Abstract—The navigation and mobile positioning paradigm has been an active research topic in the past two decades. The proliferation of mobile devices equipped with inertial sensors able to detect phone motion lead to an increase in research and product development in the field of personal locating technologies. Within this scenario, this paper proposes a combined technique for the estimation of user’s position in indoor environments, which only relies on sensors of a commercial smartphone, without the active aid of external infrastructures. The smartphone’s video camera is used to identify known keypoints, named anchors previously identified and geo-referenced in the building map. A SURF algorithm is employed for quick and effective detection of image features independently from the viewpoint. The accelerometers and the compass of the smartphone are used to estimate the amount of steps and the direction in which the user is moving by performing statistical data processing. Heading data is compensated taking into account the relative position of the smartphone with respect to the user. Experiments have been carried out in a real scenario showing the effectiveness of the proposed solutions.

I. INTRODUCTION

The problem of mobile positioning and navigation in dynamic and unknown indoor environments has been under study over the last two decades. The importance of knowing accurately and reliably location information of a mobile phone inside a venue has drastically increased due to the wide number of emerging location-aware technologies and solutions, context-aware applications, indoor emergency response scenarios (e.g., urban fire or earthquakes), augmented reality applications, mobile guides for big buildings like hospital, university campus, big small and so on.

A classification of indoor navigation approaches relies on whether additional external infrastructures are needed or not. According to the first category, different technological solutions have been proposed, such as: WiFi trilateration, RFID tags/readers, AGPS, Bluetooth transmitters/receivers [1]. The mentioned solutions can provide a good level of accuracy but are affected by high costs (a pre-installed sensing infrastructure is assumed to be in place), limiting their impact. To the other category belong inertial systems, which represent cheaper and simpler solutions but often less accurate thus requiring periodical recalibration to limit location estimation errors.

In this paper we focus on the second category to limit the costs for installing and maintaining the external infrastructure. Specifically, we consider the adoption of an Inertial Navigation System (INS) that is a self-contained navigation system using inertial detectors. Indeed, the location technique is based on the measurement provided by all the sensing and computing technologies available in a common smartphone: accelerometers and gyroscopes sensors, internet connection and high-resolution cameras.

Interests in such a positioning technology increase because many commercial mobile phones are equipped with these motion sensors thus are able to provide key information like: position, orientation and velocity of a moving object/user through direct measurements, also in indoor environment.

The typical scenario is the following: a user wishes to move from place A to place B in an unknown indoor environment without getting lost. He initially takes a photo of a geo-referenced 2D-bar code to acquire the map of the building and set his initial position, then starts walking. As described in our previous implementation [12], the only requirements are a modern smartphone equipped with accelerometer, digital compass, built-in camera and WiFi connection. The data read from phone sensors, combined with a map of the building and a known starting point, allows for user movement tracking.

The estimation process is based on dead reckoning thus is affected by cumulative error. To overcome this limitation the system needs a periodical recalibration achievable by using geo-referenced bi-dimensional tags hanged on the walls of the building in combination with plane homographic technique to derive additional information about the relative orientation and distance of the user to the reference point (2D-bar code).

The position estimation process is crucial in indoor application and the error can vary a lot depending on how the localization problem is faced. Systems based only on inertial sensors are affected by bigger error than infrastructure based approach, thus the calibration process has to be as precise as possible. To further limit the positioning error in the previous system [12], we propose an improved combined approach, for reducing the positioning error during motion and location fixes.

During motion, the data from the accelerometers and compass is filtered over a certain amount of time in order to compensate potential errors regarding the actual movement of the user and the motion direction. The accumulating errors...
are also a consequence of changes in the phone’s relative position to the user’s body, therefore a position model of the smartphone has been developed for applying the appropriate corrections to both direction and displacement.

For a periodic position fix, an image-based localization system making use of the built-in camera is employed. By developing local feature detection, description and matching between a query image, acquired by the user, and a database containing a collection of geo-referenced images related to the chosen environment, the user’s position can be accurately fixed.

The proposed solution is based on the SURF (Speed-up robust features), which allows for a quick and effective detection of image features without being affected by the user’s viewpoint. The INS system can be recalibrated (position fix) by taking photo of anchor points (nodes with a known position) present in the indoor environment.

The paper is organized as follows. The following section describes briefly the state of the art related to indoor navigation system. The background for our work is then provided in section III. The fourth section gives an overview of the whole system and describes the proposed solution. Experimental results are presented in the fifth section. Finally, in the last section, we draw conclusions and present the planned future work.

