Benchmarking Inertial Sensor-Aided Localization and Tracking Methods

Daniel Kurz* Sebastian Lieberknecht† Selim Benhimane‡

metaio GmbH

Figure 1: We present a methodology to evaluate inertial sensor-aided feature descriptors, such as GAFD [17] (left) and GREFD [16] (center), on benchmark datasets, that do not contain any inertial sensor measurements. At the example of metaio's template tracking benchmarking set [18], we show how synthesizing inertial sensor measurements from the ground truth poses enables the evaluation of such methods at virtually no extra cost while providing comparable results to using real inertial sensor data, which was validated using the setup in the right photo.

Abstract

This paper investigates means to benchmark methods for camera pose localization and tracking that in addition to a camera image make use of inertial sensor measurements. In particular the direction of the gravity has recently shown to provide useful information to aid vision-based approaches making them outperform vision-only methods. Obviously, it is desirable to benchmark the performance of such methods and to compare them with state-of-the-art approaches, but to the best of our knowledge, all publicly available benchmarking datasets unfortunately lack gravity information.

We present different simple means to generate one's own benchmarks for inertial sensor-aided localization and tracking methods and most considerably show how existing datasets, that do not have inertial sensor data, can be exploited. We demonstrate how to evaluate Gravity-Aligned Feature Descriptors (GAFD) and Gravity-Rectified Feature Descriptors (GREFD) on an existing benchmark dataset with ground truth poses. By synthesizing gravity measurements from these poses we achieve similar results to using real sensor measurements at significantly less effort. Most importantly, the proposed procedure enables the comparison with existing evaluation results on the same data. The paper concludes with a requirements analysis and suggestions for the design of future benchmarking datasets for localization and tracking methods.

1 Introduction and Motivation

A fundamental task in video-see-through Augmented Reality (AR) is camera pose localization and tracking. To be able to render virtual 3D content precisely registered with a real object visible in a camera image, the position and orientation of the camera with respect to this object must be known. Classical computer vision methods aim to determine the camera pose based on information in the camera image only. For such methods it does theoretically not make any difference if the camera moves or the object moves or both move. They only consider the relative transformation between the camera and the real object. In many applications, however, it is known that only the camera moves while the object does not. This in particular applies for handheld AR applications, when the real object is the entire environment or part of it.

If this is known to be the case, inertial sensors attached to the camera can be used to aid computer vision-based methods. Both relative and absolute orientation measurements can be useful and applied in different ways. While relative values, such as the rotation rate measured with a gyroscope, can be used to improve frame-to-frame tracking, e.g. by providing better priors [5], absolute measurements, such as the direction of the gravity, also allow for improving the initialization of a localization and tracking system.

Different approaches to vision-based camera localization and tracking have been proposed in the literature that make use of inertial sensors. In this paper, we will focus on two approaches to aid feature detection, description and matching by measuring the direction of the gravity, that are highly suitable for handheld AR. Gravity-Aligned Feature Descriptors (GAFD) [17] can be applied for static and (close to) vertical surfaces, such as building façades, cf. figure 1 (left). They use the direction of the gravity as canonical feature orientation instead of the dominant gradient direction around the feature as in standard rotation-invariant approaches. Thereby discriminative power between similar features at different orientations is improved. For (close to) horizontal surfaces, such as a magazine lying on a table, we recently proposed Gravity-Rectified Feature Descriptors (GREFD) [16] that rectify the camera image based on the measured gravity prior to the detection and description of features, see figure 1 (center). This results in an improved matching precision, particularly under steep viewing angles.

After discussing related work in section 2, we will explain simple means to create benchmark tests for the feature description methods explained above that include gravity measurements. In section 4, we will explain how to exploit existing benchmark schemes with given ground truth poses to enable the evaluation of inertial sensor-aided approaches. For future benchmarking datasets, it makes sense to include inertial sensor data and in general as much information as available. This is discussed in section 5 in more detail. Eventually, the final section 6 concludes the paper.

*e-mail: daniel.kurz@metaio.com
†e-mail: sebastian.lieberknecht@metaio.com
‡e-mail: selim.benhimane@metaio.com

2 Related Work

Visual tracking has been an active research area since decades. Over the years, a variety of methods were proposed, such that eventually common benchmarks were introduced where ground truth concerning the camera pose or image homographies were given. Such a benchmark can be based either on synthetically rendered images or on images captured by a real camera. For example, Baker and Matthews [3] used synthetic image warping to compare four template tracking algorithms by varying the warping amplitude and noise level of the image.

While creating synthetic images has the advantage that every parameter involved can be set with very high precision, the creation of realistic images is a challenging task as there are numerous effects involved in the physical imaging process, like non-linear sensitivity to light of the camera sensor, motion blur, photon-based lighting, sensor noise, discretization, non-synchronous pixels due to rolling shutter, blooming or limited color depths. In the following, we thus concentrate on datasets generated from real image data, which computer vision methods are ultimately designed for.

