

A tracking solution for mobile augmented reality based on sensor-aided marker-less tracking and panoramic mapping

L. Yu · S. K. Ong · A. Y. C. Nee

Received: 15 July 2014 / Revised: 4 November 2014 / Accepted: 21 December 2014 /
Published online: 4 January 2015
© Springer Science+Business Media New York 2015

Abstract This paper proposes a tracking system for outdoor augmented reality (AR) on handheld devices based on an integration of vision tracking and on-device sensor measurement. To deal with the unpredictable and complex visual information in an outdoor environment, two tracking schemes are proposed for both near-field and far-field tracking scenarios. A sensor-aided binary descriptor is combined with an intensity-based tracking algorithm to deliver a 3D tracking system for fronto-parallel planar surfaces in near-field tracking. In far-field tracking, a sensor-guided panoramic tracking and mapping approach is proposed which allows a creation of the panorama of distant scenes on the fly with camera rotation motion to be tracked at the same time. This implementation allows near real-time creation of panoramic maps on-device; therefore, the users are able to tag information on the training target instantly.

Keywords Mobile augmented reality · Feature matching · Tracking and registration · Panoramic mapping

1 Introduction

To track the pose of a camera and register the virtual information to the real environment is a crucial step in an augmented reality (AR) pipeline. The requirements of a convincing tracking solution for large-scale outdoor applications include efficiency, accuracy, portability and scalability. Real-time performance and sub-pixel accuracy are the essential requirements. In addition, light-weighted hardware devices are in demand since outdoor applications are expected to be carried around easily and scaled up to a certain level.

Solutions for outdoor AR tracking in early days relied solely on the GPS to localize the camera position and inertial sensors to measure the orientation. However, the precision and update rate of GPS is not sufficient for accurate tracking. Inertial sensing also suffers from the problem of error drifting and measurement distortion due to electro-magnetic fields in the environment. Systems based on GPS and inertial sensors might achieve acceptable precision

L. Yu · S. K. Ong (✉) · A. Y. C. Nee
Mechanical Engineering Department, National University of Singapore, 9 Engineering Drive 1,
Singapore 117576, Singapore
e-mail: mpeongsk@nus.edu.sg

for registering the position of interest in 2D when the targets are far away from the observer, but results in large errors when the target to track gets closer and 3D tracking is a requirement. Advanced computer vision techniques can provide solutions for 3D tracking with sub-pixel accuracy. However, the challenge in vision-based systems is the uncontrolled condition of the environment that they are operating in. This requires the vision-based tracking systems to rely on natural characteristics or visual cues of the environment to perform tracking.

Many recent works propose hybrid tracking solutions combining three types of tracking to fulfil the requirement of efficiency and accuracy. Smartphones which are integrated with the on-board Micro-electromechanical systems (MEMS) based sensors, GPS and cameras, have become a popular platform for mobile AR applications. Its mobility and portability makes it an ideal platform to be deployed in outdoor applications. However, the embedded hardware sensors provide poorer performance compared to stand-alone devices. The computational power of mobile devices also has limitation on the complexity level of the algorithms that could be run. Exploiting all the sensing devices on the smartphones to achieve the required efficiency and accuracy is the main concern of the researchers.

The proposed solution integrates the sensor measurement and the state-of-the-art vision algorithms to achieve adequate tracking result. The proposed solution attempts to make best use of the static visual information and provide a possible solution for tracking in large-scale outdoor area. It takes two types of static visual cues into consideration, namely, fronto-parallel textured planar surfaces and distant scenes. Fronto-parallel planar surface, such as road signs, advertisement boards and building facades, are widely distributed and easy to find in the real world. These objects also contain planar patterns with rich corner features and large intensity gradient that are favoured by the proposed 3D tracking method. Another type of tracking target is the distant scenes viewed relatively far away from the users. Distant scenes, such as landmarks, provide not only the visual cues for tracking but also geographical significance in some cases. A rotational-only panoramic mapping and tracking solution is proposed to deal with far-field targets. Mapping of panorama requires a pure-rotational motion of the camera, which cannot be strictly followed by a handheld device. However, the transitional error is much smaller than the rotation motion since the target is far enough, so it can be ignored. The main contributions of this work are (1) a feature descriptor with sensor information generation for efficient keypoint matching, (2) 3D tracking solution based on illumination-invariant intensity-based optimization for near-field scenes, (3) real-time 2D panoramic mapping and tracking for far-field scenes, and (4) a complete implementation of the tracking solution on a handheld device for outdoor augmented reality.

The remaining of the paper first reviews some related work in the area. Next, a discussion of the design of the tracking solution is presented in details in Sections 3, 4, and 5. Section 3 introduces a novel approach of feature detection and matching by integrating the sensor measurement with an efficient binary descriptor at the algorithmic-level to deliver a descriptor matching algorithm. Section 4 discusses the 3D tracking approach by building on top of the proposed descriptor matching algorithm, a direct intensity-based tracking module. Section 5 details the design of the panoramic mapping and tracking approach for distant scenes. Experimental results and discussion are given in Section 6. A final conclusion is presented in Section 7.

2 Related work

Tracking is a popular sub-domain in AR research. One of the pioneer research works is the “Touring Machine” by Feiner et al. [5]. This tracking system integrates various types of

sensors, including GPS and magnetic compasses, for navigation and registration of tracking objects. With the advance of computer vision-based tracking algorithms, several vision-based tracking solutions have been reported [11, 14, 19], which have been demonstrated to be more accurate than sensor-based tracking. The advantage of vision-based tracking is its high adaptability to unprepared environments. For example, the point cloud-based tracking system presented [19] is able to build multiple 3D maps for an unknown environment online and the marker-less tracking schemes [11] can perform 3D registration based on image visual cues extracted from camera frames. However, vision algorithms have a higher requirement for computational power, which prevents their use in low-power and portable devices for outdoor applications. Hybrid approach that combines various types of sensing techniques is a promising approach to achieve efficient and accurate tracking results, and yet can run on devices with limited computational power. Some research works on hybrid tracking schemes and mobile-platform implementations are reviewed in detail in the remaining of this section.

In hybrid tracking, systems that combine various types of sensing techniques, inertial sensors have been demonstrated to benefit feature detection and matching in terms of speed and accuracy in some cases. Bleser and Stricker [3] explored several visual-inertial fusing models and developed a sensor-fused 3D model-based tracking system. Schall et al. [17] presented their outdoor AR system using a multi-sensor fusion approach. Both their works require vision tracking and sensor measurement to be performed separately. The data from the two approaches are fused using filters. In some other researches, the inertial measurements are tightly coupled with the feature tracking algorithms. The system developed by Hwangbo et al. [8], uses instantaneous rotation obtained from inertial sensors to update the parameter of a tracking motion model. The approach is based on the Kanade-Lucas-Tomasi (KLT) feature tracking algorithm. Since KLT-based tracking converges only in a small convergence region, it fails easily when the camera motion is too large. Measurements from gyroscopes are used to update the template warping so that the chance of convergence is increased. Kurz [9] proposed to align the orientation of local feature descriptors with the gravity. Their approaches are effective for detecting and matching of congruent and near-congruent features in static scenes. Calonder et al. [4] proposed an efficient binary descriptor Binary Robust Independent Elementary Features (BRISIF), which is sufficiently light-weight to run real-time on devices with low computational power. The main problem with the BRIEF descriptor is that it is not rotation-aware. This paper proposes a feature descriptor extended from the BRIEF descriptor that is rotated accordingly with the device orientation to achieve a rotational-invariant performance.

