HAL
open science

An Optical Tracking System based on Hybrid Stereo/Single-View Registration and Controlled Cameras

Guillaume Cortes, Eric Marchand, Jérôme Ardouin, Anatole Lécuyer

To cite this version:

Guillaume Cortes, Eric Marchand, Jérôme Ardouin, Anatole Lécuyer. An Optical Tracking System based on Hybrid Stereo/Single-View Registration and Controlled Cameras. IEEE/RSJ Int. Conf. on Intelligent Robots and Systems, IROS'17 , Sep 2017, Vancouver, Canada. pp.6185-6190. hal-01562327

HAL Id: hal-01562327

https://inria.hal.science/hal-01562327v1

Submitted on 13 Jul 2017

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

An Optical Tracking System based on Hybrid Stereo/Single-View Registration and Controlled Cameras

Guillaume Cortes[1], Eric Marchand[2], Jérôme Ardouin and Anatole Lécuyer[3]

***Abstract*— Optical tracking is widely used in robotics applications such as unmanned aerial vehicle (UAV) localization. Unfortunately, such systems require many cameras and are, consequently, expensive. In this paper, we propose an approach to considerably increase the optical tracking volume without adding cameras. First, when the target becomes no longer visible by at least two cameras we propose a single-view tracking mode which requires only one camera. Furthermore, we propose to rely on controlled cameras able to track the UAV all around the volume to provide 6DoF tracking data through multi-view registration. This is achieved by using a visual servoing scheme. The two methods can be combined in order to maximize the tracking volume. We propose a proof-of-concept of such an optical tracking system based on two consumer-grade cameras and a pan-tilt actuator and we used this approach on UAV localization.**

I. Introduction

In recent years, research interest in robot localization has grown rapidly. Tracking systems are required to provide information to plan a trajectory or to assist human operations. This is for example the case, among many others, for medical robotics or unmanned aerial vehicles (UAV) localization. Such tracking systems could be based on ultrasound, magnetic, inertial or optical sensors. Nowadays it is usual to rely on optical tracking (such as Optotrak, Vicon, Optitrack, etc) for surgery [1], virtual reality [2] or UAV localization [3]. In this paper we propose a method that intends to extend the tracking volume of optical tracking systems using cameras mounted on robot heads. As a proof-of-concept, we will consider an UAV localization application.

Among these devices, outdoor UAV localization is commonly performed with on-board inertial and/or GPS localization that can be coupled with vision-based techniques [4] [5] [6]. They provide monocular-based techniques for motion estimation such as SLAM technique or optical-flow estimation. Those visual estimations are then merged with GPS or IMU data with a Kalman filter. These techniques are mainly used for outdoor UAV flying. A few low-cost tracking systems are available for indoor UAV localization where GPS data may not be available. For indoor localization [7] proposed an ultrasound technique that merges the data with IMU measurement in structured environment. Nevertheless ultrasound localization can introduce delay and the echos can disrupt the signal. To provide a more faithful signal [8] proposed a laser-based localization but due to the limited laser range and field of view the UAV could sometimes be lost. To overcome data loss from the previous approaches and from monocular vision, [3] introduced a stereo vision localization. However this technique restricts the tracking volume to the overlapping area of both camera views. Many industrial actors like Vicon use additional optical sensors to cover a larger multi-view volume (volume covered by at least two cameras). Indeed Vicon or Optitrack tracking systems are built with numerous sensors to provide a motion capture environment for UAV localization at an expensive price. Pan-tilt cameras were introduced by [9] in a motion capture system to increase the volume covered by multi-view localization while limiting the number of sensors. The system performs 3D reconstruction and feature-based pose estimation. However this study does not describe any camera control algorithms. Furthermore, the number of considered cameras may prevent the use of such approach for low-cost tracking systems.

In this paper, which extends our previous work [10], we present an approach which intends to maximize the tracking volume of optical tracking systems. It could also overcome some occlusion problems by relaxing the current constraints on camera positioning and multi-view requirements. As such, since adding cameras can be expensive and is not always possible (due to the lack of space) we propose not to use additional ones.

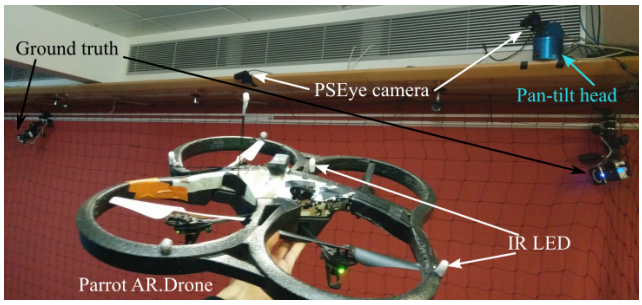


Fig. 1: The optical tracking prototype based on our approach.

