Image-based localization for an indoor VR/AR construction training system

IMAGE-BASED LOCALIZATION FOR AN INDOOR VR/AR CONSTRUCTION TRAINING SYSTEM

Ludovico Carozza, Frédéric Bosché, Mohamed Abdel-Wahab
School of the Built Environment, Heriot-Watt University, Edinburgh, UK

ABSTRACT: *Virtual /Augmented Reality (VR/AR) technologies have been increasingly used in recent years to support different areas of the construction industry. Their simulation capabilities can enable different construction stakeholders to evaluate the impact of their choices not only on the built environment, but also with regard to the correct execution of operational procedures. Training providers, such as Further Education (FE) colleges, can also enhance their trainee's experience through the simulation of realistic construction contexts whilst eliminating health and safety risks. Current approaches for the simulation of learning environments in Construction, such as Virtual Learning Environment (VLEs), provide limited degree of interactivity during the execution of real working tasks. Whilst immersive approaches (e.g. CAVE-based) can provide enhanced visualization of simulated environments, they require complex and expensive set-up with limited practical interaction in real construction projects context.*
This paper outlines a localization approach employed in the development of an Immersive Environment (IE) for Construction training, cheaper than CAVE-based approaches and which has the potential to be rolled-out to the FE sector for maximizing the benefit to the construction industry. Pose estimation of the trainee is achieved by processing images acquired by a monocular camera integral with his head while performing tasks in a virtual construction environment. Realistic perception of the working environment and its potentially hazardous conditions can thus be consistently delivered to the trainee through immersive display devices (e.g. goggles).
Preliminary performance of the localization approach is reported in the context of working at heights (which has a wide applicability to a range of construction trades, such as scaffolders and roofers), whilst highlighting the potential benefits for trainees. Current limitations of the localization approach are also discussed suggesting directions for future development.

KEYWORDS: *Image-based, localization, VR/AR, and construction training*

1. INTRODUCTION

Applications of Virtual/Augmented Reality (VR/AR) to the Architecture, Engineering and Construction (AEC) sector have been gaining considerable attention from the industrial and academic community for their inherent simulation capabilities in different contexts, such as: enhanced project visualization and design review (particularly with BIM) (Bosché et al. 2012, Bassanino 2010, Woodward et al. 2010), on-site information retrieval (Yeh et al. 2012), and plant operatives training (Wang et al. 2004). However, construction trades (such as scaffolders, roofers, painter and decorators, etc.) have not yet benefited from training in simulated work environments by using VR/AR technology. The key benefit of using VR/AR for construction trades training is that it can create a realistic learning environment for training, e.g. working at height, without exposing trainees or instructors to any health and safety risks (Abdel-Wahab, 2012). It can provide immediate interaction with realistic environments, through real-time feedback to ensure that trainees consistently perform to the required standards.

VR systems currently considered in construction education and training are essentially based on Virtual Learning Environments (VLEs), like Moodle and Blackboard (Abdel-Wahab, 2012), or CAVE-type 3D immersive environments. Simulated environments are developed for VLEs that are essentially educational video games with totally simulated environment that the trainee interacts (and learns) with. This type of virtual training results in a limited degree of interactivity and immersion, two aspects deemed of great importance in the construction sector context (Abdel-Wahab 2012, Dalgarno et al. 2010).

On the contrary, immersive learning environments can provide potential learning benefits such as: immersive spatial representation, by providing enhanced visualization of the virtual environment through 3D immersive displays; and immediate interaction with realistic environments, through a direct and immediate feedback reflecting consistently the actions performed by the trainee (e.g. trainee's movements). These enhanced functionalities can find application in realistic training in operational procedures, e.g. construction equipment operation (Wang et al. 2004), and safe simulation of hazardous conditions, e.g. working at heights.

3D immersive environments currently investigated and used for construction training are mainly CAVE-type environments (CruzNeira et al. 1992; Lau et al. 2007). These require the set-up of dedicated facilities with significant impact on complexity and costs. These facilities include, among others, the installation of entire plant simulators with dedicated hardware for interaction, or the set-up of huge video screens and dedicated projectors, with impact on installation costs and energy consumption. ACT-UK (ACT 2009) is an example of CAVE-type environment that was developed for construction management training (deVries et al. 2004).

The work presented here is conducted in the context of the Immersive Controlled Environment (ICE) project, which aims at simulating a real construction workplace for real construction activities conducted by the trainee without any risk of injury (*immersive experience),* as well as assessing the trainee's performance (*controlled experience*). As part of this project, an alternative 3D immersive environment is presented that leverages significant advancements recently made in localization and visualization technologies, in particular: portable immersive display, and image-based localization systems.

Novel portable immersive display systems, such as (Vuzix), offer the immersive functionality of CAVE-type systems for a fraction of the cost. Our system is based on the use of such display systems. However, the disadvantage of such systems is that they require robust approaches for tracking the movement of the head of the user. This localization functionality plays a key role in providing the trainee with a consistent and realistic spatial perception of the simulated virtual scenario.

