

Real-time target tracking using a PTZ camera with Delay-compensated IMM-UKF and Nonlinear Model Predictive Control

Le Van Khanh, Luong Thu Phong

Abstract: Tracking a maneuvering target with a Pan-Tilt-Zoom (PTZ) camera is a visual-servo problem in which state estimation, observation geometry, and actuator dynamics are tightly coupled. In real-time systems, image data are often processed asynchronously and with significant latency, so a measurement may reach the estimator after the physical time at which the image was captured. This paper develops an IMM-UKF consisting of two CV and CT-Singer models combined with an NIS-based adaptive Markov mechanism. Maneuver evidence is bounded, smoothed in measurement time, and mapped to a row-stochastic transition matrix. For out-of-sequence measurements, Rollback-Replay is extended to restore simultaneously the kinematic states, covariance matrices, model probabilities, adaptive state, and Markov matrix at the past time; the delayed measurement is updated using the correct Pan/Tilt geometry and focal length at the image-capture time, after which subsequent measurement events are reprocessed in true chronological order. The corrected target state is mapped into a reference sequence compatible with the PTZ actuator dynamics and supplied to the NMPC controller. In the optimization problem, the target position on the image plane is predicted directly through the nonlinear LOS-pixel relationship; image error is optimized together with dynamic tracking error, mechanical limits are retained as hard constraints, and the requirement to keep the target within the field of view is formulated as a soft constraint to preserve feasibility during aggressive target maneuvers or actuator saturation. Zoom is controlled independently by a gain-scheduled PID controller that uses the apparent target size to balance magnification and field-of-view reserve. The architecture is designed to address in a unified manner three difficulties specific to real-time PTZ systems: target motion-mode changes, delayed visual measurements, and the risk of losing the target from the field of view.

Keywords— PTZ camera; real-time target tracking; Adaptive IMM-UKF; adaptive Markov matrix; Rollback-Replay; out-of-sequence measurement; NMPC; field-of-view constraint; visual servoing; Gain-Scheduled PID.

I. INTRODUCTION

In robotics and automated surveillance systems, camera-based target tracking is an important visual-servo problem in which image information is used directly or indirectly to control the motion of the observation mechanism [1]-[3]. The target position in the image depends not only on the target motion state but also on the instantaneous Pan/Tilt angles and camera focal length, while these quantities

continuously change under the action of the control loop. Therefore, estimation error, actuator dynamics, and observation geometry are directly coupled in the closed loop [1]-[5].

The difficulty becomes pronounced when the target maneuvers rapidly. Interacting Multiple Model (IMM) methods make it possible to combine multiple kinematic hypotheses and update model probabilities according to observation information [10], [11], [13], [14]. When the measurement depends nonlinearly on the state, the UKF is an appropriate choice for propagating statistics through the measurement function without requiring local Jacobian-based linearization [15], [16]. However, IMM performance depends substantially on the model-transition probability matrix. If these probabilities are kept fixed, the filter may react slowly when the motion mode changes; conversely, if they are adjusted directly from an instantaneous innovation, the transition matrix can easily be dominated by noise or by bias caused by measurement latency.

Another characteristic of real-time vision systems is the asynchronous nature of the processing chain. Images must pass through acquisition, detection, data transmission, and processing before reaching the estimator; therefore, a measurement may arrive after the physical time at which the image was captured. This is the Out-of-Sequence Measurement (OOSM) problem, which has been studied in target-tracking theory [17], [18]. For a PTZ camera, the impact of OOSM is not limited to the difference between the target state in the past and the current state. Each pixel measurement is also associated with a specific Pan/Tilt configuration and a specific focal length at the capture time. If a delayed measurement is processed using the camera geometry at its arrival time, the timing mismatch is injected directly into the filter innovation.

This study develops an IMM-UKF with two CV and CT-Singer models in the absolute line-of-sight (LOS) angular space. The two transition probabilities $CV \rightarrow CT$ and $CT \rightarrow CV$ are parameterized directly from the adaptive state, so the Markov matrix is row-stochastic by construction. The adaptive state and the Markov matrix are treated as historical internal variables of the estimator. They are updated after each measurement and are associated with the physical timestamp of that measurement rather than with the packet reception time.

