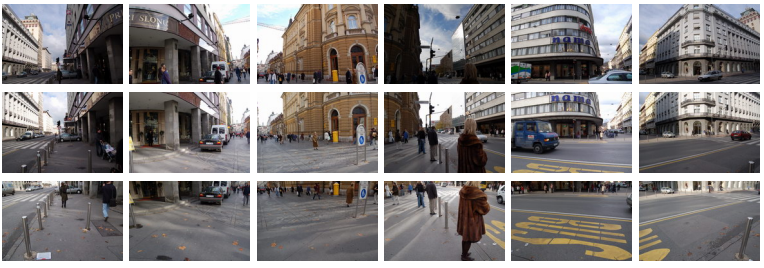

Fig. 3. The 18 images shot from standpoint T02 are shown. Using 6 orientations and 3 tilt angles a field of view of $360° \times 100°$ is covered.

Fig. 4. Three out of 17 query images for Nikon D50, Sony DSC-W1 and SE W800. In the first column are the most similar reference images.

In conclusion the vision-based positioning system works quite well, however more effort should be put into improving the five positioning strategies that we use to position and orient the query image w.r.t. the reference images. The major improvement of our vision-based positioning system over the system of [5] is that we employ a novel high-dimensional feature matching method of [6]. By using this feature matching method we are able to identify with much higher accuracy than the existing methods (e.g., [32]) which reference images depict the same scene as the query image. Additionally, the weighting of correspondences provided by the method of [6] increases the probability of correct estimation of geometric relations. It is also worth to mention, that the number of reference images to be searched could be cut down substantially if a coarse initial position would be provided by non-vision sensors such as GPS, WiFi or GSM. The speed-up thus achieved would be linear with the number of reference images to be related to the query image.

Table 1. Comparison of positioning accuracy for Nikon D50, Sony DSC-W1 and SE W800 cameras. All results are presented in meters.

Image	Q01	Q02	Q03	Q04	Q05	Q06	Q07	Q08	Q09
Nikon D50	8.7	31.8	NA	8.2	NA	NA	4.3	1.7	5.0
Sony DSC-W1	9.4	31.8	NA	8.2	NA	NA	4.3	1.7	5.0
SE W800	6.1	23.8	NA	8.2	NA	NA	4.3	1.7	5.0

mean error	**median error**	Q10	Q11	Q12	Q13	Q14	Q15	Q16	Q17
7.8	**4.9**	4.7	4.7	3.7	7.7	5.5	2.0	2.0	18.9
11.2	**6.6**	4.7	10.4	3.7	37.7	4.6	14.7	2.0	18.9
14.8	**6.2**	26.8	NA	3.7	50.2	6.2	36.1	2.0	18.9

D50-Q01 T01-3U T34-3U

Fig. 5. The query image Q01 (left) is matched to two reference images and therefore the position could be calculated by means of triangulation

Fig. 6. Query images that could not be positioned (bottom row) and the most similar, manually selected images among reference images (top row)

4 Continuous Positioning by Vision and Dead Reckoning

In the previous section we presented single snapshot based positioning. In this section we aim to combine the vision based positioning system with an incremental positioning system to allow continuous localization. This also allows the comparison to the evaluation of continuous positioning by different combinations of GPS, WiFi & DR (Sec. 5) in a real world scenario, which employs the same pedestrian dead reckoning device. In our evaluation we consider also the impact of the frequency of vision snapshots to the overall accuracy of the system.

For the fusion of vision-based snapshot positioning and relative motion estimation, whose accuracy degrade over time, we employ a Kalman filter. For the prediction of the state vector for any given timestep, the time and the estimated displacement since the last snaphot are employed, while the variance in the update by vision is considered as static value, derived from the average accuracy of the vision results.