Several techniques can be considered to resolve this localization problem, and we propose to use an image-based approach applied on a video sequence acquired from a monocular camera integral with the trainee's head (i.e. with the display goggles). The acquired images are registered with respect to a three-dimensional visual model of the training room, acquired in advance once and for all. The trainee's head pose is then employed to deliver any virtual construction environment in a consistent manner through the display goggles. Compared to other localization approaches, image-based approaches offer significant advantages in terms of set-up complexity and cost. As a result, the overall system that we propose has the potential to deliver highly immersive, consistent and realistic VR/AR experiences at a fraction of the cost of CAVE-type systems.

This paper focuses on the head localization functional component. The proposed approach is described and its performance assessed in terms of accuracy, robustness and processing time. This assessment ultimately determines what the future of the proposed approach can be with regard to VR/AR systems.

This paper is structured as follows. In Sect. 2, current technologies proposed for localization are briefly reviewed, highlighting advantages and drawbacks of different approaches. In Sect.3, we describe our localization approach in the context of the ICE project, emphasizing the strategies devised to cope with the most important localization issues. The experimental assessment of the performance of our approach is presented and discussed in Sect. 4. Benefits and limitations of our approach are discussed in Sect.5, along with directions for future improvements.

2. BACKGROUND

Accurate and robust localization (i.e. estimation of position and orientation) of the viewpoint within the environment is crucial to provide consistent interaction with the user. Mainly two general approaches are applied to localize and track objects (head of the user, reference points of a tool, etc.) within its environment: inside-out (ego-motion or direct) and outside-in (indirect) approaches. According to these approaches, information provided by different sensors mounted internally or externally, respectively, to the entity to be tracked is processed. Global and local position systems (GPS, WIFI), environment sensors (RFID), as well as Inertial Measurement Systems and vision-based systems, constitute the main state-of-the-art technologies employed to this purpose (Feng Zhou et al. 2008). Integration of different technologies (acoustic, magnetic, inertial, optical, etc.) within hybrid approaches has also been proposed to cope with the drawbacks of single approaches by exploiting complementary performances (Feng Zhou et al. 2008, Ligorio and Sabatini 2013). Desirable features of 6-DOF localization methods include, among others, coverage and range limitation, robustness to environment interferences (magnetic, visual occlusions, etc.), robustness to fast motion dynamics, and absence of drift for long range paths.

In the context of CAVE training approaches, hybrid systems are usually employed. Inertial-ultrasonic hybrid tracking (Intersense 2002), as well as multi-camera tracking of optical fiducial markers tracking, based on LED (Welch. et al. 2001) or reflective beacons (Pintaric and Kauffmann 2007) are currently among the solutions employed. These systems require on-purpose set-up and calibration to achieve accurate localization – at the

expense of several thousand dollars.

Vision-based approaches are nowadays standing out due to low cost and widespread availability of digital cameras. Broadly speaking, issues in terms of robustness to motion patterns (abrupt scale and view-point changes, lighting conditions, blurring, etc.) and time responsiveness (i.e. latency) still have a significant impact, so that requirements in computational resources and scalability to large environments are in general of crucial importance (Dong et al. 2009). Moreover, important issues concern relocalization after tracking failures and error drift, especially for closed loop sequences. State of the art vision-based methods are based on *global* localization approaches (Skrypnyk and Lowe 2004) that conceptually overcome these issues, but often at the expense of a greater computational burden. The works in (Lim et al. 2012, Dong et al. 2009) rely on the a-priori visual knowledge of a three-dimensional model to establish 2D-3D correspondences between the image domain and a three-dimensional reference frame. For each processed frame, *camera/space resectioning algorithms* (Hartley and Zisserman 2003) are employed to determine from these correspondences the absolute 6-DOF pose of the camera in the three-dimensional reference frame. Generally, for computational reasons, matching of sparse visual descriptors, rather than recognition of geometric structure (e.g. lines/shapes) is employed concurrently with other strategies, like space partitioning (Tingdahl et al. 2012, Carozza et al. 2012) and keyframe selection (Dong et al. 2009) to prune the search space and speed up the process.

3. OVERVIEW OF THE PROPOSED SYSTEM

We propose a vision-based head localization approach that is motivated by the works of Carozza et al. (2012), and Lim et al. (2012), employing the a-priori visual knowledge of a three-dimensional model to establish 2D-3D correspondences between the image domain and a three-dimensional reference frame. The training immersive environment we devised consists of a training room, conveniently covered with textured images, for example with posters (Fig. 1).

Fig. 1: Panoramic image of the training room whose walls have been conveniently covered with textured pictures.

