Robust 6-DOF Immersive Navigation Using Commodity Hardware

L. Carozza*
School of the Built Environment
Heriot-Watt University

F. Bosché†
School of the Built Environment
Heriot-Watt University

M. Abdel-Wahab‡
School of the Built Environment
Heriot-Watt University

Abstract

In this paper we present a novel visual-inertial 6-DOF localization approach that can be directly integrated in a wearable immersive system for simulation and training. In this context, while CAVE environments typically require complex and expensive set-up, our approach relies on visual and inertial information provided by commodity hardware, *i.e.* a consumer monocular camera and an Inertial Measurement Unit (IMU).

We propose a novel robust pipeline based on state-of-the-art image-based localization and sensor fusion approaches. A loosely-coupled sensor fusion approach, which makes use of robust orientation information from the IMU, is employed to cope with failures in visual tracking (*e.g.* due to camera fast motion) in order to limit motion jitters. Fast and smooth re-localization is also provided to track position following visual tracking outage and guarantee continued operation. The 6-DOF information is then used to render consistently VR contents on a stereoscopic HMD. The proposed system, demonstrated in the context of Construction, runs at 30 fps on a standard PC and requires a very limited set-up for its intended application.

Keywords: 6-DOF navigation, visual-inertial tracking, natural walking, commodity hardware

1 Introduction

Recent advances in the simulation capabilities of VR systems have stirred up the interest for immersive simulation and training in different fields. Indeed, immersive environments can be used to simulate varying operative scenarios, so that the user can experience critical situations and interact with them without being exposed to health and safety hazards. In this context, navigation of virtual environments can be achieved by assigning the user's viewpoint through input devices (*e.g.* joystick, 3D mouse), and/or by tracking the natural walking (head) of the user. In particular, the latter type of interface is intuitively more natural and can provide, by involving visual, vestibular and proprioceptive systems, a more consistent and comfortable perception of the explored environment. Benefits in terms of increased spatial awareness, presence, and reduced cybersickness have been emphasized in different works [Chen et al. 2013].

Accordingly, the *localization* stage, which aims at estimating in real time the position and orientation of the user's head during his movements, is of critical importance. Existing VR applications have emphasized how *robustness*, *accuracy* and *precision*, as well as real-time performance, still represent crucial open issues [Welch and Foxlin 2002]. Similarly, complexity in system set-up, range scalability, as well as cost effectiveness are deemed relevant criteria for the design and the assessment of tracking systems. CAVE systems [Cruz-Neira et al. 1992], which represent the standard for 3D immersive environments, often implement head tracking by tracking IR markers through multiple cameras/sensors [Welch and Foxlin 2002], with the rendered scene projected on wide screens surrounding the user. Existing commercial systems employed in such environments require dedicated facilities, on-purpose calibration and set-up procedures, with significant impact on the overall complexity and cost.

For these reasons, recent research efforts have aimed to reach a good trade off between acceptable performance and overall system complexity and cost. In particular, recent developments in computer vision, HMD and other technologies are paving the way for a wide diffusion of commodity devices that can be integrated into systems that are robust and very cost-effective. For example, a conceptual demonstration of the potential use of consumer hardware (Nintendo's Wii games console) in optical tracking has been presented in [Hay et al. 2008].

In this work an *inside-out* tracking approach, relying on the complementary action of visual and inertial tracking, is proposed. The 6-DOF pose of the trainee's head is estimated by robustly integrating visual information acquired by a monocular camera and inertial data provided by an Inertial Measurement Unit (IMU) integral with a stereoscopic HMD (Fig. 1). The main contribution consists in a novel localization pipeline conceived to cope with fast changes in motion patterns and limit drift and jitter effects, so to minimize system outage and provide consistent user experience. 6-DOF global localization is initially achieved through image registration with respect to a 3D *map* of visual descriptors of the training room, built off-line in advance using Structure from Motion (*SfM*). A feature tracking strategy exploiting spatio-temporal contiguity among consecutive video frames is employed to track the pose in real time preserving robustness over prolonged periods. In addition, orientation data provided by the HMD's IMU at high rate (1 kHz) are jointly employed to estimate the pose, acting as the main sensor when visual tracking fails. A loosely-coupled sensor fusion strategy is used in order to filter all the data and stabilize trajectory.