On this basis, the Rollback-Replay mechanism is extended to be consistent with the complete Adaptive IMM. Rollback-Replay not only corrects the state trajectory but also reconstructs the model-transition history of the Adaptive IMM before the current state is passed to the

Manuscript received June 29, 2026.

Le Van Khanh M.Sc. is with Hanoi, Vietnam (e-mail: levankhanhth@gmail.com, <https://orcid.org/0000-0002-6376-1769>)

Luong Thu Phong PhD.Eng. is with Hanoi, Vietnam (e-mail: luongthuphongmpei@gmail.com)

control layer. With complete checkpoints and event sequences, the procedure is algorithmically equivalent, apart from numerical error, to supplying the same set of measurements to the same Adaptive IMM-UKF in their correct chronological order.

At the control layer, the corrected target state is not passed directly to the controller as a simple angular trajectory. The IMM prediction sequence is mapped into a dynamic reference compatible with the Pan/Tilt servo model, including angle, angular velocity, and reference velocity command while accounting for the actuator time constant. The Pan/Tilt controller is formulated as an NMPC. The equivalent Pan/Tilt servo model is linear when the time constants are fixed; the nonlinearity of the optimization problem originates from the PTZ observation geometry itself. At each step of the prediction horizon, the target state and camera state are projected into pixel coordinates through the nonlinear LOS-pixel relationship. Thus, the NMPC directly optimizes error on the image plane rather than only angular tracking error. Travel, velocity, velocity-command, and command-rate limits are maintained as hard constraints. Meanwhile, the requirement to keep the target within the safe field-of-view region is formulated as a nonlinear soft constraint with slack variables, allowing the optimization problem to remain feasible when the target maneuvers aggressively.

The Zoom axis is controlled independently to avoid increasing the dimension of the NMPC problem. Rather than requiring the target distance and the true object size, the Zoom controller directly uses the normalized bounding-box size in the image. The desired zoom level is adjusted according to LOS velocity and image error: when the target maneuvers strongly or approaches the image boundary, the FOV is widened to increase observation reserve; when the target is stable near the optical center, the camera can increase magnification. The PID gains are scheduled within bounded ranges according to the tracking condition. The actual FOV and focal length after the Zoom action are fed back to the NMPC in the next cycle, so the Pan/Tilt and Zoom branches remain decoupled at the control level while being directly coupled through observation geometry.

The novelty of the study lies in the consistent treatment of the relationship among model adaptation, delayed measurements, PTZ observation geometry, and predictive control within the same closed loop.

1. The Adaptive Markov mechanism has a strictly structured probabilistic form. The NIS of the CV branch is converted into bounded maneuver evidence, smoothed in measurement time, and mapped directly to the CV→CT and CT→CV transition probabilities.

2. The Rollback-Replay mechanism is consistent with the complete Adaptive IMM. When processing OOSMs, the filter simultaneously restores the kinematic state, covariance, model probabilities, adaptive state, and Markov matrix.

3. A predictive-control layer specifically tailored to PTZ cameras is developed. The target-state sequence is mapped into references compatible with the actuator dynamics; the target position in the image is predicted directly using nonlinear geometry; image error is optimized together with

dynamic error, mechanical constraints remain hard, and the FOV is incorporated as a soft constraint.

The architecture is evaluated in a simulation environment with maneuvering targets, delayed measurements, and camera dynamic limits. In addition to angular and pixel errors, the evaluation considers estimator consistency, model probabilities, the ability to maintain the target within the FOV, observation-constraint violation, control-signal smoothness, and computation time in order to clarify the individual contribution of each mechanism in the PTZ closed loop.

II. GENERAL ARCHITECTURE AND DYNAMIC MODEL OF THE CAMERA SYSTEM

A. General system architecture

The visual-control system is organized as a real-time hierarchical architecture. The image stream is processed and estimated and then introduced into the target-tracking control loop according to the principles of visual servoing and image-based predictive control [1]-[3], [21], [24]. The architecture consists of five main functional blocks:

- 1. Signal acquisition and interaction block,
- 2. Visual-processing block,
- 3. State-estimation block,
- 4. Control block,
- 5. Actuator block.

The visual-processing block extracts the target pixel coordinates on the image plane from the real-time video stream. The measurement data are then transferred to the IMM-UKF estimator to infer the target kinematic state. The estimated state is subsequently passed to the NMPC control block to generate optimal Pan/Tilt commands for the camera.