In an off-line process, pictures of the room are acquired and processed in a 3D reconstruction pipeline, resulting in a 3D map of visual features (that are mainly extracted from the textured images). It is important to note that our approach does not require the on-purpose setup of fiducial markers with specific configuration (e.g., known spatial distribution, visual pattern, markers' optical reflectivity, etc.), which can be complex and time-consuming, nor does it require the calibration of multiple cameras. Instead, the physical boundary of the immersive room simply needs to be covered with randomly positioned textured images that just need to be at an appreciable scale with respect to the camera field of view and the room size.

During on-line operations, the trainee is equipped with a monocular camera, integral with his head, i.e. with the immersive display goggles. In our system, the camera points towards the rear or side to reduce occlusion issues, and captures images of the room boundaries in real-time. For each image, visual features are extracted and matched against those contained in the map of 3D features created offline. The resulting matches enable the calculation of the position of the camera, and subsequently of the trainee's head. The estimated head pose is employed to render in the VR goggles carried by the trainee the corresponding view of the construction virtual environment experienced.

The off-line process for reconstructing the three-dimensional map of visual features of the room is described in Section 3.1. Then, the on-line process for localizing the head of the trainee/user within the room is described in Section 3.2.

3.1 Off-Line Reconstruction Stage

This stage aims at creating a 3D model of the visual features of the training room, whose boundary has been covered with pictures (see Fig. 1). For this, a set of overlapping pictures of the training room is acquired from different viewpoints. We call this set of pictures the *reconstruction set*. We then perform the 3D reconstruction of the SIFT (Lowe 2004) features, extracted from the reconstruction set of pictures, through sparse bundle adjustment – we use the Structure-from-Motion *Bundler* framework (Bundler 2006) described in (Snavely et al. 2007). As a result, 3D coordinates of SIFT features, as well as estimations of the camera intrinsic (focal length and distortion coefficients for a pin-hole camera model) and extrinsic (i.e., position and orientation) parameters are achieved (for more details, see Bundler 2006). The resulting 3D point cloud is filtered using a thresholding minimum number of cameras contributing to the reconstruction of each point, *num_count*, in order to select only the "best reconstructed" reference points.

SIFT are currently considered as the best visual descriptors in terms of robustness (Gauglitz et al. 2011), but at the expense of a considerable computational effort for matching pairs of features. This seriously limits image processing speeds, and subsequently in our case VR quality performance during on-line operations. Following an approach similar to the one adopted by Lim et al. (2012), we exploit the achieved SIFT reconstruction, which has already reconstructed robust salient features, to compute more efficient descriptors for the corresponding 3D cloud. The following process is used: For each reconstruction image, SURF keypoints (Bay et al. 2008) are extracted. Then, the reconstructed camera pose is exploited to compute the 2D re-projections of the 3D points on the reconstruction image. Finally, for each SURF keypoint, the 3D point with the closest re-projection (within a search radius *distance_threshold*) is associated to the corresponding SURF descriptor, thus obtaining a *database of 3D referenced SURF descriptors,* hereinafter called *map*. This database is filtered again according to a minimum number of reconstructing cameras (this threshold being set to *num_count/3),* so that features with low repeatability are discarded. We note that all SURF descriptors matched to a given 3D point should be close in the descriptors' vector space, that is with distances (Euclidean, in the case of SURF) having a low-mean and low-variance distribution. This aspect has been investigated at a preliminary level, with our experiments showing mean distance values in the range of [0.18; 1.12] and distance variances in the range [0.01; 0.93]. Therefore, we propose to keep for a 3D point a unique descriptor with associated *global feature strength* as repeatability score, calculated by averaging the corresponding descriptors and their strengths respectively. This strategy aims at reduce the size of the database while preserving repeatable and distinctive features for better on-line performance.

3.2 On-Line Localization Stage

During on-line operations, i.e. in our context during a construction training session, the system processes the image sequence acquired live and in real-time from a camera mounted integral to the trainee's head (i.e. to the display goggles). We call this new set of camera images the *target sequence*. Different strategies are employed to estimate the pose of the camera from correspondences between the visual features extracted from each target image and the map created off-line. These are described below.

3.2.1 Initialization

When the first frame is processed, or the pose is completely lost, there is no clue about the camera pose, which hence needs to be initialized. SURF features are extracted from the processed target image and their descriptors matched with the N strongest SURF descriptors of the map (N=500, in our tests), organized in a k-d tree indexing structure (Skrypnyk and Lowe 2004) to improve efficiency. For this, efficient approximate nearest neighbor interrogation of the map is employed, followed by a ratio test (α=0.65) to prune false matching. This process ultimately retrieves, for each 2D keypoint extracted from the target image, the 3D coordinates of the "best" matching point in the map, leading to a set $S(x_{2d}, X_{3d})$ of one-to-one matching coordinates. The pose of the camera, that is the rotation matrix R and the camera centre position C, is then robustly determined employing RANSAC filtering followed by Levenberg-Marquardt optimization over the resulting set of inliers (Hartley and Zissermann 2003). To this purpose, the re-projection error is employed as cost function (see Snavely et al. 2007 Appendix 1 for more details), with the camera intrinsic parameters known from the off-line stage or other camera calibration methods (Bouguet 2004).

