# Visual Tracking vs Optical Tracking in Computer-Assisted Intervention

Ricardo Simões, Carolina Raposo, João P. Barreto, Philip Edwards, and Danail Stoyanov

***Abstract*—Computed-Assisted Intervention (CAI) aims to safely guide the surgeon during surgical interventions, which typically relies on Optical Tracking (OT) systems to provide the location of tools and instruments in a global reference frame, in real-time. Despite being very accurate, the existing OT systems have two main drawbacks: the difficulty of preserving lines-of-sight and the very high initial capital investment.**

**We propose a new Visual Tracking (VT) system that effectively overcomes these issues by making use of 3D visual markers and an inexpensive monocular camera that can be located relatively close to the patient's anatomy. Besides having these advantages, the new VT system also facilitates the navigation procedure by providing the guidance information in real-time using Augmented Reality.**

**A thorough experimental evaluation demonstrates the validity of our approach, which is as accurate as the state-of-the-art OT system Optotrak Certus and better than Polaris Spectra [1]. It is also significantly easier to use since the requirement of the existence of a line-of-sight is always satisfied. Results obtained on an experiment that mimics a common procedure in the OR, as well as preliminary cadaver trials, confirm that the proposed VT system is clinically viable, making it clear that this is an important advance in the literature of tracking for CAI.**

***Keywords*—3D Registration, Computed-Assisted Intervention, Optical Tracking, Visual Tracking**

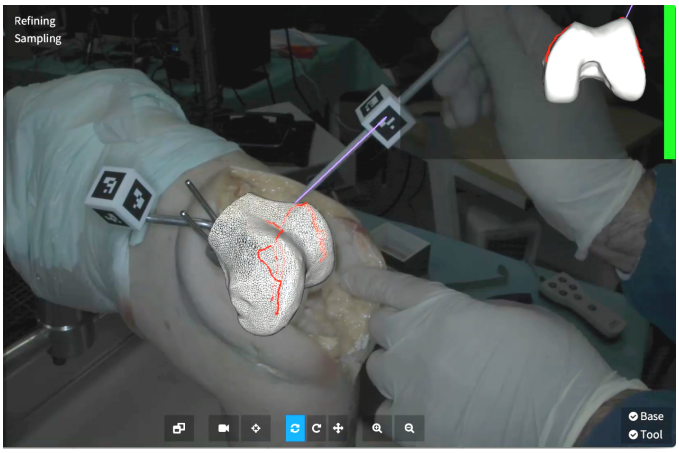



Fig. 1: The proposed VT system applied in a Total Knee Replacement (TKR) procedure performed on a cadaver (video in supplementary material). The system allows the registration of a pre-operative model with the bone and its subsequent overlay in the image using AR.

## I. Introduction

Computed-Assisted Intervention (CAI) aims to safely guide the surgeon during surgical procedures and relies on the existence of an accurate 3D model of the patient, typically obtained through CT-scan or MRI [2], [3], [4], [5], [6], [7]. This pre-operative 3D model is used in the offline stage of the CAI procedure for planning the surgery. The online stage consists in the intra-operative navigation, where the computer guides the surgeon to execute the procedure as planned.

Intra-operative navigation often requires the real-time localization of tools and instruments with respect to a reference frame rigidly attached to the targeted bone or organ. Also, in order to use the information from the offline planning stage, the 3D model of the bone may need to be overlaid with the actual bone. To accomplish these tasks, it is common to use Optical Tracking (OT) systems that consist of a stationary tower equipped with at least two infrared (IR) cameras (base station) for tracking a set of fiducial markers that are rigidly attached to an instrument or bone [8], [9], [10]. There are two types of markers commonly used by OT systems: active and passive. Active markers emit IR light which is detected by the tower, whereas passive markers are made with highly reflective material to reflect IR light emitted by the tower, allowing their identification. Both types of markers present similar accuracies, with the active markers being slightly more accurate [8] but needing a clear line-of-sight for being identified. Unambiguous active marker identification requires multiplexed IR LEDs for only one marker to be seen by the camera system at a time. The position of each marker is estimated by simple triangulation, allowing the 3D pose of the object of interest to be computed in the reference frame of the tower [11].

Another approach for 3D pose estimation is to use electromagnetic tracking (EMT). EMT systems produce low intensity electromagnetic fields inducing electrical current in the probes' sensors. The fields are time multiplexed and produced with alternating current to facilitate probe detection. Position and orientation are determined by analyzing the obtained field strength in the probe and the direction of the generated magnetic field, respectively [12].

In most existing surgical navigation solutions, the surgeon starts by attaching a marker to the patient or targeted organ (world marker) and then uses a calibrated probe that is instrumented with another marker for pinpointing anatomical

Ricardo Simões, Carolina Raposo and João P. Barreto are with the Institute of Systems and Robotics, University of Coimbra, Portugal. Carolina Raposo and João P. Barreto are also with Perceive3D, Portugal.

Philip Edwards and Danail Stoyanov are with the Centre for Medical Image Computing, University College London, UK

landmarks. The tracking system provides the location of these landmarks in the world marker reference frame, allowing the pre-operative model to be registered with the patient. This provides a safe guidance to the surgeon throughout the medical procedure since the pose of instruments with respect to the patient can be determined in real-time.

Despite providing reliable 3D information in real-time, the presented tracking systems have several drawbacks that hamper the broad dissemination of navigation systems. The first is that both OT and EMT require a significant investment in capital equipment due to the necessity of acquiring a base station. Also, and regarding OT, another important disadvantage is the fact that it requires that the base station has a clear line-of-sight to all the markers, constraining both the layout of the Operating Room (OR) and the movements of the medical team members. Although EMT does not have this line-of-sight requirement, it has the problem of being affected by the presence of metal objects or other electrical devices, providing less accurate measurements than OT [13].

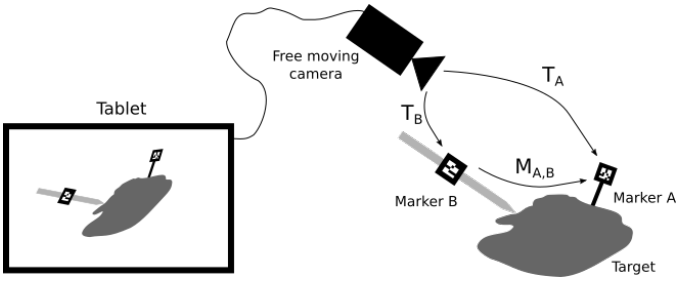
This paper proposes an image processing pipeline for detecting markers that can be used as an alternative tracking system which has the advantage of overcoming the problems inherent to the aforementioned schemes. The idea consists in attaching recognizable visual markers to the bone and tools, and using a monocular camera that can be freely moved to estimate their relative pose in 3D. The marker that is rigidly attached to the bone works as an absolute reference since all the measurements are made with respect to it. This visual tracking (VT) system provides real-time 3D information of the bone surface, allowing its registration with pre-operative models. Also, the usage of augmented reality (AR) facilitates the navigation process by providing a more intuitive guidance. Figure 1 illustrates these features by showing the VT system in operation in the Total Knee Replacement (TKR) procedure performed on a cadaver specimen.