The effectiveness of our system (whose hardware cost is around 500$) is demonstrated in the context of natural navigation in VR scenarios for Construction training.

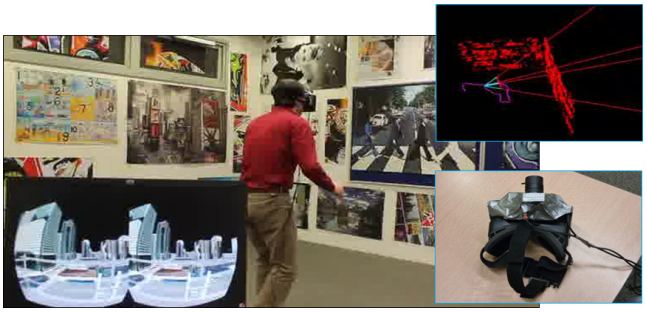

Figure 1: *Illustration of the main components of the proposed immersive sytem.*

*e-mail: L.Carozza@hw.ac.uk
†e-mail:F.N.Bosche@hw.ac.uk
‡e-mail:M.Abdel-Wahab@hw.ac.uk

2 Related Works

In view of the considerations above, we focus on two on-line localization methods, *i.e.*vision-based global localization and inertial tracking, and their integration due to their complementary advantages.

Research on vision-based approaches has recently focused on landmark and model-based (*e.g.* CAD [Bleser and Stricker 2008]) global localization methods [Oskiper et al. 2011; Zhu et al. 2008]. Global localization approaches estimate the camera pose from 2D-3D correspondences of image features, extracted from the current image, with a set of 3D landmarks of the scene, encoded in a database of visual descriptors. This approach does not suffer from error accumulation (drift) and allows robust relocalization. However, limited matching accuracy, resulting from image poor quality (*e.g.* motion blur) and 3D reconstruction errors, can result in jitter. Different strategies have been adopted for implementing robustly and efficiently all the stages involved, *i.e.* scene encoding and database construction [Zhu et al. 2008; Lim et al. 2012], feature detection, description and matching [Gauglitz et al. 2011]. However, these methods do not overcome the inherent lack of robustness of vision-based localization methods to fast motion blur.

Inertial and vision-based approaches can benefit from each other, providing backup solutions in case of dropout of one of the two, or aiding each other. A considerable number of works (*e.g.* [Oskiper et al. 2011; Bleser and Stricker 2008; Aron et al. 2007]) have discussed different strategies to combine visual and inertial information. However, how to optimally fuse those data so to reach a good trade-off between complexity, computational load and overall robustness still represents an open issue [Bleser and Stricker 2008; Oskiper et al. 2011]. Several approaches rely on vision-based tracking as the main strategy that is then supported by inertial data when visual information is unreliable [Aron et al. 2007] (*e.g.* fast motion, occlusion, poor scene modelling). Alternatively, systems mainly relying on inertial tracking can be aided by guided visual registration. In [Bleser and Stricker 2008], the use of different kinds of inertial models has been investigated, also discussing the impact of integration of accelerometer data for position tracking. However, in general positional initialization requires a semi-automatic procedure, and due to the inherent dead reckoning effect, filter divergence must be robustly detected and handled.

3 Key Stages of the Proposed Approach

The proposed method relies on two fundamental stages. First, an *off-line visual reconstruction stage* is performed in advance, once and for all, to encode the visual structure of 3D natural landmarks present in the scene into a database of visual descriptors, or *map*, as described in Sect. 4. During on-line operations, an *hybrid localization approach* couples in an Extended Kalman Filter (EKF) framework the robust high-rate orientation data from the IMU with visual information from landmark matching and frame-to-frame tracking to robustly estimate the head's pose. Specific strategies are proposed to detect failures in visual tracking and relocalize the system, preserving real time performance, as detailed in Sect. 5.

