Visual-Inertial 6-DOF Localization for a Wearable Immersive VR/AR System

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

We present a real-time visual-inertial localization approach directly integrable in a wearable immersive system for simulation and training. In this context, while CAVE systems typically require complex and expensive set-up, our approach relies on visual and inertial information provided by consumer monocular camera and Inertial Measurement Unit, embedded in a wearable stereoscopic HMD. 6-DOF localization is achieved through image registration with respect to a 3D map of descriptors of the training room and robust tracking of visual features. We propose a novel efficient and robust pipeline based on state-of-the-art image-based localization and sensor fusion approaches, which makes use of robust orientation information from IMU, to cope with camera fast motion and limit motion jitters. The proposed system runs at 30 fps on a standard PC and requires very limited set-up for its intended application.

1 Introduction

Recent advances in the simulation capabilities of VR/AR systems have stirred up the interest for immersive simulation and training in different fields. Realistic perception of virtual environments, as well as consistent augmentation of real world with information tags or virtual objects, are of great importance in different sectors, *e.g.* enhanced project visualization and design review or computer-aided surgery in medical training. 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 order to deliver realistic and interactive user experience, the *localization* stage, which aims at estimating in real time the position and orientation of the trainee's head, is of crucial importance. CAVE systems, which represent the standard for 3D immersive environments, often implement head tracking by tracking IR markers through multiple cameras/sensors, with the rendered scene projected on wide screens surrounding the user. Different commercial systems, requiring dedicated facilities, on-purpose calibration and set-up procedures, are generally employed in such environments, 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 (robustness, accuracy, precision) and overall system complexity and cost. In particular, recent developments in HMD technologies are paving the way for the integration of commodity consumer devices into systems that can be robust and very cost-effective (potentially a small fraction of the cost of conventional CAVE approaches).

In this work an *ego-motion* approach, relying on the complementary action of visual and inertial tracking, is proposed. The 6-DOF pose of the trainee's head is estimated using visual information acquired by a monocular camera and inertial data provided by an high rate (1 kHz) 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.

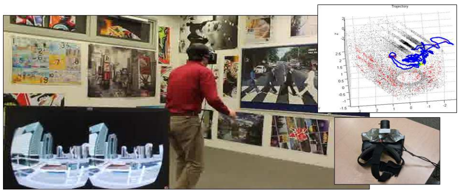

Figure 1: Illustration of the main components of the proposed immersive system.

2 Key stages of the Proposed Approach

The proposed method relies on two fundamental stages. First, an *off-line visual reconstruction stage*, performed in advance once and for all, encodes the visual structure of 3D landmarks present in the scene into a database of visual descriptors, or *map*. A multi-descriptor implementation of this stage allows flexibility during on-line operations. During on-line operations, an *on-line hybrid localization approach*, which 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, is employed to robustly estimate the head's pose. The main features of the two stages are summarized in the following paragraphs.

2.1 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 achieved using the *Bundler* framework [4]. 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 [2] is adopted to compute more efficient descriptors simultaneously preserving a good trade-off with robustness. Two different kinds of visual features, respectively SURF and BRISK descriptors, have been comparatively evaluated due to their different robustness and computational efficiency. The 3D map is then filtered according to an average repeatability score and number of reconstructing cameras.

2.2 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:

- In the `INITIALIZATION` mode, the absolute pose of the camera is determined from scratch through a visual *global*

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

matching approach. A set of query descriptors, computed for N_{extr} keypoints extracted from the current image, is matched with the descriptors of the whole scene map through *fast approximate nearest neighbor* search. Given the set of 2D-3D correspondences, the absolute camera pose is estimated by using the 3-point algorithm within a RANSAC framework for robust geometric verification [1].

After the very first initialization, an on-the-fly "hand-eye" calibration of the camera-IMU system, performed on a batch of $\{I,\Gamma\}_t$ pairs acquired during the first N_{calib} frames, permits to refer the inertial measures to the global pose. Our approach is similar to [3], but it can be applied directly on-line (details are omitted due to lack of space).