Since VT uses only a small camera located relatively close to the markers and that can be freely moved, instead of a large base station, the issues related with the preservation of lines-of-sight in the OR and high initial capital investment are automatically overcome. Also, the vast experimental evidence provided in this paper demonstrates the robustness and accuracy of VT, as well as its superiority with respect to a commonly used OT system, the NDI Polaris Spectra. Experiments also show that the proposed VT system is as accurate as the NDI Optotrak Certus, which is the state-of-the-art OT system [9], [14].

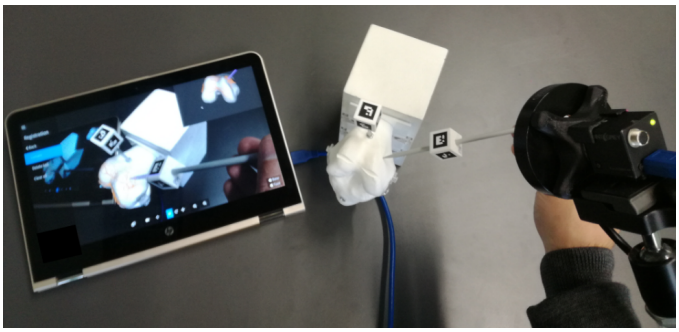
In summary, the article proposes a new alternative for obtaining the pose of 3D objects in the OR with important advantages in terms of accuracy, usability and overall cost. The contribution can be a game changer in CAI and Medical Robotics, overcoming many of the difficulties in usability and cost that explain their current low penetration (less than 5%).

### A. Overview

We start by overviewing the VT concept in Section II. Section III describes a standard pipeline for detecting and estimating the pose of visual markers and performs a first study on accuracy under different lighting conditions. Section IV shows how to improve the image processing pipeline to operate with good accuracy across different circumstances. The pipeline is extended to use 3D objects with multiple fiducial markers in Section V. Finally, Section VI reports a set of experiments that compare the accuracies of the VT and OT systems.




(a) VT concept scheme with relations between camera and fiducial markers.



(b) Proposed VT system applied in a TKR simulated scenario.

Fig. 2: Illustration of the proposed VT system, with (a) depicting a free moving camera connected to a tablet/laptop that observes Marker A, which is rigidly attached to the target object, and a touch probe instrumented with Marker B. (b) shows this setup being applied to a simulated scenario of TKR.

## II. Visual Tracking concept

The VT pipeline starts by detecting two visual markers, with one being rigidly attached to the bone, to work as the world marker (WM), and the other being instrumented in a touch probe, referred to as the tool marker (TM). Since it is assumed that the intrinsic parameters of the camera are known as well as the 3D coordinates of each marker, their poses w.r.t. the camera reference frame can be determined using homography [15], in case of planar markers, or Perspective-n-Point (PnP) [16] estimation, otherwise, by using the corresponding 2D points in the image. Whenever the WM and the TM are simultaneously visible in the image, the pose of the TM in WM coordinates can be computed by using as anchor the camera reference frame. Supposing that the touch probe is calibrated, i.e., the 3D coordinates of its tip w.r.t. the TM are known, it is possible to reconstruct 3D points that are pinpointed in the bone surface with the probe and represent them in WM coordinates. This brings an important advantage which is the fact that even if

the tracking process is interrupted, all the previously acquired information is restored when the WM becomes visible.

By using a high frame rate HD camera (30-60 fps) and a regular laptop or tablet, it is possible to perform the tracking and 3D reconstruction of points in real-time, as is done in OT systems. However, there is an important difference w.r.t. OT, which is the size of the required equipment. In VT, the camera can move freely, which means that it can either be placed in an holder or in the light over the operating field, it can be hand-held by the assistant, be in a robotic arm or even be in the surgeon's forehead, opening the way for future integration of AR headsets such as HoloLens™ (Microsoft Corporation, USA)) [17], [18], [19].

The VT concept is illustrated in Figure 2 as a scheme (2(a)) and being applied to a simulation of a TKR procedure (2(b)). It is object of patent application by the University of Coimbra (UC) and is being developed by a spin-off company, Perceive3D [20].

## III. 3D POSE ESTIMATION OF VISUAL PLANAR MARKERS: EVALUATION OF A STANDARD IMAGE PROCESSING PIPELINE

### A. Overview of the image processing pipeline

Our first attempt to detect visual markers is performed using a standard implementation of ALVAR [21], which is a library for virtual and augmented realities that allows the detection, identification and pose estimation of fiducial markers. Although there exist other solutions with code publicly available (e.g. ARToolkit [22]), our preliminary analysis showed that ALVAR is the one with the best trade-off between accuracy and computational complexity.

The camera is assumed to be calibrated with its radial distortion properly modelled [23]. We use 15mm square-shaped markers with binary codes that enable unique identification, similar to the ones used in calibration [24] and AR [25]. The binary codes are selected in order to assure rotation invariance and to maximize the Hamming distance. ALVAR takes as input an image and outputs both the 3D pose with respect to the camera and the corner points $\mathbf{P}_j, j = 1, \ldots, 4$ of each detected fiducial marker.

ALVAR starts by applying adaptive binarization to the image, and then performs blob detection for finding putative locations for the markers. Lines are fitted to the edges of each blob and the ones that yield sets of 4 lines are potential markers. Then, line intersection is performed for finding the corner points. The four 2D corners and the corresponding known 3D coordinates are used for estimating an homography $\mathsf{H} = \mathsf{R}+\mathbf{t}\mathbf{n}^\mathsf{T}/d$, where $\mathsf{R}$ is the rotation, $\mathbf{t}$ is the translation and $\mathbf{n}, d$ are the plane parameters (normal and distance to origin). Since the 3D coordinates of the marker corners are known, this becomes a model-to-view homography [26], having the simplified form $\mathsf{H} = \lambda\,[\mathbf{r}_1 \quad \mathbf{r}_2 \quad \mathbf{t}]$, where $\mathbf{r}_1$ and $\mathbf{r}_2$ are the first and second columns of $\mathsf{R}$, and $\lambda$ is a scalar. From this homography, $\mathsf{R}$ and $\mathbf{t}$ can be estimated in a unique manner. First, $\lambda$ is determined from $\lambda = ||\mathbf{h}_1||$, with $||\mathbf{h}_1||$ being the L2-norm of the first column of $\mathsf{H}$. Then, $\mathbf{r}_1$, $\mathbf{r}_2$ and $\mathbf{t}$ come in a straightforward manner from the columns of $\lambda^{-1}\mathsf{H}$. The full rotation matrix is determined by $\mathsf{R} = [\mathbf{r}_1 \quad \mathbf{r}_2 \quad \mathbf{r}_1 \times \mathbf{r}_2]$.