To achieve this goal, the contributions of this paper are:

- When the tracked target is no longer visible by at least two cameras, we occasionally enable switching to a single-view tracking mode using 2D-3D registration algorithms (Figure 2b). The localization accuracy remains precise enough although more noisy.
- The proposition of a control scheme which allows the camera to remain centered on the target (Figure 2c). It allows to track the target through a larger volume. Thus, the multi-view registration can be achieved as much as possible. This is carried out by a visual servoing process with cameras mounted on a pan-tilt head.

[1]Realyz, Inria Rennes, France g.cortes@realyz.com
[2]Université de Rennes 1, IRISA, France marchand@irisa.fr
[3]Inria Rennes, France anatole.lecuyer@inria.fr

Our approach is then based on two complementary methods (Figure 2d) which allow an efficient tracking process with a minimal number of cameras.

II. Hybrid Stereo/Single-view Tracking

Our goal is to propose an approach that intends to maximize the optical tracking volume using a minimal number of cameras. This approach is composed of two methods that can be combined. In the following we consider a state-of-the-art stereo optical system composed of two cameras. Nevertheless everything can be transposed to multi-view systems [11] composed of N sensors. In this section we consider stationary cameras and we introduce our first method based on single-view tracking to increase the global tracking volume.

In this method the tracking can be performed with two techniques depending on if the stereo tracking is available. If stereo is available then the localization is performed with a registration between 3D points and a 3D model (3D-3D registration) otherwise it is performed with a registration between 2D points and a 3D model (2D-3D registration). As in many optical tracking devices, the system requires an off-line calibration process to determine the internal parameters of the cameras and their relative positions. Then, the on-line real-time tracking is performed.

Single-view tracking is generally constrained to use an external motion capture system to define the marker structure and the reference frame [12]. By using a stereo mode we are relieved of using external systems. The stereo mode enables reconstructing the targets' points and defining their structure. Moreover the reference frame $\mathscr{F}_w$ is set with a specific target that defines its y and x-axis.

A. Off-line system calibration

The calibration of the system is performed in two steps:

First, each camera is calibrated to determine its intrinsic parameters (focal length, principal point and distortion coefficients). Intrinsic calibration of the cameras is achieved by using a calibration chessboard and estimating the parameters with an algorithm based on [13].

The second step of the calibration process determines the essential matrices, ${}^{c'}\mathbf{E}_c$, relating each pair of cameras (c, c'). The essential matrices can be decomposed to recover the pose of camera c in camera c' frame, ${}^{c'}\mathbf{M}_c$ as follows:

$$ {}^{c'}\mathbf{E}_c = {}^{c'}\mathbf{R}_c[{}^{c'}\mathbf{t}_c]_\times \quad \text{with} \quad {}^{c'}\mathbf{M}_c = \begin{pmatrix} {}^{c'}\mathbf{R}_c & {}^{c'}\mathbf{t}_c \\ \mathbf{0}_{1\times 3} & 1 \end{pmatrix} \tag{1} $$

where $[{}^{c'}\mathbf{t}_c]_\times$ is the skew-symmetric matrix of vector ${}^{c'}\mathbf{t}_c$ and ${}^{c'}\mathbf{R}_c$ is a rotation matrix. Then by determining a reference frame $\mathscr{F}_w$, the pose ${}^{w}\mathbf{M}_c$ of each camera in the reference frame is computed. The essential matrix estimation is based on the normalized 8-points algorithm with RANSAC [11].

B. On-line real-time stereo tracking

The on-line real-time stereo tracking performs the localization of a target in the reference frame whenever the target is visible by both cameras. It first requires a 2D feature extraction to determine the position of the markers in the different images. Then the 2D features are correlated and triangulated to perform a 3D-3D registration.

1) 2D Feature extraction: The feature extraction determines the position of the bright markers of the target on the different camera images. A recursive algorithm is used to find the different sets of connected bright pixels before computing the barycenter of each set that defines the blob's positions. Once the blobs' positions are retrieved, they are corrected by taking into account the radial and tangential lens distortions.

2) 2D Feature correlation: The points from one image are associated with their corresponding points in the other images. This is possible by using the epipolar constraint that states that two corresponding image points $\mathbf{x}_c$ and $\mathbf{x}_{c'}$ related by ${}^{c}\mathbf{E}_{c'}$ should fulfill $\mathbf{x}_c^\top {}^{c}\mathbf{E}_{c'}\mathbf{x}_{c'} = 0$ where ${}^{c}\mathbf{E}_{c'}$ is the essential matrix computed through extrinsic calibration. This equation constraints the point $\mathbf{x}_{c'}$ to lie on the line directed by the vector $\mathbf{L}_c = \mathbf{x}_c^\top {}^{c}\mathbf{E}_{c'}$. Thus $\mathbf{L}_c$ is the epipolar line of $\mathbf{x}_c$ on the second image.