Some research works have demonstrated the tracking for AR on mobile devices based on efficient computer vision algorithms [10, 21]. In [10], a point-and-shoot tagging system is introduced for vertical and horizontal surfaces on a mobile phone. The approach is based on perspective patch recognition instead of feature matching. The system utilizes the sensors on the mobile device to recognize the orientation of the surface as well as rectify the captured image into a fronto-parallel view. Users are allowed to add and insert virtual objects on-device by simply pointing the phone to the target. In [21], Scale-invariant Feature Transform (SIFT) [12] and *Fern* [13] feature matching algorithms are modified to run on mobile devices. The feature matching approach is further integrated with a patch tracker in their work to achieve an efficient and accurate tracking result. To achieve sub-pixel accuracy and robust tracking performance, a direct-intensity based tracking method is proposed in this paper. The efficient second-order minimization (ESM) [2] method is used to solve the optimization problem. In the authors' approach to accelerate the performance of the optimization process, a sub-grid model is proposed. A fast and easy-to-operate reference image selection and training process allows the user to tag information on a suitable target on-device instantly.

To track the scenes in a reasonable distance, six degrees-of-freedom (DoF) tracking is not strictly required. Gammeter [7] proposed an AR tracking system that tracks the camera motion in 2D translation and in-plane rotation only, with the assumption that the tracking target's relative movement to the camera is orthogonal to the camera's view direction. Working together with a server-side recognition method, this tracking system is able to track city landmarks with acceptable errors. Another scene-based tracking method by Wagner restricted the motion of the camera in pure rotations and performs the tracking and the panoramic mapping at the same time [20]. Tracking against the map created on the fly results in high accuracy. The authors' approach adapts the idea of panoramic mapping and tracking. The sensor information is used to guide the user to create the map. In tracking the camera motion against the map, the abovementioned descriptor is used for matching before a motion refinement step by minimizing the reprojection errors of the features.

3 Sensor-aided feature matching

Local feature detection and matching is an essential step in a vision-based tracking system. A sensor-aided BRIEF descriptor is proposed for the feature detection and matching in the authors' algorithm. BRIEF descriptor is created as a binary vector based on the intensity difference test of a set of paired pixels with a predefined distribution around the detected feature point. The distribution of the test pairs can be random or with a designed sampling pattern. Experiments in [4] indicate that an isotropic Gaussian distribution around the centre of the patch can give the best result in terms of recognition rate. In comparison to some other popular fast-to-compute descriptors as Speeded-Up Robust Features (SURF) [1] and SIFT [12], BRIEF is demonstrated to be more efficient. In addition to efficiency, BRIEF descriptor yields higher recognition rates in the cases where invariance to scale change and in-plane rotation are not required.

In the outdoor AR scenario, the keypoint features are assumed to be static – from fronto-parallel planar surfaces. In this case where the coarse orientations of the camera are predictable with inertial sensors, the binary test patterns can always be aligned with the gravity through a simple in-plane rotation. The Oriented FAST and Rotated BRIEF (ORB) descriptor is designed to be rotational-aware [16]. It estimates the orientations of the Features from Accelerated Segment Test (FAST) keypoint features [15] from the intensity centroids of the features. Instead of estimation from image intensity, the orientation could be measured from the sensors on device as proposed in this paper. The BRIEF binary test pairs are obtained simply from a randomly generated Gaussian distribution instead of trained from a large binary test pool. The approach of utilizing sensor measurements simplifies the implementation of the feature descriptor. In this section, the formulation of the BRIEF descriptor is revisited. Based on that the rotation of the test pattern using a given rotation angle and rotation estimation from hardware sensors on a mobile device are discussed.

3.1 Platform and sensor measurement

An Android device is chosen as the host platform. The device has built-in MEMS sensors, including accelerometer, gyroscope and magnetometer. Although the low-cost sensors in the devices suffer from noise and drift when functioning individually, a sensor fusion scheme could provide a sound estimation of the orientation of the device. Android SDK provides an orientation estimation function from the fusing of the accelerometers and magnetometer (Ac/Mg). The drawback of using combined accelerometer-magnetometer measurement is the jittering noise

mainly because that the magnetic field is highly affected by the devices in the environment. The orientation estimation from the Ac/Mg is fused with the gyroscope measurement using a simple complementary filter so that non-jittering and drift-free orientation measurement can be obtained. Figure 1 shows the comparison of the Ac/Mg sensor measurement and the filtered measurement during the measurement period when the device is hand held. Rapid orientation changes are deliberately added. The orientation measurements about three axes are plotted separately in Fig. 1. In Fig. 1, the horizontal axis is the time interval and the vertical axis is the orientation change measured with respected to a pre-defined reference. The filtered motion indicated by the red line exhibits a smoother performance than the non-filtered motion as shown in Fig. 1.

The global and local coordinates of the device adopted are shown in Fig. 2; the X-Y-Z coordinate system indicates the world coordinates attached to the earth, while the x-y-z coordinate system indicates the body coordinates attached to the device. The orientation of the device in the world coordinate is described by the azimuth (rotation about Z), pitch (rotation about X) and roll (rotation about Y). The rotations are set as zero when the two coordinates are coincident.

The device is constrained to be working in the landscape mode as illustrated in Fig. 2. To best utilize the gravity information of features from vertical surfaces, the camera is expected to be held such that the screen plane is roughly perpendicular to the x-axis. Therefore, the rotation angle with respect to the gravity is the pitch angle (rotation about X-axis), which is passed to the vision-tracker to rotate the test pattern.

3.2 BRIEF Descriptor with rotation-awareness

The BRIEF descriptor is constructed from a set of binary intensity test τ on a smoothed image patch \mathbf{p} . A single binary test can be defined in Eq. (1) where $\mathbf{p}(\mathbf{a})$ and $\mathbf{p}(\mathbf{b})$ are the intensities of the pixel at position a and position b in the image patch \mathbf{p} .

$$\tau(\mathbf{p}; \mathbf{a}, \mathbf{b}) := \begin{cases} 1 & \mathbf{p}(\mathbf{a}) < \mathbf{p}(\mathbf{b}) \\ 0 & \mathbf{p}(\mathbf{a}) \geq \mathbf{p}(\mathbf{b}) \end{cases} \quad (1)$$

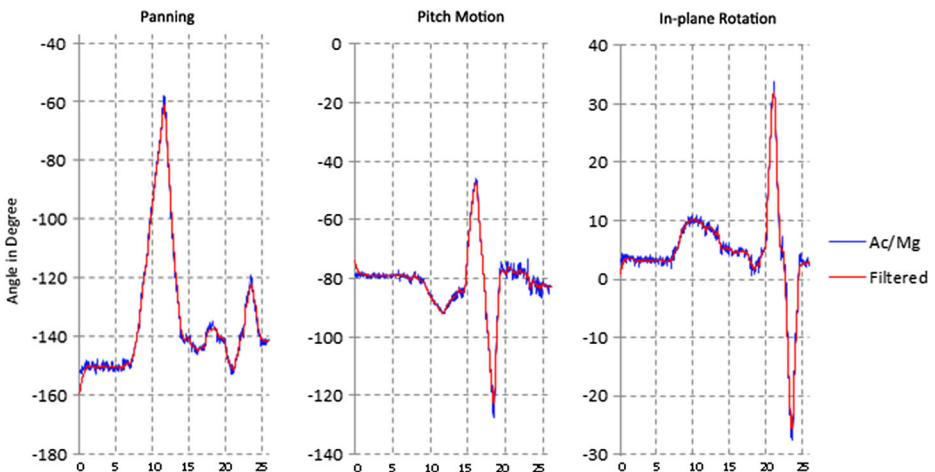


Fig. 1 Comparison of sensor measurement in motions about 3-axis

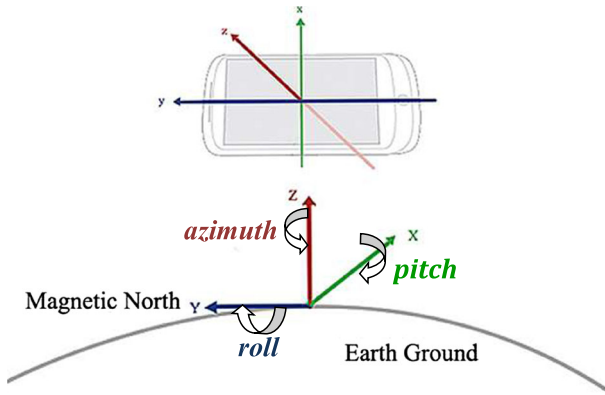


Fig. 2 World coordinates X-Y-Z and body coordinates x-y-z adopted by the device

A feature descriptor for a keypoint is a bit string as the response to the intensity difference of these n pairs of binary tests. The n tests are sampled from an isotropic Gaussian distribution, which is denoted as matrix \mathbf{G} . \mathbf{G} is formed of $2n$ test points (n test pairs, x and y denote the position of the pixel in the image plane coordinate).