B. Extended dynamic model of the pan/tilt mechanism

The Pan/Tilt mechanism is modeled as a first-order dynamic system that accounts for actuator inertia; this form is suitable for NMPC formulations with velocity and input-rate constraints [19], [20]. To be consistent with the NMPC controller, the motor velocity commands are included directly in the camera state vector:

$$X_c = [\theta, \phi, \omega_p, \omega_t, u_p, u_t]^T, \quad (1)$$

where θ, ϕ are the camera Pan/Tilt angles, ω_p, ω_t are the actual angular velocities of the two axes, and u_p, u_t are the velocity commands sent to the actuators.

The optimized system input is the derivative of the velocity command:

$$a_c = [\dot{u}_p, \dot{u}_t]^T = [a_p, a_t]^T \quad (2)$$

The continuous-time camera dynamics are described by the following differential equations:

$$\begin{cases} \dot{\theta} = \omega_p, \quad \dot{\phi} = \omega_t, \\ \dot{\omega}_p = \frac{u_p - \omega_p}{T_p}, \quad \dot{\omega}_t = \frac{u_t - \omega_t}{T_t}, \\ \dot{u}_p = a_p, \quad \dot{u}_t = a_t. \end{cases} \quad (3)$$

where T_p, T_t are the inertia time constants of the Pan and Tilt

mechanisms. This model enables the NMPC to optimize the rate of change of the velocity command directly while incorporating input and input-rate constraints in the same optimization problem [19], [22]-[24]. The physical constraints of the camera are:

$$\begin{cases} \phi_{\min} \leq \phi \leq \phi_{\max} \\ |\omega_p| \leq \omega_p^{\max}, |\omega_t| \leq \omega_t^{\max} \\ |u_p| \leq u_p^{\max}, |u_t| \leq u_t^{\max} \\ |a_p| \leq a_p^{\max}, |a_t| \leq a_t^{\max} \end{cases} \quad (4)$$

These constraints are used directly in the NMPC problem to ensure that the control commands do not exceed the mechanical limits of the camera.



Figure 1. PTZ camera model and frame of reference.

C. Object detection and image processing

The output of the image detection and tracking module is represented by an object bounding box, which is a common form of representation in modern object detection and tracking algorithms [7]–[9]. The measurement vector determines the target position in the pixel coordinate space:

$$Z_m = [u_m, v_m]^T. \quad (5)$$

where the coordinates of the target center (u_m, v_m) on the image plane are determined by $u_m = (x_{\min} + x_{\max})/2$, $v_m = (y_{\min} + y_{\max})/2$, with $x_{\min}$, $y_{\min}$, $x_{\max}$ and $y_{\max}$ being the bounding box enclosing the object.

From (5), the angular deviations in the absolute Line-of-Sight (LOS) space are determined and directly mapped through trigonometric equations [4]–[5]:

$$\begin{cases} \Delta\phi_m = \arctan\left(\frac{e_{v,m}}{f_y(t_m)}\right), \\ \Delta\theta_m = \arctan\left(\frac{e_{u,m}}{f_x(t_m)}\right), \end{cases} \quad (6)$$

where $e_{u,m} = u_m - c_x$ and $e_{v,m} = v_m - c_y$ are the relative position deviations on the image plane from the optical center (c_x, c_y) to the target. From (5) and (6), the absolute angular measurement of the target in the camera's base coordinate frame is determined:

$$\begin{cases} \tilde{\theta}_{T,m} = \text{wrap}_\pi(\theta_c(t_m) + \Delta\theta_m), \\ \tilde{\phi}_{T,m} = \phi_c(t_m) + \Delta\phi_m. \end{cases} \quad (7)$$

where $\text{wrap}_\pi(\cdot)$ is the operator that normalizes the angle to the interval $(-\pi, \pi]$.

The measurement covariance matrix in the pixel space is modeled according to the detection confidence [10]:

$$R_{uv,m} = \begin{bmatrix} \sigma_{u,m}^2 & 0 \\ 0 & \sigma_{v,m}^2 \end{bmatrix}, \quad (8)$$

where $\sigma_{u,m} = \frac{\sigma_{u,0}}{\max(\rho_m, \rho_{\min})}$, $\sigma_{v,m} = \frac{\sigma_{v,0}}{\max(\rho_m, \rho_{\min})}$, $\sigma_{u,0}$ and $\sigma_{v,0}$ are confidence-related constants.