II. RELATED WORK

For smartphone-only implementations, the dead-reckoning method based on step counters and heading, integrating the data over time to estimate the current user position is presented in [2]. Adaptive Kalman filters and activity based map matching are used to improve the position estimate.

An evolution of this basic approach is presented in [3], using a dual approach: simple step detection and step heading estimation combined with matching detected steps onto the expected route from the source to the destination using sequence alignment algorithms. Instead of a more general localization problem, the localization problem is solved on a specified route, allowing compensating for inaccuracies and offering accurate turn-by-turn directions.

The impact of the phone’s position with respect to the user is approached in [4] by employing two complementary methods. The first method is to employ an accelerometer and magnetometer. The accelerometer gives a reference direction for gravity, and the component of the magnetic field perpendicular to gravity gives a reference for magnetic north. The benefit of this method is that each measurement is perpendicular to gravity gives a reference for magnetic north. As a result, the INS system can be recalibrated (position fix) by taking a photo of anchor points (nodes with a known position) present in the indoor environment.

The Magnetometer is an approach robust to the changes of illumination, distinctiveness, and robustness, yet can be computed and compared much faster. This is achieved by relying on integral images for image convolutions, by building on the strengths of the leading existing detectors and descriptors and by simplifying these methods to the essential.

The system in [7] can provide a user’s view direction with its location by comparing a query image captured by the user and panoramas in a database. The mentioned solution uses a camera equipped with GPS to get a rough location and minimize the search area, thus works only in outdoor environment. Each panorama in the database is obtained capturing 18 images at 20 degree intervals in order to avoid distortion and finally merge them. The left side and right side of the generated image is 0 degrees, which means north. As a pre-processing, SURF features’ database of each panorama is generated. Then, during the online phase, SURF features of a query image is computed and matched with SURF features of each panorama individually in order to retrieve the user’s location. For providing a user’s view direction the homographic matrix between the query and the selected panorama is calculated.

A low-cost method of real time positioning using the phone’s camera phone is presented in [8],[9]. User location is determined by detecting unobtrusive known fiduciary markers around a building. Indoor navigation is allowed by the continuous scanning of environment in real time (15 Hz or more) in search of strategically placed fiduciary markers. The system is only based on paper markers (square markers or frame markers) and static digital maps. No additional infrastructure is needed.

A fast image matching algorithm in which visual attention system is used to guide local interest point detection represented with SURF descriptor is described in [10]. One-to-one symmetric search is performed on descriptors to select a set of matched interest point pairs.

Outlying false matches are identified and filtered out while remaining pairs are weighted by their saliency. Weights are summed up yielding a similarity score. Images are considered to be near-duplicates if similarity score exceeds a certain threshold.

An approach to locate a walking pedestrian in urban area by a camera image of first-person vision is proposed in [11]. Reference points on a path have been registered with other first-person vision images taken by someone in advance at different time and day. First a conventional bag of features approach with SURF is applied in order to find out the top candidate image among reference images. Then, the authors define seven matching criteria that derive from the property of first-person vision to verify the SURF key pairs so as to tell if the selected is acceptable and to reject false matching.

III. BACKGROUND THEORY

A. Acceleration sensors

The integrated accelerometers in modern smartphones are tri-axial devices which allow the detection of accelerations forces in $m/s^2$ along the X, Y and Z axes. The values of the acceleration are positive or negative based on the direction in
which the phone is moving and based on the position of the phone. When the device is lying still in a flat position, the accelerations on the three axes are:

\[
\begin{bmatrix}
a_x \\
a_y \\
a_z \\
\end{bmatrix} = \begin{bmatrix}
0 \\
0 \\
-g \\
\end{bmatrix}
\]  

(1)

where \( g \) is the earth’s gravitation. In fact, using the presence of gravity distributed on the three axes, the orientation of the device can be calculated using the modulus of the accelerations \( |a_m| \) given by

\[
|a_m| = \sqrt{a_x^2 + a_y^2 + a_z^2}
\]

(2)

and the following equations:

\[
\alpha = \cos^{-1}\left(\frac{a_x}{a_m}\right), \quad \beta = \cos^{-1}\left(\frac{a_y}{a_m}\right).
\]

(3)

The angles \( \alpha \) and \( \beta \) are the angles which the device is forming with the X and Y axes. The scaled values of \( \alpha \) and \( \beta \) give the angles for roll and pitch, as illustrated in figure 1. For the smartphone lying still along the Y axis on a flat horizontal surface, the pitch and roll are equal to zero, changing when moving from 0° to 360°.