$$\mathbf{G} = \begin{pmatrix} x_1, x_2, \dots & \dots, x_{2n} \\ y_1, y_2, \dots & \dots, y_{2n} \end{pmatrix} \tag{2}$$

To steer the BRIEF test pattern to the orientation of the camera detected by the inertial sensors, a rotation is applied to each point in \mathbf{G} . The rotated test set is $\mathbf{G}_\theta = \mathbf{R}_\theta \mathbf{G}$. The feature descriptor at an image patch \mathbf{p} , constructed at a particular rotational angle θ , from the test set \mathbf{G} is formulated as:

$$f_{n,\theta}(\mathbf{p}) := \sum_{1 \leq i \leq n} 2^{i-1} \tau(\mathbf{p}; [\mathbf{G}_\theta]_{2i-1}, [\mathbf{G}_\theta]_{2i}) \tag{3}$$

Descriptor for each keypoint extracted from the input images is presented as feature $f_{k,\theta}$. k represents the dimension of the binary descriptor, usually 128, 256 or 512, and θ is the detected rotational angle as mentioned earlier. FAST keypoints are first extracted from images and a set of rotated binary descriptors is generated for the extracted keypoints. Brute-force search based on Hamming distance between descriptors generated from two images is used for image matching. Figure 3a depicts the rotations of the test pattern at θ of 40° and of 80° . Each line segment indicates a pair of binary tests with the ends representing the location of the test points \mathbf{a} and \mathbf{b} on the image patch \mathbf{p} . The purpose of rotating the test pattern is essentially to align the pattern to the image patch as close to the alignment at zero in-plane rotation as illustrated in Fig. 3b. As the BRIEF descriptor has modest tolerance to affine transformation, the in-plane rotation angle is not strictly constrained.

4 Near-field 3d tracking solution

A robust tracking solution requires that image registration errors are within sub-pixel level. The feature-based detection-and-tracking approach is sensitive to noise and can often cause jittering.

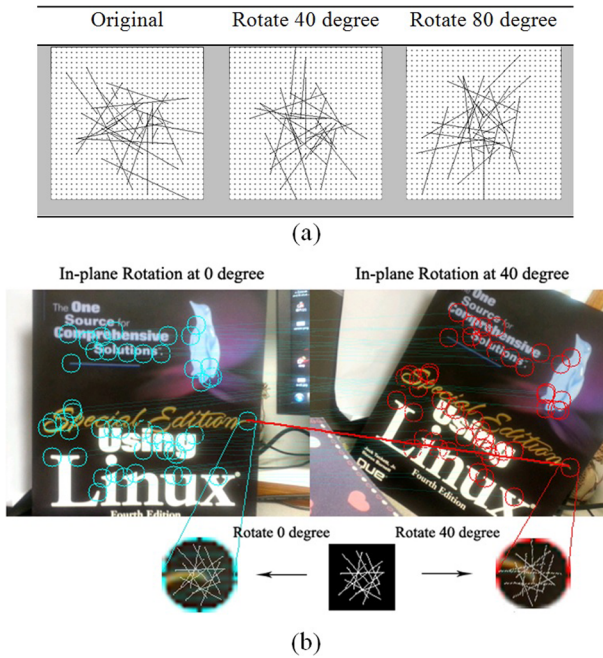


Fig. 3 Rotation of the test pattern applying over the image patch

In the proposed near-field tracking system, the estimation of the camera pose from feature-based tracking is used as an initial guess of an iterative image alignment method. In this case, the pixel intensities of the template image and that of the camera are required as the input.

Figure 4 illustrates the proposed tracking scheme. When tracking starts, the system extracts keypoints from every input frame and matches them against the reference image using the

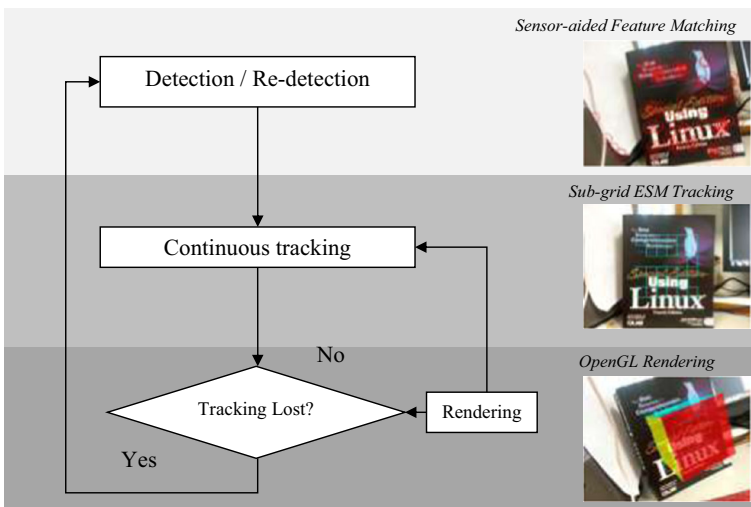


Fig. 4 Steps in the proposed textured planar surface-based AR tracking system

sensor-aided feature matching method. After a successful match is observed, the ESM tracking method starts to serve its role to continue tracking. In the situation of tracking loss, the system goes back to the detection stage and performs detection again.

Image alignment is modelled as a non-linear optimization problem to minimize the sum of squared differences between the intensity of the template image and the one of the warped camera images. It can be solved iteratively providing a rough estimation of the warping parameters as an initial guess. A recently developed algorithm ESM is used to solve the non-linear optimization problem. Compared to the widely used iterative methods, such as Gauss-Newton and Levenberg-Marquardt (LM), the ESM has been shown to have a higher convergence rate. The sub-grid ESM Tracking approach with a proposed illumination model to counter illumination variance in the outdoor lighting condition is discussed in the remaining part of this section.

4.1 Sub-grid ESM tracking

The formation of the intensity-based tracking problem is first discussed. Let I^* denote the template image and I is the frame image that needs to be warped with a transformation to the template. Suppose pixel p_i from the frame image can be warped to the pixels p_i^* in the template image by a warping function \mathcal{W} . If the points in the template image form a planar surface, a homography transformation $H(x)$ can be used to represent the warping. The homography transformation contains eight parameters denoted by a vector x . For a set of n pixels p^* from a selected region of the template image, a vector $y(x)$ of the image difference from the template image and the warped frame image can be defined as:

$$y(x) = [y_1(x) \ y_2(x) \ \dots \ y_n(x)]^T \tag{4}$$

where

$$y_i(x) = I \left(\mathcal{W} \langle \hat{H} \rangle \circ \mathcal{W} \langle H(x) \rangle (p_i^*) \right) - I^* (p_i^*) \tag{5}$$

Here the warping transformation is defined as an initial estimation $\mathcal{W} \langle H(x) \rangle$ composited with an incremental warping $\mathcal{W} \langle \hat{H} \rangle$.

