Gimbal Control for Vision-based Target Tracking

Rita Cunha[‡], Miguel Malaca[‡], Vasco Sampaio[‡], Bruno Guerreiro[‡],
Paraskevi Nousi[†], Ioannis Mademlis[†], Anastasios Tefas[†], Ioannis Pitas[†]
[‡]Institute for Systems and Robotics (ISR/LARSyS), Instituto Superior Técnico, Lisbon, Portugal
[†]Department of Informatics, Aristotle University of Thessaloniki, Thessaloniki, Greece

***Abstract*—This paper addresses the problem of controlling the orientation of a 3-axis gimbal that is carrying a cinematography camera, using image measurements for feedback. The control objective is to keep a moving target of interest at the center of the image plane. A Region-of-Interest (ROI) that encloses the target's image is generated through the combination of a visual object detector and a visual object tracker based on Convolutional Neural Networks. These are specially tailored to allow for high frame rate performance with restricted computational power. Given the target's ROI, an attitude error in the form of a rotation matrix is computed and a attitude controller is designed, which guarantees convergence of the target's image to the center of the image plane. Experimental results with a human face as the target of interest are presented to illustrate the performance of the proposed scheme.**

***Index Terms*—vision-based control, visual object detection, visual object tracking, attitude control**

I. Introduction

Gimbals are typically used in aerial vehicles to attenuate vibrations and stabilize the camera in the presence of angular motion of the vehicle and other disturbances, through inertial stabilization [1]–[3]. In addition, gimbals effectively augment the Field of View (FOV) of the camera through angular motion, which can either compensate for or be combined with translational motion to achieve the desired shooting objective. This ability to mitigate the FOV constraint becomes even more critical when tracking of moving targets is involved [4]. In fact, physical, autonomous target tracking using a camera/gimbal combination is an extremely important functionality for vision-enabled robotic systems, e.g., in autonomous cinematography/intelligent shooting applications [5]–[8].

Achieving the goal of pointing the camera towards a target can be divided into two tasks. The first concerns visual object detection and tracking. This module provides the image measurements that are used as input to the second module, which is responsible for controlling the gimbal so that the camera optical axis points in the desired direction.

In this paper, we propose to address the gimbal control problem as a problem of attitude tracking on the Special Orthogonal Group $\mathbb{SO}(3)$, using rotation matrices to represent the gimbal attitude [9], [10]. One of the contributions of the paper is the definition of a reference rotation matrix to be tracked, and ensuing gimbal controller, that can be directly expressed as functions of the image measurements, with no need depth estimation. The reference rotation matrix is defined as a function of the relative position between the target and the camera and keeps the camera horizontally aligned. The standard structure for an attitude controller on $\mathbb{SO}(3)$ is then adopted, based on the resulting error rotation matrix. We then show that this error matrix can be expressed directly as a function of the image measurements. To achieve horizontal alignment, body-fixed measurements of the gravitational acceleration, typically provided by accelerometers, are also required.

The paper is structured as follows. Section II summarizes the related work in terms of visual object detection and tracking, as well as gimbal control strategies. Sections III and IV formulate the problem and present a nonlinear control law for the gimbal control. Section V summarizes the strategy used for obtaining the visual estimates of the object location in the image frame, while Section VI present experimental results for the overall algorithm. Finally, Section VII presents some concluding remarks.

The research leading to these results has received funding from the European Union's European Union Horizon 2020 research and innovation programme under grant agreement No 731667 (MULTIDRONE).