### B. Evaluation Setup

The proper functioning of a VT system requires that high accuracy under a relatively large nearby volume is achieved, as well as resilience to different lighting conditions. In this section, we perform a set of experiments to assess the performance of our proposed VT system under these different circumstances.

Figure 3 illustrates the experiment that tries to mimic the operating conditions in the OR for the case of the camera being in a holding arm or hand-held by the assistant. The test consists in freely moving a ruler with 4 distinct markers in a workspace in front of the camera. The camera continuously acquires video and the system computes the relative pose between all visible markers. The estimated relative poses are compared against the ground truth (GT) and the errors in translation and rotation are analysed.

In more detail, for each frame $i$, the motion errors are computed as follows. Let $\mathsf{T}^i_k, k = 1, \ldots, 4$ be the pose estimated for marker $k$ in frame $i$. For all possible pairs of markers (6 in total), the relative pose $\mathsf{M}^i_{A,B}$ between markers $A$ and $B$ comes from

$$\mathsf{M}^i_{A,B} = {\mathsf{T}^i_B}^{-1}\mathsf{T}^i_A, \quad A, B = 1, \ldots, 4 \wedge A < B. \tag{1}$$

The residual transformation in frame $i$ between markers $A$ and $B$, $\mathsf{E}^i_{A,B}$, is computed using the GT transformation:

$$\mathsf{E}^i_{A,B} = \mathsf{GT}^{-1}_{A,B}\mathsf{M}^i_{A,B}, \tag{2}$$

and its rotation and translation components give the motion errors. The translation error is simply the norm of the translation component of $\mathsf{E}^i_{A,B}$ divided by the known distance between the markers, in order to provide an analysis that is independent of the distance. The rotation error is given by the angle of its rotation component after transforming it into an axis-angle representation using Rodrigues' Formula [27].

In a real operation scenario the distance between markers can range from a few centimetres to about 15cm. In order to simulate this, we placed the markers in the ruler in a straight line, each being at a distance of 50 mm from its adjacent ones, so that the GT transformation $\mathsf{GT}$ between all markers is known. This allows us to independently consider three datasets: S1 (50 mm) with 3 measures per frame, S2 (100 mm) with 2 measurements per frame, and S2 (150 mm) with 1 measurement per frame.

Also, in a real operation scenario, the work volume can roughly range from 100 to 250 mm in depth from the camera. Thus, acquisition was made at 3 different depths: D1 at 100 mm, D2 at 175 mm and D3 at 250 mm, with slant being consistently applied in all directions, i.e., the ruler did not strictly move in fronto-parallel planes. For each depth, the volume of operation was roughly $200 \times 200 \times 100$mm, as depicted in the diagram of Figure 3(a).

This setup enables us to study in detail how the pose estimation accuracy varies both with the distance between the markers and the distance to the camera.

Since the work volume is considered relatively near to the camera, we need to focus the lens at a finite distance, while keeping good depth of field. A camera with short focal

length is adequate, not only because of its wide field of view (FOV), but also because it tends to have better depth of field than cameras with larger focal length. Therefore, we used a Point Grey Flea3$^{TM}$ (FLIR Integrated Imaging Solutions Inc.) camera, with resolution of 1920x1080 pixels (HD) working at 30-60fps and equipped with a 1.4/3.5mm lens in the experiments. However, and in order to have good image quality, the tuning of exposure/gain is also critical, which puts challenges in terms of balancing noise while assuring invariance to light conditions.

In order to evaluate these trade-offs, we ran tests under the following different lighting conditions (Figure 3(b)):

- NATURAL: diffuse daylight illumination
- ARTIFICIAL: artificial ambient illumination
- OR: strong incident light as the one often used in the OR

with different camera settings (exposure/gain parameters):

- AUTO: automatic camera parameters
- MANUAL 1: ambient adjusted parameters (high gain to compensate low ambient light)
- MANUAL 2: OR adjusted parameters (low gain to compensate strong incident light)
- RING: Ring Light adjusted parameters (similar to OR). In this case, there is an extra source of light which is a ring placed around the camera lens.

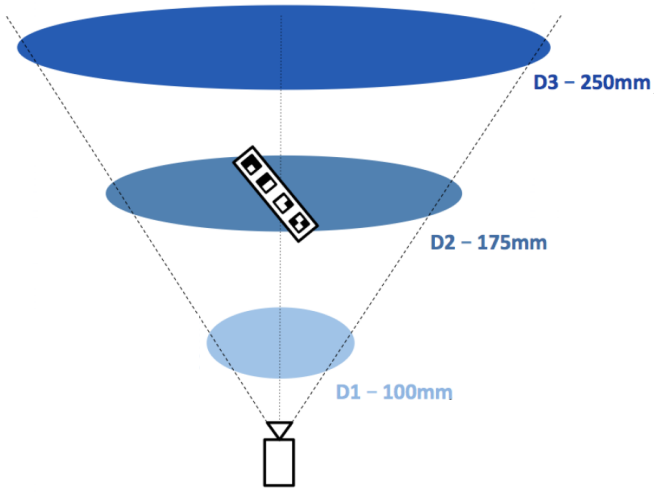
### C. Experimental Evaluation

The initial set of experiments only considers the first three camera settings AUTO, MANUAL 1 and MANUAL 2, since the ring light was only included in a later stage (Section IV). The obtained motion errors are shown in the first three rows of Figure 4, as well as the detection rates, which are given by the ratio between the number of markers detected in all frames and the total number of markers ($4 \times N$, with $N$ being the number of frames). For each plot, different boxplots and bar graphs correspond to different depths of acquisition (D1, D2 and D3 in Figure 3(a)) in order to enable an analysis as a function of the distance of the ruler to the camera. The best results obtained for each light condition are surrounded by a green box.

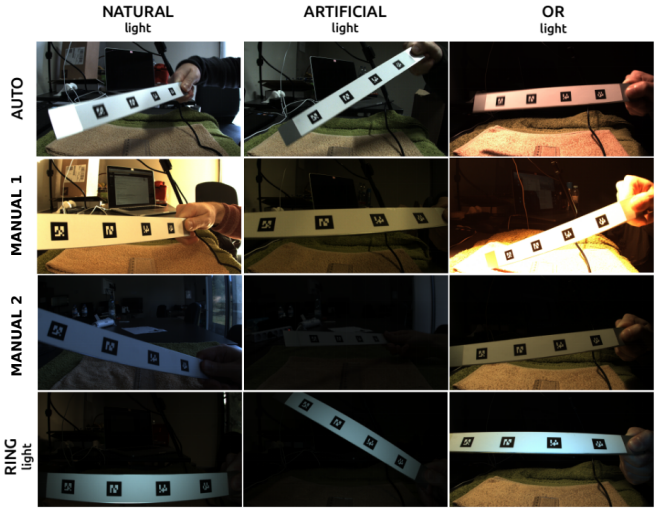
The presented motion errors demonstrate that, as expected, higher accuracies are obtained for the camera setting that corresponds to the light condition, i.e, for illuminations NATURAL and ARTIFICIAL the best results were obtained with setting MANUAL 1 and for OR illumination the most appropriate setting is MANUAL 2. In these cases, high accuracies were achieved, with the third quartiles being around 1.5% in translation and 1° in rotation. In the remaining cases, errors over 2.5% in rotation and 1.5° in translation are achieved. The AUTO camera setting is the one that provides the lowest accuracies, independently of the light condition. This can be explained by the fact that the variation of exposure/gain leads to out of focus situations, making the results quite unpredictable. In general terms, it can also be observed that motion errors tend to slightly increase with the distance to the camera, which is expected since the markers appear smaller in the image.