III. STATE ESTIMATION WITH AN INTERACTING MULTIPLE-MODEL UNSCENTED KALMAN FILTER

To handle both the nonlinearity of the image measurement and target maneuvers, this study employs an IMM-UKF, in which the IMM combines multiple kinematic models and the UKF performs nonlinear estimation through the Unscented Transform [10]–[16]. Compared with the local linearization used in the EKF, the UKF is widely used for nonlinear systems because it propagates statistics through sigma points without directly computing the Jacobian matrix of the nonlinear model [15], [16]. The CV and CT-Singer models describe, respectively, approximately uniform motion and maneuvering motion with turning/acceleration, consistent with kinematic models commonly used for maneuvering-target tracking [10], [12], [14]. State interaction, mixing, and model-probability updates follow the standard IMM framework for systems with Markovian model switching [11], [13]. The NIS index is used to assess the statistical consistency of the innovation; under Gaussian-noise assumptions, it can be compared with a χ^2 threshold corresponding to the measurement dimension [10].

A. Kinematic modeling with IMM

To estimate the target state independently of camera motion and reduce the effect of pixel-coordinate oscillations, the IMM-UKF estimator is constructed in LOS space. The two kinematic models use the same state structure so that the interaction and state-mixing processes can be performed directly within the IMM [10], [11], [13], [14]. The target-state vector is defined in the six-dimensional LOS space as:

$$X_T = [\theta_T, \phi_T, v_T, \psi_T, \omega_T, a_T]^T \quad (9)$$

where, θ_T , ϕ_T are the azimuth and elevation angles of the target in the camera's base coordinate frame; v_T is the total angular velocity of the target; ψ_T is the motion heading angle of the target; ω_T is the angular turn rate; and a_T is the tangential acceleration of the LOS speed v_T .

The two models are propagated over the sigma-point set to satisfy real-time requirements:

Constant-velocity (CV) model: The CV model describes motion for which the magnitude and direction of LOS velocity vary negligibly over a short time interval [10], [14].

$$\begin{cases} \dot{\theta}_T = v_T \cos \psi_T, \dot{\phi}_T = v_T \sin \psi_T, \\ \dot{v}_T = a_T, \dot{\psi}_T = \omega_T, \\ \dot{\omega}_T = -\lambda_{\omega}^{CV} \omega_T, \dot{a}_T = -\lambda_a^{CV} a_T. \end{cases} \quad (10)$$

The two coefficients λ_{ω}^{CV} and λ_a^{CV} are chosen to be sufficiently large to drive ω_T and a_T close to 0, reflecting the assumption of no turning and no acceleration in the CV model. Compact form:

$$\dot{X}_T = f_{CV}(X_T) + w_{CV},$$

where $w_{CV} \sim N(0, Q_{CV})$ represents the process noise.

Maneuvering model with acceleration (CT-Singer): To describe a target that simultaneously changes the LOS motion direction and speed magnitude, the maneuvering model is constructed using a coordinated-turn structure combined with a Singer-type acceleration component [12], [14].

$$\begin{cases} \dot{\theta}_T = v_T \cos \psi_T, \dot{\phi}_T = v_T \sin \psi_T, \\ \dot{\omega}_T = -\frac{1}{\tau_{\omega}} \omega_T, \dot{a}_T = -\frac{1}{\tau_a} a_T, \\ \dot{v}_T = a_T, \dot{\psi}_T = \omega_T, \end{cases} \quad (11)$$

where τ_{ω} and τ_a are the time constants characterizing the decay of the turn rate and tangential acceleration, respectively. Compact form:

$$\dot{X}_T = f_{CT}(X_T) + w_{CT},$$

where $w_{CT} \sim N(0, Q_{CT})$ represents the process noise.

Adaptive Markov model-transition matrix: The Markov model-transition matrix is a core component of the IMM and describes the transition probabilities among the kinematic hypotheses over time.