II. Related work

A. Gimbal control

Gimbal control has been extensively studied in the past, as an integral part of the so-called Inertial Stabilized Platforms (ISPs) used to stabilize the line of sight of a sensor mounted on a platform, which is possibly moving and rotating, relative to a target or an inertial reference frame. The range of sensors and applications is wide and includes cameras, telescopes, antennas, and weapon systems, to name a few [11]. A excellent introduction to the topic can be found in the special issue from the IEEE Control Systems Magazine dedicated to ISP technology [3], [11]. The typical gimbal control system, which is also the one adopted in this paper, has a inner-outer loop structure, with a high bandwidth inner loop system that tracks angular rate commands and rejects disturbances based on rate gyros measurements and a lower-bandwidth outer loop system that is responsible for the actual pointing and tracking [1], [3], [11]. Thus, the pointing and tracking performance is limited by the dynamics the inner-loop system, which is primarily determined by the bandwidth of the gyroscopes and actuation system and disturbance effects, i.e. the motion of the platform itself.

This paper is primarily concerned with the outer-loop control system, which receives image measurements and computes angular rate commands to correct the image displacements. Unlike the work [1], which defines an error directly in the image plane, we construct an error rotation matrix and define an attitude tracking controller directly on $\mathbb{SO}(3)$. Recent work has been dedicated to this problem to provide almost global asymptotic tracking solutions [9], [10]. More recently global solutions based on hybrid control techniques [12] or with

guarantees of finite-time convergence [13] have also been proposed, all of which can be applied to the problem gimbal control under the form proposed in this paper.

B. Visual object detection and tracking

In general, deep Convolutional Neural Networks (CNNs) have been excelling continuously on various challenging visual analysis tasks and competitions. Deep, feed-forward neural models have been successfully trained and deployed on such tasks, partly due to the public availability of large annotated datasets and partly due to the continuous development of increasingly more powerful GP-GPUs. One approach to achieving real-time performance with restricted computational hardware is to use one-stage deep neural detectors, structured around the concept of "anchors". These detectors, such as Single-Shot Detector (SSD) [14] and You Only Look Once (YOLO) [15] simultaneously regress the pixel coordinates of visible object ROIs (in the form of spatial offsets from the predefined anchors) and assign them class labels. The first part of these models is typically a base feature extracting network, such as AlexNet, VGG-16, MobileNet v1 or Inception v2, that has been pre-trained on a ImageNet or COCO dataset for whole-image classification tasks.

SSD [14] is a single stage multi-object detector, meaning that a single feed forward pass of an image suffices for the extraction of multiple bounding boxes with coordinate and class information and no region proposal occurs internally. In [16], SSD was used as a meta-architecture for single stage object detection and compared against region-based detectors. Among the findings of that work, was that SSD with MobileNets and Inception V2 for the feature extraction step provided the best time performance at the cost of lower detection precision, as evaluated on the challenging COCO dataset.

Correlation filter-based 2D visual target tracking algorithms are suitable for real-time applications [17], especially on embedded systems that tend to have limited computational resources. A correlation filter tracker regresses the representations of all possible object template translations to a Gaussian distribution. The original ROI object template is regressed to its peak. Due to the circulant structure of the template representations, the regression problem can be solved in the Fourier domain, thus accelerating the learning and testing processes of the tracker.

The success of CNNs in various visual analysis tasks, has led to their adoption for visual tracking. SiamFC [18] is one such CNN-based tracker, trained as a fully convolutional siamese network, which performs cross correlation between the features extracted from the target and a candidate region to find the new position of the target.

III. Problem Formulation

Let $\{W\}$ denote the world reference frame with origin fixed in the environment and East-North-Up (ENU) orientation. Consider also two additional reference frames, the camera reference frame $\{C\}$ with z-axis aligned with the optical axis but with opposite sign and the target reference frame $\{T\}$ attached to the moving target of interest (see Fig. 1).

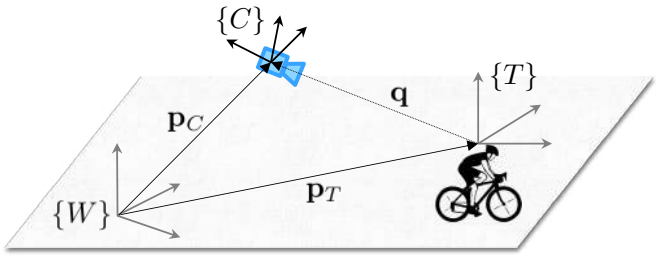

Fig. 1. Problem setup and notation.