If less than *min_num_inliers* inliers are found, the pose is rejected, the corresponding frame skipped and the system remains in *Initialization* mode for processing the following target image. On the other hand, if the pose estimation is successful, the system switches to *Tracking* mode.

3.2.2 Tracking

Once the pose has been initialized successfully, for the subsequent frames the pose is estimated from 2D-3D correspondences achieved by performing *feature tracking* between consecutive frames. More in detail, given the set $x_{2d}(t-1)$ of image locations of the matched SURF features for the last successfully computed frame at time (t-1), the Lucas-Kanade-Tomasi tracker (LKT, Lucas et al. 1994) is employed to estimate their locations for the current image at time *t*, $x_{2d}(t)$. In general, tracking approaches permit to significantly speed up the 2D-3D matching stage by exploiting spatio-temporal continuity, so that local motion fields, sufficiently small for consecutive frame, can be quickly estimated from local image analysis. On the other hand, in the presence of large camera displacements due to abrupt motions, *Relocalization* is required to recover from lost poses. To identify such situation and perform relocalization, the same strategy as above could be applied: if less than *min_num_inliers* inliers are found, the *Tracking* is considered unsuccessful and the method re-enters the *Initialization* (i.e. relocalization) stage, so that the tracker can be reinitialized with new features to track for the subsequent frames.

However, *Initialization* is a time-expensive matching process that should be employed as little as possible. Therefore, a more robust tracking strategy has been put in place that reduces potential frequency of *Initialization*. In order to track a sufficient number of features, otherwise often lost after few frames, and also to prevent potential ambiguity in estimating pose due to uneven distribution of the tracked features, a measure of skewness for the spatial distribution within the image plane is evaluated at each frame. A lattice of T=16 cells is built on the current image and an occupancy map is computed with each cell assigned a score calculated as the number of keypoints located within it over a maximum number of features to track, F (we use F=160). As a measure of skewness of the scores' distribution, the Cyhelský's skewness coefficient is considered:

$$S = (C\_L - C\_R)/T$$

where *C_L* and *C_R* are the number of scores below and above the expected score (for a uniform distribution) 1/T.

If S exceeds the threshold S_TH (set to 0.65, in our tests), *tracker resetting* (*Reset* mode) is triggered for the subsequent frame, that is new features are added to the feature tracker by re-projecting on the 2D image plane the 3D features contained in the frustum of the previous successfully computed pose. In addition, local non-maxima suppression of the re-projected keypoints is performed in order to retain widespread strong features. In this way, a higher number of features with a more widespread spatial distribution can be tracked, with benefits in terms of less frequent relocalizations (*Initialization* or *Reset* mode) and robustness of the estimated pose. In fact, this tracking strategy can be even more robust with respect to the matching strategy, for example in presence of global illumination changes or blurring artifacts in less textured areas (presenting few weaker features to match). Tracking strategies are inherently prone to drift, and this strategy has also the aim of curbing this effect by performing periodic adjustment triggered by "poor quality" (i.e. poor feature distribution) of pose estimation.

In addition, in order to smooth the resulting camera trajectories, the computed poses are filtered using Extended Kalman filtering. The filter is initialized in the *Initialization* state and the pose is tracked accordingly while the system is in Tracking mode. The dynamic model adopted in this work is similar to the one described in (Davison et al 2007, Tingdahl et al. 2012), due to its trade-off between simplicity (only linear and angular velocities are modeled) and smoothing performance. To improve the numerical stability of the resulting dynamic system, preconditioning with a scaling factor λ=10 is applied to the 3D coordinates of the matched points of the map (i.e. the *measures vector* of the EKF), so to avoid that small values could lead to filter divergence (Perea et al. 2007). EKF results are rejected as unreliable if the changes in orientation are too severe or the residuals increase, indicating possible divergence, in which case *Initialization* is conducted.

4. EXPERIMENTAL RESULTS

In this section we present the results of several experiments. In particular, we focus on the performance of the localization approach for on-line sequences acquired for two different training rooms, for which both the room set-up as well as the motion patterns are different. The walls of the two rectangular rooms (ROOM1, and ROOM2, hereinafter) have been previously covered with textured posters with different pattern and size (see Fig. 1 for ROOM2), with different spatial arrangement for the two rooms, so to cover almost all the room perimeter and guarantee visual distinctiveness in the different parts of the room.

Both off-line reconstruction and on-line localization stages have been performed using sequences of images from videos of the rooms acquired by a hand-held digital camera (Panasonic CCD DMC-TZ6, 640 x 480, 30 fps

MJPEG). The intrinsic camera parameters estimated by the Bundler framework during the off-line reconstruction stage are also used in the on-line experiments, which simulates the trainee's movements according to different motion patterns.