4.1 Setup

The setup used for our experiments in this and the next section consists mainly of standard, of-the-shelf hardware. As main device for our recordings we use a tablet PC, which allows easy marking of the ground truth while walking for recording. For the experiments on multi-sensor positioning in the next Sec. 5 the computer is equipped with a suitable PCMCIA Wireless Network Card for scanning 802.11b/g networks (Prism 2 chipset) and connected via Bluetooth to a Holux GR236 GPS receiver (SiRFIII), placed on the shoulder of the user for best possible reception. For gathering relative position information we use a Dead Reckoning Compass (DRC) from Vectronix [19], which measures the displacement of the user with an electronic compass and step estimation by accelerometer and outputs an estimated position in geographic coordinates once a second over a wired connection. We decided to record the standard output of positions near the geographic point of origin in a stretch, not using the feature of setting the device to an initial position for adapted output to the current geographic height. As we're interested in the relative displacement per time interval in meters for the fusion, this allows for a more flexible evaluation. The DRC is placed on the recommended position on the back of the user, near the body's center of gravity. The so-called body height parameter, which influences the internal step length estimation, is calibrated to the user, and the magnetic offset adjusted to the local declination of the earth magnetic field. No individual optimization for the single traces or part of traces was done for a realistic setting. For the body offset angle a value of zero degree was used, as expected by the position of the sensor on the body. Good fixation of a DR sensor and a static body offset parameter can be expected to work better in realistic scenarios than current techniques for online correction [17]. For the recording we used an adopted version of the Placelab stumbler, which was extended to support the DR device and for creation of timestamped logfile entries by clicks on the map when passing ground truth points.

4.2 Experiments

For the evaluation we analyzed 7 different traces in an area of the inner city of Ljubljana. For ground truth we decided to use the 47 exactly measured points (34 reference & 13 query image points), which have a mean distance of about 20m (Fig. 2). Additional ground truth points were introduced for the pedestrian underpass (P) and behind the entry to an alleyway (E). Roofed areas with partially occluded view to the sky are marked with R. Each trace consists of a

different set of these points and we recorded timestamps when passing them. The length of the traces varies from 0.6km to 1.75km, with a total length of 7.4km. These traces are outdoors, with only short parts with obstructed view to the sky. To simulate normal pedestrian behavior, we walked constantly and did not stop on each ground truth point, except for a small portion including the query points for the visual positioning. The photo shooting on query points was simulated by stopping and turning, while the actual photos were taken separately with the 3 different cameras (Fig. 4) on the same day and commonly used for all traces.

4.3 Results

As motivated earlier this second mode of vision-based positioning incorporates the relative position estimate from DRC to allow continuous positioning like typically provided by GPS devices. For this we evaluate the same traces as for the multi-sensor positioning (Sec. 5) but employ different sub-sets of image-based query points in combination with dead reckoning. A Kalman filter is used to fuse the information from the vision-based positioning and the movement estimate from the DRC device. More specifically, whenever a trace reaches one of the included query points a correction is achieved by vision-based positioning and between these points the position estimate solely relies on the DRC-estimate. Note that the correction is only available in case of a successful positioning of the query image (cf. Table 1) For traces not starting on a query point an initial error of 10m is assumed. In Table 2 we show results for the median error over all traces for the different cameras. The highest overall accuracy (9.5m to 10.3m depending on camera type) is reached when using all available query points. Reducing the number of query points results in less accurate position estimates: using (on average) only every 2nd (12.4m to 12.9m), every 3rd (12.6m to 14.4m) or every 4th query point (14.1m to 14.9m). Fig. 7 shows a representative example. This reduction in accuracy however is quite low given that the average distance between the query points is increased from about 100m to 330m. Further increasing the average distance between query points to 269m and 455m mostly results in a larger variance of the achievable accuracies (11.2m to 18.6m). In general the results show, that Kalman filter based fusion of vision-based positioning and the dead reckoning device provides the user with an accurate position

Table 2. Results of vision-based positioning with dead reckoning for different query points available on the traces and the used cameras.

Available QPs	Nikon D50	Sony DSC-W1	SE W800	avg. QP distance
All	9.5m	10.2m	10.3m	94.5m
every 2nd (avg)	12.4m	12.9m	12.5m	172.7m
every 3rd (avg)	12.6m	13.1m	14.4m	279.9m
every 4th (avg)	14.1m	14.7m	14.9m	319.3m
Q04,Q07,Q13	11.9m	13.1m	14.7m	283.6m
Q04,Q09,Q13	11.2m	12.9m	14.0m	269.0m
Q04,Q13	14.9m	18.3m	18.6m	455.3m