Concerning the detection rate, it can be seen that for light conditions NATURAL and ARTIFICIAL, the highest rates are achieved with setting MANUAL 1, being only slightly superior than those obtained with the AUTO setting. For OR illumination, all settings provide good detection rates.

In conclusion, the MANUAL 1 and MANUAL 2 camera settings yield considerably smaller errors than the AUTO mode, for all light conditions except for the combination MANUAL 1/OR where overexposure is observed. This indicates that constant tuning clearly leads to better accuracy. However, and despite our efforts, it is difficult to have a tuning that works well under all circumstances (see Figure 3(b)).




(a) A ruler with 4 markers is moved in front of the camera at different depths, in order to simulate the possible situations of a real operation scenario and assess how the distance to the camera affects the estimation accuracy.




(b) Images acquired with different illumination conditions and camera settings.

Fig. 3: (a) Scheme of a four-marker ruler for image acquisition at multiple distances, and (b) a set of example images of the four-marker ruler for all the possible combinations of illumination conditions and camera pre-sets.

## IV. Improving accuracy and robustness to illumination conditions

Despite the improvements obtained with the tuning of camera settings, results for NATURAL/ARTIFICIAL and OR light conditions are still far from the desired accuracies. Moreover, a VT system would greatly benefit from having a camera setting solution that provides accurate results for all lighting conditions.

With the intent of obtaining such a solution, we propose modifications to the image processing pipeline, as well as the inclusion of an extra source of illumination that is attached to the camera. This extra source of light serves to homogenize the illumination conditions, regardless of the amount and type of external illumination present in the scene. It is a strong LED ring light that surrounds the lens of the camera and a new tuning of parameters is considered, where we set the exposure/gain parameter to a low value. The last row of Figure 3(b) shows examples of images acquired under the ring light conditions (new light + new tuning), demonstrating its effectiveness in making acquisition conditions more homogeneous, across all types of light conditions.

### A. Improving Homography Estimation and Factorization

We propose a new pipeline for estimating the pose of a planar marker that includes several modifications when compared to the method described in Section III-A.

The input image is processed by ALVAR, that outputs a set of point locations corresponding to the corners of the marker. Then, the standard 4-point algorithm is used for estimating the homography, which is decomposed into two rigid motions $\mathsf{T}_1$ and $\mathsf{T}_2$ using the solution proposed by Bartoli *et al.* [26]. Next, the best motion hypothesis is selected as the one that yields the smallest reprojection error. Whenever the marker approaches a fronto-parallel configuration, the provided motion hypotheses become similar and it often occurs that the incorrect one provides the smallest reprojection error. In this case, temporal information is used, i.e., for a given frame, the pose of a marker is chosen as the one that is closest to the pose of the same marker in the previous frame. If the current frame is the first being processed or the system failed to detect the marker in the previous frame, the rigid motion that yields the smallest reprojection error is chosen. Moreover, if the difference between the average reprojection errors lies below a pre-defined threshold, which usually happens when both solutions are inaccurate, the motion hypotheses for that frame are discarded and the next one is processed. Finally, the selected pose is refined using photoconsistency, as presented in [28].

### B. Experimental Evaluation

Due to the poor performance of camera setting MANUAL 1 in OR illumination and camera setting MANUAL 2 in NATURAL and ARTIFICIAL illumination conditions, we do not further evaluate those combinations. Thus, for the second set of experiments, besides the new RING setting, we only consider the BEST MANUAL camera setting: MANUAL 1 for illumination conditions NATURAL and ARTIFICIAL and MANUAL 2 for the OR light condition. The fourth and fifth rows of Figure 4 present results for all these combinations of camera setting/light condition obtained with the new proposed image processing pipeline.

The advantages of the new pipeline become evident after comparing the results obtained with the BEST MANUAL camera setting (row 4 of Figure 4) with the results obtained in the first experiment (first three rows of Figure 4), where a dramatic improvement in accuracy while keeping very good detection rates is observed.

Concerning the new camera setting RING, results show that an effective normalization of the light conditions is achieved since similar accuracies are obtained for all the different illumination situations. Unlike the previous case where high accuracies where only obtained if the camera setting was adjusted to the existing light condition, in this case, the low gain/exposure parameter together with the ring light homogenizes the surrounding illumination, making the ruler properly visible in the image and thus facilitating detection and improving pose estimation accuracy. Besides normalizing the illumination conditions, the RING camera setting also typically leads to higher accuracies than BEST MANUAL.

Detection rates obtained with the new image processing pipeline reveal an overall slight decrease when compared to the ones obtained in the first experiment due to the new step of discarding poor solutions. However, for all study cases, the values are approximately 85% or higher, being sufficiently good for a practical application.

Moreover, the new proposed camera setting succeeds in normalizing the light conditions. This means that when using a ring light attached to the lens, we can work with constant camera pre-sets. In case we do not want to use the ring light, variable pre-sets must be employed.

## V. Extension to Multi-Markers - 3D Multiple planar markers

In this section, we extend the image processing pipeline to work with 3D fiducial markers. These 3D markers consist of cubes with edge length of 22mm and a 15mm squared planar marker in each face. The primary goal of using 3D fiducials is to improve usability by enabling the user to freely move the camera and/or instrumented tool while maintaining visibility. Another advantage is that more accurate pose estimations can be achieved due to the frequent fact that more than one face belonging to the same fiducial marker is observed simultaneously.

### A. Pipeline

In order to estimate the pose of a 3D fiducial marker, it is necessary that all the measurements performed on the object are represented in a common coordinate system, i.e., the 3D marker is calibrated. In the particular case of the fiducials being cubes, 4 points composing a square are detected in each face and their 3D coordinates must be represented in the same reference system. This can be accomplished by defining one of the faces as the base and then finding the transformation