Moreover, a virtual model (a scaffold new a brick wall, in our example) was aligned manually with the room maps after the reconstruction stage, so that it can be rendered during the on-line stage according to the estimated camera pose for each of the processed video frames.

Tests were performed off-line on the video sequences on a Dell Aurora Alienware PC (Intel i7-3280 @ 3.6GHz, 8GB RAM). Videos with results are available at http://www.ice.hw.ac.uk/. To assess the performance of our approach, visual evaluation of the rendered views under the motion patterns has been performed initially to assess qualitatively the robustness of our approach. Furthermore, the *reprojection error* has been used as measure of accuracy. Time performance of the localization process has also been considered in view of addressing future latency issues that can potentially affect user experience.

4.1.1 Experiments in ROOM1

For the off-line reconstruction stages 265 video frames of the room acquired from different viewpoints have been used, yielding a map of 1348 SURF features. For the on-line stage, a video sequence of 1149 frames (duration 38 s) has been acquired moving around the room following a smooth motion pattern, still presenting compression artifacts and changes in scales. In Fig. 2 the mean reprojection error and the processing time for each frame are reported.

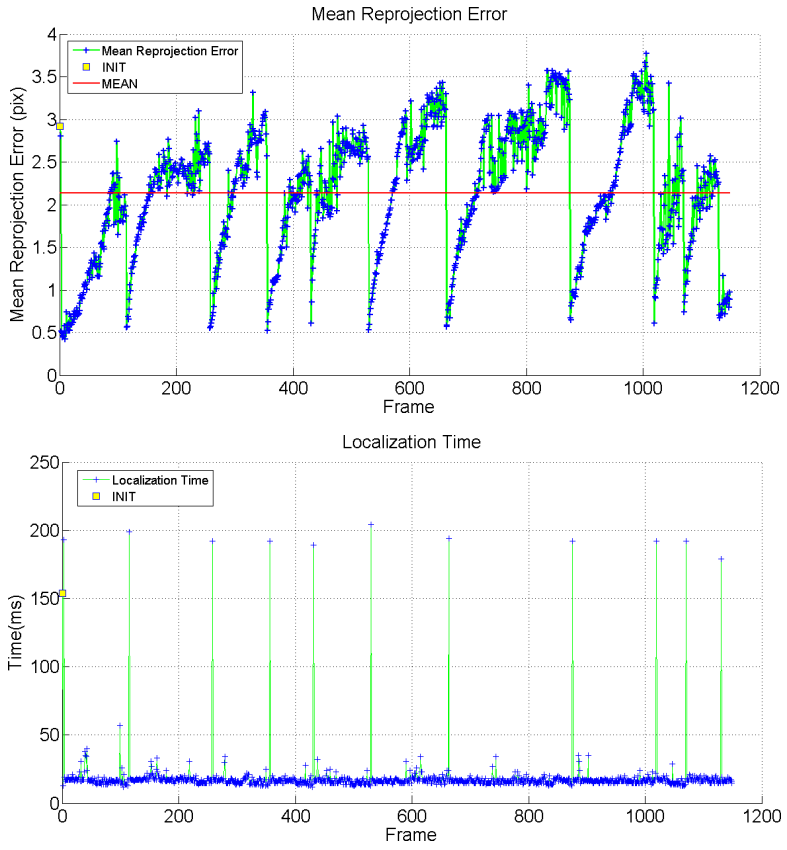

Fig. 2: Mean reprojection error (top) and processing time (bottom) for the localization stage for test in ROOM1.

As it can be noticed, tracking introduces some drift effect, adjusted periodically by the tracker reinitialization (*Reset* mode). The magnitude of the tracking errors (about 2 pixels on the average) is comparable with the ones obtained by other vision-based Virtual Reality applications (Klein and Drummond 2003). Visual estimation shows consistent views of the virtual environment for all the frames, with near real-time performance (25-30 fps) and some latency introduced by the occurrences of the more time-expensive *Reset* mode. It must be noticed that this is

mainly due to a current inefficient implementation of this stage, that can be improved by taking advantage of space partitioning techniques (see Sect. 5).

4.1.2 Experiments in ROOM2

A map of 2266 features has been achieved from 153 video frames of ROOM2 through the off-line reconstruction stage. Results related with a video of 2415 frames (1.20 min) are shown in Fig. 3.