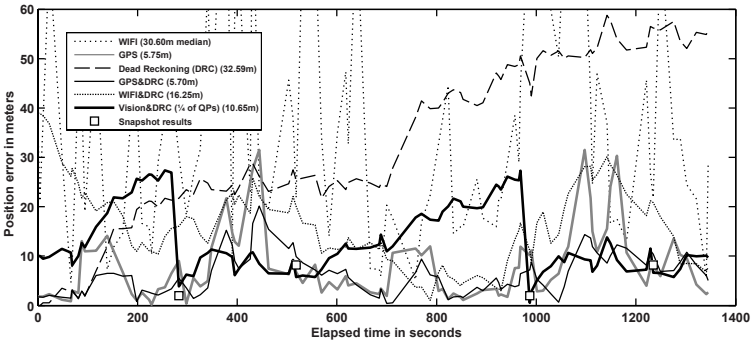

Fig. 7. Comparison of different sensors and combinations for Trace 6 (1743m). This is the longest trace and starts in the park (south) and uses reference points counterclockwise for 2 rounds (c.f. Fig. 2).

estimate even when using only few snapshots on the trace. As we will see in Sec. 5, the accuracies are often better than WiFi & DR-based positioning and in many cases comparable to or slightly worse than GPS. In Fig. 7 we show results for the longest trace with a set of every 4th querypoint available. Also included are the results from multi-sensor positioning (Sec. 5) for comparison.

5 Multi-sensor Positioning

For comparison to the vision-based positioning we evaluate the different non-vision-based positioning techniques, and combinations thereof in the same scenario. In our experiments we employ location information from GPS, WiFi and DR and combine them in different combinations to a single position estimate. We decided to build on the open source Placelab framework [13] for our experiments, as it offers a flexible basis with built-in WiFi positioning and fusion of location information of different sensors with a particle filter [33]. As an important aspect the framework aims less for precision than on coverage for the WiFi part. Instead of tedious reference measurements, which are unrealistic for a greater outdoor area, it makes use of public available maps of access point, to which users can easily contribute by wardriving/warwalking.

5.1 Setup and Experiments

As introduced in Sec. 4, we use for the recording mostly of-the-shelf hardware and an adopted version of the placelab stumbler, which was extended to support the dead reckoning device as additional sensor and to support marking of the ground truth points with timestamped log entries. As a database for access points maps we used the freely available and growing *wigle.net* database with currently over 9 million mapped access points.

For the evaluation of positioning by different combinations of GPS, WiFi and Dead Reckoning we extended the Placelab framework [13] to fuse the relative

Table 3. Comparison of positioning accuracy for Multi-Sensor based positioning. All results are presented in meters.

	Length	**Single Sensors**								
	(m)	**WiFi**			**GPS**			**Dead Reckoning**		
		mean	med	max	mean	med	max	mean	med	max
1	866.2	30.0	22.3	88.0	11.7	10.4	29.4	19.9	21.6	35.7
2	828.5	36.5	29.5	124.6	12.5	11.1	33.9	15.2	14.8	26.2
3	598.4	56.2	48.1	117.8	9.1	6.2	53.2	13.1	13.6	18.5
4	689.9	39.3	33.1	93.0	10.4	8.4	29.0	26.6	29.6	37.9
5	999.2	57.7	50.2	154.3	10.2	7.3	28.6	20.2	24.2	29.7
6	1743.8	36.3	30.6	122.9	7.8	5.7	31.5	32.6	26.9	58.8
7	1647.7	51.5	38.6	211.9	8.2	6.7	21.2	16.4	13.0	40.7
		43.9	**36.1**	**211.9**	**10.0**	**8.0**	**53.2**	**20.6**	**20.5**	**58.8**
	Length	**Particle Filter**								
	(m)	GPS&WiFi&DRC			GPS&DRC			WiFi&DRC		
		mean	med	max	mean	med	max	mean	med	max
1	866.2	10.1	10.1	27.4	10.0	8.9	25.9	19.0	13.1	56.3
2	828.5	11.2	9.5	28.2	11.2	9.8	27.7	18.6	18.6	27.8
3	598.4	7.3	4.6	23.5	7.1	4.8	23.5	23.6	23.8	39.1
4	689.9	10.3	7.6	31.3	10.1	7.2	32.2	12.6	7.8	47.1
5	999.2	7.4	6.2	21.1	7.6	6.5	23.1	9.9	9.1	20.1
6	1743.8	7.1	5.9	20.8	6.1	5.7	20.2	16.2	15.4	39.0
7	1647.7	5.8	4.9	15.0	5.6	5.0	14.7	16.0	15.1	30.2
		8.5	**7.0**	**31.3**	**8.2**	**6.9**	**32.2**	**16.6**	**14.7**	**56.3**

motion estimate of the dead reckoning device with the absolute position estimate of WiFi and GPS. The GPS and WiFi measurements are used in the sensor model of the particle filter to update the weight of the particles. From the dead reckoning device we extract the relative displacement per interval, and use this information in a specially built motion model for the particle filter, to take motion information of the user and the typical error of the dead reckoning device into account.