Fig. 1. Axes and rotation angles of a smartphone

B. Magnetic Sensor

The azimuth \( \gamma \) from figure 1 represents the angle between the magnetic north and the Y axis of the smartphone with the display heading up. Just like a digital compass, the values for the azimuth are between 0° and 359°, with 0° for the magnetic north, 90° for east and so on. The azimuth value returned by the magnetic sensor of a smartphone is highly susceptible to electromagnetic interference and is also quite unstable for typical devices, emerging the need for a periodic recalibration, which can be performed by rotating the smartphone in a 8-like pattern [1].

C. SURF

SURF [5]-[6] is an efficient scale and rotation invariant interest point detector and descriptor. It allows for quick and effective feature detection even against different image transformations like image rotation, scale illumination and small viewpoint changes.

Much of the performance increase can be attributed to the use of an intermediate image representation, known as the Integral Image that can be rapidly computed from an input image [13]. This section shows a brief summary of its construction process.

1. Interest point detection

SURF is a Hessian matrix based interest point detector. It searches for blob-like structure at locations where the determinant of this matrix is maximal. Given a point \( X = (x, y) \) in an image \( I(x,y) \), the Hessian matrix \( H = (X, \sigma) \), as function of both space \( X \) and scale \( \sigma \), is defined as follows:

\[
H(X, \sigma) = \begin{bmatrix}
L_{xx}(X, \sigma) & L_{xy}(X, \sigma) \\
L_{xy}(X, \sigma) & L_{yy}(X, \sigma)
\end{bmatrix}
\]

(4)

where \( L_{xx}(X, \sigma) \) refers to the convolution of the second order Gaussian derivative \( \frac{\partial^2 g(\sigma)}{\partial x^2} \) with the image at point \( X = (x, y) \) and similarly for \( L_{yy}(X, \sigma) \) and \( L_{xy}(X, \sigma) \). These derivatives are known as Laplacian of Gaussians. The approximated determinant of the Hessian represents the blob responses at location \( X = (x, y) \) in the image. In order to detect interest points over different scale a non maxima suppression in a 3 x 3 x 3 neighborhood is applied. To do this each pixel in the scale-space is compared to its 26 neighbors, comprised of the 8 points in the native scale and the 9 in each of the scales above and below. Finally the maxima of the determinant of the Hessian matrix are then interpolated in both space and scale to sub-pixel accuracy.

2. Interest point description

The SURF descriptor describes the distribution of pixel intensities within a scale dependent neighborhood of each interest point detected by the Fast-Hessian. Integral images in conjunction with Haar wavelets are used in order to increase robustness and decrease computation time. Haar wavelets are used to find gradients in the x and y directions. The first step in descriptor’s extraction consists of fixing a reproducible orientation based on information from a circular region around the interest point. Then, a scale dependent window aligned to the selected orientation is constructed and a 64-dimensional vector (SURF descriptor) is extracted from it. The dominant orientation is estimated by calculating the sum of all responses within a circle segment covering an angle of \( \pi / 3 \) around the origin. At each position, the two summed x and y responses are used to form a new vector.

The longest vector defines the orientation of the interest point. The first step for the extraction of the descriptor is to construct a square region aligned with the selected orientation around the interest point. It contains the pixels which will form entries in the descriptor vector and is of size 20\( \sigma \), where \( \sigma \) refers to the detected scale. A further division into 4x4 regular sub regions is performed within each Haar wavelets
of size $2\sigma$, calculated for $5 \times 5$ regularly spaced sample points. Hence, each sub-region has a four dimensional descriptor vector, thus concatenating this for all $4 \times 4$ sub-regions a descriptor vector of length 64, invariant to different image transformation is obtained.

3. Descriptor Matching

The descriptor matching is performed by implementing the so called One to One algorithm [10]. Given two sets of descriptors $\{P\}$ and $\{Q\}$ extracted from a pair of images $(I_1, I_2)$, it returns pairs of closest descriptors using an Euclidean metric $\rho(P, Q)$.

IV. THE PROPOSED SYSTEM

A. INS architecture

Among the wide number of indoor navigation solutions, we propose a system capable to localize a user on the basis of the capabilities of a modern smartphone equipped with camera, digital compass, accelerometer and WiFi connection. The only external infrastructure is given by some 2-dimensional barcodes positioned in key points.