3) Triangulation: The triangulation process allows to recover a 3D point from its projections into several image planes. The computation of the 3D point depends on the relative position between the cameras that may vary when using controlled cameras. In practice, triangulation algorithms such as the mid-point or DLT [14] are adapted to determine the optimal 3D point.

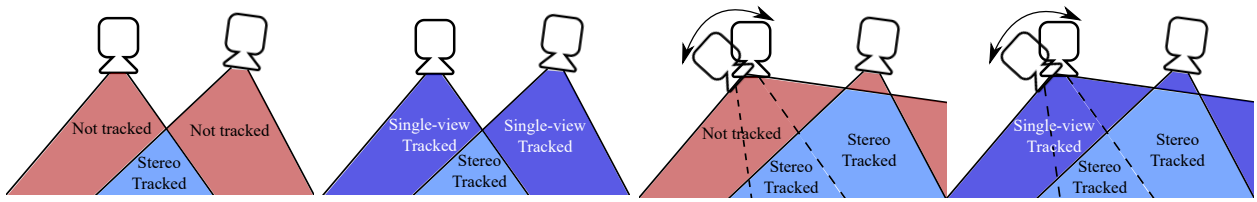
4) Registration: The final step of real-time stereo tracking recovers the pose (position and orientation) of the target in the reference frame (e.g. [15]). If the target is visible from several views then the registration matches a 3D point cloud to a 3D model. First the pose ${}^{c}\mathbf{M}_o$ of the target in the camera frame is estimated. This is achieved by minimizing the error between the 3D reconstructed points ${}^{c}\mathbf{X}_i$ (in the camera frame) and their corresponding 3D points ${}^{o}\mathbf{X}_i$ (in the object frame) transferred in the camera frame through ${}^{c}\mathbf{M}_o$. By denoting $\mathbf{q} = ({}^{c}\mathbf{t}_o, \theta u)^\top$ a minimal representation of ${}^{c}\mathbf{M}_o$, the problem is reformulated:

$$ \hat{\mathbf{q}} = \arg\min_{\mathbf{q}} \sum_{i=1}^{N} ({}^{c}\mathbf{X}_i - {}^{c}\mathbf{M}_o {}^{o}\mathbf{X}_i)^2. \tag{2} $$

The problem is solved by initializing the pose, ${}^{c}\mathbf{M}_o$, with a linear solution, based on [16], and refining it with a non-linear Gauss-Newton estimation. The algorithm presented above assumes that the matching between the ${}^{c}\mathbf{X}_i$ and the ${}^{o}\mathbf{X}_i$ is known as in [9].

5) Transformation to reference frame: Once ${}^{c}\mathbf{M}_o$ is estimated, the pose ${}^{w}\mathbf{M}_o$ of the target in the reference frame can be recovered with ${}^{w}\mathbf{M}_o = {}^{w}\mathbf{M}_c {}^{c}\mathbf{M}_o$. ${}^{w}\mathbf{M}_c$ defines the pose of the camera in the reference frame and will vary with the controlled cameras. An additional calibration process will then be required and will be explained in section III together with camera control algorithms.

6) Filtering: An optional low-pass filtering process can be performed at the end of the registration. Filtering will help reduce noise and prevent drop outs but may add a little latency. As an example we have implemented a predictive Kalman filter [17] with constant acceleration state and position measurements.



(a) State-of-the-art (b) Stereo/single-view method (c) Controlled camera method (d) Combination

Fig. 2: Comparison between (a) stereo optical tracking systems, (b) our single-view tracking method, (c) our controlled camera method (illustrated here with one mobile camera) and (d) both methods.

C. Single-view tracking mode

The single-view tracking mode is used to localize the target when it is visible by only one camera. Thus, the tracking is still carried out when the target is out of the field of view of almost all the cameras. Since the target is visible by only one camera, steps II-B.2 and II-B.3 can not be processed. Thus the localization is directly performed from the 2D extracted features (section II-B.1).

The transformation ${}^{c}\mathbf{M}_o$ between the 2D projected points $\mathbf{x}_i$ (in the image frame) and their corresponding 3D target points ${}^{o}\mathbf{X}_i$ (in the target frame) is estimated:

$$\hat{\mathbf{q}} = \arg\min_{\mathbf{q}} \sum_{i=1}^{N} (\mathbf{x}_i - \mathbf{\Pi}\,{}^{c}\mathbf{M}_o\,{}^{o}\mathbf{X}_i)^2 \tag{3}$$

where $\mathbf{\Pi}$ is the projection matrix and $\mathbf{q}$ is a minimal representation of ${}^{c}\mathbf{M}_o$. The problem is solved by initializing the pose, ${}^{c}\mathbf{M}_o$, with a linear solution and refining with a non-linear Gauss-Newton estimation [15].