Let $M_m \in \{CV, CT\}$ denote the target kinematic mode at measurement time t_m . The probability of transition from model i at the current step to model j at the next step is defined as:

$$p_{ij,m} = \Pr(M_{m+1} = j | M_m = i, I_m), \quad (12)$$

where I_m denotes the information set available after the measurement at t_m has been processed. Row i of the Markov matrix represents the transition-probability distribution from model i , with $\sum p_{ij,m} = 1$.

In a conventional IMM, the transition probabilities p_{ij} are usually selected as fixed values from prior information. This choice is appropriate when the motion characteristics vary slowly, but it may reduce the response speed of the IMM

when the target abruptly changes from nearly uniform motion to a maneuvering phase. Therefore, in this study, the transition matrix is adaptively adjusted according to the statistical mismatch of the CV model.

At each time with a valid measurement, the NIS of the CV branch is defined as:

$$\varepsilon_m^{CV} = (r_m^{CV})^T (S_m^{CV})^{-1} r_m^{CV} \quad (13)$$

where r_m^{CV} is the innovation vector of the CV model and S_m^{CV} is the corresponding innovation covariance matrix. When the CV model is consistent with the actual motion, the noise is approximately Gaussian, and the covariance model of the filter is consistent, the NIS approximately satisfies:

$$\varepsilon_m^{CV} \sim \chi_d^2,$$

where $d=2$ is the dimension of the pixel measurement vector.

To modify the Markov probabilities, the NIS is converted into a maneuver-evidence variable bounded in $[0,1]$:

$$\xi_m = \text{sat}_{[0,1]} \left(\frac{\varepsilon_m^{CV} - T_L}{T_H - T_L} \right), \quad T_L < T_H, \quad (14)$$

where the thresholds T_L and T_H are selected from quantiles of the χ_d^2 distribution:

$$T_L = F_{\chi_d^2}^{-1}(q_L), \quad T_H = F_{\chi_d^2}^{-1}(q_H), \quad 0 < q_L < q_H < 1.$$

When $\varepsilon_m^{CV} \leq T_L$, $\xi_m = 0$, corresponding to no strong maneuver. When $\varepsilon_m^{CV} \geq T_H$, $\xi_m = 1$, corresponding to a strong maneuver.

An adaptive state variable s_m , $0 \leq s_m \leq 1$ is used to reduce fluctuations of ξ_m :

$$\beta_m = 1 - \exp\left(-\frac{t_m - t_{m-1}}{\tau_s}\right), \quad s_m = (1 - \beta_m)s_{m-1} + \beta_m \xi_m, \quad (15)$$

where $\tau_s > 0$ is the time constant of the adaptation process.

From s_m , the two transition probabilities between the two models are:

$$\begin{aligned} p_{CV,CT,m} &= p_{12}^{\min} + (p_{12}^{\max} - p_{12}^{\min})s_m, \\ p_{CT,CV,m} &= p_{21}^{\min} + (p_{21}^{\max} - p_{21}^{\min})(1 - s_m) \end{aligned} \quad (16)$$

subject to $0 < p_{12}^{\min} < p_{12}^{\max} < 1$, $0 < p_{21}^{\min} < p_{21}^{\max} < 1$.

The adaptive Markov model-transition matrix is constructed directly as:

$$\Pi_m = \begin{bmatrix} 1 - p_{CV,CT,m} & p_{CV,CT,m} \\ p_{CT,CV,m} & 1 - p_{CT,CV,m} \end{bmatrix} \quad (17)$$

where $p_{ij,m} \in [0,1]$, $\Pi_m \mathbf{1} = \mathbf{1}$.

IMM state interaction: Consider the posterior states after the previous measurement $\hat{X}_{m-1}^{(i)+}$, $P_{m-1}^{(i)+}$, the posterior model probabilities $\mu_{i,m-1}^+$ and the transition matrix $\Pi_{m-1} = [p_{ij,m-1}]$. each destination model j , the prior model probability is predicted as:

$$\bar{\mu}_{j,m} = \sum_{i=1}^M p_{ij,m-1} \mu_{i,m-1}^+ \quad (18)$$

The mixing probability from model i to the initial condition of model j is:

$$\mu_{ij,m-1}^{\text{mix}} = \frac{p_{ij,m-1} \mu_{i,m-1}^+}{\bar{\mu}_{j,m}}, \quad i, j = 1, \dots, M. \quad (19)$$

with $\sum_i \mu_{ij,m-1} = 1$.