The configuration of $\{C\}$ with respect to $\{W\}$ is denoted by $({}^W\mathbf{p}_C, R_C) \in \mathbb{SE}(3)$, where ${}^W\mathbf{p}_C \in \mathbb{R}^3$ is the position of the origin of $\{C\}$ expressed in $\{W\}$ and $R_C \in \mathbb{SO}(3)$ is the rotation matrix from $\{C\}$ to $\{W\}$, $\mathbb{SO}(3)$ denotes the Special Orthogonal Group of order (3) and $\mathbb{SE}(3)$ the Special Euclidean Group of order three. Similarly, the configuration of $\{T\}$ with respect to $\{W\}$ is denoted by $({}^W\mathbf{p}_T, R_T) \in \mathbb{SE}(3)$.

A simplified kinematic model for the gimbal angular motion is adopted, which can be described by

$$\dot{R}_C = R_C S(\boldsymbol{\omega}) \tag{1}$$

where $\boldsymbol{\omega}$ denotes the angular velocity and the operator $S : \mathbb{R}^3 \mapsto so(3)$ maps vector in $\mathbb{R}^3$ to skew-symmetric matrices, such that for $\mathbf{a} = \begin{bmatrix} a_1 & a_2 & a_3 \end{bmatrix}'$

$$S(\mathbf{a}) = \begin{bmatrix} 0 & a_3 & -a_2 \\ -a_3 & 0 & a_1 \\ a_2 & -a_1 & 0 \end{bmatrix} \tag{2}$$

and thus $S(\mathbf{a})\mathbf{b} = \mathbf{a} \times \mathbf{b}$, where $\times$ denotes the cross-product.

Remark. *Under the assumption that the gimbal is equipped with an Inertial Measurement Unit (IMU) and a low-level controller, it is reasonable to adopt the simplified kinematic model described in* (1)*. In this case, the low-level controller receives the measurements from the IMU together with angular velocity references to be tracked and computes the actual inputs for the gimbal joint motors. Fig. 2 shows an example of such a gimbal.*

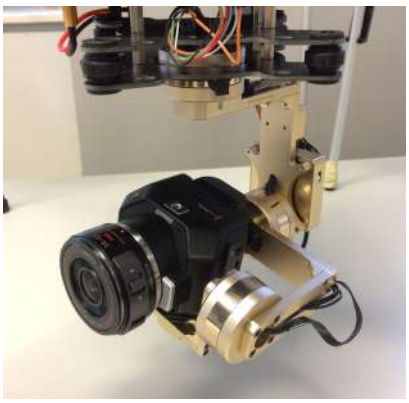

Fig. 2. 3-D Gimbal and Camera equipped with IMU and low-level controller.

It is also assumed that a gimbal and target can move freely, meaning that ${}^W\mathbf{p}_C(t)$ and ${}^W\mathbf{p}_T(t)$ are time-varying, and that the gimbal orientation can be controlled independently from this translational motion.

To complete the problem formulation, it is convenient to introduce the relative position between the target and the camera, with coordinates in $\{W\}$ given by

$$ {}^W\mathbf{q} = \begin{bmatrix} q_x & q_y & q_z \end{bmatrix}' = {}^W\mathbf{p}_C - {}^W\mathbf{p}_T. \tag{3}$$

and coordinates in $\{C\}$ given by

$$ \mathbf{q} = {}^C\mathbf{q} = R_C'\,{}^W\mathbf{q}, \tag{4}$$

where R_C' denotes the transpose of R_C, which is also its inverse, $R_C'R_C = R_CR_C' = I_3$. For convenience, we drop the superscript ^{C} for vectors expressed in $\{C\}$ and keep the superscript ^{W} for those expressed in $\{W\}$.