Equation (3) provides several solutions when three 3D points are considered. Thus we have embedded at least four non-coplanar markers on the considered target (Figure 1) so that the 2D-3D registration gives one solution. The 2D-3D registration algorithm presented above assumes that the matching between the $\mathbf{x}_i$ and the ${}^{o}\mathbf{X}_i$ is known. In our implementation of the approach, the matching is carried out using brute force. Once ${}^{c}\mathbf{M}_o$ is estimated, steps II-B.5 and II-B.6 can be performed as for stereo tracking.

III. Increasing Optical Tracking Volume with Controlled Cameras

Our second method enables to control the cameras to keep the target (constellation) in the cameras field of view (Figure 2c). Even if only one camera can move, the stereo volume increases and the single-view tracking volume becomes even larger (Figure 2d).

A visual servoing process controls the camera so that the target projection is close to the image center. The motion of the cameras is automated by mounting them on robots. Using a camera mounted on a robot requires an off-line calibration process to determine the position of the camera frame, $\mathscr{F}_c$, in the robot's end-effector frame, $\mathscr{F}_e$, which is required to recover the position of the camera in the reference frame, $\mathscr{F}_w$, and perform pose estimation.

A. Off-line controlled camera calibration

The controlled camera calibration process recovers the pose ${}^{e}\mathbf{M}_c$ of the camera in the end-effector frame of the robot [18] which is constant. Indeed, in practice, when fixing a camera to a robot one has to know ${}^{e}\mathbf{M}_c$ which is needed to compute, ${}^{w}\mathbf{M}_{c(t)}$, the pose of the camera in the reference frame at each instant t. ${}^{w}\mathbf{M}_c$ is required to compute the pose of the constellation in the reference frame and to determine the essential matrix for stereo reconstruction. Indeed for a pair of cameras c and c', the essential matrix, ${}^{c'}\mathbf{E}_{c(t)}$, can be deduced from the transformation ${}^{c'}\mathbf{M}_{c(t)}$ (equation (1)) which is computed as:

$${}^{c'}\mathbf{M}_{c(t)} = {}^{c'}\mathbf{M}_w\,{}^{w}\mathbf{M}_{c(0)}\,{}^{c(0)}\mathbf{M}_{e(0)}\,{}^{e(0)}\mathbf{M}_{e(t)}\,{}^{e(t)}\mathbf{M}_{c(t)}. \tag{4}$$

Matrix ${}^{c'}\mathbf{M}_w$ is known by the previously made extrinsic calibration. Same goes for ${}^{w}\mathbf{M}_{c(0)}$ since the extrinsic calibration is made at $t = 0$. Matrix ${}^{e(0)}\mathbf{M}_{e(t)}$ which represents the transformation of the end-effector frame at instant t in the end-effector frame at instant 0 varies but is known by odometry measurements. Thus the only unknown in equation (4) is ${}^{e}\mathbf{M}_c = {}^{c(0)}\mathbf{M}_{e(0)} = {}^{e(t)}\mathbf{M}_{c(t)}$ (see Figure 3).

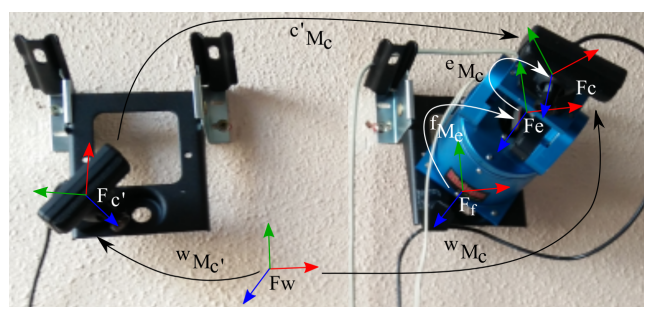


Fig. 3: Cameras configuration with two cameras and one pan-tilt head.

To obtain ${}^{e}\mathbf{M}_c$ we used a stationary 4 points target (or a calibration chessboard) and estimated its single-view pose for different positions of the robot's end-effector frame. Figure 4 illustrate the calibration setup for two positions of the the robot's end-effector frame e1 and e2 that lead to two positions of the camera c1 and c2. Since the target frame, $\mathscr{F}_o$, and the robot reference frame, $\mathscr{F}_f$, are fixed ${}^{f}\mathbf{M}_o$ is constant and given by:

$${}^{f}\mathbf{M}_o = {}^{f}\mathbf{M}_{e1}\,{}^{e}\mathbf{M}_c\,{}^{c1}\mathbf{M}_o = {}^{f}\mathbf{M}_{e2}\,{}^{e}\mathbf{M}_c\,{}^{c2}\mathbf{M}_o \tag{5}$$

where for each position i the transformation ${}^{f}\mathbf{M}_{ei}$ is given by the robot configuration and the transformation ${}^{ci}\mathbf{M}_o$ can be estimated through single-view registration (PnP algorithm).