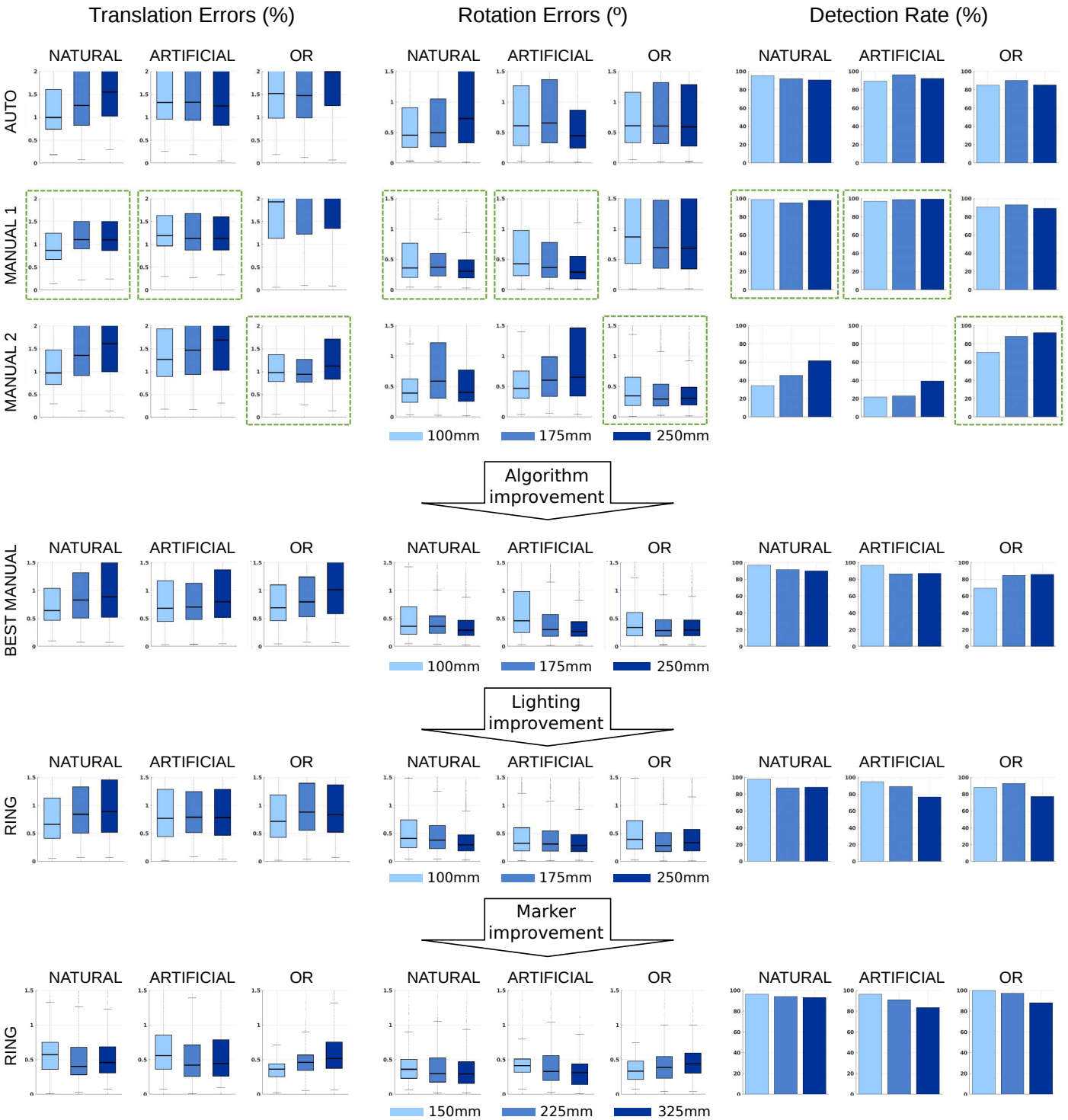



Fig. 4: Motion errors and detection rates of the different stages of the proposed VT system for different lighting conditions, camera settings and working distances. Different shades of blue corresponding to different working distances.

between each of the other faces and the base. For this, we start by acquiring a dataset where we move the cube in front of the camera such that all its faces are visible at least once. Then, we use the VT pipeline proposed in Section IV to estimate the pose of each face in camera coordinates. For each frame, we compute the poses between all pairs of visible markers

and feed their rotation components to the method presented in [29] for robustly performing relative rotation averaging. The translation components of all relative poses are represented in the reference frame of the base marker using these estimated average rotations and their average is computed by solving a linear system of equations. This allows the corners of any face to be represented in the coordinate system of the base marker.

Thus, at each frame, for a cube with $F$ visible markers, $F = 1, \ldots, 3$, its pose is estimated using $4F$ 2D-3D correspondences. When only one face is visible ($F = 1$), the pipeline estimates the pose using the algorithm proposed in Section IV. The largest value for $F$ is 3 because that is the maximum possible number of simultaneously visible faces.

The proposed image detection pipeline for 3D markers, that is only employed when more than one face is visible, starts by detecting and identifying each visible marker independently. Then, the 3D coordinates of each identified face are represented in the base marker reference frame using the calibration information and the corresponding 2D coordinates of the extracted corners yield 2D-3D correspondences. These matches are used as input to a PnP [16] algorithm for estimating the pose of the 3D marker. As a final step, the initial pose is refined with photoconsistency on all detected markers simultaneously.

### *B. Experimental Evaluation*

This section presents results to assess the performance of the pose estimation pipeline for 3D fiducial markers. For this, we built a new ruler with two cubes rigidly attached at a distance of approximately 200 mm. Differently from the single marker approach, the GT relative pose between the two cubes is not known so we created a pseudo-GT by acquiring a large set of images with both 3D markers visible in different poses, and in a controlled environment with good illumination. For each frame, the VT system presented in this section allows the estimation of the relative pose between the two markers. The robust rotation averaging method and subsequent translation averaging scheme used for calibrating the cubes are used for estimating a relative pose to be considered as pseudo-GT.

Similarly to the previous sections, the ruler was moved in front of the camera at depths $D1 = 150mm$, $D2 = 225mm$ and $D3 = 325mm$, and the results w.r.t. the pseudo-GT are presented in the last row of Figure 4. Results are only shown for the RING setting because the conclusions are similar for all light conditions and camera settings.

Comparing with the results from the previous section, a significant improvement both in motion estimation accuracy and rate of detection is observed, despite the working depths being larger (up to 325mm instead of 250mm). This indicates that besides the great advantage of removing the limitations of visibility, using 3D markers enables the system to achieve higher accuracies at longer distances from the camera, allowing the working volume to be increased. However, please take into account that the large decrease in motion errors is partially due to the usage of a pseudo GT and not a real GT. So, more important than the error values is their spreading, and results show that the error distributions are significantly narrower than the ones in the penultimate row of Figure 4, indicating that higher accuracies were achieved.

Another important aspect to highlight is that besides leading to higher accuracies, the RING camera setting significantly increases the detection rate in the OR illumination condition. This can be explained by the fact that the strong OR light creates shadows in the cube itself, making the area of the image where it appears to be very dark, precluding the detection of the marker. Excepting that case, all detection rates are around 90% or above, being about 10 percentage points higher than the ones in the penultimate row of Figure 4, evincing the increase in visibility provided by the 3D markers.