In a typical scenario a user needs to move from place A to place B in an unknown indoor environment. The initial position of the user is retrieved by scanning and decoding a geo-referenced datamatrix (2D barcode) placed aside the map of the floor with the embedded phone’s camera. The maps with the barcode are assumed to be hanged on the wall at the interest points. Based on the URL encoded in the datamatrix, the application downloads from a dedicated server the digital indoor vector map for the specific floor together with the initial position of the user on the map (corresponding to the point where the user stands when scanning the datamatrix). The user’s initial position is more precisely defined in terms of distance and orientation angle from the reference QR code using plane homographic techniques. When the user starts walking, the application draws step by step the position of the user, as a continuous line, over the downloaded map of the building floor. The application tracks the number of steps taken by the user based on the numerical values returned by the smartphone’s accelerometers [12].

The heading is retrieved by considering the output of the magnetometer. Taking in consideration that the magnetometer retrieves the magnetic north with respect to the phone’s current orientation which might be diverse form the walking direction of the user, the need for a compensation of the heading arises. This compensation is performed by analyzing the position of the phone with respect to the user starting from an initial known position.

On the basis of the corrected heading and the number of steps taken, the application deduces that user is near to some anchor points and suggests him to recalibrate the system in order to reduce the position error. Thus the user takes a photo of the closest anchor point, sends it to the server and waits for the response that will show the most probable position in the map building on the phone’s display.

B. Heading correction

While the step counter presented in [12] and based on the modulus of the output of the triaxial accelerometer produces satisfactory results, problems arise for accurately determining the heading of the user. These problems arise due to the fact that the output of the magnetic sensor is related to a smartphone in a flat position, heading in the same direction as the user. If the smartphone gets in a different position, for example used for talking on the phone or placed in a pocket, the change in position will be erroneously intended as a heading change.

To compensate the heading of the smartphone for position changes relative to the user, we developed a position classifier based on the interpretation of the pitch, roll and relative azimuth values as defined in the third section. For simplifying the classifier, we assumed only 90° rotations for each of the three axes, resulting in 8 possible positions for each axe (4 for the positive values and 4 for the negative values), for a total of 24 positions. The classification is based on the values of pitch and roll, with a tolerance of +/- 30°. An excerpt of the 24 positions is presented in table 1, for a smartphone held with the screen vertically in front of a standing user, vertically on the left and right side and laterally rotated.

<table>
<thead>
<tr>
<th>Pos.</th>
<th>$\alpha$</th>
<th>$\beta$</th>
<th>$\gamma$</th>
</tr>
</thead>
<tbody>
<tr>
<td>60°&lt; $\alpha$&lt; 120°</td>
<td>330°&lt; $\beta$&lt; 30°</td>
<td>$\gamma_{ref}$+330°&lt; $\gamma$&lt; $\gamma_{ref}$+30°</td>
<td></td>
</tr>
<tr>
<td>150°&lt; $\alpha$&lt; 210°</td>
<td>330°&lt; $\beta$&lt; 30°</td>
<td>$\gamma_{ref}$+120°&lt; $\gamma$&lt; $\gamma_{ref}$+60°</td>
<td></td>
</tr>
<tr>
<td>330°&lt; $\alpha$&lt; 30°</td>
<td>330°&lt; $\beta$&lt; 30°</td>
<td>$\gamma_{ref}$+120°&lt; $\gamma$&lt; $\gamma_{ref}$+60°</td>
<td></td>
</tr>
<tr>
<td>150°&lt; $\alpha$&lt; 210°</td>
<td>60°&lt; $\beta$&lt; 120°</td>
<td>$\gamma_{ref}$+120°&lt; $\gamma$&lt; $\gamma_{ref}$+60°</td>
<td></td>
</tr>
</tbody>
</table>

The reference azimuth $\gamma_{ref}$ is recorded in the moment when the user is initially scanning the geo-referenced datamatrix and is presumed known, taking into consideration that the user has to stand in front of the data matrix to perform the scan.

Starting from this point, the heading of the user is calculated based on the actual azimuth values given by the smartphone magnetometer corrected by values corresponding to the calculated position of the smartphone.

C. Image-based calibration

The periodic position fix is addressed by developing a local feature detection, a description and a matching algorithm between a query image (acquired in real time by the user) and a database containing a collection of geo-localized images.
The entire process can be basically divided in offline and online phases. The offline phase, specific to each building, has to be executed only once (or when new anchor points need to be introduced in the indoor environment), resulting in the creation of a database. The data acquisition block can be seen as a sort of calibration: a certain amount of anchor points/locations will be chosen, depending on the size and layout of the building. At each of these locations, a subset of \( n \) photos from fix distance and different direction is taken in order to maximize the probability to have a true match. Once collected, every image is processed with the SURF algorithm to extract significant features, which are then coded in a descriptor vector. The created database represents a collection of anchor points at different locations in the building, taken under various illumination condition (light on/off) and from different viewpoints (frontal or lateral view). In particular we choose as anchor point internal/external door and gate, lighting system and air conditioning system.