5.2 Results

In our evaluation we analyze the achievable accuracies of GPS, WiFi and DR in our scenario, alone and in meaningful combinations. Detailed results on the different traces are given in Table 3. GPS and WiFi-based positioning enable the user to determine absolute position, while DR can only determine relative position of the user.

GPS. As our traces are situated outdoors with a view to the sky, a valid GPS signal is available throughout the entire traces, with some short exceptions in a pedestrian underpass. Positioning by GPS alone performs well in this setting with an average median error of 8m over the 7 traces and median maximum error of 29.4m.

It has to be considered for these results, that the position of the GPS device on the shoulder of the user was a best case scenario, and obviously not common in daily use. Additionally, the used GPS device based on a SiRF III chipset, has

shown an explicitly higher sensitivity than devices based on earlier chipsets and semi-professional GPS device for surveying. Signal loss was encountered only shortly in traces that include a pedestrian underpass, but in this situation the measured error increased to a maximum of 53.2m.

WiFi. For the positioning by WiFi the test area had not the best premises, due to an uneven distribution of the access points or even areas with no active access points at all (e.g. shopping street in the north), as shown in Fig. 2. In the case-study we used the freely available beacon databases of the wigle.net community after recording and adding separately gathered warwalking data of approximately 2h in the test area and neighboring areas. Thus, for a realistic scenario, the employed beacon database is fully independent from the traces used in evaluation.

In our traces we reached an average reception of 3.57 known access points on the ground truth points, varying between 3.05 to 3.92. Positioning based on triangulation of the available beacons yielded an average median error of 36.1m over the 7 traces, varying from 22.3m to 50.2m. The maximum error encountered, which has to be taken into account when using only WiFi for priming of e.g. vision-based positioning, varies for the traces between 81m and 210m, with a median of 122m.

Dead Reckoning. The DRC device was fixed on the belt at the recommended position on the back of the user and adjusted to the user (see Sec. 4.1). No individual optimization for the single traces or part of traces was done for a realistic setting. For the evaluation of DR alone we determined the error which would occur in the traces of our scenario, given a perfect positioning in the beginning. For a more general accuracy estimation of the positioning by DR in our scenario we look how the accumulated error changes over distance. For the evaluation of DR alone the DRC output of absolute coordinates was transformed to the local coordinates of the test area, while for the fusion with other sensor data only the relative translation per interval was used in the motion model of the particle filter.

When assuming a perfect initial positioning and continuation by DR alone, we determined that for our traces the smallest maximum error was 18m on the shortest trace (598m), while the biggest error was 58m was on the longest trace(1743m). For all traces the maximum error encountered was less than 5.5% of trace length with a median of 3.2%. Trace 6 (Fig. 7) shows a typical error accumulation. In some of the other traces, e.g. 1 and 2, the accumulated error decreases strongly at some points of time by chance due to counteracting errors. While this decreases the overall error of DR for this approach to evaluation, it is no advantage for the fusion with other sensors, which need an accurate estimate of the users displacement over shorter time periods.

GPS & Dead Reckoning. The fusion of GPS and DR shows the best accuracy, further improving the accuracy of GPS. For the fusion of GPS and DR device data we employ an adopted particle filter, which uses the relative position information in the motion model. On average we reach median error of

6.9m in 7 traces, an improvement of 1.1m over positioning with GPS only, and all traces show a better accuracy than 10m in median. Also the maximum error of GPS is significantly reduced due to the position updates by DR. For example Trace 3 encountered a short phase of difficult reception in a roofed area and signal loss in a pedestrian underpass. In this case, using GPS alone results in a maximum error of 53.2m due to holding of the last known position, whereas the combination with DR shows only a maximum error of only 23.5m. In Fig. 7 the resulting error over time for Trace 6 is shown, including also GPS and DR error over time. In this trace the maximum error is decreased from 31.5m to 20.2m in a difficult reception situation in roofed area, while the median error is only slightly improved.