The proposed method is supposed to work in a sufficiently textured environment, but without particular constraints about the scene's geometrical structure (*e.g.* not necessarily planar [Aron et al. 2007]). It has been assessed in a room whose walls have been covered with posters with a random layout, so without requiring installation of calibrated landmarks with specific configuration which can be complex and time-consuming. The system also does not require the calibration of multiple cameras. This test environment, even if of limited size, presents most of the challenging issues common to the localization problem in general contexts. Moreover, exploration of large virtual environments can still be achieved by means of techniques like *redirected walking* [Williams et al. 2007].

4 Off-Line Reconstruction Stage

Given an input sequence of images of the scene taken from different viewpoints, a sparse 3D reconstruction (point cloud) based on SIFT features is initially performed using the *Bundler* SfM framework [Snavely et al. 2008]. Due to the computational effort required by SIFT, which would affect time performance during on-line operations, an approach similar to the one employed in [Lim et al. 2012] is adopted to compute more efficient descriptors, with the aim of simultaneously preserving a good trade-off with robustness. Two different approaches in terms of detection, description and matching of visual features have been considered. The first approach is based, for both detection and description, on the widely used SURF features, which offer a good trade-off between robustness and computational performance. The second approach makes use of *binary* features for both detection, using ORB keypoints , and description, employing BRISK. In particular, binary features have the advantage of providing very fast detection as well as efficient computation and matching of compact descriptors, with comparable robustness for most common situations (see [Heinly et al. 2012] for a comparative evaluation).

5 On-Line Localization Stage

During on-line operations, the *global* pose of the user's head is estimated at each time instant t from synchronized pairs of images ($I(t)$) and IMU data ($\Gamma(t)$), $\{I,\Gamma\}_t$, according to different modes, detailed in the following paragraphs.

5.1 Pose Initialization

In the `INITIALIZATION` mode, the absolute pose of the camera is determined from scratch through a *visual matching* approach. A set of query descriptors is computed for N_{extr} keypoints extracted from the current camera image and matched with the descriptors of the whole scene map through *fast approximate nearest neighbor* search. Given the set $S_M(t)$ of the 2D-3D correspondences, the absolute camera pose is estimated by using the 3-point algorithm [Haralick et al. 1994] within a RANSAC framework for robust geometric verification.

After the very first initialization, a camera-IMU "hand-eye" calibration procedure is also performed. The calibration matrix, referring the inertial measures to the global reference frame, is estimated from $\{I,\Gamma\}_t$ pairs according to the classical hand-eye calibration equation. However, in our system only the rotational component of the calibration matrix needs to be estimated, since IMU accelerations are not directly employed for pose estimation (this rapidly resulting in positional drift). Accordingly, the centripetal component can be neglected. This permits to combine on-the-fly the simplified calibration equations for a batch of $\{I,\Gamma\}_t$ orientation pairs, acquired during the first N_{calib} frames, and solve the resulting system in a least square sense.

5.2 Tracking

Once successfully initialized, the system enters the `TRACKING` mode, where pose tracking is performed by fusing the visual and inertial data in an EKF framework.

As far as the visual information is concerned, the global matching stage may not always be sufficiently reliable (*e.g.* during fast mo-