## VI. Visual tracking Vs Opto-Tracking

This section presents a thorough experimental validation of the accuracy and usability of our proposed VT system and two state-of-the-art OT systems: the NDI Optotrak Certus [1] and the NDI Polaris Spectra optical system. Although it is reported that the Optotrak Certus has a 3D accuracy of 0.1mm for each of the active markers in its official website [1], recent studies show that it reaches RMS errors of approximately 0.4mm when using a tool for registration with a set of at least four of these markers [30]. Also, the NDI Spectra reports errors of 0.25mm for each marker in the official website [31], but recent studies indicate errors around 0.9mm for a medical tool [30].

Two different experiments were performed, with the first having the intent to assess the accuracy of the tracking systems and the second serving to mimic a procedure that is common in CAI. All the experiments pass by using a calibrated touch probe to reconstruct 3D points on the surface of known objects. The touch probe used with Certus comprises 4 active optical markers and the one used with Spectra has 5 passive markers. Both touch probes have a cube with a planar visual marker in each face attached to them so that they are identified and tracked by our proposed VT system. The intent of using the same tool to acquire 3D data is to ensure fairness in the experiments.

In order to be able to reconstruct 3D points using any tracking system, the touch probe must be calibrated, i.e., the location of its tip w.r.t. its marker must be known. To accomplish this, a first step of calibrating the visual and optical (both passive and active) markers is performed where their relative poses are estimated. For the cube, the calibration method described in Section V-A is employed. For finding the relative poses between the individual optical markers, the tool is moved in front of the base station and NDI's proprietary software is used. After having the calibration of the 3D markers, the location of the tool tip can be determined by moving the tool around a fixed pivot point and considering another static cube as the world marker. We use VT to estimate the poses between the tool marker (TM) and the world marker (WM), and then estimate the pivot point in TM coordinates, providing the full calibration of the touch probe. An identical approach is used for calibrating the tool from the optical markers using NDI's proprietary software.

The remainder of this section describes the two different experiments that were performed, and reports the accuracies obtained with our proposed VT system and the NDI systems Optotrak Certus and Polaris Spectra. First, we measure the

distance between points with known locations on the faces of a quadrangular pyramid. Lastly, we simulate a procedure in the OR, where we have the 3D model of a knee bone and perform the registration of trajectories acquired on its surface with the tool.

### A. Measurement of distance

In the first experiment, the quality of all three tracking systems is assessed by measuring the distances between points at known locations and comparing with the ground truth distances. For this, we use a quadrangular pyramid with 19 small holes in each face, making a total of 76 points, to which we attach a visual and an optical markers to work as global reference frames. Keeping the object static, we place the tool tip in each hole to obtain its 3D position using each tracking system. Figure 5(a) shows the experiment being performed with Certus and Figure 5(b) shows the point reconstruction process using both the Spectra OT system and our proposed VT system. Borders in red and green correspond to Certus and Spectra, respectively.

During acquisition, we noticed that both OT systems require a very good visibility of the markers in order to provide a proper measurement, which was sometimes hard to accomplish due to line-of-sight occlusions. Despite our efforts in always placing the touch probe in the line-of-sight of the base station, it was impossible to do so in some of the points of the pyramid, which were not reconstructed.

In this experiment, we compute the 2850 distances between all possible pairs of points and define as error the absolute difference between the estimated distance and the GT distance. Since the experiment comprised two trials, a total of 5700 distances was computed for each acquisition procedure. The distributions of errors for VT, Certus and Spectra are shown in Figure 5(c), from left to right respectively. The circle in each boxplot represents the RMS error. The figure also shows the detection rate for all three tracking systems.

The first conclusion to be drawn is that Spectra reveals the worst accuracy results, meaning both median and RMS values are worse than for the other two systems under analysis. VT and Certus present a similar accuracy, with the latter being slightly more precise, i.e. the dispersion of errors is larger for VT. However, both systems achieve high accuracy, with median errors below 0.5mm and differing only approximately 0.15mm between each other.

Regarding the detection rates, VT significantly outperforms Certus and Spectra. This is caused by the fact that the NDI systems require a clear line-of-sight between the tower and the markers, which cannot always be satisfied. In this experiment, the two faces of the pyramid that are closest to the tower were properly seen by the OT systems, while the remaining faces did not have good visibility, and thus many of their points could not be reconstructed. Certus, having active markers, demonstrates a better detection rate than Spectra but both systems are deficient in certain acquisition conditions. The proposed VT system overcomes this issue not only by using 3D markers that can be detected independently of the viewing direction but also because the camera can move freely, unlike the towers of the OT systems. This lack of visibility can be the reason for Certus slightly outperforming VT in terms of accuracy since the points that are farther away from the tower, and thus more difficult to reconstruct, are not included in the statistics.

In summary, this experiment demonstrates that besides being comparable in accuracy with Certus, our VT system presents much better usability and higher practicality, having a detection rate of 100%. When compared to the most widely used system in the OR, Spectra, VT is not only more practical but also significantly more accurate, proving to be a viable alternative to the state-of-the-art OT systems.

### B. Registration

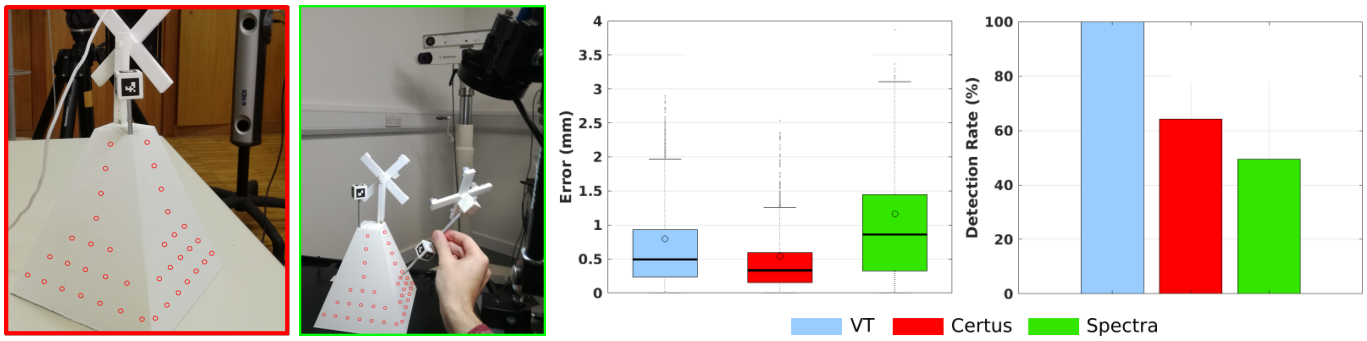
This experiment attempts to mimic a common CAI procedure performed in the OR, for instance in TKR, where the patient's knee is registered with a pre-operative model obtained through CT-scan or MRI, allowing navigation during the medical intervention.

To accomplish this, we printed a 3D model of a knee to work as the real bone, to which we attached a visual and an optical markers, as in the previous experiment, to work as world markers (WM). Also, a calibrated tool is used to reconstruct 3D points on the surface of the bone using the VT and the OT systems. The printed model of the knee and the data acquisition setup are shown in Figure 6(a).