Mikolajczyk and Schmid [20] used still images to compare affine region detectors. Their dataset consists of eight sets of six images each and was used by many others (e.g. [9, 4, 22]). The homographies that relate the images of each set were computed from manual correspondences which were afterwards refined automatically.

Moreels and Perona [21] generated a database for 3D objects based on a turntable setup using a static stereo camera. One image pair was captured for each 5° rotation of the turntable.

None of the aforementioned datasets was designed for handheld AR, and consequently they do not embody the effects of a rather unconstrained 6 DoF motion on the live image acquisition of a camera. Zimmerman, Matas and Svoboda [23] published a dataset which partly relied on a handheld camera. Ground truth homographies were created by manually defined correspondences for all images. To our knowledge, this was the first dataset that can be used to quantitatively evaluate the precision and accuracy of detection and tracking methods given real images from a handheld camera.

Similarly, Gauglitz et al. [13] used a setup that relies on color-coded balls on the plane of the target to estimate ground truth homographies, which allows for automatic image alignment as long as all four balls are visible. The setup was used to evaluate different combinations of feature descriptors and detectors and also feature orientation assignment strategies [14] based on image intensities.

Lieberknecht et al. [18] presented a dataset where the ground truth camera frames are based on an mechanical measurement arm which operates at sub-millimeter precision. Despite the necessity of having fiducials visible in part of the sequences for the synchronization between captured poses of the arm and captured images from the camera mounted on its end effector, no constraint is imposed on the motion of the camera or lighting conditions. The dataset uses eight targets belonging to four different texturedness levels and consists of five similar types of motions. Each target and motion is represented by 1200 images. The dataset has originally been used to evaluate four detection and tracking methods and was later used by other researchers, e.g. [10, 19, 11].

Gruber et al. [15] presented a dataset which contains the ground truth pose data from a mechanical measurement arm, an outside-in infrared tracking system and a coordinate measurement system. The benchmark is called "City of Sights" as it consists of 3D paper models of popular real buildings which can be easily rebuilt to allow others to e.g. perform qualitative evaluations or also extend the dataset given appropriate reference tracking systems.

The very recently published dataset by Chen et al. [7] was used to evaluate an accompanying image-based landmark identification algorithm. The dataset was recorded on the streets of San Francisco and consists of 1.7 M images created from 150k panoramic images. For the panoramic images, a multitude of additional measurements was recorded such as spherical LIDAR data, GPS, inertial measurements and distance measurements. There are further 800 geo-tagged query images captured with mobile phones. The authors assume that there is a coarse inertial estimate in the form of 90° steps (e.g. landscape or portrait mode) and use a fixed feature description alignment.

On a separate track, there are also specialized projects online which deal with the e.g. variable exposure or panorama video [12] or multi-sensor data originating from robotic SLAM systems [1]. However, up to now there is no dataset available which consists of real images from a handheld camera that include high-resolution readings from inertial sensors.

3 Creating Simple Benchmarks

The goal of a benchmark for localization and tracking is, given a camera pose determined by a tracking system, to be able to tell if it is correct or not. The most reliable method is to compare the determined pose with a ground truth pose.

As presented in the previous section, one way to gain ground truth poses for real camera images is to use an additional tracking system which is considered very accurate and reliable to determine the camera pose. For instance, a mechanical measurement arm is mounted to the camera in [18] measuring ground truth poses. While this procedure provides very accurate results, it is in general very time-consuming and requires both expensive hardware and accurate calibration.

In the following, we will take a look at the benchmark methods used in [16] which are less costly and discuss their pros and cons. The captured data in these evaluations comprises camera images or sequences of camera images taken with a mobile phone and the corresponding measurement of the gravity vector for each image.

3.1 Fully Automatic Pose Verification

It is possible to benchmark camera poses without any ground truth data by evaluating an error function which is assumed to correspond to the accuracy of the pose. In [16], feature-based localization of a planar template using GAFD or GREFD is evaluated based on the Zero-mean Normalized Cross Correlation (ZNCC) between the reference template and the current camera image warped with the homography which led to the computed pose. The ZNCC values can eventually either be used as a continuous quality measure or be thresholded to decide if a localization was correct or not.

This method is very convenient, as it does not require any manual work nor any special hardware. It can be used to evaluate huge amounts of images and corresponding poses very efficiently. But unfortunately, it can be unreliable in particular cases. An obvious example are repetitive structures in the environment. If the method is for instance supposed to determine the camera pose with respect to one particular window at a building façade and there are many similar looking windows, it is impossible for this method to tell if the correct window was chosen as reference or not.