tion) or efficient for real-time requirements. Frame-to-frame tracking can provide more robustness and precision, since it exploits spatio-temporal continuity between consecutive image frames, but it can lead to long term drift. A framework based on the Kanade-Lucas-Tomasi (KLT) tracker [Shi and Tomasi 1994], capable to handle moderate translations, has been employed. The tracker is initialized with the 2D locations of the keypoints of the set $S_M(t)$ obtained during pose initialization/relocalization. However, as the camera moves, keypoints get lost. Accordingly, a robust procedure to *update* the tracker has been implemented to ensure prolonged tracking. To identify when the tracker should be updated, a spatial skewness coefficient γ is computed for each frame using the (sub)set of successfully tracked keypoints, $S_T(t)$. For calculating γ, the image frame is divided into a lattice of $L = 4 \times 4 = 16$ cells, called *frame keypoint occupancy map*, and for each occupancy map cell, C, the density score $\rho = |S_T \cap C| / |C|$ (where $|.|$ returns the set count) is compared with the expected score for a uniform distribution, $\rho_{uni} = 1/L$. Given the number of cells with score below ρ_{uni}, N_b, and the number of cells with score above ρ_{uni}, N_a, we finally calculate $\gamma = (N_b - N_a)/L$. If γ falls below $\gamma_{min} = 0.65$, the tracker is *re-initialized* by uniformly sampling a maximum number $k_1 = 160$ 3D points of the map within the camera frustum. These points are then projected on the image plane, thus providing again a uniform set of 2D-3D correspondences, *i.e.* keypoints, to be tracked in the subsequent frames. For each successfully tracked frame, pose estimation is performed as described in Sect. 5.1, but employed just to filter the outliers and return a set of robust 2D-3D correspondences $S_I(t)$ to be fed into the sensor fusion framework for pose filtering.

This solution is very efficient, but it can fail in case of image degradation (fast motions), or occlusions. In these cases, the system enters the `TRACKING_IMU` mode that relies on the IMU data alone. Among different possible strategies, we have chosen to assume the position fixed during complete visual outage, and frequently invoke the `RELOCALIZATION` (Sect. 5.3). The intent of this approach is to limit the time interval of visual outage and accordingly positional drift.

5.2.1 Sensor Fusion

Different approaches have been adopted in the literature for the design of the EKF stage for visual-inertial systems [Bleser and Stricker 2008; Aron et al. 2007]. The "constant velocity, constant angular velocity" model has been widely used as a simplified linear motion model involving position and orientation first derivatives while treating accelerations as noise. However, this kind of models can often lead to divergence or wrong convergence [Perea et al. 2007], due to the poor modeling of real motions and to non-linearities, so that positional and rotational variables are not uncoupled. To cope with these issues, we use a loosely-coupled approach, which relies, by means of the calibration matrix, on the *global* orientation robustly estimated by the IMU. The constant velocity model is simplified so that the state $\mathbf{x}$ contains only the positional variables $\mathbf{x} = [\mathbf{p}\ \dot{\mathbf{p}}]$. The measurement equation employs the *measured* 2D-3D correspondences from the inliers set $S_I(t)$ and the *predicted* projections $\mathbf{m}^-$ of the 3D points on the image plane, according to the camera projective model (with intrinsic parameters and lens distortion estimated during the off-line stage). In this way, the effect of non-linearities and error coupling is reduced, generally leading to more stable pose estimates.

The EKF fails if an excessive state variation or increase in residuals (divergence) are detected, bringing the system to the `TRACKING_IMU` mode.

Table 1: *On-line Sequences*

Map	Seq.	#F(mins)	$\#F_{loc}$	$\#F_{IMU}$	T_M(ms)	T_T(ms)
SURF	SEQ1	7200 (4)	2600 (36%)	4600 (64%)	286 ± 15	21 ± 5
SURF	SEQ2	3600 (2)	2660 (74%)	940 (26%)	299 ± 22	19 ± 3
BRISK	SEQ1	7200 (4)	4634 (64%)	2566 (36%)	142 ± 28	25 ± 6
BRISK	SEQ2	3600 (2)	2858 (80%)	742 (20%)	130 ± 27	20 ± 3

The number of frames (#*F*) and duration (mins), the number of frames localized by the sensor fusion approach ($\#F_{Loc}$), and in the `TRACKING_IMU` mode ($\#F_{IMU}$), together with related timings (in ms, mean $\pm$ std.dev.) for the matching (T_M) and tracking (T_T) stages, are summarized for the two sequences processed according to different visual features (SURF/BRISK).