During the online phase, as shown in figure 2, a user who wants to know his current position collects an image of his surrounding on the basis of the anchor points proposed by the system and sends the captured query image to the database for localization upon the map. To perform positioning, the algorithm investigates how similar the query image is to each image in the database by extracting and comparing its features with those of other images in the database. Finally, the application returns the image of the most probable location fixes the user position on the display.

V. EXPERIMENTAL RESULTS

A. System Implementation

The proposed algorithm was tested on an iPhone 3 GS with a 600 MHz ARM cortex Processor and a 3MP built-in camera. The feature extraction functionalities have been implemented making use of the OpenCV library. All tests were run on an Apple Mac Book Pro Intel Core 2 Duo machine with 2.4 GHz, 4 GB. The software implementations of both heading corrections and SURF algorithm were able to run in parallel with a CPU load that remained under 50%.

B. Heading Corrections

The tests for evaluating the feasibility of the position model and the corrections on the heading were performed using a reference circular path. This known path was run ten times holding the smartphone with its Y axis straight towards the walking direction. For a number of 16 points on this path, the azimuth was calculated as the mean azimuth from the 10 runs.

Subsequently, the known path was covered in a similar manner but moving the smartphone with respect to the user’s body in a series of 5 previously known typical positions: front chest pocket, side pocket, rear trouser pocket, left ear, right ear. The changes in position of the smartphone were performed at known instants of time, in order to be able to synchronize the values with the test runs. The known path was covered performing the same position changes for 10 consecutive times.

Table 2 shows the mean azimuth errors in degrees for the 16 known points on the path. The 5 grayed columns represent the points where the position changes took place. The azimuth values in these columns are the corrected ones based on the method presented in the fourth section.

As it can be noticed the azimuth errors for the points with no smartphone position change are less than 10°, while the errors for position changes are higher, but still in an acceptable range, not more than 18°.

<table>
<thead>
<tr>
<th>Point</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
</tr>
</thead>
<tbody>
<tr>
<td>Error (°)</td>
<td>5</td>
<td>10</td>
<td>18.2</td>
<td>7.5</td>
<td>20</td>
<td>6.2</td>
<td>5.5</td>
<td>13.4</td>
</tr>
</tbody>
</table>

Table 2: Azimuth Errors for a Known Path

C. SURF Algorithm Data Set

We performed the tests in a building of the Campus of the University of Cagliari, where 35 anchor points have been chosen. For each of these anchors we captured 3 photos, at a distance of around 3 meters and with 3 different viewpoints: 0° (frontal view), + 45° and – 45°. Thus we have a database made of a total of 105 images. The number of features could vary a lot among images, from several hundreds to few thousands, as shown in the following test images.

![Query Image](image1.png) ![Image 61](image2.png) ![Image 62](image3.png) ![Image 63](image4.png)

Fig. 3. Query image (left) and selected images from the data base
The first test was carried out on an image with 65 features belonging to the air conditioning system category as shown in figure 3.

The graph in figure 4 shows the number of features that match between the query image and each photo in the entire database.

The blue line shows the results when no threshold value in the One to One algorithm is applied. The yellow, green and red lines correspond respectively to TH=0.5, TH=0.7 and TH=0.9. As it can be noted, if we consider as query image a generic view of image number 63 (with 65 features), the highest number of correct matching features has been obtained for TH=0.5 and TH=0.7. For the mentioned threshold values, we find three maxima in correspondence of images 61, 62 and 63 in the DB. The graph in figure 4 shows how the algorithm correctly selects the references image and finally identifies the most matched image as the query image (number 63).

VI. CONCLUSIONS AND FUTURE WORK

In this paper we presented an indoor localization solution that use only the capabilities of a modern PDA equipped with a high resolution built-in camera, internet connection, motion and magnetic sensors, an image recognition system and a map with several geo localized images of the building. The proposed prototype is based on SURF algorithm for feature extraction and description, the One to One algorithm for descriptors matching and on processing of accelerometer and magnetometer data for counting steps and calculating heading. Several tests were carried out and the results are promising.

Future developments will consist in the integration of a plane homographic technique to improve the real time behavior of the entire project. These developments will allow for a better estimation of the user position in terms of view angle and distance from anchor point. In addition, more refined processing techniques for motion and heading data will be employed and tests to assess the overall error of the joint implementation of the SURF algorithm and the accelerometer data. Other future developments will consist in optimizing the software implementation of the algorithms in order to reduce the CPU load and memory usage.

REFERENCES