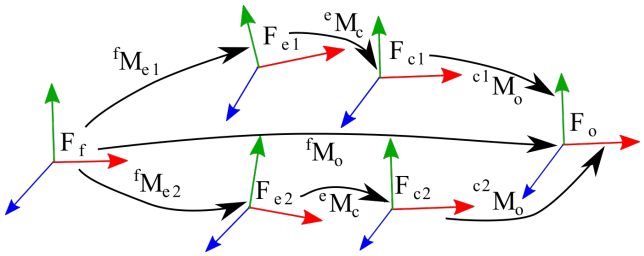


Fig. 4: Frame configuration for controlled camera calibration with 2 camera positions.

From equation (5) one can separate the rotation and translation parts of the transformations [18] to obtain two solvable equations:

$$1.\ \mathbf{A}_R {}^e\mathbf{R}_c = {}^e\mathbf{R}_c \mathbf{B}_R \quad \text{and} \quad 2.\ \mathbf{A}_t {}^e\mathbf{t}_c = {}^e\mathbf{R}_c \mathbf{b}_t \tag{6}$$

where $\mathbf{A}_R$, $\mathbf{B}_R$, $\mathbf{A}_t$ and $\mathbf{b}_t$ are computed from the measurements and are respectively two rotation matrices, a matrix and a column vector.

Equation (6.2) can be solved for ${}^e\mathbf{t}_c$ with a least square linear method once the solution ${}^e\mathbf{R}_c$ of equation (6.1) is found [19]. The solution involves converting the system to a linear least square system by using a different representation of the rotations. For a rotation $\mathbf{R}$ of angle θ and unit axis $\mathbf{u}$, the vector $\mathbf{p}_R = 2sin(\theta/2)\mathbf{u}$ is defined and equation (6.1) can be rewritten as:

$$Skew(\mathbf{p}_A + \mathbf{p}_B)\mathbf{x} = \mathbf{p}_B - \mathbf{p}_A. \tag{7}$$

However $Skew(\mathbf{p}_A + \mathbf{p}_B)$ has rank 2 so at least 3 positions of the camera are required to solve the system. Finally the angle θ and the unit axis $\mathbf{u}$ can be extracted from $\mathbf{x}$ to recover ${}^e\mathbf{R}_c$ [18].

B. Controlling camera displacements: visual servoing

To achieve the control of the camera, we consider a visual servoing scheme [20]. The goal of visual servoing is to control the dynamic of a system by using visual information provided by one camera. The goal is to regulate an error defined in the image space to zero. This error, to be minimized, is based on visual features that correspond to geometric features. Here we consider the projection of the center of gravity of the constellation $\mathbf{x} = (x, y)^\top$ that we want to see in the center of the image $\mathbf{x}^* = (0, 0)^\top$ (coordinates are expressed in normalized coordinates taking account of the camera calibration parameters).

Considering the actual pose of the camera $\mathbf{r}$ the problem can therefore be written as an optimization process:

$$\widehat{\mathbf{r}} = \arg\min_{\mathbf{r}}((\mathbf{x}(\mathbf{r}) - \mathbf{x}^*)^\top(\mathbf{x}(\mathbf{r}) - \mathbf{x}^*)) \tag{8}$$

where $\widehat{\mathbf{r}}$ is the pose reached after the optimization process (servoing process). This visual servoing task is achieved by iteratively applying a velocity to the camera. This requires the knowledge of the interaction matrix $\mathbf{L_x}$ of $\mathbf{x(r)}$ that links the variation of $\dot{\mathbf{x}}$ to the camera velocity and which is defined as:

$$\dot{\mathbf{x}}(\mathbf{r}) = \mathbf{L_x}\mathbf{v} \qquad \text{with} \qquad \mathbf{L_x} = \begin{pmatrix} xy & -(1+x^2) \\ 1+y^2 & -xy \end{pmatrix} \tag{9}$$

where $\mathbf{v}$ is the camera velocity (expressed in the camera frame). $\mathbf{L_x}$ is given is given for the specific case of a pan-tilt camera that will be considered in the paper.

This equation leads to the expression of the velocity that needs to be applied to the robot. The control law is classically given by:

$$\mathbf{v} = -\lambda \mathbf{L_s}^+(\mathbf{x}(\mathbf{r}) - \mathbf{x}^*) \tag{10}$$

where λ is a positive scalar and $\mathbf{L_x}^+$ is the pseudo-inverse of the interaction matrix. To compute, as usual, the velocity in the joint space of the robot, the control law is given by [20], [21]:

$$\dot{\mathbf{q}} = -\lambda \mathbf{J_x}^+(\mathbf{x}(\mathbf{r}) - \mathbf{x}^*) \quad \text{with} \quad \mathbf{J_x} = \mathbf{L_x}{}^c\mathbf{V}_e{}^e\mathbf{J}(\mathbf{q}) \tag{11}$$

where $\dot{\mathbf{q}}$ is the robot joint velocity and ${}^e\mathbf{J}(\mathbf{q})$ is the classical robot Jacobian expressed in the end effector frame (this Jacobian depends of the considered system). ${}^c\mathbf{V}_e$ is the spatial motion transform matrix [20] from the camera frame to the end-effector frame.