3.2 Manual Ground Truth Generation

Another option to gain ground truth poses is by manually defining the position of certain points of the reference model in every camera image. For planar objects the transformation an object undergoes when imaged can be fully described with a homography, which is a (3×3) matrix. This transformation can be computed with a closed-form solution given only four corresponding points. Based on this homography, not only a computed pose but also point correspondences between reference and current images can be easily validated, as we did in [16] to benchmark feature descriptors by their matches. For arbitrarily shaped (i.e. non-planar) objects, at least four correspondences need to be manually defined per image for the computation of a unique ground truth pose.

While this approach provides in theory all information needed for reliable benchmarks, it is in practice only applicable for small datasets. Not only is the manual definition of points, e.g. by clicking on them, very time-consuming but also it is tedious and therefore error-prone. Ideally, all manually defined correspondences need to be manually double-checked which makes this procedure impractical for image sets containing thousands of images. In addition to that, motion blur, defocus, partial occlusions and camera poses for which particular reference points are not visible in the camera image make the precise manual definition of their position in the image impossible.

4 Exploiting Existing Benchmarks

As can be seen in section 2, a variety of benchmark datasets already exist. However, none of them contains inertial sensor data making an evaluation of inertial sensor-aided methods impossible. Even the very recently published dataset by Chen *et al.* [7] containing several hundred images captured with mobile phones having GPS information, does not provide inertial sensor data. For the publicly available dataset by metaio [18], there exist ground truth poses for every frame gained with a measurement arm attached to the camera, but no gravity information. Creating new sequences with a similar setup does not only require expensive hardware but also takes a lot of effort. Most importantly, it will result in new sequences that do not enable comparison with methods that were tested on the original dataset. If it was possible, we would ideally not take new sequences but instead record the inertial sensor measurements for the already existing sequences to ensure comparability.

4.1 Proposed Benchmarking Method

We propose to synthesize inertial sensor measurements from the ground truth pose provided in existing benchmark datasets. Knowing the camera pose allows for transformation of any vector in a world coordinate system, including an arbitrarily defined gravity vector, into the camera coordinate system. The transformed vector can then be used in the same way an algorithm would use the gravity vector measured with inertial sensors. In this paper, we evaluate this procedure at the example of metaio's template benchmark [18]. We study the effect of both GAFD [17] and GREFD [16] in comparison with regular rotation-invariant feature descriptors which try to assign a repeatable orientation to every feature based on local image gradients.

4.2 Comparison With Real Sensor Measurements

In order to validate of the proposed method, it is essential to compare the synthetic gravity vectors used in this approach with gravity vectors measured with real inertial sensors. Therefore, we rebuilt the setup used in [18] but replaced the industrial camera by a mobile phone (Apple's iPhone 4) which in addition to a camera is equipped with inertial sensors. The interested reader is referred to [18] for details on the calibration procedure of the setup.

Using the calibrated setup, we record a sequence of 500 frames showing the "Isetta" template and six fiducial markers located on a horizontal surface (as in figure 1) while alternating view points and camera angles. The images have a resolution of (480 × 360) pixels and are recorded at approximately 25 Hz. With every camera image, we store a time-stamp and the corresponding gravity vector measured with inertial sensors on the phone. In parallel, a PC that is connected to the measurement arm, stores corresponding ground truth camera poses.

As in the original paper, the synchronization of ground truth poses and camera images is done based on the detected corners of the markers visible in the camera image. After synchronization, the residual of the reprojected corners is 0.94 pixels which makes the sequence usable as ground truth data.

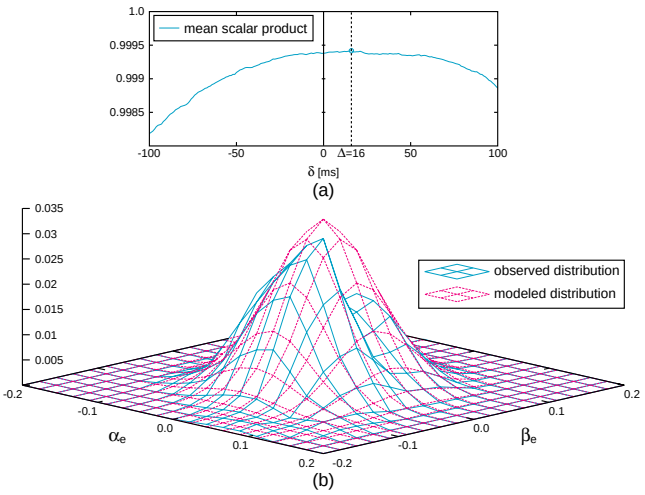

Figure 2: Correlation between the measured and the ground truth gravity vectors as a function of the delay between the two (a). The observed error distribution of the real gravity measurements can be modeled with a Gaussian distribution (b).