Given the initial estimated homography $H(x)$, the template image I^* , the frame image I , and the set of pixels from the selected region p^* , the compositional incremental warping \hat{H} can be solved iteratively through minimizing the image difference $y(x)$. To address this problem, the ESM is developed as a non-linear optimization procedure with a convergence rate similar to the second-order methods, but with an efficiency of the of the Hessians. The image difference is approximated as (for $x=x_0$):

$$y(x_0) \approx y(0) + \frac{1}{2} (J(0) + J(x_0)) x_0 \tag{6}$$

$J(0)$ and $J(x_0)$ are the current Jacobian and the reference Jacobian respectively. Using the homography Lie algebra parameterization, an approximated J_{esm} matrix can be used to replace the current Jacobians:

$$J_{esm} = \frac{1}{2} (J(0) + J(x_0)) = \frac{1}{2} (J_I + J_{I^*}) J_w J_G \tag{7}$$

The Jacobian for ESM tracking update is based on the gradient of the template patch J_I , and the warped patch of the current image J_I and two constant matrices J_w and J_G . The

gradient of the image patch is obtained using the Sobel filter. Let $\overline{J_x}$ and $\overline{J_y}$ be the average of J_I and J_I with respect to x and y in 2D image plane. The Jacobian J_{esm} is parameterized as:

$$J_{esm} = \begin{bmatrix} \overline{J_x} & \overline{J_y} & \overline{J_x x} & \overline{J_y y} & \overline{J_x x} - \overline{J_y y} & -\overline{J_x x} - \overline{J_y y} & -\overline{J_x x^2} - \overline{J_y xy} & -\overline{J_x xy} - \overline{J_y y^2} \end{bmatrix} \tag{8}$$

The update for the Homography x_0 can be calculated by solving the non-linear least square optimization problem formulated as

$$\Delta x = (J_{esm}^T J_{esm})^{-1} J_{esm}^T y_i(x) \tag{9}$$

Iteratively x_0 is estimated and used to update the homography \hat{H} . Formulated with Lie Algebra, the homography update is obtained using the matrix exponential function:

$$\hat{H} \leftarrow \hat{H} e^{\sum_{k=1}^8 x_k A_k} \tag{10}$$

where A is a basis of the Lie Algebra. Detailed formulation is stated by Benhimane and Malis [2]. In the actual implementation, the selected region of the template image is divided into sub-grids for tracking, instead of using all the pixels in the image. Tracking of the image is based on a 144×144 pixels square area positioned at the centre of the template image. This large square is divided into 6×6 sub-grids, which is 24×24 pixel each. The average image gradient within each sub-grid is computed, and only sub-grids with gradients above 10 grey levels per pixel are used in tracking. The criterion is selected because the image gradient is used to compute the Jacobian matrices in ESM. Experiment shows that image regions with low gradients do not contribute much to ESM convergence, therefore these regions are not considered.

4.2 Illumination model

The image alignment based on the minimization of intensity error as formulated in Eq. (5) is based on the validity of the assumption that the image intensity of the frame image is consistent with one of the template images, such that the intensity error is purely due to object pose change. In the practical situation, illumination is rarely constant especially in an outdoor environment. Silveira and Malis proposed an illumination model to validate the intensity constancy assumption [18]. In their work, the parameters are estimated together with the motion parameters within the ESM iterations, such that the number of parameters is greatly increased. For an efficient implementation on a mobile device, the illumination model is proposed to estimate the parameters directly from the warped and reference images. This is feasible as the pose in the previous frame is close to the current image at the start of ESM. Suppose $p_{i,j}$ is the i^{th} pixel in the sub-grid j in the frame image, $I(p_{i,j})$ denotes its intensity as stated above. Let m_j and d_j be the mean and standard deviation of the pixel intensities in the sub-grid j in the warped frame image, and m_j^* and d_j^* be the corresponding values for the template image, the modified pixel intensity $I'(p_{i,j})$ is computed as:

$$I'(p_{i,j}) = (d_j^*/d_j) (I(p_{i,j}) - m_j) + m_j^* \tag{11}$$

Therefore, Eq. (5) is modified as

$$y_{i,j}(x) = I'(\mathcal{W}\langle \hat{H} \rangle \circ \mathcal{W}(H(x)) (p_{i,j}^*)) - I'(p_{i,j}^*) \tag{12}$$

4.3 Reference image selection

To select a valid reference image as the template for tracking, two criteria need to be fulfilled. The first criterion is that a minimum number of sub-grids with sufficient intensity gradient should be guaranteed. To guide the user in selecting surfaces with good image gradients, only sub-grids with sufficient gradient are rendered during the selection process. This provides an indication of the suitability of a surface for ESM tracking. The second requirement is the movement of the camera with respect to the surface. This is because the homography decomposition method adopted here results in two possible solutions for the normal. As the normal vector is defined in the reference camera frame in the formulation of homography, the correct normal vector that remains constant when the sideways motion is greater than 0.5 % of the perpendicular distance between the camera and the plane is selected as the correct normal. Since it is assumed the planar surfaces to be tracked in the outdoor environment are vertically distributed, it is convenient for the user to hold the camera parallel to the surfaces and move sideways. As a further check, tracking is continued with a virtual 3D object augmented onto the planar surface after the normal vector is determined. This allows the user to check the accuracy visually.

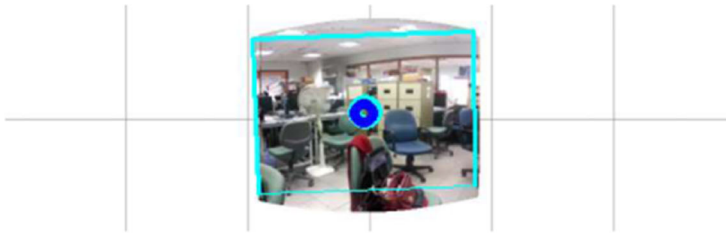
5 Far-field panoramic mapping and tracking

A panoramic mapping and tracking approach is proposed to track and tag information efficiently on the far-field scenes. In the outdoor scenario where the user holds the handheld device to point at the targeted scenes, the translation of the handheld device is very small compared to the distance between the camera and the scene. Therefore, it can be assumed as a rotational-only motion of the camera by ignoring the parallax error. This approach is first explored by Wagner [1]. Similar to this approach, the map into sub-grids is divided for real-time tracking and pose recovery. The difference is that the proposed sensor-aided feature matching algorithm is used to perform the pose estimation, and the sensor measurement from the device to guide the mapping and tracking process. In this section, the sensor-aided mapping and tracking algorithm is first explained, followed by a detailed discussion of the projection surface selection, motion estimation and map rendering.

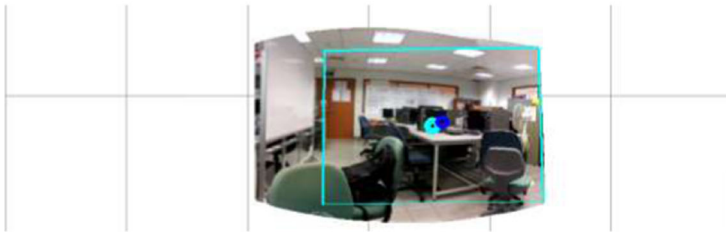
5.1 Sensor-aided mapping with grids

In the proposed system, a 180° panoramic map with size of 960 in width and 300 in height is rendered. The input camera frame is taken as 320×240. The map is divided into grids with size of 180×150 as shown in Fig. 5a. It is assumed that the user holds the device in such a way that the image plane is aligned with the gravity direction. The first input frame is placed at the centre of the map, with its orientation information from the sensor recorded. The orientations of the following camera frame input are calculated relative to this first measurement. Since the orientation is measured in each frame, the camera heading can be calculated in the map coordinate. As shown in Fig. 5b and c, the camera heading is highlighted with a dark blue dot. This is a rough indication of the camera heading on the map which provides guidance for users to move the device as close as to a pure-rotational motion.