C. Registration

When using controlled cameras, the registration is carried out either with stereo or single-view registration algorithms after updating the pose ${}^w\mathbf{M}_{c(t)}$ of the camera in the reference frame at instant t through equation (4). At instant $t = 0$ the system was calibrated so every parameter of the system is known at position $c(0)$ of the camera. If stereo tracking is available, once ${}^w\mathbf{M}_{c(t)}$ is obtained the essential matrix is computed with equations (4) and (1) and the pose registration is performed as in Section II-B. Otherwise the registration is performed as in Section II-C.

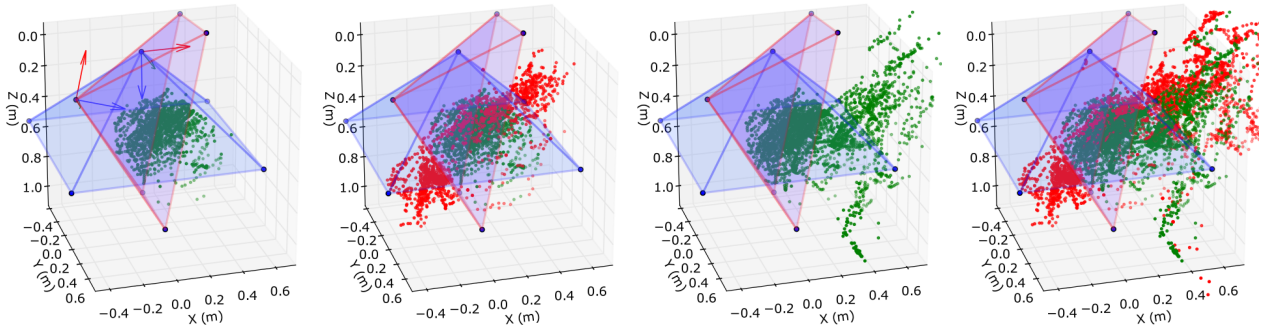
IV. Results and Performance

We designed a proof-of-concept of our method. Some tests and comparisons were performed on the designed system. In the following we present our prototype and the different tests that were carried out.

A. Proof-of-concept

As presented in our previous work [10], we have tested our approach on a wall-sized virtual reality display. In this paper we propose a proof-of-concept of our approach for UAV localization in indoor environments (*see accompanying video*).

The tracking system (Figure 1) is composed of two Sony PSEye cameras providing 320x240 images at a 150Hz refresh rate. The cameras were modified with short focal length lenses (2.1mm) providing a final field-of-view of 87° by 70°. An infrared band-pass filter was added to each lens. One camera is mounted on a TracLabs Biclops pan-tilt robot. A Parrot AR.Drone 2.0 was tracked. Four non-coplanar active infrared LEDs were rigidly attached to the UAV (Figure 1). As active markers were used, the cameras did not have to be equipped with infrared LED rings. A diffuser was added to each LED to provide isotropic light diffusion.



(a) State-of-the-art (b) Single-view tracking (c) Controlled camera (d) Combination

Fig. 5: Tracking volume of our methods compared to state of the art stereo optical tracking (our system with two stationary cameras). Green points were computed with stereo tracking and red one with single-view tracking. The pyramids illustrate the fields of view of the two cameras used by the system (red: a stationary camera, blue: a controlled camera).

B. Volume gain

The optical tracking volume of our approach was compared with a state-of-the-art stereo tracking (our system with two stationary cameras and without single-view tracking). To visualize the tracking volume of the different solutions tracking data was computed through the entire volume. First it was computed for state-of-the-art stereo tracking with two cameras (Figure 5a) then using single-view tracking. Thus the single-view tracking was active when the stereo was not available. Several poses were computed with single-view tracking at both sides of the stereo space. In our case (Figure 5b) there were almost as many stereo registrations as single-view registrations so we estimated a volume gain of around 100%. A third test was to activate the controlled camera mounted on the pan-tilt head (blue cone in Figure 5) and compute stereo tracking data as depicted in Figure 5c. Finally the two methods were merged and, by using the controlled camera and the single-view tracking, a far larger tracking volume was obtained (Figure 5d).