Adopting a pin-hole camera model, the 2-D image pixel coordinates $\mathbf{y} \in \mathbb{R}^2$ of the point with 3-D coordinates $\mathbf{q} = \begin{bmatrix} q_x & q_y & q_z \end{bmatrix}^T \in \mathbb{R}^3$ expressed in $\{C\}$ are given by

$$ \begin{bmatrix} \mathbf{y} \\ 1 \end{bmatrix} = A\frac{1}{q_z}\mathbf{q}, \tag{5}$$

where $A \in \mathbb{R}^{3\times 3}$ is the matrix of camera intrinsic parameters.

In this framework, the control objective can be defined as follows.

Problem 1. *Consider the gimbal system described by* (1) *and camera model described in* (5)*. Using the image measurements* $\mathbf{y}_i$*, define a control law for the angular velocity input* $\boldsymbol{\omega}$ *such that* $R_C(t)$ *asymptotically converges to a desired orientation* $R_C^*(t)$*, which guarantees that the target of interest is centered in the image plane, i.e.*

$$ \begin{bmatrix} \mathbf{y}^* \\ 1 \end{bmatrix} = A\begin{bmatrix} \mathbf{0} \\ 1 \end{bmatrix} \tag{6}$$

IV. Gimbal Control

As defined in Problem 1, the control objective amounts to providing a control law for the angular velocity input $\boldsymbol{\omega}$ such that the rotation matrix $R_C(t)$ asymptotically converges to $R_C^*(t)$. Given the model defined in (1), this matches exactly the problem of attitude tracking on $\mathbb{SO}(3)$, which has been treated extensively in the literature (see, for example, [9], [10] and references therein).

For completeness and assuming that R_C^* is given, we present a tracking controller law for a kinematic model evolving on $\mathbb{SO}(3)$ and show that it is asymptotically stabilizing. Then, we show how to construct R_C^* and the corresponding error matrix, which guarantee that the image of the target is centered in the image plane.

A. Attitude Tracking on $\mathbb{SO}(3)$

Working directly on $\mathbb{SO}(3)$, the orientation error can be defined as

$$ R_e = R_C'R_C^* \in \mathbb{SO}(3) \tag{7}$$

Taking the time derivative of R_e, we obtain the error system

$$ \dot{R}_e = -S(\boldsymbol{\omega})R_e + R_eS(\boldsymbol{\omega}^*) \tag{8}$$

To define a control law for $\boldsymbol{\omega}$, consider the Lyapunov function given by

$$ V(R_e) = \mathrm{tr}(I_3 - R_e) \tag{9}$$

where tr denotes the trace and I_3 the three by three identity matrix. Straightforward computations show that V is a positive definite function of R_e, meaning that $V(R_e) \geq 0$ and $V(R_e) = 0$ if and only if $R_e = I_3$. After some algebraic manipulations, the time-derivative can be written as

$$ \dot{V} = (\boldsymbol{\omega} - \boldsymbol{\omega}^*)'S^{-1}(R_e - R_e') \tag{10}$$

Thus, the control law

$$ \boldsymbol{\omega} = -kS^{-1}(R_e - R_e') + \boldsymbol{\omega}^* \tag{11}$$

guarantees that $\dot{V}$ is a negative semi-definite function of R_e and $R_e = I_3$ is an asymptotically stable equilibrium point of (8). Further analyses shows that in fact $R_e = I_3$ is almost globally asymptotically stable, meaning that the system converges to the desired equilibrium except for a zero measure set of initial conditions. For details, the reader is referred to [9], [10]. In what follows, it is assumed that the desired rotation matrix R_C^* is static or slowly time-varying, such that $\boldsymbol{\omega}^* \approx 0$ and $\boldsymbol{\omega}$ becomes

$$ \boldsymbol{\omega} = -kS^{-1}(R_e - R_e'). \tag{12}$$

B. Application to Vision-based Gimbal Control

To define the desired camera orientation R_C^*, we start by noting that when the camera is aligned with the target $\mathbf{q}^* = {R_C^*}'{}^W\mathbf{q} = \|{}^W\mathbf{q}\|\mathbf{e}_3$, where $\mathbf{e}_3 = [0\ 0\ 1]'$, or equivalently, the z-axis of the camera and ${}^W\mathbf{q}$ have the same direction (see Fig. 1). This observation together with the fact that the camera should be horizontally aligned suggest the follow expression for R_C^*