Now that we have camera images with corresponding gravity measurements and ground truth poses, we aim to compare the real gravity measurements $\mathbf{g}_{sensor}(t)$ with the corresponding ground truth vector $\mathbf{g}_{gtruth}(t) = -{}^{c}\mathbf{R}_{w}(t)\mathbf{z}_{w}$. As the print-out is located in a horizontal orientation, gravity corresponds to the the negative z-axis of the world coordinate system ($-\mathbf{z}_w$) which needs to be transformed to the camera coordinate system by the rotational part of the ground truth pose ${}^{c}\mathbf{R}_{w}$. Once we know the characteristics of the real data, we are able to synthesize gravity vectors $\mathbf{g}_{synth}(t)$ that behave comparably to the real measurements.

The degradation model we use comprises a noise term and a delay between the moment a camera image was taken and the point in time where the corresponding gravity vector was measured. Since the camera images were used for synchronization, they are in sync with the ground truth poses and therefore with the ground truth gravity vectors. Since this is not necessarily the case for real sensor measurements, we first measure the temporal offset between the real gravity measurements and the ground truth gravity vectors. To this end, we compute the delay Δ, for which the mean of the scalar product between the measured and the ground truth gravity vectors over all frames has its global maximum.

$$\Delta = \arg\max_{\delta} \left(\frac{1}{n} \sum_{t} \left(\mathbf{g}_{sensor}(t)^{\top} \cdot \mathbf{g}_{gtruth}(t+\delta) \right) \right)$$

Figure 2a plots this correlation as a function of the delay δ. As can be seen, for the device we used, the maximum is reached at Δ = 16 ms, which will be used as offset in the following.

After synchronization of the real measurements and the ground truth, we aim to quantify the distribution of the error between the noisy real measurements and the close to perfect ground truth. We therefore parametrize the normalized gravity vectors with two angles (α_s, β_s) and (α_g, β_g) which are computed as

$$\mathbf{g}_{sensor} = \begin{bmatrix} \cos\alpha_s \cos\beta_s \\ \sin\alpha_s \cos\beta_s \\ \sin\beta_s \end{bmatrix} \text{ and } \mathbf{g}_{gtruth} = \begin{bmatrix} \cos\alpha_g \cos\beta_g \\ \sin\alpha_g \cos\beta_g \\ \sin\beta_g \end{bmatrix}.$$

The distribution of the angular differences between sensor measurements and ground truth data $\alpha_e = (\alpha_s - \alpha_g)$ and $\beta_e = (\beta_s - \beta_g)$ is plotted in figure 2b and can be modeled with a two-dimensional

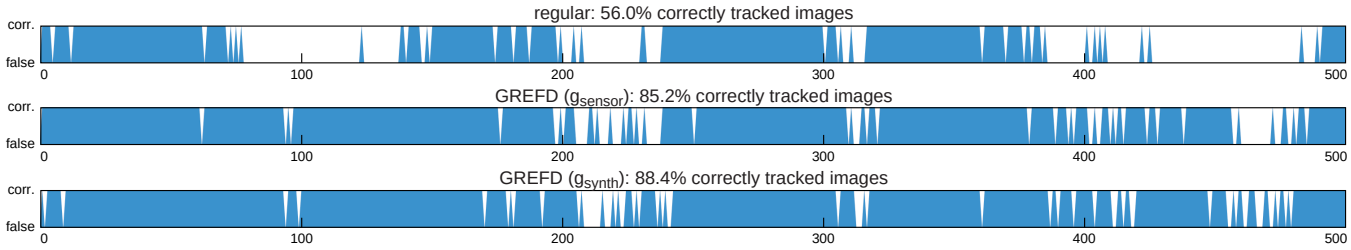

Figure 3: Distribution of correctly tracked images in a sequence of 500 frames showing the "Isetta" template using regular feature descriptors, real gravity measurements in GREFD($\mathbf{g}_{sensor}$) and the proposed method in GREFD($\mathbf{g}_{synth}$). The latter two methods provide comparable results.

Gaussian distribution parametrized with expected value $\mu = (0,0)$ and variance $\sigma^2 = 0.00092$. We use the BoxMuller transform [6] to compute noise with the modeled distribution of the real inertial sensors to synthesize gravity vectors as

$$\mathbf{g}_{synth}(t) = \begin{bmatrix} \cos(\alpha_g(t+\Delta)+X)\cos(\beta_g(t+\Delta)+Y) \\ \sin(\alpha_g(t+\Delta)+X)\cos(\beta_g(t+\Delta)+Y) \\ \sin(\beta_g(t+\Delta)+Y) \end{bmatrix}$$

where X and Y are two independent random variables with a normal distribution of standard deviation σ.