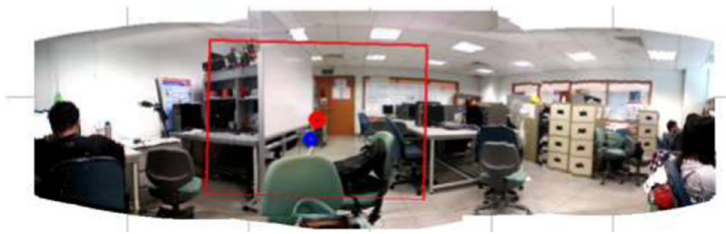
To ensure an accurate mapping, the input frame is matched against the nearest map grids that have already been filled. The spatial adjacency between the input frame and the generated map grids are first estimated using sensors. For frame which overlapped area on the filled map



(a) First image is placed at the centre, the map is divided into grids.



(b) Light blue highlights the view of the current camera tracked, dark blue indicates the sensor-measured camera heading.



(c) When tracking is lost, the camera view is frozen at the last tracked view and highlighted red. The sensor continues to update the heading.

Fig. 5 Screenshots in real-time performance of panoramic mapping and tracking. **a** First image is placed at the centre, the map is divided into grids. **b** Light blue highlights the view of the current camera tracked, dark blue indicates the sensor-measured camera heading. **c** When tracking is lost, the camera view is frozen at the last tracked view and highlighted red. The sensor continues to update the heading

grids is large enough, an accurate orientation is computed using the matched features and refined by LM optimization. The algorithm is summarized as follows:

Sensor-aided Mapping

- 1) The map is divided into grids with each grid assigned an orientation information
- 2) First frame input: set as the centre image
- 3) Input frame: compared orientation measured by inertial sensors with the pre-defined orientation of each grid to decide the nearest grids
 - If the nearest grids are filled and the overlapping area is large enough, proceed to step 4)
 - Otherwise continue to run step 3)

- 4) Match the input frame with the nearest grid and calculate the orientation of the frame
- 5) Warp input frame and render on the map, update the map

5.2 Projection surface

The cylindrical projection surface is selected in the design. Cylindrical mapping allows a full 360° variation of the horizontal rotational angle. However, it limits the pitch angle variation because the map becomes more compressed at the top and the bottom. In the proposed system, the range of pitch angle constraint is acceptable because vision cues in the sky or on the ground are not vital. As shown in Fig. 6, (x, y) denotes the point on the frame image, and (u, v) the point on the projection surface. (x, y) indicates the pixel position in width and height-direction respectively, while (u, v) indicates the pixel position in angular direction and the longitudinal axis respectively. The mapping between them can be formulated as either forward projection (Eq. (13)), i.e., project the point on the frame image point forward onto the cylindrical surface, or backward projection, Eq. (14), i.e., project the point on cylindrical surface backward onto the frame image, where f is the focal length and s is the radius of the cylinder that can be set to $s=f$ to minimize the distortion of image warping

$$u = \text{stan}^{-1} \frac{x}{y} ; v = s \frac{y}{\sqrt{x^2 + y^2}} \tag{13}$$

$$x = f \tan \frac{u}{s} ; y = \frac{v}{s} f \sec \frac{u}{s} \tag{14}$$

5.3 Motion estimation

The motion is assumed to be a fixed-position three DoF rotation. In this case, the rotation of the camera R can be estimated directly from the homography H between two images providing the camera intrinsic matrix K :

$$R = K^{-1}HK \tag{15}$$

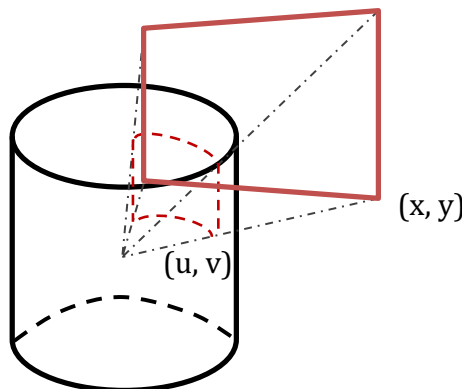


Fig. 6 Project frame image onto the cylindrical projection surface

The rough estimation of homography is based on the aforementioned sensor-aided BRIEF descriptor matching. To accelerate the matching process, the estimated pose of the previous frame is used to filter out some keypoints since it is assumed that the camera is not moving too fast. To filter out the non-matched keypoints, first the detected keypoints in the coming frame are warped back to the reference image using the inverse of the previously estimated \mathbf{H} . A matching mask is created with a predefined search area for each keypoint in the reference image. If the warped keypoints reside inside in the search area, matching between them is conducted, otherwise the keypoint is discarded. This approach not only reduces the matching time, but also increases the correctness of Random Sample Consensus (RANSAC) estimation [6].

When the camera undergoes a fixed-axis rotational motion \mathbf{R} , the current frame can be related to the previous frame using the homography transformation \mathbf{H} , which can be extracted using Eq. (15). To achieve sub-pixel accuracy, the rotational angles are refined by minimizing the sum of the squared distance errors of all the visible feature points that are re-projected to the backward projection of the nearest sub-grid of the map. The origin is pre-assigned at the centre sub-grid of the map. Next, let \mathbf{R}_j be the rotation of the camera from the origin to the orientation where the backward projection of the j^{th} sub-grid of the map is obtained, and $\hat{\mathbf{R}}$ be the rotation of camera from the origin to the orientation where the input frame image is taken. The re-projected pixel coordinates \mathbf{p}'_j on the sub-grid image of the featured pixels \mathbf{p}_i on the frame image is formulated as:

$$\mathbf{p}'_j = \mathbf{K} \mathbf{R}_j \hat{\mathbf{R}}^T \mathbf{K}^{-1} \quad (16)$$

The sum of squared re-projection error is:

$$e = \sum |\mathbf{p}_j - \mathbf{p}'_j|^2 \quad (17)$$

In this case, only feature points that are visible to the sub-grid image after the re-projection are considered. This is a non-linear least squares problem which is solved using the LM algorithm. The resulted $\hat{\mathbf{R}}$ is used in the next compositing step as well as in the tracking of the camera heading.

5.4 Efficient image composition

Given the rotation of the camera, the camera frame could be forward projected to the projection surface and blended with the panoramic map. To blend the coming frame image to the map efficiently, an image composition approach is proposed to ensure real-time performance. In the compositing step, two image masks are updated. The first image mask is a mask of the generated map on the cylinder surface and the second mask is a forward projected mask of the current processing frame to the surface using the estimated rotations. Since the masks are stored as binary images, the nearest-neighbour interpolation can be used for image warping. Only the pixels outside the overlapping area of the two masks need to be filled. These pixel positions on the projection surface are backward projected on the camera frame to an inter-pixel position. Their colour values are evaluated through bilinear-interpolation of the neighbouring pixels on the frame image. No blending technique is applied to save computational power. The experimental results show that lighting change has little effect on the compositing result if the camera is moved slowly.

6 Experimental results

6.1 Sensor-aided feature matching evaluation

To evaluate the matching performance of the rotated BRIEF descriptor, images with synthetic in-plane rotations were tested. The textured image in Fig. 7a is transformed with in-plane rotations ranged from -90° to $+90^\circ$ with an interval of 10° . Each rotated image is matched against the original image. 500 keypoints are extracted from each of them. The matching rate is calculated using the method stated in [4]. As shown in Fig. 7b, the matching rate for the BRIEF descriptor remains acceptable at rotation change of ± 10 degrees but decreases dramatically when change is beyond ± 20 degrees. The pre-known rotation is shown to be able to facilitate the matching process to guarantee a matching rate above 70 % over the rotation range from -90 to $+90^\circ$.

6.2 Evaluation of surface detection robustness

Real-time performance on mobile devices is evaluated using four video sequences with the following characteristics: (i) camera motion with rotation and tilt, (ii) occlusion, (iii) camera pointing away and back, and (iv) scale change and noise data.