While already improving the accuracy and maximum error in an outdoor setting with only short phases of reception outage, the biggest advantage could be seen in scenarios mixed with longer indoor periods, e.g. shopping malls. The use of the DR additionally to GPS is able to reduce the average error to less than 50% from 46.5m to 21.5m for a mixed inner city/shopping mall example, and has reduced the maximum error from 138.3m to 39.0m in a different scenario not shown in this paper.

WiFi & Dead Reckoning. The combination of WiFi positioning and DR yields a significant improvement in our experiments. With a median error of 14.7m this combination is in the range of twice the error of GPS and combinations including GPS. Also the maximum error is heavily reduced compared to WiFi-positioning, with a maximum of 56.3m and median of 39.0m. This is even comparable to results of GPS positioning, and could represent a sound basis for the priming of vision-based technologies without the use of GPS.

GPS & WiFi & Dead Reckoning. The combination of all 3 sensors did not yield further improvements over the use of GPS and DR in our scenario. This can be explained by the high accuracy of the latter position methods, which results in slightly lower accuracy when using also the less accurate WiFi positioning in the particle filter. Due to the high weight of GPS and DR in the Particle Filter the error is only 0.1m higher and shows a similar characteristic to the GPS&DR, and could offer extended availability when the reception of GPS cannot be guaranteed for longer periods.

6 Conclusion and Outlook

This paper analyzed and extended the current state of the art in vision-based positioning for outdoor positioning and provided a comparison to other positioning technologies such as GPS and WiFi on a large and challenging image data set. Accuracies obtained for single snapshot based positioning are in the order of 5-6.5m depending on the employed camera. By extending single snapshot based positioning with a dead reckoning compass, continuous localization is possible with accuracies in the order of 10-15m which is only slightly worse than standard GPS and better than typical WiFi-based positioning. We also described various

novel combinations for multi-sensor-based positioning such as WiFi & Dead Reckoning which significantly improve accuracies w.r.t. WiFi alone.

Vision-based positioning has gained interest not only in the research community but also for commercial application. Even though the obtained accuracies are promising there are several issues to be solved before vision-based positioning will be a commodity service available on any camera-equipped hand-held device. Current devices have neither the processing nor the storage capabilities to perform vision-based positioning on the device, therefore the query images have to be sent for remote processing. Also the amount of images that need to be considered depends on the size of the uncertainty in the position before vision-based positioning. In that sense it is advantageous to possess a rough first position estimate to reduce processing time. Nevertheless due to the high commercial interest (e.g. from digital map providers) and with further improvement in fast image database indexing, vision-based positioning may soon become a reality.

Acknowledgements. This research has been supported in part by: Research program Computer Vision P2-0214 (RS), EU FP6-004250-IP project CoSy, EU MRTN-CT-2004-005439 project VISIONTRAIN, and EU FP6-511051 project MOBVIS.

References

1. Hightower, J., Borriello, G.: Location Systems for Ubiquitous Computing. Computer 34(8), 57–66 (2001)
2. Hazas, M., Scott, J., Krumm, J.: Location-Aware Computing Comes of Age. Computer 37(2), 95–97 (2004)
3. Cheng, Y.C., Chawathe, Y., LaMarca, A., Krumm, J.: Accuracy Characterization for Metropolitan-scale Wi-Fi Localization. In: MobiSys (2005)
4. Varshavsky, A., Chen, M.Y., de Lara, E., Froehlich, J., Haehnel, D., Hightower, J., LaMarca, A., Potter, F., Sohn, T., Tang, K., Smith, I.: Are GSM Phones THE Solution for Localization? In: WMCSA 2006 (2006)
5. Zhang, W., Košecká, J.: Image based localization in urban environments. In: International Symposium on 3D Data Processing, Visualization and Transmission, pp. 33–40 (2006)
6. Omerčević, D., Drbohlav, O., Leonardis, A.: High-Dimensional Feature Matching: Employing the Concept of Meaningful Nearest Neighbors. In: ICCV (2007)
7. Johansson, B., Cipolla, R.: A system for automatic pose-estimation from a single image in a city scene. In: Proc. of International Conference on Signal Processing, Pattern Recognition, and Applications (2002)
8. Robertson, D., Cipolla, R.: An image-based system for urban navigation. In: BMVC, pp. 819–828 (2004)
9. Zhang, W., Košecká, J.: Hierarchical building recognition. Image and Vision Computing 25(5), 704–716 (2007)
10. Yeh, T., Tollmar, K., Darrell, T.: Searching the web with mobile images for location recognition. In: CVPR, vol. 2, pp. 76–81 (2004)
11. Bahl, P., Padmanabhan, V.N.: RADAR: An In-Building RF-based User Location and Tracking System. In: IEEE Infocom 2000, IEEE Computer Society Press, Los Alamitos (2000)