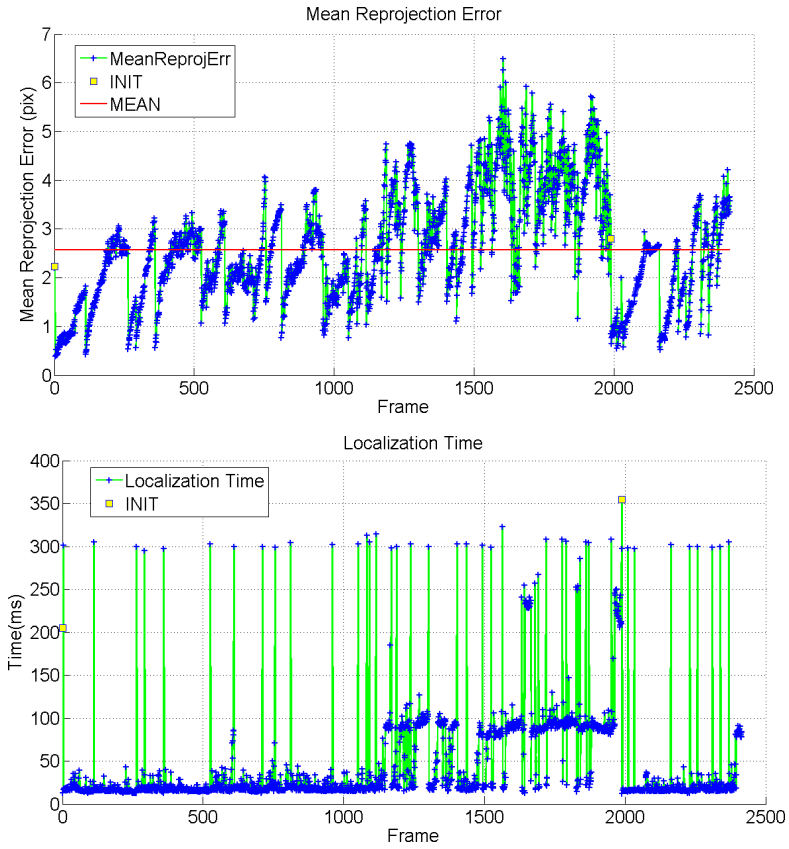

Fig. 3: Mean reprojection error (top) and processing time for the localization stage (bottom) for test in ROOM2.

Performance is comparable with the one obtained for the test in ROOM1 for almost all the video sequence. We note though that frames 1500 to 1718 rely on a low textured area (the room's ceiling) for which poor reconstruction was obtained (due to very few posters installed). In this case, the tracking strategy presents a drift that is then followed by the *Initialization* at frame 1988. A more efficient room set-up as well as space partitioning strategy can limit the impact of this drawback, as discussed in Sect. 5.

To illustrate the application of the our localization approach to the devised virtual training system in construction, the rendered views corresponding to four estimated camera poses from the estimated trajectory in ROOM2 are shown in Fig. 4.

5. CONCLUSIONS

Immersive spatial representation, using AR/VR, can enhance the learning experience of construction trainees by providing a simulated construction site environment whilst eliminating H&S risks, such as working from heights. The benefit accrued from this approach is that trainees can focus on mastering the task at hand in a safe environment and potentially enhance their performance. In this context, a vision-based localization approach, to be employed as a key component of a 3D Immersive Controlled Environment (ICE) for training in Construction, has been presented in this work. Due to recent developments of vision-based technologies, a significant advantage of

the proposed approach is that it has the potential of delivering immersive experience of a training environment at a fraction of the cost of CAVE-type systems.