Using the above, we create synthesized gravity vectors that are comparable in terms of noise and delay to those provided by the inertial sensors of an iPhone 4. In order to validate that using these vectors also provides comparable results on the template localization performance, we ran our method on the above mentioned image sequence in three different configurations. First, we use regular feature descriptors to receive a reference performance measure. We then run the method again using GREFD and the real gravity vectors measured with inertial sensors (GREFD($\mathbf{g}_{sensor}$)). Finally, we run the test again using synthesized gravity vectors (GREFD($\mathbf{g}_{synth}$)). The results are shown in figure 3 and clearly confirm that GREFD performs similarly when using the proposed method compared with real sensor measurements both in terms of distribution and sum of the ratio of correctly tracked images, i.e. images with less 10 pixels reprojection error at the template corners.

In the following evaluation we will use both ground truth gravity vectors and those vectors synthesized as explained above and study the impact of the artificial degradation on the localization of planar templates.

4.3 Gravity-Aligned Feature Descriptors

Gravity-Aligned Feature Descriptors (GAFD) are designed for static and (close to) vertical surfaces, such as building façades, TV screens or billboards.

Therefore, we define the synthetic gravity vector $\mathbf{g}_{GAFD}$ as if the template was located on a vertical surface in an upright orientation. This means that the gravity vector corresponds to the negative y-axis ($-\mathbf{y}_w$) of the world coordinate system associated to the printout of the template, cf. figure 4. Given the ground truth pose, this vector is transformed to the camera coordinate system by multiplying it with the rotational part of the pose ${}^c\mathbf{R}_w$ and is then used as gravity vector. As explained above, we use a degradation model adding noise and a delay to the ground truth gravity vector in order to create synthesized gravity vectors with similar properties as real gravity measurements. The feature description algorithm finally projects the 3D gravity vector onto the image plane for every feature to compute its orientation.

The features and their orientation using GAFD are illustrated in the second row of figure 5 for some exemplary frames of the dataset.

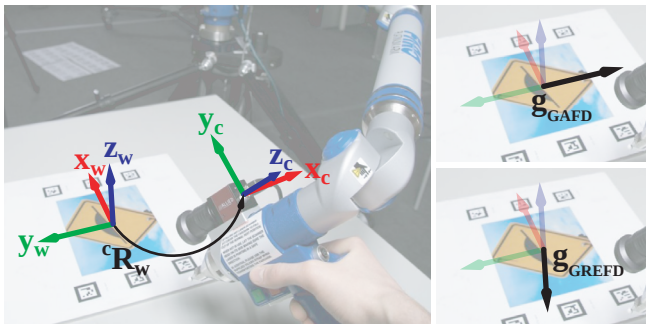

Figure 4: Visualization of the involved coordinate systems and the synthesized gravity vectors for the evaluation of GAFD and GREFD.

4.4 Gravity-Rectified Feature Descriptors

The purpose of Gravity-Rectified Feature Descriptors (GREFD) is to improve the description of features corresponding to physical points located on horizontal surfaces.

Therefore, we consider the template to be oriented horizontally for this evaluation and consequently compute the gravity vector $\mathbf{g}_{GREFD}$ as the negative z-axis ($-\mathbf{z}_w$) of the world coordinate system, as illustrated in the bottom right of figure 4. This vector is then again transformed into the camera coordinate system using the known ground truth pose and artificial degradation is applied resulting in a plausible synthesized gravity vector. Finally, this vector is used to rectify the camera image before detecting and describing features in it. The support regions of individual GREFD are displayed in figure 5 in the bottom row. Note, that the white quads correspond to squares in the gravity-rectified images.

4.5 Results and Discussion

The template localization system we use first detects, describes and matches local image features from the reference image and every query camera image using a custom 48-dimensional feature descriptor. Based on these matches, it then tries to estimate a homography between the reference image and the corresponding area in the current image using PROSAC [8]. If successful, this homography is eventually refined using the Inverse Compositional [2] image registration method before computing the camera pose. All algorithms we use were optimized to run in real-time on handheld devices such as consumer mobile phones.

For the evaluation results presented in this paper, we chose two templates from [18] which we believe correspond to those kinds of templates that are frequently used in handheld AR applications and therefore are particularly relevant. The "Stop" template represents objects with a low texturedness while the "Isetta" has a normal density of texture comparable with many real-life objects. The results for all five sequences per template using regular rotation-invariant feature descriptors, GAFD and GREFD are displayed in figure 6.