$$ \begin{aligned} R_C^* &= \begin{bmatrix} -\dfrac{S({}^W\mathbf{q})^2\mathbf{e}_3}{\|S({}^W\mathbf{q})\mathbf{e}_3\|} & \dfrac{S({}^W\mathbf{q})\mathbf{e}_3}{\|S({}^W\mathbf{q})\mathbf{e}_3\|} & \dfrac{{}^W\mathbf{q}}{\|{}^W\mathbf{q}\|} \end{bmatrix} \\ &= \begin{bmatrix} * & \frac{q_y}{\sqrt{q_x^2+q_y^2}} & * \\ * & \frac{-q_x}{\sqrt{q_x^2+q_y^2}} & * \\ \frac{\sqrt{q_x^2+q_y^2}}{\sqrt{q_x^2+q_y^2+q_z^2}} & 0 & \frac{q_z}{\sqrt{q_x^2+q_y^2+q_z^2}} \end{bmatrix} \end{aligned} \tag{13}$$

where it is assumed that the camera is always above the target, i.e. $q_z > 0$, but not directly above the target, i.e. $[q_x\ q_y] \neq 0$.

Proposition 1. *Let* $\mathbf{y}^* \in \mathbb{R}^2$ *denote the image coordinates of the origin of* $\{T\}$ *when* $R_C = R_C^*$*. If* R_C^* *is given by* (13) *then* $[\mathbf{y}^{*\prime}\ 1]' = A[\mathbf{0}'\ 1]'$ *and the camera is horizontally aligned.*

Proof. According to (5) and (4), we can write

$$ \begin{bmatrix} \mathbf{y} \\ 1 \end{bmatrix} = A\frac{1}{\mathbf{e}_3'\mathbf{q}}\mathbf{q}. \tag{14}$$

If $R_C = R_C^*$, it immediately follows from (13) that $\mathbf{q}^* = {R_C^*}'{}^W\mathbf{q} = \|\mathbf{q}\|\mathbf{e}_3$ and thus

$$ \begin{bmatrix} \mathbf{y}^* \\ 1 \end{bmatrix} = A\frac{\mathbf{q}^*}{\mathbf{e}_3'\mathbf{q}^*} = A\begin{bmatrix} \mathbf{0} \\ 1 \end{bmatrix} \tag{15}$$

To show that the camera is horizontally aligned, it suffices to note that the y-axis of the desired camera frame is orthogonal to the z-axis of the world reference frame $\{W\}$, i.e., $\mathbf{e}_3' R_C^* \mathbf{e}_2 = 0$, where $\mathbf{e}_2 = \begin{bmatrix} 0 & 1 & 0 \end{bmatrix}'$. □

Next, we show that the orientation error in the form of R_e can be reconstructed directly from the image measurements $\mathbf{y}$ and the gravitational vector expressed in $\{C\}$. Consequently, for static or slowly time-varying relative positions between the camera and gimbal, the control law $\boldsymbol{\omega}$ can be expressed as a function of the image measurements and accelerometer measurements denoted by $\mathbf{a}$.

Lemma 1. *Assume that the matrix of intrinsic parameters A is known and let $\mathbf{y} \in \mathbb{R}^2$ denote the image coordinates of the origin of $\{T\}$. Then, the error rotation matrix R_e can be written as*

$$R_e = \begin{bmatrix} -\dfrac{S(\mathbf{q})^2 \mathbf{a}}{\|S(\mathbf{q})\mathbf{a}\|} & \dfrac{S(\mathbf{q})\mathbf{a}}{\|S(\mathbf{q})\mathbf{a}\|} & \dfrac{\mathbf{q}}{\|\mathbf{q}\|} \end{bmatrix} \tag{16}$$

where $\mathbf{q}/\|\mathbf{q}\|$ is given by

$$\mathbf{q}/\|\mathbf{q}\| = A^{-1} \begin{bmatrix} \mathbf{y} \\ 1 \end{bmatrix} / \| A^{-1} \begin{bmatrix} \mathbf{y} \\ 1 \end{bmatrix} \| \tag{17}$$