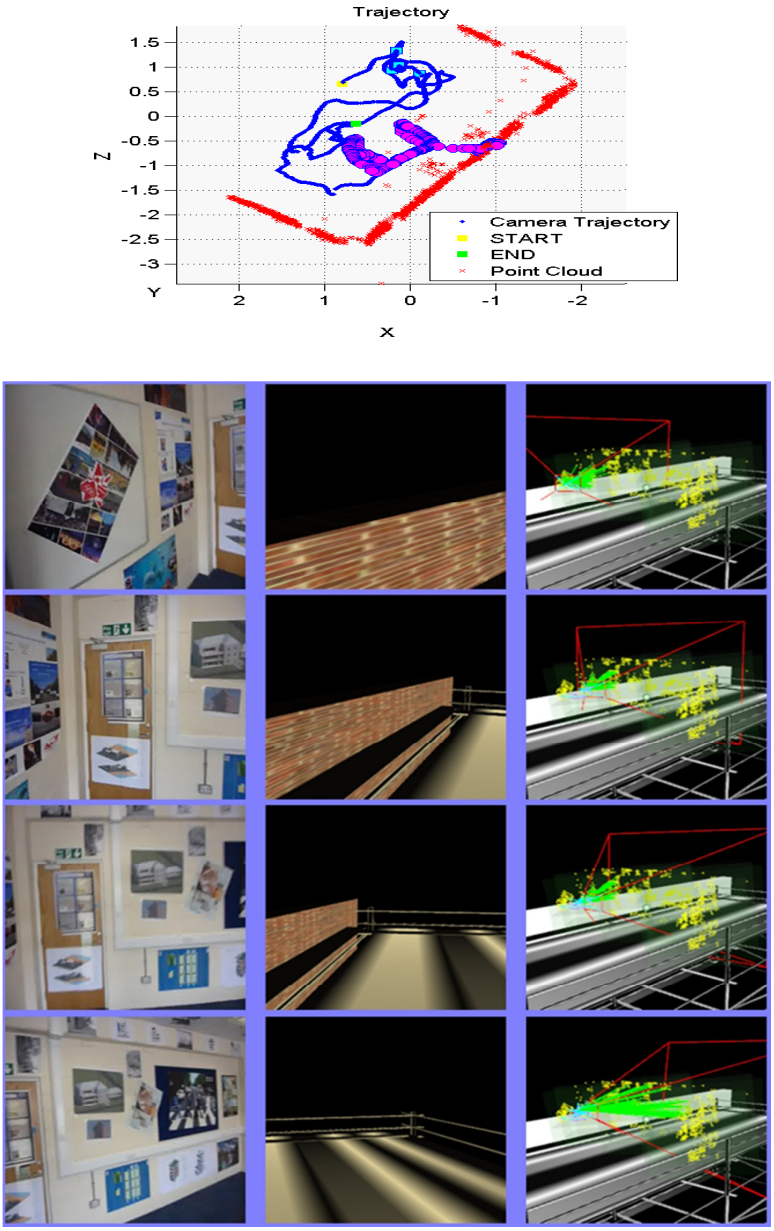

Fig. 4: Top: top view of the estimated camera trajectory for ROOM2. Magenta circles correspond to a trait of the camera path covering poor textured areas (room's ceiling). Bottom: rendered views corresponding to four estimated camera poses (cyan squares on the trajectory).

The approach consists of two stages. The first off-line reconstruction stage performs the reconstruction of a 3D visual map of the training room, covered with textured pictures (posters), from a sequence of images. The second stage aims at estimating the trainee's position and orientation within the training room during the operational stage, by registering images, acquired by a camera integral to the trainee's head, with respect to the 3D visual map. Several strategies have been devised to preserve robustness at an acceptable frame rate, including a feature tracking strategy. The performance of our method has been assessed on real video sequences of a training room, showing promising results in terms of robustness and time performance. Moreover, their analysis has shown the main current limitations, suggesting feasible strategies to overcome them. In particular, our analysis has yielded

the following conclusions:

- The current setup of the training room includes poorly textured areas that have led to failed camera localization. To reduce the risk of such situation, the training room should be sufficiently covered, with textured pictures in almost all its parts. The spatial analysis of the visual map obtained from the reconstruction stage and the use of an octree, to select spatially distributed visual features during the on-line phase (as proposed in (Tingdahl et.al 2012, Carozza et al. 2012)), could also be employed to improve robustness.
- The current implementation of our approach can be directly optimized in order to speed up significantly the process toward real-time performance. The use of an octree can speed up the retrieval of strongest features distributed in the camera frustum during the matching and tracking phases, eliminating for example the need of the slow non-maxima suppression stage employed at the moment. Furthermore, GPU processing could further accelerate many of the on-line processes, especially matching and pose computation.
- The current vision-based approach could benefit from integration with INS systems (Ligorio and Sabatini 2013). The INS could be used to provide initial estimates of the head pose at a high frame rate, which would be beneficial, especially in the case of fast motions.

6. ACKNOWLEDGEMENT

We are grateful to the Construction Industry Training Board (CITB) for funding this project. The views and opinions expressed by the authors of this publication are those of the authors and do not necessarily reflect those of CITB.

7. REFERENCES

Abdel-Wahab M.S. (2012) Rethinking apprenticeship training in the British construction industry, *Journal of Vocational Education & Training*, Vol. 64, No. 2, 145-154.

ACT (2009), Advanced Construction Technology Simulation Centre, Coventry UK, http://www.act-uk.co.uk/.

Bassanino M., Kuo-Cheng W., Jialiang Y., Khosrowshahi F., Fernando T., Skjærbæk, J. (2010), The Impact of Immersive Virtual Reality on Visualisation for a Design Review in Construction, *Information Visualisation (IV), 2010 14th International Conference*, 585-589.

Bay H., Ess A., Tuytelaars T. and Van Gool L. (2008) SURF: Speeded Up Robust Features, *Computer Vision and Image Understanding (CVIU)*, Vol. 110, No. 3, 346—359.

Bosché F., Tingdahl D., Carozza L., Van Gool L. (2012), Markerless vision based Augmented Reality for enhanced project visualization, *ISG*ISARC2012*.

Bouguet J. I. (2004) Camera Calibration Toolbox for MATLAB. www.vision.caltech.edu/bouguetj/calib_doc/

Bundler (2006) http://phototour.cs.washington.edu/bundler/.