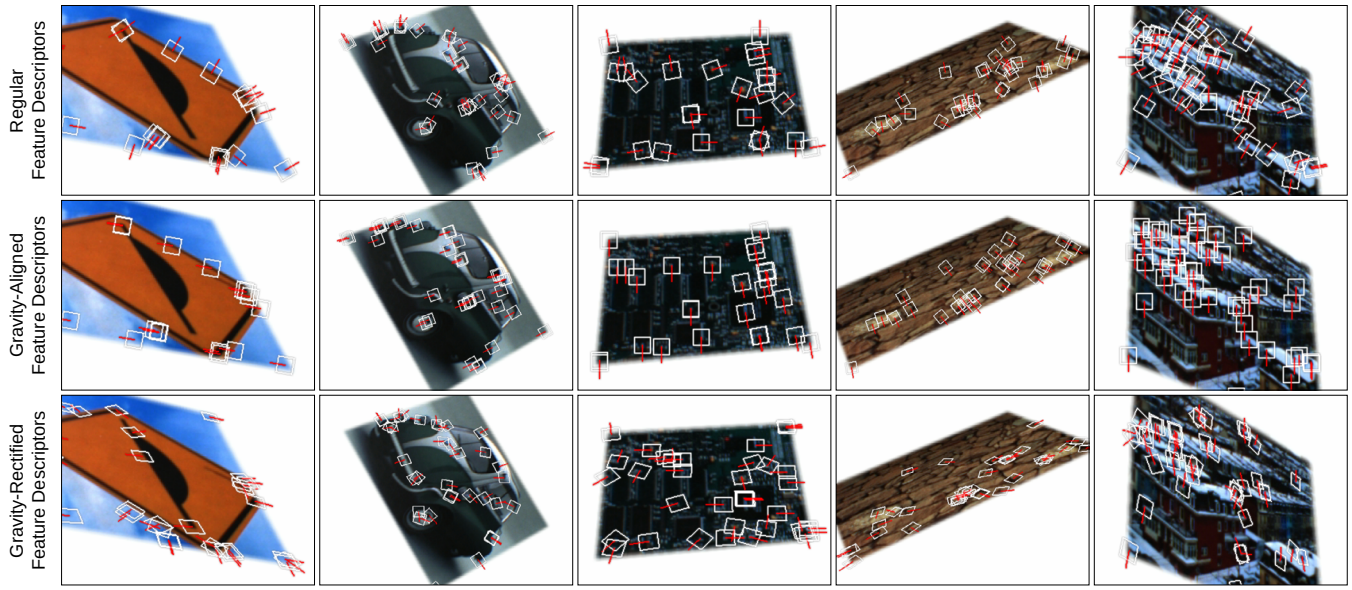

Figure 5: Visualization of the feature descriptor support regions (white) and their orientations (red) for regular descriptors (top row), GAFD (central row) and GREFD (bottom row) in different images of the used benchmark dataset. Note, that some of the images have been cropped for illustration purposes only.

The latter two methods were tested both with the ground truth gravity vectors ($\mathbf{g}_{gtruth}$) and the artificially degraded vectors ($\mathbf{g}_{synth}$) which simulate the inertial sensors in a mobile phone. We plot the ratio of images in which the localization was successful for each sequence and each method used. Similarly to [18], we consider the template as being correctly localized if the average reprojection error of the four template corners using the estimated camera pose is below 10 pixels.

Over all image sequences, the number of correct frames increases when using the inertial sensor-aided feature descriptors compared to regular ones. We observe, that artificial degradation of the ground truth vectors only has a marginal effect on the performance of both GAFD and GREFD. Therefore, the following analysis will only consider the results of the two methods using ($\mathbf{g}_{synth}$).

The relative increase in images where the template could be localized correctly when using Gravity-Aligned Feature Descriptors is on average 35.91% which is comparable to the results in [16] where real sensor measurements of the gravity vector have been used on a significantly smaller dataset. The most significant improvements using GAFD can be observed for the "Stop" template, which has a good deal of similar looking features at different orientations at the borders of the sign. As GAFD outperforms regular feature descriptors for all sequences in the test, it is clearly useful and universally applicable whenever dealing with (close to) vertical surfaces.

The use of GREFD increased the overall ratio of correctly localized images less significantly. As also discovered in the original work, these descriptors only make sense for steep camera angles while they sometimes even perform worse than regular feature descriptors for camera poses close to perpendicular to the template. On average over all image sequences used, the angle between the principal axis of the camera and the template normal is only 33.00°. In fact, GREFD only provides satisfying results in the first sequence of both templates, which is the "Angle" sequence. Here, the average angles between the template normal and the camera are 52.80° and 46.89° making them by far those sequences with the steepest angles. Again, the results confirm those gained earlier using real gravity measurements and benchmarking methods explained in 3.

An interesting characteristic of the proposed method is, that it allows to directly compare GAFD with GREFD. Note, that this would not be possible using real sensor measurements as GAFD works for (close to) vertical surfaces only while GREFD is designed for (close to) horizontal surfaces. Using the proposed scheme to synthesize gravity vectors allows for an arbitrary definition of the gravity vector in the world coordinate system. Therefore we are able to treat the same image sequence as if it was capturing a horizontal template for GREFD or a template which is located at a vertical surface in an upright orientation for GAFD.

5 Suggestions for the Design of Future Benchmarking Datasets

Looking at real applications and state-of-the-art computer vision technology for (handheld) Augmented Reality, it is clearly not sufficient to provide a model of the scene and images or video sequences with corresponding ground truth camera poses to benchmark localization and tracking methods.