Proof. Assume that the accelerometers from the gimbal's IMU approximately measure the gravity vector expressed in $\{C\}$, i.e. $\mathbf{a} = g R_C' \mathbf{e}_3$, where g is the gravitational acceleration. Then, using (7) and (13) we can write (16), noting that all columns of R_e must have norm 1 and thus knowing $\mathbf{q}$ and $\mathbf{a}$ up to scale factor is enough to compute R_e. Assuming that A is known, then

$$\frac{\mathbf{q}}{q_z} = A^{-1} \begin{bmatrix} \mathbf{y} \\ 1 \end{bmatrix} \tag{18}$$

and (17) can be readily obtained from (18) by dividing by the norm. □

V. Visual Analysis

In this section, we describe the method adopted to detect the target on the image plane and obtain estimates of $\mathbf{y}$. The presented system combines a visual object detector and a visual object tracker, while it demands cinematographic shot requirements to have been pre-specified (e.g., desired target position on video frame). It receives an uncompressed video frame from the camera in real-time and generates 2D positions of the tracked targets as ROIs (in pixel coordinates). The tracker is initialized and periodically validated by the detector. The produced target 2D position is then employed to calculate the current *visual control error*, according to the desired shot specifications. This error, which constitutes the final visual analysis system output, is simply the deviation of the current ROI on-frame position from the desired one (in pixel coordinates).

SSD, with MobileNet v1 as a base feature extractor and an input image resolution of 192×192 pixels, was selected as a detector. It was found in [19] to offer a great trade-off between speed and accuracy, since it achieves a processing rate of 22 frames per second (FPS) on embedded AI hardware (nVIDIA Jetson Tegra X2).

Additionally, a more lightweight version of SiamFC was developed, by introducing a depth factor α, similar to the one used by MobileNets [16] to improve their speed. More specifically, the number of filters in each layer of the siamese architecture is multiplied by $\alpha \in (0, 1]$, thus producing a lighter network which requires fewer operations per layer. Let n_l denote the number of filters for layer $l = 1, \ldots, 5$ for the five layers of AlexNet, the base feature extractor of SiamFC. The developed SiamFClite tracker is parameterized by $\alpha \sum_{l=1}^{5} n_l$ filters, as opposed to the $\sum_{l=1}^{5} n_l$ of the original SiamFC.

VI. Simulation and Experimental Results

In order to test both the visual detection and the control algorithm, experiments were conducted with a camera mounted on a 3-axis gimbal, considering a human face as the target of interest. The camera can stream video on a 620x480 scale at a frame rate of 30 fps and its calibration matrix is given by

$$A = \begin{bmatrix} 609.219929 & 0 & 320 \\ 0 & 821.601329 & 240 \\ 0 & 0 & 1 \end{bmatrix}$$

The gimbal, shown in Fig. 2, is equipped with an IMU and a low-level controller from BaseCam Electronics, which is ready to receive commands in the form of Euler angle rates $\dot{\boldsymbol{\lambda}}$. These can be readily computed from the angular rate commands $\boldsymbol{\omega}$ defined in (12) by applying the standard transformation between the two.

In the experiments the face detector and tracker produced a visual control error in form of the image coordinates $\mathbf{y}$ also at 30 fps, which were used to feed the gimbal controller. An example experiment is illustrated in Figures 3-5 and the corresponding video is available at [1]. As shown in Fig. 3, whenever the subject moves, the gimbal rotates to compensate for that motion and guarantee convergence to the center of the image.