C. Comparison with Vicon's optical tracking

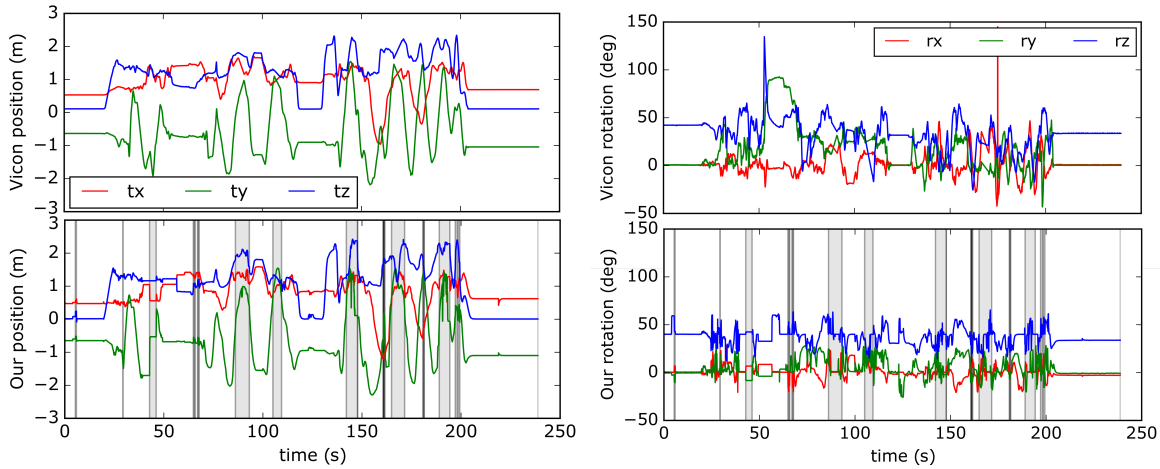
Our tracking system was installed in a room also equipped with a Vicon optical tracking system composed of 8 Bonita 10 (1024×1024 at 250Hz) and 4 Vero v1.3 (1280×1024 at 250Hz) cameras. With such installation we were able to compare the performances of our approach with the Vicon's performances. Nevertheless we did not try to overtake Vicon. Our goal was to provide a larger tracking volume at low cost. In the following a qualitative comparison between the two systems is introduced.

1) Pose estimation: The pose of a flying UAV was estimated, at each instant, with our system and with the Vicon tracker. Figure 6a illustrates the variations for the three components of vector ${}^{w}\mathbf{t}_o$ while Figure 6b illustrates the three components of the Rodrigues representation of ${}^{w}\mathbf{R}_o$. The grey zones define the moments when our tracking was performed using single-view localization. At these instants the UAV was visible by only one of the cameras in our system. Since the calibration was made separately and the systems were not synchronized the error between both measurements is not of interest and a qualitative comparison of the pose is proposed.

2) Jitter: Jitter was measured by leaving the UAV at a stationary position and recording its pose during 7000 measurements without filtering process. The UAV was placed at around 2.5m of the cameras. Figure 7 illustrates the spatial distribution of the reconstructed positions. With Vicon measurements the 95% confidence radius of the distribution lies at 0.18mm. For our stereo tracking it lies at 0.86mm. The measurements with the single-view tracking are more noisy since the 95% confidence radius lies at 10.2mm. Nevertheless this noise is reduced when getting closer to the image frame of the camera. Some tests were carried out at 60cm of the camera and the 95% confidence radius lied at 1.4mm. These uncertainties are mainly oriented along the depth axis of the camera frame and are affected by the spatial resolution in the image.

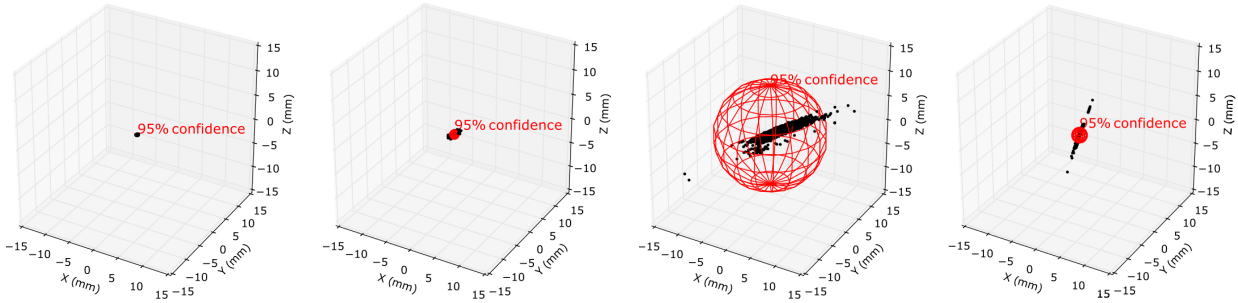
D. Discussion

Our approach is based on two complementary methods which were implemented, tested and compared. Using controlled cameras together with single-view registration enabled to considerably increase the tracking volume up to 100% and more. Considering the Vicon as ground truth, our system shows good performances for such a low-cost device. Nevertheless the calibration of both systems was made independently and they were not synchronized. Thus a quantitative comparison was unfortunately not possible. Since the Vicon tracker was composed of 12 high-resolution cameras (1MP) to cover all of the required volume it could be interesting to compare our performances with the ones of a Vicon tracker composed of at most 3 cameras. Regarding our proof-of-concept, several improvements could be obtained on the hardware components. We used wide-angle lenses to maximize the visibility but such lenses induce a loss in resolution that can degrade the feature extraction and increase jitter. It could be interesting to test our approach with standard lenses. Higher quality sensors (e.g. high-resolution cameras) and/or hardware synchronization could also be used to reduce jitter and increase tracking stability and accuracy, but at a higher cost.