12. Ekahau. Online http://www.ekahau.com
13. LaMarca, A., Chawathe, Y., Consolvo, S., Hightower, J., Smith, I., Scott, J., Sohn, T., Howard, J., Hughes, J., Potter, F., Tabert, J., Powledge, P., Borriello, G., Schilit, B.: Place Lab: Device Positioning Using Radio Beacons in the Wild. In: Gellersen, H.-W., Want, R., Schmidt, A. (eds.) PERVASIVE 2005. LNCS, vol. 3468, Springer, Heidelberg (2005)
14. Skyhook Wireless. Online http://www.skyhookwireless.com/
15. Navizon: Peer-to-Peer Wireless Positioning. Online http://www.navizon.com/
16. Judd, T.: A Personal Dead Reckoning Module. In: ION GPS 1997 (1997)
17. Macheiner, K.: Performance Analysis of a Commercial Multi-Sensor Pedestrian Navigation System. Master's thesis, IGMS, Graz University of Technology (September 2004)
18. Ladetto, Q., Merminod, B.: Digital Magnetic Compass and Gyroscope Integration for Pedestrian Navigation. In: 9th Saint Petersburg International Conference on Integrated Navigation Systems (2002)
19. Vectronix. Online http://www.vectronix.ch/
20. Randell, C., Djiallis, C., Muller, H.L.: Personal Position Measurement Using Dead Reckoning. In: Fensel, D., Sycara, K.P., Mylopoulos, J. (eds.) ISWC 2003. LNCS, vol. 2870, Springer, Heidelberg (2003)
21. Gabaglio, V.: GPS/INS Integration for Pedestrian Navigation. Astronomisch-geodätische Arbeiten in der Schweiz, vol. 64 (2003) ISBN 3-908440-07-6
22. Jirawimut, R., Ptasinski, P., Garaj, V., Cecelja, F., Balachandran, W.: A Method for Dead Reckoning Parameter Correction in Pedestrian Navigation system. Instrumentation and Measurement 52(1), 209–215 (2003)
23. Kourogi, M., Kurata, T.: Personal Positioning based on Walking Locomotion Analysis with Self-Contained Sensors and a Wearable Camera. In: ISMAR 2003 (2003)
24. Beauregard, S., Haas, H.: Pedestrian Dead Reckoning: A Basis for Personal Positioning. In: WPNC 2006 (2006)
25. Bertram, J.E.A., Ruina, A.: Multiple Walking Speed-Frequency Relations are Predicted by Constrained Optimization. Journal of Theoretical Biology 209(4) (2001)
26. Mikolajczyk, K., Tuytelaars, T., Schmid, C., Zisserman, A., Matas, J., Schaffalitzky, F., Kadir, T., Van Gool, L.: A Comparison of Affine Region Detectors. IJCV 65(1-2), 43–72 (2005)
27. Mikolajczyk, K., Schmid, C.: A Performance Evaluation of Local Descriptors. PAMI 27(10), 1615–1630 (2005)
28. Matas, J., Chum, O., Urban, M., Pajdla, T.: Robust wide-baseline stereo from maximally stable extremal regions. Image and Vision Computing 22(10), 761–767 (2004)
29. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. IJCV 60(2), 91–110 (2004)
30. Hartley, R., Zisserman, A.: Multiple View Geometry in Computer Vision, 2nd edn. Cambridge University Press, Cambridge (2004)
31. Nister, D.: An efficient solution to the five-point relative pose problem. IEEE PAMI 26(6), 756–777 (2004)
32. Tuytelaars, T., Van Gool, L.: Content-based image retrieval based on local affinely invariant regions. In: Huijsmans, D.P., Smeulders, A.W.M. (eds.) VISUAL 1999. LNCS, vol. 1614, pp. 493–500. Springer, Heidelberg (1999)
33. Hightower, J., Borriello, G.: Particle Filters for Location Estimation in Ubiquitous Computing: A Case Study. In: Davies, N., Mynatt, E.D., Siio, I. (eds.) UbiComp 2004. LNCS, vol. 3205, Springer, Heidelberg (2004)