In this experiment, we acquired trajectories using Certus, Spectra and VT, mainly containing points in the superior zone of the condiliar region, which is enclosed in a purple line in Figure 6(a), and, for each trajectory, performed the registration with the virtual model, which is a dense point cloud, using the algorithm proposed in [32] for curve-vs-surface registration. The acquisition was performed by keeping the knee static and freely moving the camera, in order to mimic a real operation scenario. Figure 6(b) depicts the steps of trajectory acquisition and registration, with the result of the latter being shown using Augmented Reality (AR) by overlaying the registered virtual model with the bone.

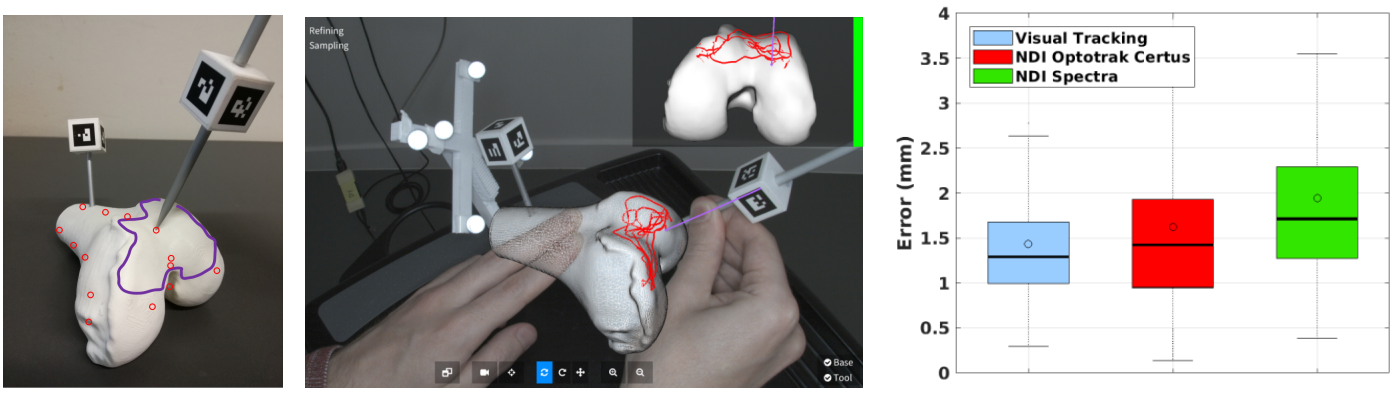
The registration result allows the virtual model to be represented in the WM's reference frame. Using this information, it becomes possible to compute the distance between 3D points reconstructed on known locations of the knee and the corresponding points on the virtual model. Our printed knee has small holes in pre-defined locations, whose 3D coordinates in the virtual model are known. We reconstruct the points corresponding to the 23 landmarks shown as small red circles in Figure 6(a) by placing the tool tip in each hole and, for each registration result, we compute the distance between every reconstructed point and the corresponding 3D point in the virtual model. This is the concept of Target Registration Error (TRE) [14] that is commonly used to measure accuracy in computer and robotic systems.

The distribution of distances is given in Figure 6(c), as well as the RMS value in colored circles. It can be seen that the VT system is the most accurate one, yielding a narrower distribution of distances and a lower median value than the competing systems. Similarly to the previous experiment, Spectra presents




(a) Test with Certus. (b) Test with Spectra. (c) Distance errors and detection rates for all possible combinations of two points in the pyramid.

Fig. 5: Experiment for the measurement of distance between all possible pairs of points reconstructed on the surface of a quadrangular pyramid. (a) and (b) show the point locations as red circles. (c) depicts the distribution of errors and the detection rates for all three tracking systems.




(a) Dry knee model. (b) Trajectory acquired with VT and registration shown with AR. (c) Point to model distance distribution boxplots.

Fig. 6: Assessment of the registration accuracy using trajectories acquired with the proposed VT system and the competing OT systems. (a) A dry knee model is shown with the area of acquisition highlighted, as well as the locations of the reconstructed points. (b) The VT system being applied for performing the registration of a 3D curve with a pre-operative model. (c) Registration errors obtained with all three systems.

the worst performance with the reconstructed points being typically more than 1.5mm away from the registered model.

The results of this experiment are coherent with the previous ones, and the higher accuracy of VT can be explained by the fact that more points are reconstructed than using OT that requires clear lines-of-sight, thus leading to more complete reconstructions of the bone surface and consequently better registrations. This confirms the high accuracy and usability of our proposed VT system, and proves that it is a valid alternative to the commonly used NDI systems, Certus and Spectra.

## VII. Conclusion

This paper proposes a new VT system to be used in CAI as a replacement for the commonly used OT and EMT systems that allow intra-operative navigation by locating in real-time the tools and instruments w.r.t. a global reference frame.

Our system requires only the usage of visual markers that are attached to the bone and tools, and a monocular camera that can be freely moved. The system is capable of working with visual markers that can either be planar or 3D patterns. This brings important advantages w.r.t. the existing tracking systems, namely the ease of preserving lines-of-sight, the possibility of using AR to facilitate the navigation process and the fact that it does not require a high initial capital investment.

The thorough experimental validation presented in this paper allows us to draw important conclusions. The first is that using 3D markers instead of planar markers is highly advantageous in terms of usability because visibility is always maintained and the pose estimation accuracy is equally high. Moreover, the comparison with the state-of-the-art OT systems Optotrak Certus and Polaris Spectra, showed that besides being much easier to use, due to the facility of preserving lines-of-sight,

our system is more accurate than the most used OT system in the OR, the Polaris Spectra, and presents similar accuracies to the gold-standard tracking system, the Optotrak Certus. The experiments also show that the VT system's accuracy is not affected by movement of the markers during data acquisition, meaning that the camera can be freely moved. This is important not only because of visibility issues, but also because, during navigation, it allows the visualization of the patient's anatomy from different viewpoints. As long as the marker that is attached to the bone is visible, all the information from the planning can be overlaid on the targeted organ. This analysis, together with the preliminary tests performed on a cadaver, demonstrate that the VT system proposed in this paper is a viable alternative to the tracking systems that are currently used in CAI and thus an important advance in the literature.

## ACKNOWLEDGMENT

The authors thank the Portuguese Science Foundation and COMPETE2020 program for generous funding through project VisArthro (ref.: PTDC/EEIAUT/3024/2014). This paper was also funded by the European Unions Horizon 2020 research and innovation programme under grant agreement No 766850.

## REFERENCES

[1] NDI, "Ndi optotrak certus." [Online]. Available: https://www.ndigital.com/msci/products/optotrak-certus/

[2] F. Cenni, A. Timoncini, A. Ensini, S. Tamarri, C. Belvedere, V. D'angeli, S. Giannini, and A. Leardini, "Three-dimensional implant position and orientation after total knee replacement performed with patient-specific instrumentation systems," *Journal of Orthopaedic Research*, vol. 32, no. 2, pp. 331–337, 2014.