5.3 Relocalization

The `RELOCALIZATION` stage is implemented similarly to the `INITIALIZATION` stage, but performing feature matching only for map points contained within an *expanded* camera frustum, *i.e.* by considering a camera sensor with width and length both twice larger than the nominal one. The resulting pose is then filtered through the EKF.

`RELOCALIZATION` is best invoked when the IMU measures a *quasi-static condition*, *i.e.* the norm of the (gravity-compensated) acceleration vector and the angular rotation are below the thresholds $M_{acc} = 0.2$ (m/s^2) and $M_{rot} = 0.3$ (rad/s), respectively. Indeed, in this condition images are likely to be more stable, which aids the visual relocalization.

If `RELOCALIZATION` fails, the system remains in `TRACKING_IMU` mode for up to a maximum of N_{lost} consecutive failures, after which `INITIALIZATION` is invoked.

6 Experimental Results

The wearable immersive system consists of a PtGrey FireFlyMV camera (30 fps, 640×480), equipped with a varifocal lens (3 - 8 mm), mounted integrally with an OculusVR Rift HMD. Tests were performed in a rectangular room (Fig. 1) on an area of 3.75 m $\times$ 5.70 m. In order to assess properly the on-line performance of our approach, an approach similar to the method used in [Oskiper et al. 2011] is employed. A dense virtual model of the room has been reconstructed by re-meshing a laser point cloud and registered with the map's 3D point cloud. In this way the views of the *virtual room*, rendered according to the estimated pose, can be visually compared to the acquired images that constitute an indirect ground truth. We report the results related to two on-line sequences whose details are summarized in Table 1. The system performed live at approximately 30 fps on average using a Dell Aurora Alienware PC.

6.1 Test 1 – Free motion

The sequence SEQ1 (4 mins), containing multiple motion patterns (2 looping paths, rotation on approximately fixed position, fast motions), is analyzed in the following. In Fig. 2, images acquired by the camera are shown next to the rendered views of the room model at four different time instants, initially with good visual agreement (Fig. 2, first two columns). During the subsequent fast motions, the SURF method enters the `TRACKING_IMU` stage, which means that it does not capture the positional variations. Because SURF matching is slow, relocalization using SURF cannot be invoked too frequently in order not to impact time performance. As a result, the system is more prone to positional drift, which is quite noticeable after a prolonged outage of the visual tracking stage (Fig. 2, third and fourth column). In contrast, the relocalization by BRISK matching can be invoked more frequently without affecting too negatively the time performance, limiting the risk of prolonged outage

Figure 2: *Test 1: real camera images (top), and rendered views of the virtual room for BRISK (center) and SURF (bottom) for four sample time instants (columns).*

of the visual tracking stage. Indeed by using BRISK visual agreement (Fig. 2) is still good after relocalization (third column), with limited drift even after a long tracking period (fourth column).

6.2 Test 2 – Looping path

The sequence SEQ2 (2 mins) is a looping-path sequence and it is analyzed to evaluate the accuracy (in particular drift) of our method. The system is initially lifted from a predefined location, then head's free rotations are performed at different velocities (also pointing to untextured areas). The user keeps the waist fixed, but still limited translations of the head (shaking, bending) are performed, before returning to the predefined starting position. The 3D loop closure error is 0.09 m for the BRISK method, and 0.13 m for the SURF method.

7 Application

The proposed system is intended for simulating hazardous working conditions (such as working at heights) in vocational training in Construction. The video accompanying this paper shows two 6-DOF navigation experiments for a user immersed within a virtual model of a scaffold with an approximate height of 10 meters, overlooking a city model. In particular, the different working stages of the tracking system during natural walking within the training room are shown together with the camera stream, for a free path presenting fast motions (the system is hand-held) and for a looping path, respectively.