- Once successfully initialized, the system switches to `TRACKING` mode. Pose tracking is performed by fusing the visual and inertial data in an EKF framework. As far as the visual information is concerned, a frame-to-frame tracking framework based on Kanade-Lucas-Tomasi (KLT) tracker is employed. A robust procedure for the *re-initialization* of the tracker, based on a spatial skewness coefficient γ of a *keypoint occupancy map*, has been implemented (see [1]), so that a sufficient number of spatially distributed keypoints is preserved over long periods of tracking. 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. In case of fast motion and/or image blur, or occlusions, the visual tracking approach can fail. 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 keep fixed the position during complete visual outage, and frequently invoke the `RELOCALIZATION`. The intent of this approach is to limit the time interval of visual outage and accordingly positional drift. This selected strategy is based on the observation that it is likely that users do not translate significantly while simultaneously rotating their head fast. Inertial and visual data are initially processed *separately*, so to provide a robust real-time estimation of the *orientation* and a set of visual inliers, respectively, and are then fused together in the EKF to determine the *position*. This approach results in increased overall stability, decoupling cross errors due to nonlinearities that can lead to divergence, as well as efficiency by guiding the visual search.

- When unreliable poses are detected, the system enters the `RELOCALIZATION` mode, which employs a fast guided visual matching (just within an *expanded* camera frustum) to recover effectively the pose, relying in the meantime on IMU data alone (`TRACKING_IMU`). If the system is unable to recover the pose for N_{lost} consecutive frames, it enters again the `INITIALIZATION` mode.

3 Experimental Results

The wearable immersive system consists of a PtGrey FireFlyMV camera (30 fps, 640 × 480) mounted integrally with an OculusVR Rift HMD. Tests were performed in a rectangular room (3.75 m × 5.70 m), with walls covered with differently textured posters arranged according to a random layout.

In order to assess properly the on-line performance of our system, the following approach 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.

Here we present results related to a sample sequence (4 mins), containing multiple motion patterns (2 looping paths, rotation on approximately fixed position, fast motions). The system performed live at approximately 30 fps on average on a Dell Aurora Alienware PC. In Fig. 2, four sample images acquired by the camera are shown next to the rendered views of the room model. It can be seen that, using BRISK, visual agreement is still good after relocalization (third column), showing very limited drift even after a long tracking period (fourth column). In contrast, by using SURF, since relocalization cannot be invoked too frequently in order not to impact time performance, the system is more prone to positional drift after a prolonged outage of the visual tracking stage (Fig. 2, third and fourth column). By analyzing a similar sequence (4 mins), with the user free to walk but returning three times to the same predefined location, the average loop closure errors were found to be 0.18 m for BRISK and 0.88 m for SURF.

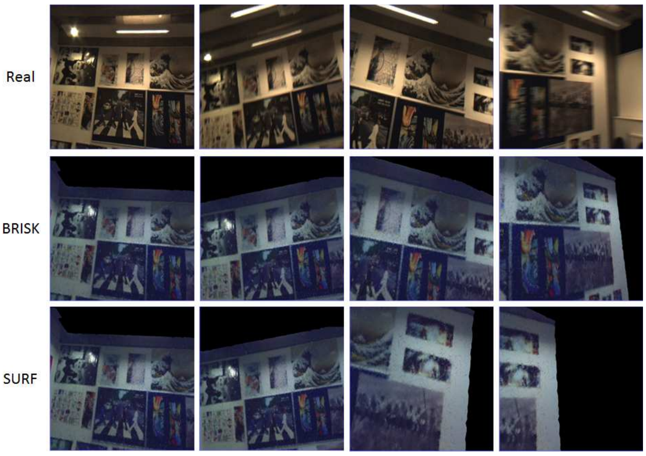

Figure 2: Real camera images (top), and rendered view of the virtual room for BRISK (center) and SURF (bottom), for four sample time instants (columns).

4 Conclusion

Our live experiments show a good overall consistency for different motion patterns; the role of fast and frequent relocalization proved to be crucial in limiting drift and jitter effects. This was emphasized by assessing the performances of two different visual approaches, based on BRISK and SURF, in different conditions. These considerations have prompted us to follow two main directions for future work. A robust procedure for the selection of stable and accurate landmarks from the pre-built database, maximizing feature repeatability and distinctiveness, can significantly improve the performance of the re-localization stage. Moreover, the development of a method for robust integration and interleaving of global matching and visual tracking, both aided by IMU information, can have a further significant impact on limiting drift and jitter effects.

References

[1] L. Carozza, F. Bosché, and M. Abdel-Wahab. Image-based localization for an indoor VR/AR construction training system. In *CONVR 2013*.

[2] H. Lim, S. N. Sinha, M. F. Cohen, and M. Uyttendaele. Real-time image-based 6-DOF localization in large-scale environments. In *CVPR*, pages 1043–1050, 2012.

[3] J. Lobo and J. Dias. Relative pose calibration between visual and inertial sensors. *IJRR*, 26(6):561–575, June 2007.

[4] N. Snavely, S. M. Seitz, and R. Szeliski. Modeling the world from internet photo collections. *IJCV*, 80(2):189–210, Nov. 2008.