The mixed initial state of branch j is determined by:

$$\bar{X}_{0,m-1}^{(j)} = \sum_{i=1}^M \mu_{ij,m-1}^{\text{mix}} \hat{X}_{m-1}^{(i)}. \quad (20)$$

Let $d_{ij} = \hat{X}_{m-1}^{(i)} - \bar{X}_{0,m-1}^{(j)}$. The mixed covariance matrix is then:

$$\bar{P}_{0,m-1}^{(j)} = \sum_{i=1}^M \mu_{ij,m-1}^{\text{mix}} \left(P_{m-1}^{(i)+} + d_{ij} d_{ij}^T \right) \quad (21)$$

From the mixed initial condition, the UKF for each model is propagated over the exact interval Δt_m :

$$\left(\hat{X}_m^{(j,\text{pred})}, P_m^{(j,\text{pred})} \right) = P_j^{\text{UKF}} \left(\bar{X}_{0,m-1}^{(j)}, \bar{P}_{0,m-1}^{(j)}, \Delta t_m \right) \quad (22)$$

B. Nonlinear measurement equations

The IMM-UKF measurement model is formulated as a nonlinear projection function from the target LOS space to pixel-coordinate space, based on the pinhole camera model and processed by the UKF [4], [5], [15], [16].

The nonlinear measurement equation at time t_m is:

$$Z_m = h(X_{T,m}) + r_m, \quad (23)$$

where, $h(X_{T,m})$ is the nonlinear measurement function and

$r_m \sim N(0, R_{uv,m})$ is the measurement noise.

From (5) and (6), the angular deviations in LOS space are calculated using the trigonometric function:

$$\begin{cases} \Delta\theta_m = \text{atan2}(\sin(\theta_{T,m} - \theta_{C,m}), \cos(\theta_{T,m} - \theta_{C,m})), \\ \Delta\phi_m = \text{atan2}(\sin(\phi_{T,m} - \phi_{C,m}), \cos(\phi_{T,m} - \phi_{C,m})). \end{cases} \quad (24)$$

The expected pixel coordinates $(\hat{u}, \hat{v})$ are then determined as:

$$\begin{cases} \hat{u} = c_x + f_x \cdot \tan(\Delta\theta_m), \\ \hat{v} = c_y - f_y \cdot \tan(\Delta\phi_m). \end{cases} \quad (25)$$

Equation (20) represents the local perspective projection between LOS angular deviation and image pixel coordinates, where the focal length and optical center are the intrinsic camera parameters [4], [5].

The complete nonlinear measurement equation of the system is written as:

$$Z_m = h(X_{T,m}, \Theta_C(t_m), f_x(t_m), f_y(t_m)) + r_m, \quad (26)$$

where, $\Theta_C(t_m) = [\theta_{C,m}, \phi_{C,m}]^T$ denotes the camera Pan/Tilt angles at image-capture time, and the measurement function is:

$$h(X_{T,m}) = \begin{bmatrix} c_x + f_x \cdot \tan(X_{T,m}[0] - \theta_{C,m}) \\ c_y - f_y \cdot \tan(X_{T,m}[1] - \phi_{C,m}) \end{bmatrix}$$

The innovation covariance matrix obtained by projecting the sigma points through (26) is used both for the UKF update and for computing the NIS of the CV branch according to (13).

C. Delay compensation mechanism

Assume that the system is operating at the current time t_k , while a visual measurement vector Z_m was captured at time $t_m < t_k$ and only reaches the estimator after a delay. If this measurement is updated directly at t_k , the time mismatch can increase the innovation and NIS, thereby not only biasing the kinematic state but also directly affecting the adaptive Markov-matrix mechanism [17], [18].

In this study, the Rollback-Replay mechanism is extended to be consistent with the complete internal state of the Adaptive IMM. At each stored time t_i , the full estimator state is defined as:

$$S_i = \left\{ \left(\hat{X}_{T,i}^{(j)}, P_{i,i}^{(j)} \right)_{j=1}^M, \mu_i, \Pi_i, S_i, t_{z,i}^{\text{last}} \right\}, \quad (27)$$

where $t_{z,i}^{\text{last}}$ is the timestamp of the latest valid measurement event that has been assimilated into the current Adaptive IMM state. To distinguish events having the same timestamp, each checkpoint also stores q_z^{last} and the committed key $\kappa_z^{\text{last}} = (t_z^{\text{last}}, q_z^{\text{last}})$. This quantity changes only after a valid measurement update; prediction does not change the historical measurement boundary.