A feature-rich textured map is selected as a target for detection so that a wide variation in feature appearance is introduced in the test. Each video sequence contains 400 frames at size of 320×240 pixels. The frames are scaled at a level of 2 and targeted keypoints are constrained to below 300 per scale to guarantee a real-time processing rate. Pose estimation is assumed to be successful when more than 15 inliers are detected using RANSAC estimation. The test is executed on a mobile phone with a quad-core 1.5 GHz CPU and 1 GB RAM. For all four video sequences, the processing time and the pose estimation result for each frame are recorded. Figure 8 shows the sample frames from each video sequence and the recognition rate, which is the ratio of the successful-detected frames over the total frames.

For sequence (i) and (ii), the performance of the tracker is relatively good. The tracking is interrupted only for a small number of frames. The processing time is shorter for sequence (ii) because the most time-consuming process is matching using a brute-force search, which depends heavily on the number of keypoints. Fewer number of keypoints are available in sequence (ii) in the case of occlusion. Test for sequence (iii) shows that tracking could be picked up quickly when the camera is moving to the target. When the camera is moving away,

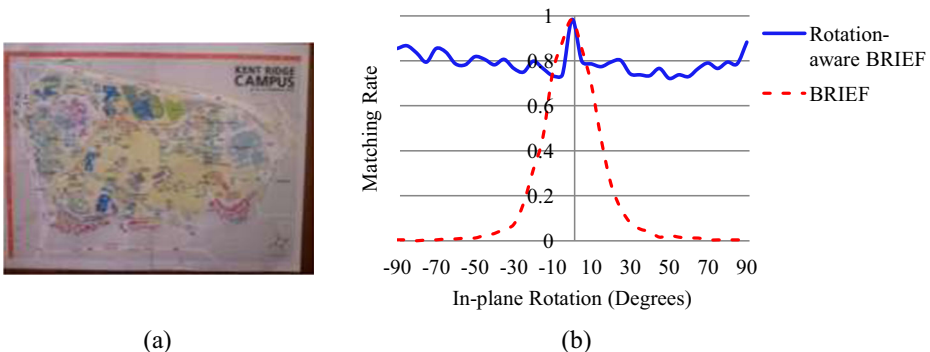


Fig. 7 Matching rate of rotation-aware BRIEF and BRIEF under synthetic rotations

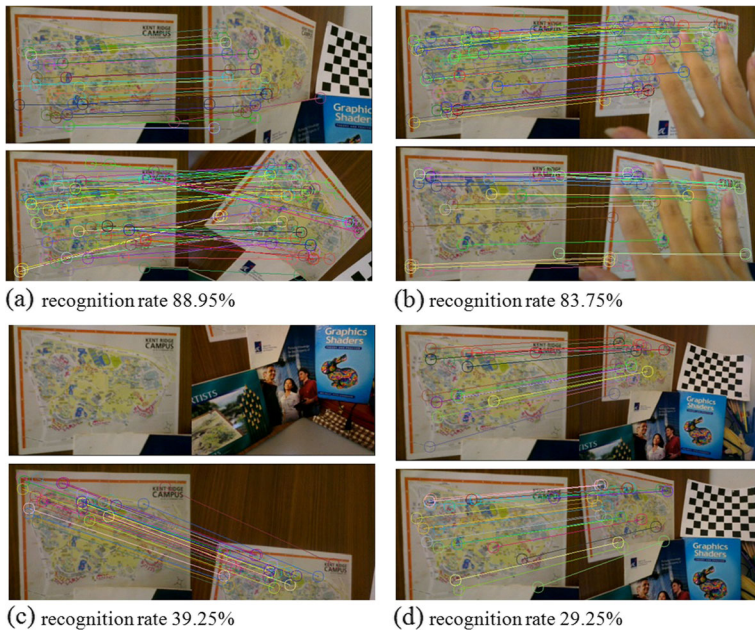


Fig. 8 Sample frames from the video sequences for evaluation of recognition rate **a** sequence (i), **b** sequence (ii), **c** sequence (iii), and **d** sequence (iv)

tracking is lost while the processing time drops accordingly. The performance of the tracker for sequence (iv) is not very good as the target number of keypoints is constrained to 300 for efficiency, which reduces the discriminative capability of the tracker. The overall performance of the matching algorithm based on the four tests is efficient, maintaining a frame rate between 20 and 30 fps. The tolerance for camera rotation and tilt, scaling and occlusion is high, and detection is high when a clear target is presented.

6.3 3D tracking performance evaluation

The robustness of 3D tracking is evaluated and compared with an approach that is similar to the tracking method proposed in [20]. The Patch Tracker method in [20] uses a reference image as the only data source similar to the ESM-based method proposed in this paper. Instead of using pixel intensities, keypoints are used for optimization. Optimization is performed to minimize the re-projection error of keypoints extracted from the reference image to the current frame. The patch size of the Patch Tracker is selected at 15 pixels and the search window is set at 41 pixels, so as to achieve similar frame rate for the tracker as the proposed intensity-based tracking approach.

Four video sequences are used for evaluation in this test. Video sequences with different surfaces and properties are used for tracking evaluation. These surfaces include (i) feature-rich poster, (ii) logo with noisy background, (iii) building facades, and (iv) reflective surface. A few frames from these sequences are shown in Fig. 9. The robustness of the tracking is evaluated as the ratio of successful tracked frames over total frames in the sequence. To simulate fast motion of the camera, the test video sequences are sped up in the experiments. Figure 10 shows the robustness of the proposed intensity-based tracking with the feature-based tracking



Fig. 9 Sample frames from the test video sequences

using the Patch Tracker tested over four video sequences with five different speeds. Overall, the ESM-based tracking out-performs the feature-based Patch Tracker. For sequence (i) in which the features are rich and with high distinctiveness, both methods perform well. The Patch Tracker performs well because of the good feature information available. For sequence (iv) in which reflective surface is detected, the intensity-based method performs better than the feature-based method.

The proposed tracking method experiences less jittering in real-time as compared to feature-based tracking. The x, y position estimation over time for a portion of sequence (iv) is plotted in Fig. 11 for both methods. This test video clip starts with a slow motion of the camera, followed by an introduction of a larger pose change and faster movement. It can be observed from the plots that the proposed method experiences smoother changes over time than the

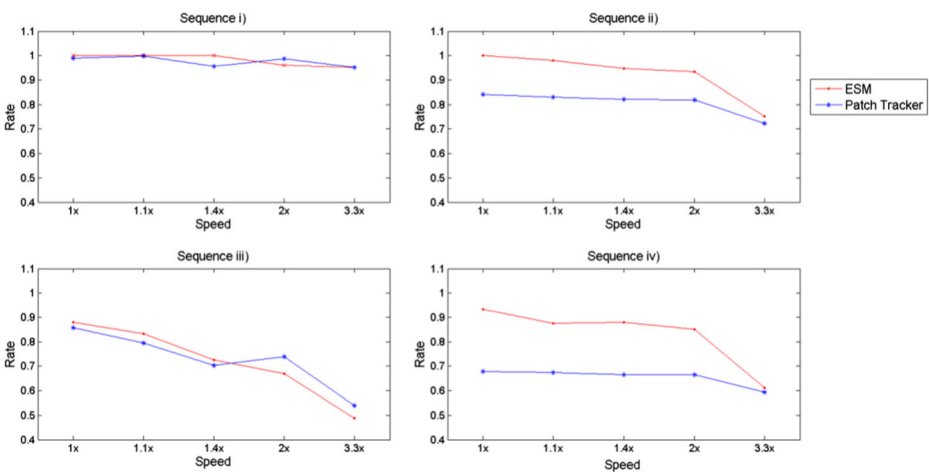


Fig. 10 Tracking rates comparison for ESM-based tracking and patch tracker on video sequences in Fig. 9 with five speeds