(a) Position (b) Orientation

Fig. 6: Components of the estimated pose of the tracked UAV at each instant with Vicon and with our system.



(a) Vicon (b) Our stereo tracking (c) Our single-view tracking (d) Close-range single-view

Fig. 7: Positional jitter comparison between our system and the Vicon's.

V. Conclusion

We proposed two complementary methods to maximize the optical tracking volume with a limited number of cameras and applied it to indoor UAV localization. First we proposed to rely on a single-view registration when the multi-view registration is not available. The second method allows to control the cameras to track the target through the entire volume. Controlling the cameras brings more liberty to position them and both methods help providing a larger tracking volume (up to 100%) without adding expensive sensors.

References

[1] R. H. Taylor et al. "An image-directed robotic system for precise orthopaedic surgery," *IEEE TRA*, 10(3):261–275, 1994.

[2] T. Pintaric and H. Kaufmann, "Affordable infrared-optical pose-tracking for virtual and augmented reality," *IEEE VR*, pp. 44–51, 2007.

[3] Y. M. Mustafah, A. W. Azman, and F. Akbar., "Indoor uav positioning using stereo vision sensor," 2012, pp. 575–579.

[4] J. Engel, J. Sturm, and D. Cremers, "Camera-based navigation of a low-cost quadrocopter," *IEEE IROS*, pp. 2815–2821, 2012.

[5] L. V. Santana, et al, "A trajectory tracking and 3d positioning controller for the AR drone quadrotor," *IEEE ICUAS*, pp. 756–767, 2014.

[6] K. E. Wenzel, A. Masselli, and A. Zell, "Automatic take off, tracking and landing of a miniature uav on a moving carrier vehicle," *JIRS*, 61(1):221–238, 2011.

[7] J. F. Roberts, et al. "Quadrotor using minimal sensing for autonomous indoor flight," in *EMAV*, 2007.

[8] S. Grzonka, G. Grisetti, and W. Burgard, "Towards a navigation system for autonomous indoor flying," in *IEEE ICRA*, 2009, pp. 2878–2883.

[9] K. Kurihara, S. Hoshino, K. Yamane, and Y. Nakamura, "Optical motion capture system with pan-tilt camera tracking and realtime data processing," *IEEE ICRA*, pp. 1241–1248, 2002.

[10] G. Cortes, E. Marchand, J. Ardouin, and A. Lécuyer, "Increasing optical tracking workspace of vr applications using controlled cameras," *IEEE 3DUI*, 2017.

[11] R. Hartley and A. Zisserman, *Multiple view geometry in computer vision*. Cambridge university press, 2003.

[12] S. Vogt, A. Khamene, F. Sauer, and H. Niemann, "Single camera tracking of marker clusters: Multiparameter cluster optimization and experimental verification," *IEEE ISMAR*, p. 127, 2002.

[13] Z. Zhang, "A flexible new technique for camera calibration," *IEEE PAMI*, 22(11):1330–1334, 2000.

[14] R. I. Hartley and P. Sturm, "Triangulation," *Computer vision and image understanding*, 68(2):146–157, 1997.

[15] E. Marchand, H. Uchiyama, and F.Spindler, "Pose estimation for augmented reality: a hands-on survey," *IEEE TVCG*, 2016.

[16] K. S. Arun, T. S. Huang, and S. D. Blostein, "Least-squares fitting of two 3-d point sets," *IEEE PAMI*, no. 5, pp. 698–700, 1987.

[17] Y. Bar-Shalom and X.-R. Li, *Estimation and Tracking, Principles, Techniques, and Software.* Boston: Artech House, 1993.

[18] R. Y. Tsai et al, "A new technique for fully autonomous and efficient 3d robotics hand/eye calibration," *IEEE TRA*, 5(3):345–358, 1989.

[19] Y. C. Shiu and S. Ahmad, "Calibration of wrist-mounted robotic sensors by solving homogeneous transform equations of the form AX= XB," *IEEE TRA*, 5(1):16–29, 1989.

[20] F. Chaumette and S. Hutchinson, "Visual servo control, part i: Basic approaches," *IEEE Robot. Autom. Mag.*, 13(4):82–90, December 2006.

[21] E. Marchand, F. Spindler, and F. Chaumette, "Visp for visual servoing: a generic software platform with a wide class of robot control skills," *IEEE Robot. Autom. Mag.*, 12(4):40–52, December 2005.