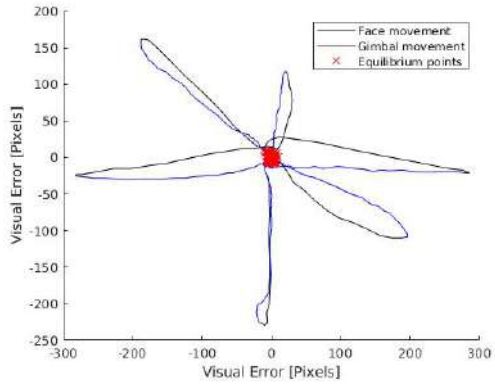

Fig. 3. Position of the face in the image

Figure 4 shows the time evolution of the X and Y pixel coordinates of the visual control error, together with the values of the Lyapunov function V. As expected, the image deviates from the center and V increases whenever there is motion of

[1] http://users.isr.ist.utl.pt/~rmac/videos/gimbal_test.mp4

the target, because the feedforward term $\boldsymbol{\omega}^*$ is not present in the control law. As soon as the target stops moving, the image and the value of V quickly converge (in approximately 1s) back to the origin.

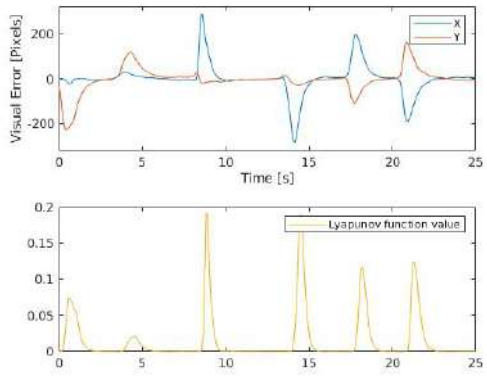

Fig. 4. X and Y visual error and respective Lyapunov function value

To further describe the experiments, Fig. 5 shows the time evolution of the gimbal orientation represented in the form of roll, pitch, and yaw Euler angles. As expected, when the target motion is either horizontal or vertical, the angular motions are approximately decoupled and the roll angle remains constant and close to zero. When the motion is in an oblique direction, which occurs at $t = 18s$, all angles change to guarantee convergence of the target image to the origin, although the change in the roll angle continues to be negligible.

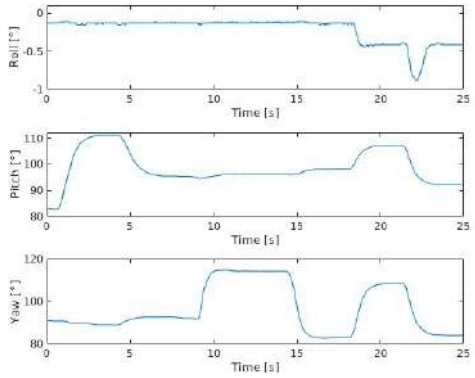

Fig. 5. Gimbal orientation in Roll, Pitch and Yaw

VII. Conclusions

In this paper, we addressed the problem of controlling the orientation a gimbal-mounted camera to point at a target of interest. The proposed solution combines a fast and reliable deep learning visual object detector and tracker, suited for low computational power implementation, with an attitude controller that is based on image accelerometer measurements and guarantees convergence of the target image to origin of the image plane. Experimental results have shown the effectiveness of the proposed solution, involving human face detection and tracking. Future work will focus on more dynamic scenarios, involving different targets and more aggressive camera and target motions.

References

[1] Z. Hurak and M. Rezac, "Image-based pointing and tracking for inertially stabilized airborne camera platform," *IEEE Transactions on Control Systems Technology*, vol. 20, no. 5, pp. 1146–1159, Sept. 2012.

[2] F. Königseder, W. Kemmetmüller, and A. Kugi, "Attitude estimation using redundant inertial measurement units for the control of a camera stabilization platform," *IEEE Transactions on Control Systems Technology*, vol. 24, no. 5, pp. 1837–1844, Sept. 2016.

[3] M. K. Masten, "Inertially stabilized platforms for optical imaging systems," *IEEE Control Systems Magazine*, vol. 28, no. 1, pp. 47–64, Feb. 2008.