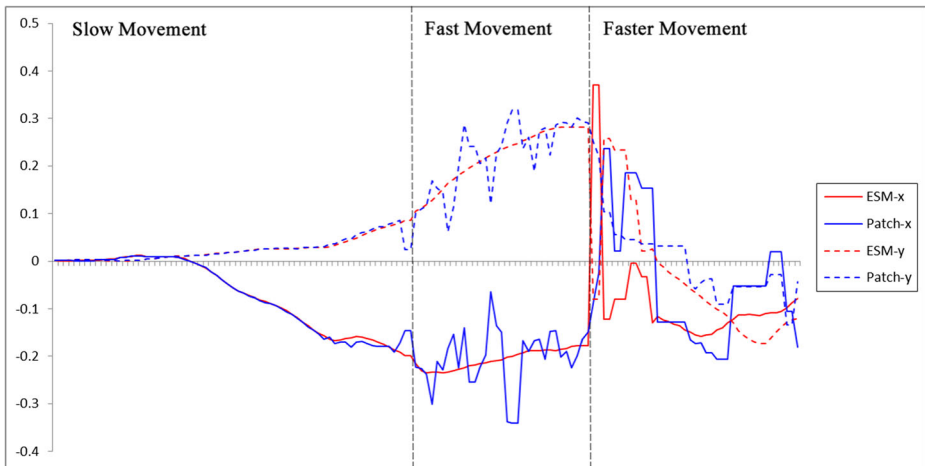


Fig. 11 Position estimation at X/Y-directions using ESM-based tracking and patch tracker

feature-based method. There are jitters in the estimation of the x, y position using the Patch Tracker when a faster movement of the camera is encountered. However, it is observed that the ESM-based method also fails to perform well during a faster movement of the camera.

To evaluate the proposed illumination model against the discrete illumination model [18], the model accuracy is used for evaluation. The model accuracy is measured using the Root-Mean-Square (RMS) pixel intensity error obtained after ESM has converged. The pixel intensity errors reported are relative to the 256 grey levels of the images process, and are not normalized. The RMS pixel error, average processing time and sub-grids used per frame are shown in Fig. 12. It can be observed that the RMS pixel error of the proposed illumination model is lower and shows less variation.

6.4 3D tracking efficiency evaluation

The performance of the near-field 3D tracking is evaluated on the set of video sequences as shown in Fig. 9. These surfaces, which have a rich set of corner features and large intensity gradient, are commonly found in the real world. To enable the scale-

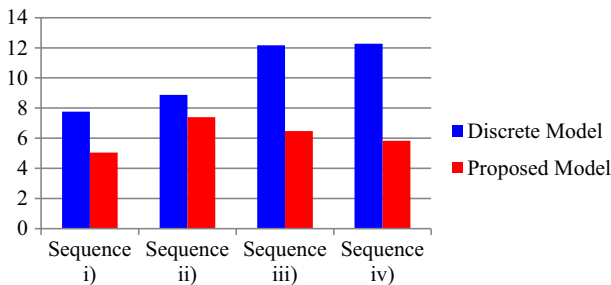


Fig. 12 Comparison of the RMS pixel error for the discrete and proposed illumination models on video sequences in Fig. 9

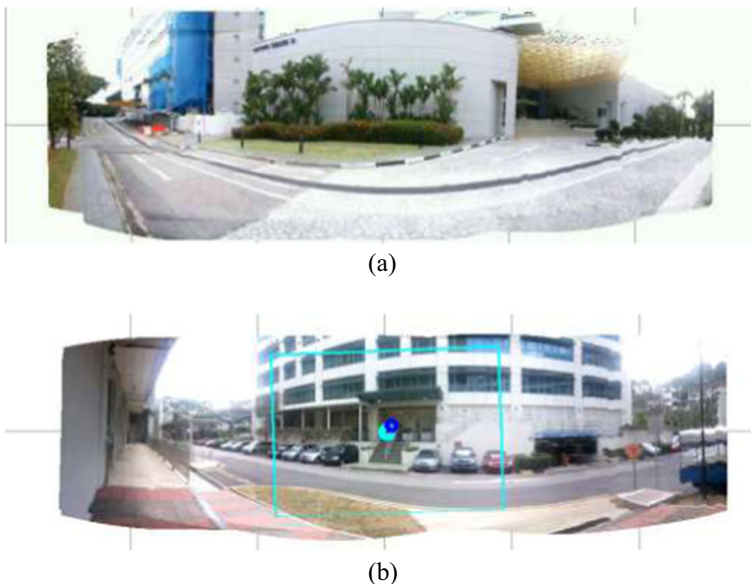
Table 1 Efficiency evaluation of the tracking system

	Time in milliseconds (300 keypoints, 2 levels of scales)			
	Mean	Standard deviation	Min	Max
Detection	69.07	33.82	10.35	163.67
Tracking	32.33	18.11	2.68	101.68

awareness of the feature tracking approach, a set of descriptors for the referenced image is computed on two scales with a downsizing factor of $\sqrt{2}$. 300 keypoints are targeted in each scale, which is empirically decided to be a good trade-off between the efficiency and the recognition performance. To train a reference image on the device takes around 50 milliseconds, which allow instant creation of target on-device. The speed for detection and tracking is summarized in Table 1. The number of keypoints to be detected and matched strongly affects the processing time. In practice, the frame rate of the tracking system is guaranteed at 15–30 fps.

6.5 Panoramic mapping and tracking evaluation

Tests in outdoor environment are conducted around the campus. With sufficient keypoints to be detected, mapping and tracking can be performed efficiently. Two examples are shown in Fig. 13. The results show that illumination variation can be handled well with the slow motion of the camera. The errors accumulate and become visible when the camera moves farther from the centre images. The resulted map could not compete in quality with panoramic images

**Fig. 13** Results of mapping and tracking in outdoor environment

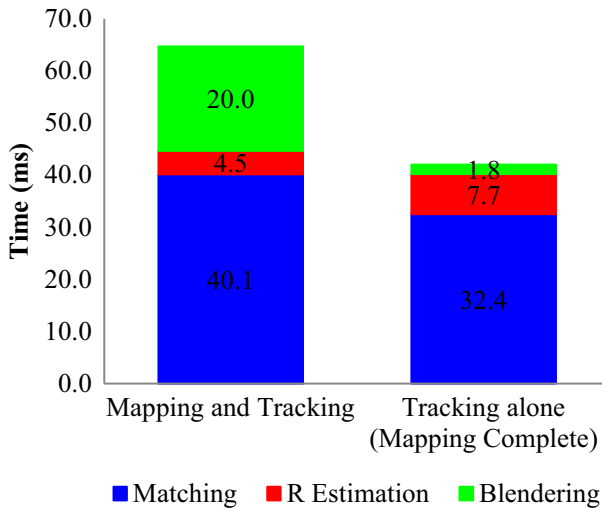


Fig. 14 Average time taken in mapping and tracking

generated offline. However, it can be used as a preview of a map that can be generated later using more robust stitching algorithms.

The efficiency of this tracking approach is evaluated as shown in Fig. 14. During the mapping and tracking stage, it takes an average of 65 milliseconds to process a frame. Map rendering takes a relatively longer time due to warping of image masks. When the map is completed, only tracking needs to be performed. The processing time is reduced to around 40 milliseconds. Matching is still the most time consuming step. With a good initial estimation from the matching result, rotation optimization (R estimation) using the LM algorithm can converge in a few milliseconds as indicated in Fig. 14.