While the preceding section explains a functional way to evaluate inertial sensor-aided feature descriptors and possibly other methods relying on the gravity without having access to real sensor data, this does not mean that future benchmarking datasets should ignore the existence of inertial sensors equipped to virtually any handheld device nowadays.

We believe that methods that make use of auxiliary data to aid computer vision algorithms will become increasingly more important and complex. Particularly in large-scale outdoor environments it is indispensable to combine different sources of information to achieve the ultimate goal of a precise camera pose estimation at any place on earth. This not only is interesting from a scientific point of view but most importantly it is critical for the long-term success of Augmented Reality browsers such as junaio[1]. In order to make a benchmarking dataset useful and applicable in the long run, we therefore propose to capture and save all available information with every image. This includes not only the data that state-of-the-art algorithms make use of but also data that we currently cannot even think of how to use it but which may be useful in the future.

[1] http://www.junaio.com

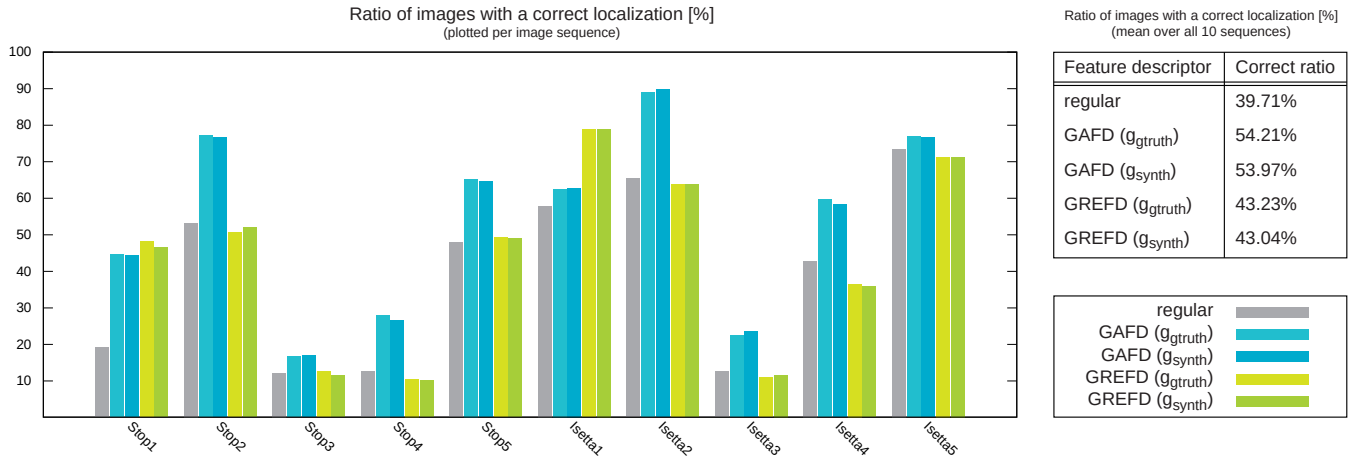

Figure 6: Results comparing GAFD and GREFD with regular rotation-invariant local feature descriptors. While GAFD outperforms regular descriptors for virtually all sequences, the performance of GREFD is strongly depending on the steepness of the camera angles used.

Useful data that should be included with future benchmarking datasets includes but is not limited to

- Device used (manufacturer, model, firmware version)
- Inertial sensor data (gravity, user acceleration, rotation rate)
- Digital compass data (heading, accuracy)
- GPS data (coordinate, altitude, accuracy)
- Camera properties (shutter time, gain, focus, etc.)
- Flash (on/off/auto, did fire/did not fire)
- Image regions for auto exposure and auto focus
- Accurate and consistent time-stamps for all the above
- Available networks (WiFi, GSM, etc.)
- Time and date
- Corresponding weather situation

Besides providing auxiliary data, it is important to use plausible hardware for the targeted field of application to ensure realistic results. As we are mainly focusing on handheld AR applications, we propose to use the latest consumer handheld devices, such as mobile phones or tablet PCs.

To qualify for outdoor scenarios, it is also critical for a tracking method to be robust against significant changes in illumination and shadows. An ideal benchmark dataset for such methods would contain camera images taken at different times of the day and different weather and light situations, as it is the case in [12].

6 Conclusions and Future Work

We presented a methodology to benchmark inertial sensor-aided localization and tracking methods using benchmarking datasets that do not provide inertial sensor data. If the ground truth poses are given, we have shown how these can be used in combination with a degradation model to synthesize gravity vectors that behave comparable to real inertial sensor measurements. We validated the approach by capturing a new image sequence with ground truth poses using a phone with built-in inertial sensors. The tested template localization method performed similarly when using synthesized and real gravity measurements. The evaluation results do not only confirm the findings about the two examined feature descriptors GAFD and GREFD gained earlier on smaller datasets, but also can be seen as a field test of the proposed approach.