[3] M. Allan, S. Thompson, M. J. Clarkson, S. Ourselin, D. J. Hawkes, J. Kelly, and D. Stoyanov, "2d-3d pose tracking of rigid instruments in minimally invasive surgery," in *International Conference on Information Processing in Computer-assisted Interventions*, 2014, pp. 1–10.

[4] S. S.-P. Hsu, J. Gateno, R. B. Bell, D. L. Hirsch, M. R. Markiewicz, J. F. Teichgraeber, X. Zhou, and J. J. Xia, "Accuracy of a computer-aided surgical simulation protocol for orthognathic surgery: a prospective multicenter study," *Journal of Oral and Maxillofacial Surgery*, vol. 71, no. 1, pp. 128–142, 2013.

[5] R. H. Taylor, A. Menciassi, G. Fichtinger, P. Fiorini, and P. Dario, "Medical robotics and computer-integrated surgery," in *Springer handbook of robotics*. Springer, 2016, pp. 1657–1684.

[6] K. A. Rodby, S. Turin, R. J. Jacobs, J. F. Cruz, V. J. Hassid, A. Kolokythas, and A. K. Antony, "Advances in oncologic head and neck reconstruction: systematic review and future considerations of virtual surgical planning and computer aided design/computer aided modeling," *Journal of Plastic, Reconstructive & Aesthetic Surgery*, vol. 67, no. 9, pp. 1171–1185, 2014.

[7] W. Barrett, D. Hoeffel, D. Dalury, J. B. Mason, J. Murphy, and S. Himden, "In-vivo alignment comparing patient specific instrumentation with both conventional and computer assisted surgery (cas) instrumentation in total knee arthroplasty," *The Journal of arthroplasty*, vol. 29, no. 2, pp. 343–347, 2014.

[8] A. D. Wiles, D. G. Thompson, and D. D. Frantz, "Accuracy assessment and interpretation for optical tracking systems," in *Proceedings of SPIE*, vol. 5367, 2004, pp. 421–432.

[9] NDI, "Technical specification sheet, northern digital inc." 2007.

[10] E. B. Strong, A. Rafii, B. Holhweg-Majert, S. C. Fuller, and M. C. Metzger, "Comparison of 3 optical navigation systems for computer-aided maxillofacial surgery," *Archives of Otolaryngology–Head & Neck Surgery*, vol. 134, no. 10, pp. 1080–1084, 2008.

[11] J. G. Webster and H. Eren, *Measurement, instrumentation, and sensors handbook: spatial, mechanical, thermal, and radiation measurement*. CRC press, 2014, vol. 1.

[12] A. M. Franz, T. Haidegger, W. Birkfellner, K. Cleary, T. M. Peters, and L. Maier-Hein, "Electromagnetic tracking in medicinea review of technology, validation, and applications," *IEEE Transactions on Medical Imaging*, vol. 33, no. 8, pp. 1702–1725, Aug 2014.

[13] J. B. Hummel, M. R. Bax, M. L. Figl, Y. Kang, C. Maurer, W. W. Birkfellner, H. Bergmann, and R. Shahidi, "Design and application of an assessment protocol for electromagnetic tracking systems," *Medical physics*, vol. 32, no. 7, pp. 2371–2379, 2005.

[14] T. Haidegger, P. Kazanzides, I. Rudas, B. Benyó, and Z. Benyó, "The importance of accuracy measurement standards for computer-integrated interventional systems," 09 2018.

[15] C. Mei, S. Benhimane, E. Malis, and P. Rives, "Efficient homography-based tracking and 3-d reconstruction for single-viewpoint sensors," *Transactions on Robotics*, vol. 24, no. 6, pp. 1352–1364, Dec. 2008.

[16] V. Lepetit, F.Moreno-Noguer, and P.Fua, "Epnp: An accurate o(n) solution to the pnp problem," *IJCV*, vol. 81, no. 2, 2009.

[17] N. Cui, P. Kharel, and V. Gruev, "Augmented reality with microsoft hololens holograms for near infrared fluorescence based image guided surgery," in *Proc. of SPIE Vol*, vol. 10049, 2017, pp. 100 490I–1.

[18] O. F. Rahman, M. Y. Nahabedian, and J. C. Sinkin, "Augmented reality and wearable technology in image-guided navigation and preoperative planning." *Plastic and reconstructive surgery. Global open*, vol. 4, no. 9, pp. e1057–e1057, 2016.

[19] A. Hamacher, S. J. Kim, S. T. Cho, S. Pardeshi, S. H. Lee, S.-J. Eun, and T. K. Whangbo, "Application of virtual, augmented, and mixed reality to urology," *International neurourology journal*, vol. 20, no. 3, p. 172, 2016.

[20] J. Barreto, "Methods and systems for computer-aided surgery using intra-operative video acquired by a free moving camera," 09 2016. [Online]. Available: http://www.google.com/patents/WO2016154557A1?cl=en

[21] ALVAR, "A Library for Virtual and Augmented Reality," http://virtual.vtt.fi/virtual/proj2/multimedia/alvar.html, 2011, [Online; accessed 08-August-2017].

[22] H. Kato, "Artoolkit," 1999. [Online]. Available: http://artoolkit.org

[23] J. Y. Bouguet, "Camera calibration toolbox for matlab," 2008. [Online]. Available: http://www.vision.caltech.edu/bouguetj/calib_doc/

[24] B. Atcheson, F. Heide, and W. Heidrich, "Caltag: High precision fiducial markers for camera calibration." in *VMV*, vol. 10, 2010, pp. 41–48.

[25] R. Munoz-Salinas, "Aruco: a minimal library for augmented reality applications based on opencv," *Universidad de Crdoba*, 2012.

[26] T. Collins and A. Bartoli, "Infinitesimal plane-based pose estimation," *Int. J. Comput. Vision*, vol. 109, no. 3, pp. 252–286, Sep. 2014. [Online]. Available: http://dx.doi.org/10.1007/s11263-014-0725-5

[27] R. M. Murray, Z. Li, S. S. Sastry, and S. S. Sastry, *A mathematical introduction to robotic manipulation*. CRC press, 1994.

[28] C. Raposo, M. Lourenço, M. Antunes, and J. P. Barreto, "Plane-based odometry using an rgb-d camera," in *BMVC*, 2013.

[29] A. Chatterjee and V. M. Govindu, "Efficient and robust large-scale rotation averaging," in *The IEEE International Conference on Computer Vision (ICCV)*, December 2013.

[30] J. Burgner, A. Simpson, J. Fitzpatrick, R. Lathrop, S. Herrell, M. Miga, and R. Webster, "A study on the theoretical and practical accuracy of conoscopic holography-based surface measurements: toward image registration in minimally invasive surgery," *Journal of Medical Robotics and Computer Assisted Surgery*, vol. 9, no. 2, pp. 190–203, 2013.

[31] NDI, "Ndi polaris spectra." [Online]. Available: https://www.ndigital.com/medical/products/polaris-family/#specifications

[32] C. Raposo and J. P. Barreto, "Systems and methods for 3d registration of curves and surfaces using local differential information," Provisional Patent 62471339, March 2017.