7 Conclusion

In this paper, a sensor-aided tracking and tagging system is proposed for outdoor augmented reality on a handheld device. A novel approach is proposed to achieve visual detection by integrating the sensor measurement with binary descriptors. Experiment conducted in this research demonstrates that sensor measurement on-device is able to facilitate the BRIEF descriptor matching in a restricted condition. The feature matching algorithm is combined with an intensity-based image alignment algorithm, named ESM tracking to deliver a near-field 3D tracking system, which can track the camera orientation based on fronto-parallel planar surfaces. The proposed sub-grids illumination model gives the tracking method the ability to handle occlusion and illumination variance. A tracking and mapping system is introduced for far-field scenes in outdoor environment. This method is developed based on the assumption that the handheld device moves in a rotation-only manner, based on the fact that the scene is relatively far-away from the view point. The system is able to create the panoramic map of the environment at the same time to track the position and orientation of the camera with respect to the generated map.

The system is tested on a campus environment with various real-world tracking targets. One of the limitations is that since the pose estimation stage in both the near-field and far-field tracking approaches is based on the feature detection and matching algorithm, the tracking fails

when there are insufficient features to be extracted from the camera view. The frame rate of the system is at 15–30 fps on average as discussed in Section 6. In order to boost the speed of the algorithm, GPU and SIMD optimization on the mobile device could be adopted in the future. For further development, the tracking system is expected to be connected to a large database server, and a real-time outdoor augmented reality system for large-area outdoor application is to be built on top of this research work.

References

1. Bay H, Tuytelaars T, Gool LV (2008) SURF: speed up robust features. *Comp Vision Image Underst (CVIU)* 110(3):346–359
2. Benhimane S, Malis E (2007) Homography-based 2D visual tracking and servoing. *Int J Robot Res* 26(7): 661–676
3. Bleser G, Stricker D (2008) Advanced tracking through efficient image processing and visual-inertial sensor fusion. In *Virtual Reality Conference 2008 (VR'08)*, NE, USA, pp. 137–144
4. Calonder MV, Lepetit V, Strecha C, Fua P (2010) BRIEF: binary robust independent elementary features. In *11th European Conference on Computer Vision*, Heraklion, Greece, pp. 778–792
5. Feiner S, Macintype B, Hollerer T, Webster T (1997) A touring machine: prototyping 3D mobile augmented reality systems for exploring the urban environment. In *Proc. Of International Symposium on Wearable Computers (ISWC)*, Cambridge, Massachusetts, 13–14 Oct, pp. 74–81
6. Fischler MA, Bolles RC (1981) Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Commun of ACM* 24(6):381–395
7. Gammeter S, Gassmann A, Bossard L (2010) Server-side object recognition and client-side object tracking for mobile augmented reality. In *2010 I.E. Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, San Francisco, USA, pp. 1–8
8. Hwangbo M, Kim J, Kanade T (2009) Inertial-aided KLT feature tracking for a moving camera. In *IEEE/RSSJ International Conference on Intelligent Robots and Systems*, 2009, St. Louis, USA, pp. 1909–1916
9. Kurz D, Benhimane S (2011) Inertial sensor-aligned visual feature descriptors. In *2011 I.E. Conference on Computer Vision and Pattern Recognition (CVPR'11)*, Colorado, USA, pp. 161–166
10. Lee W, Park Y, Lepetit V, Woo W (2010) Point-and-shoot for ubiquitous tagging on mobile phones. In *IEEE International Symposium on Mixed and Augmented Reality (ISMAR'10)*, Seoul, Korea, pp. 57–64
11. Lin L, Wang Y, Liu Y, Xiong C, Zeng K (2009) Marker-less registration based on template tracking for augmented reality. *Multimed Tools Appl J(MMTA)* 41(2):235–252
12. Lowe DG (2004) Distinctive image features from scale-invariant keypoints. *Int J Comput Vis* 60(2):91–110
13. Özyual M, Fua P, Lepetit V (2007) Fast keypoint recognition in ten lines of code. In *2007 I.E. Conference on Computer Vision and Pattern Recognition (CVPR'07)*, Minneapolis, USA, pp. 1–8
14. Reitmayr G, Drummond T (2006) Going out: robust model-based tracking for outdoor augmented reality. In *Proc. Of International Symposium on Mixed and Augmented Reality (ISMAR '06)*, Santa Barbara, 22–25 Oct, pp. 109–118, 2006
15. Rosten E, Drummond T (2006) Machine learning for high-speed corner detection. In *Proceedings of European Conference on Computer Vision*, Graz, Austria, pp 430–443
16. Rublee E, Rabaud V, Knolige K, Bradski G (2011) ORB: an efficient alternative to SIFT or SURF. In *2011 I.E. International Conference on Computer Vision (ICCV'11)*, Barcelona, Spain, pp. 2564–2571
17. Schall G, Wagner D, Reitmayr G, Taichmann E, Wieser M, Schmalstieg D, Hofmann WB (2009) Global pose estimation using multi-sensor fusion for outdoor augmented reality. In *2009 I.E. International Symposium on Mixed and Augmented Reality (ISMAR'09)*, Orlando, USA, pp. 153–162
18. Silveira G, Malis E (2007) Real-time visual tracking under arbitrary illumination changes. *Proceedings of IEEE Computer Vision and Pattern Recognition (CVPR)*, Minneapolis, pp 1–6
19. Ventura J, Hollerer T (2012) Wide-area scene mapping for mobile visual tracking. In *IEEE International Symposium on Mixed and Augmented Reality (ISMAR'12)*, Atlanta, 5–8 Nov, pp 3–12
20. Wagner D, Mulloni A, Langlotz T, Schmalstieg D (2010) Real-time panoramic mapping and tracking on mobile phones. In *2010 I.E. Virtual Reality Conference (VR'10)*, Massachusetts, USA, pp. 211–218
21. Wagner D, Reitmayr G, Mulloni A, Drummond T, Schmalstieg D (2010) Real-time detection and tracking for augmented reality on mobile phones. *IEEE Trans Vis Comput Graph* 16(3):355–368



Lu Yu received a B.Eng. degree in Mechanical and Aerospace Engineering from Nanyang Technological University, Singapore in Jul. 2010. She is a Ph.D candidate in the Augmented Reality and Assistive Technology Lab in Mechanical Engineering Department, National University of Singapore since Aug. 2010. Her current research interests include vision-based tracking technology in augmented reality, hybrid tracking solution combining CV tracking with GPS and inertial sensors, and mobile application of augmented reality in outdoor environment.



Soh Khim Ong is an associate professor in the Department of Mechanical Engineering, and Faculty Member of the NUS Graduate School for Integrative Sciences and Engineering. In November 2003, she was the first ASEAN woman selected to receive the M. Eugene Merchant Outstanding Young Manufacturing Engineer Award from the Society of Manufacturing Engineers. In August 2012, she is elected a Fellow of CIRP, The International Academy for Production Engineering, where she is the first from the Asia region and the 4th female fellow in the world to be elected. She has also received various other awards including the 2004 Singapore Youth Award for the Science & Technology Category, The Outstanding Young Person of Singapore Award for the Science & Technology Category, the 2009 Emerging Leaders Award in Academia by the US Society for Women Engineers, and the 2013 Singapore Youth Award Medal of Commendation. She was appointed by Singapore President SR Nathan to serve as a Nominated Member of the Singapore Parliament and she served as NMP from 2nd January 2005 to 19th April 2006. She is an Asia Society Asia 21 Young Leader.



A.Y.C. Nee is professor in the Department of Mechanical Engineering, National University of Singapore since 1989. He received his PhD and DEng from Manchester and UMIST respectively. His research interest is in CAD of tool, die, fixture and process planning, augmented reality applications in manufacturing, sustainable manufacturing. He is a Fellow of CIRP (International Academy for Production Engineering) and SME (Society of Manufacturing Engineers), both elected in 1990, and a Fellow of the Academy of Engineering Singapore. He served as the President of CIRP in 2012, and a recipient of SME's Gold Medal in 2014. He has published over 500 refereed journal and conference papers and 12 edited and authored books. Currently, he is the Editor-in-Chief of Springer's International Journal of Advanced Manufacturing Technology as well as serving on the editorial boards of some 20 international journals.