To enable the research community to evaluate their own inertial sensor-aided localization and tracking methods on the metaio dataset, we publicly provide the synthetic gravity vectors $\mathbf{g}_{GAFD}$ and $\mathbf{g}_{GREFD}$ of selected sequences on our research website[2] in addition to the image sequences, reference templates and intrinsic parameters of the camera.

Obviously the proposed method can be extended to synthesize further degrees of freedom, e.g. compass, in future work. In addition to absolute sensor measurements, also relative measures such as the rotation rate as measured with a gyroscope could be synthesized using the proposed technique by incorporating the ground truth poses of multiple consecutive frames. Finally, the preceding section acts as an inspiration for future work on creating new benchmarking datasets for camera pose localization and tracking algorithms. It is clearly important to have standardized and publicly available datasets to enable the comparison of tracking systems developed by research institutions and researching companies around the world. In order to cover the largest range of tracking methods possible, future datasets clearly need to be as universal as possible providing as many auxiliary information as available, including inertial sensors and beyond.

References

[1] Tracking data repository. http://www.rawseeds.org, 2011.

[2] S. Baker and I. Matthews. Equivalence and efficiency of image alignment algorithms. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2001.

[3] S. Baker and I. Matthews. Lucas-kanade 20 years on: A unifying framework. *International Journal on Computer Vision*, 56(3):221 – 255, 2004.

[4] H. Bay, A. Ess, T. Tuytelaars, and L. Van Gool. SURF: Speeded up robust features. *Computer Vision and Image Understanding*, 110(3):346–359, 2008.

[5] G. Bleser and D. Stricker. Advanced tracking through efficient image processing and visual-inertial sensor fusion. *Computer & Graphics*, 33, 2009.

[6] G. E. P. Box and M. E. Muller. A note on the generation of random normal deviates. *The Annals of Mathematical Statistics*, 29(2):610 – 611, 1958.

[7] D. M. Chen, G. Baatz, K. Köser, S. S. Tsai, R. Vedantham, T. Pylvanainen, K. Roimela, X. Chen, J. Bach, M. Pollefeys, B. Girod, and R. Grzeszczuk. City-scale landmark identification on mobile devices.

[2] http://www.metaio.com/research

In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2011.

[8] O. Chum and J. Matas. Matching with prosac - progressive sample consensus. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2005.

[9] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2005.

[10] A. Dame and E. Marchand. Accurate real-time tracking using mutual information. In *Proc. IEEE/ACM International Symposium on Mixed and Augmented Reality*, 2010.

[11] J. Dupac and J. Matas. Ultra-fast tracking based on zero-shift points. In *Acoustics, Speech and Signal Processing*, 2011.

[12] Four Eyes Lab, University of California, Santa Babara. Tracking data repository. http://tracking.mat.ucsb.edu, 2011.

[13] S. Gauglitz, T. Höllerer, and M. Turk. Evaluation of interest point detectors and feature descriptors for visual tracking. *International Journal of Computer Vision*, 94(3):335–360, 2011.

[14] S. Gauglitz, M. Turk, and T. Höllerer. Improving keypoint orientation assignment. In *Proc. British Machine Vision Conference*, 2011.

[15] L. Gruber, S. Gauglitz, J. Ventura, S. Zollmann, M. Huber, M. Schlegel, G. Klinker, D. Schmalstieg, and T. Höllerer. The city of sights: Design, construction, and measurement of an augmented reality stage set. In *Proc. IEEE/ACM International Symposium on Mixed and Augmented Reality*, 2010.

[16] D. Kurz and S. Benhimane. Gravity-aware handheld augmented reality. In *Proc. IEEE/ACM International Symposium on Mixed and Augmented Reality*, 2011.

[17] D. Kurz and S. Benhimane. Inertial sensor-aligned visual feature descriptors. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2011.

[18] S. Lieberknecht, S. Benhimane, P. Meier, and N. Navab. A dataset and evaluation methodology for template-based tracking algorithms. In *Proc. IEEE/ACM International Symposium on Mixed and Augmented Reality*, 2009.

[19] R. Megret, J. Authesserre, and Y. Berthoumieu. Bidirectional composition on lie groups for gradient-based image alignment. *Transactions on Image Processing*, 19(9):2369–2381, 2010.

[20] K. Mikolajczyk and C. Schmid. Scale and affine invariant interest point detectors. *International Journal of Computer Vision*, 60(1):63–86, 2004.

[21] P. Moreels and P. Perona. Evaluation of features detectors and descriptors based on 3d objects. *International Journal of Computer Vision*, 73(3):263–284, 2007.

[22] D. Nister and H. Stewenius. Scalable recognition with a vocabulary tree. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2006.

[23] K. Zimmerman, J. Matas, and T. Svoboda. Tracking by an optimal sequence of linear predictors. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(4):677–692, 2009.