Carozza L., Tingdahl D., Bosché F., and Van Gool L. (2012) Markerless vision-based augmented reality for urban planning, *Computer-Aided Civil and Infrastructure Engineering (CACAIE)*, to appear (published online).

Cruz-Neira C., Sandin D. J., DeFanti T. A., Kenyon R. V., and Hart J. C. (1992) The CAVE - audio visual experience automatic virtual environment, *Commun. ACM*, Vol. 35, No. 6, pp. 64-72.

Dalgarno B. and Lee M. (2010) What are the learning a affordances of 3-d virtual environments?, British Journal of Educational Technology, vol. 41, pp. 10-32.

de Vries B., Verhagen S., and Jessurun A. J. (2004) Building management simulation centre, *Automation in Construction*, Vol. 13, 679-687.

Dong Z., Zhang G., Jia J., H. Bao (2009) Keyframe-based real-time camera tracking, *IEEE 12th International Conference on Computer Vision*, 1538-1545.

Feng Zhou, H.B.-L. Duh, Billinghurst M. (2008) Trends in augmented reality tracking, interaction and display: A

review of ten years of ISMAR, *Mixed and Augmented Reality ISMAR 2008. 7th IEEE/ACM International Symposium on*, 193-202.

Gauglitz S., Höllerer T. and Turk M. (2011) Evaluation of Interest Point Detectors and Feature Descriptors for Visual Tracking, *International Journal of Computer Vision*, Vol. 94, No. 3, 335-360.

Hartley R. and Zisserman A. (2003) Multiple View Geometry in Computer Vision (2 ed.), Cambridge University Press, New York, NY, USA.

InterSense IS-900 Wide Area Precision Motion Tracker. http://www.isense.com

Klein G. and Drummond T. (2003) Robust Visual Tracking for Non-Instrumented Augmented Reality, Proc. *Second IEEE and ACM International Symposium on Mixed and Augmented Reality (ISMAR'03)*, 113-122.

Lau Y. K. H., Chan, L. K. Y. and Wong, H K. (2007). A virtual container terminal simulator for the design of terminal operation, *International Journal on Interactive Design and Manufacturing*, Vol. 1, No. 2,107–113.

Ligorio G. and Sabatini A.M. (2013) Extended Kalman filter-based methods for pose estimation using visual, inertial and magnetic sensors: comparative analysis and performance evaluation, *Sensors*, Vol. 13, No.2, 1919-41.

Lim Hyon, Sinha S.N., Cohen M.F., Uyttendaele M. (2012) Real-time image-based 6-DOF localization in large-scale environments, *2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR),* 1043-1050.

Lowe D. G. (2004) Distinctive Image Features from Scale-Invariant Keypoints. *Int. J. Comput. Vision*, Vol. 60, No. 2, 91-110.

Perea L., How J., Breger L., Elosegui P. (2007) Nonlinearity in sensor fusion: Divergence issues in EKF, modified truncated SOF, and UKF, AIAA Guidance, Navigation, and Control Conference (GNC- AIAA-2007-6514).

Pintaric T. and Kaufmann H. (2007) Affordable Infrared-Optical Pose Tracking for Virtual and Augmented Reality, *IEEE VR Workshop on Trends and Issues in Tracking for Virtual Environments,* 44 - 51.

Skrypnyk I., Lowe D.G. (2004) Scene modelling, recognition and tracking with invariant image features, *Mixed and Augmented Reality (ISMAR 2004), Third IEEE and ACM International Symposium on*, vol., no., pp.110-119.

Snavely N., Seitz S. M., Szeliski R. (2007) Modeling the World from Internet Photo Collections, *International Journal of Computer Vision*, Vol. 80, No. 2, 189-210.

Tingdahl D., De Weerdt D., Vergauwen M., Van Gool L. (2012) WEAR++: 3D Model Driven Camera Tracking On Board the International Space Station, *International Conference on 3D Imaging*.

Vuzix, http://www.vuzix.com/home/, accessed March 2013.

Wang X., Dunston P.S., Skiniewski M. (2004) Mixed Reality Technology Applications in Heavy Construction Equipment and Operator Training, *21st International Symposium on Automation and Robotics in Construction (ISARC 2004)*, 393-400.

Welch G., Bishop G., Vicci L., Brumback S., Keller K., Colucci D. (2001) High-Performance Wide-Area Optical Tracking: The HiBall Tracking System, *Presence: Teleoperators and Virtual Environments*, 2001, Vol . 10., No. 1, 1-21.

Woodward C., Hakkarainen M., Korkalo O., Kantonen T., Aittala M., Ranio K. and Khknen K. (2010) Mixed reality for mobile construction site visualization and communication, *International Conference on Construction Applications of Virtual Reality (CONVR)*, 35-44.

Yeh K., Tsai M., and Kang S. (2012) On-Site Building Information Retrieval by Using Projection-Based Augmented Reality, *J. Comput. Civ. Eng.*, Vol. 26, No. 3, 342–355.