8 Conclusion

We presented a real-time 6-DOF tracking approach based on visual-inertial sensor fusion, in the context of the development of an affordable immersive system for simulation and training. The system relies on a single camera integral with an immersive stereoscopic HMD which embeds an IMU with high sampling rate, whose cross-calibration is performed automatically on-the-fly. The different strategies employed to deal with challenging situations (fast motion, untextured areas) and limit the impact of negative factors on user experience (drift, jitter) are analyzed. Live experiments have shown an overall good consistency for different motion patterns; the role of fast and frequent relocalization has proved to be crucial in limiting drift and jitter effects. In that context, a method for robust integration and interleaving of global matching and visual tracking, both aided by IMU information, is currently under development to better filter the pose estimations minimizing the additional latency. This aims at improving the localization consistency, which is crucial to deliver a comfortable user experience.

References

ARON, M., SIMON, G., AND BERGER, M.-O. 2007. Use of inertial sensors to support video tracking. *Comput. Animat. Virtual Worlds 18*, 1 (Feb.), 57–68.

BLESER, G., AND STRICKER, D. 2008. Advanced tracking through efficient image processing and visual-inertial sensor fusion. In *IEEE VR '08*, 137–144.

CHEN, W., PLANCOULAINE, A., FÉREY, N., TOURAINE, D., NELSON, J., AND BOURDOT, P. 2013. 6DoF Navigation in Virtual Worlds: Comparison of joystick-based and head-controlled paradigms. In *ACM VRST '13*, 111–114.

CRUZ-NEIRA, C., SANDIN, D. J., DEFANTI, T. A., KENYON, R. V., AND C.HART, J. 1992. The CAVE: Audio visual experience automatic virtual environment. *Commun. ACM 35*, 6 (June), 64–72.

GAUGLITZ, S., HÖLLERER, T., AND TURK, M. 2011. Evaluation of interest point detectors and feature descriptors for visual tracking. *Int. J. Comput. Vis. 94*, 3, 335–360.

HARALICK, B., LEE, C.-N., OTTENBERG, K., AND NOLLE, M. 1994. Review and analysis of solutions of the three point perspective pose estimation problem. *Int. J. Comput. Vis. 13*, 3, 331–356.

HAY, S., NEWMAN, J., AND HARLE, R. 2008. Optical tracking using commodity hardware. In *ISMAR 2008*, 159–160.

HEINLY, J., DUNN, E., AND FRAHM, J.-M. 2012. Comparative evaluation of binary features. In *ECCV 2012*. 759–773.

LIM, H., SINHA, S. N., COHEN, M. F., AND UYTTENDAELE, M. 2012. Real-time image-based 6-DOF localization in large-scale environments. In *IEEE CVPR '12*, 1043–1050.

OSKIPER, T., CHIU, H.-P., ZHU, Z., SAMARESEKERA, S., AND KUMAR, R. 2011. Stable vision-aided navigation for large-area augmented reality. In *IEEE VR 2011*, 63–70.

PEREA, L., HOW, J., BREGER, L., AND ELOSEGUI, P. 2007. Nonlinearity in sensor fusion: Divergence issues in EKF, modified truncated SOF, and UKF. In *Proc. AIAA Guidance, Navigation, and Control Conf. 2007*.

SHI, J., AND TOMASI, C. 1994. Good features to track. In *Proceedings CVPR '94*, 593–600.

SNAVELY, N., SEITZ, S. M., AND SZELISKI, R. 2008. Modeling the world from internet photo collections. *Int. J. Comput. Vision 80*, 2 (Nov.), 189–210.

WELCH, G., AND FOXLIN, E. 2002. Motion tracking: No silver bullet, but a respectable arsenal. *IEEE Comput. Graph. Appl. 22*, 6 (Nov.), 24–38.

WILLIAMS, B., NARASIMHAM, G., RUMP, B., MCNAMARA, T. P., CARR, T. H., RIESER, J., AND BODENHEIMER, B. 2007. Exploring large virtual environments with an HMD when physical space is limited. In *Proc. 4th APGV Symposium*, 41–48.

ZHU, Z., OSKIPER, T., SAMARASEKERA, S., KUMAR, R., AND SAWHNEY, H. 2008. Real-time global localization with a pre-built visual landmark database. In *IEEE CVPR 2008*, 1–8.