[4] J. Thomas, J. Welde, G. Loianno, K. Daniilidis, and V. Kumar, "Autonomous flight for detection, localization, and tracking of moving targets with a small quadrotor," *IEEE Robotics and Automation Letters*, vol. 2, no. 3, pp. 1762–1769, July 2017.

[5] A. Torres-González, J. Capitán, R. Cunha, A. Ollero, and I. Mademlis, "A multidrone approach for autonomous cinematography planning," *Iberian Robotics Conference (ROBOT)*, 2017.

[6] I. Mademlis, V. Mygdalis, N. Nikolaidis, and I. Pitas, "Challenges in Autonomous UAV Cinematography: An Overview," in *Proceedings of the IEEE International Conference on Multimedia and Expo (ICME)*, 2018.

[7] I. Mademlis, N. Nikolaidis, A. Tefas, I. Pitas, T. Wagner, and A. Messina, "Autonomous unmanned aerial vehicles filming in dynamic unstructured outdoor environments," *IEEE Signal Processing Magazine*, vol. 36, no. 1, pp. 147–153, 2018.

[8] I. Mademlis, V. Mygdalis, N. Nikolaidis, M. Montagnuolo, F. Negro, A. Messina, and I. Pitas, "High-level multiple-UAV cinematography tools for covering outdoor events," *IEEE Transactions on Broadcasting*, 2019.

[9] N. A. Chaturvedi, A. K. Sanyal, and N. H. McClamroch, "Rigid-body attitude control," *IEEE Control Systems Magazine*, vol. 31, no. 3, pp. 30–51, June 2011.

[10] R. Cunha, C. Silvestre, and J. Hespanha, "Output-feedback control for stabilization on SE(3)," *Systems Control Letters*, vol. 57, no. 12, pp. 1013 – 1022, 2008.

[11] J. M. Hilker, "Inertially stabilized platform technology concepts and principles," *IEEE Control Systems Magazine*, vol. 28, no. 1, pp. 26–46, Feb. 2008.

[12] P. Casau, R. Cunha, R. G. Sanfelice, and C. Silvestre, "Hybrid feedback for global asymptotic stabilization on a compact manifold," in *2017 IEEE 56th Annual Conference on Decision and Control (CDC)*, Dec. 2017, pp. 2384–2389.

[13] J. Bohn and A. K. Sanyal, "Almost global finite-time stabilization of rigid body attitude dynamics using rotation matrices," *International Journal of Robust and Nonlinear Control*, vol. 26, no. 9, pp. 2008–2022, 2016.

[14] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, "SSD: Single-shot multibox detector," in *Proceedings of the European Conference on Computer Vision (ECCV)*. Springer, 2016, pp. 21–37.

[15] J. Redmon and A. Farhadi, "YOLO9000: better, faster, stronger," *arXiv preprint*, vol. 1612, 2016.

[16] J. Huang, V. Rathod, C. Sun, M. Zhu, A. Korattikara, A. Fathi, I. Fischer, Z. Wojna, Y. Song, and S. Guadarrama, "Speed/accuracy trade-offs for modern convolutional object detectors," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[17] O. Zachariadis, V. Mygdalis, I. Mademlis, N. Nikolaidis, and I. Pitas, "2D visual tracking for sports UAV cinematography applications," *Proceedings of the IEEE Global Conference on Signal and Information Processing (GlobalSIP)*, 2017.

[18] L. Bertinetto, J. Valmadre, J.F. Henriques, A. Vedaldi, and P.H.S. Torr, "Fully-convolutional siamese networks for object tracking," in *Proceedings of the European Conference on Computer Vision (ECCV)*. Springer, 2016, pp. 850–865.

[19] P. Nousi, E. Patsiouras, A. Tefas, and I. Pitas, "Convolutional neural networks for visual information analysis with limited computing resources," in *Proceedings of the IEEE International Conference on Image Processing (ICIP)*, 2018.