Each record in the historical state buffer is written as:

$$H_i = \{t_i, S_i, \Theta_C(t_i), f_i\}, \quad (28)$$

with $\Theta_C(t_i) = [\theta_C(t_i), \phi_C(t_i)]^T$, $f_i = [f_x(t_i), f_y(t_i)]^T$. In parallel with the state buffer, measurements already processed within the retrospective window are stored as events:

$$E_\ell = \{t_\ell, Z_\ell, R_{uv,\ell}, \Theta_C(t_\ell), f_\ell, q_\ell\}, \quad (29)$$

where q_ℓ is a sequence number used to determine the processing order when two events have the same timestamp. $R_{uv,\ell}$ is the noise covariance matrix of the pixel measurement at time t_ℓ . The buffer is limited by a retrospective window T_R ; With the key $\kappa_m = (t_m, q_m)$, an event is classified according to chronological order as:

$$E_m = \begin{cases} \text{in-sequence,} & \kappa_m > \kappa_z^{\text{last}}, \\ \text{duplicate,} & \kappa_m = \kappa_z^{\text{last}}, \\ \text{OOSM,} & \kappa_m < \kappa_z^{\text{last}}. \end{cases} \quad (30)$$

An OOSM is processed by Rollback-Replay only if it still lies within the retrospective window:

$$\kappa_m < \kappa_z^{\text{last}}, \quad 0 \leq t_k - t_m \leq T_R, \quad (31)$$

Therefore, the case $t_z^{\text{last}} < t_m < t_k$ is a delayed measurement that is still in sequence and is updated directly at t_m without triggering rollback t_m without triggering rollback

1. Search for the retrospective time checkpoint: When an

OOSM at t_m , is received, the latest record in the state buffer that is not later than the measurement time is selected:

$$t_b = \max \{t_i \in B_S : t_i \leq t_m\}, \quad (32)$$

where B_S is the set of timestamps stored in the buffer. The Adaptive IMM state is restored to t_b $S_{\text{filter}} \leftarrow S_b$.

Thus, the rollback simultaneously restores $\hat{X}^{(j)}$, $P^{(j)}$, μ , Π , s and t_z^{last} . This is an important difference from restoring only the kinematic state: the adaptive Markov matrix and the maneuver-memory state are returned to their correct historical values before the delayed measurement is updated.

2. Dynamic synchronization at the delayed measurement time: The camera parameters used in the nonlinear measurement function of Section 3.2 must correspond to the time at which the image was recorded t_m , rather than to the packet-arrival time t_k . If the image packet directly carries metadata at capture time, $\Theta_C(t_m)$ and f_m are used directly. If only camera telemetry at two neighboring timestamps t_b and t_{b+1} , is available, the parameters are interpolated as:

$$\begin{aligned} \theta_C(t_m) &= \text{wrap}_\pi(\theta_C(t_b) + \alpha_m \delta\theta_b), \\ \phi_C(t_m) &= (1 - \alpha_m) \phi_C(t_b) + \alpha_m \phi_C(t_{b+1}), \end{aligned} \quad (33)$$

where, $\alpha_m = (t_m - t_b) / (t_{b+1} - t_b)$, $f_m = (1 - \alpha_m) f_b + \alpha_m f_{b+1}$, and $\delta\theta_b = \text{wrap}_\pi(\theta_C(t_{b+1}) - \theta_C(t_b))$.

Because t_m generally does not coincide exactly with a state timestamp in the buffer, the IMM is propagated from t_b to immediately before the measurement update time t_m with $\Delta t_{\text{past}} = t_m - t_b$:

$$S_{m^-} = P_{\text{IMM}}(S_b, \Delta t_{\text{past}}) \quad (34)$$

where $P_{\text{IMM}}(\cdot)$ denotes the operator consisting of IMM interaction, state mixing, and conditional UKF propagation for each model [11], [13]. During intervals without a new measurement s and Π remain at the restored values, while model probabilities are predicted using the corresponding transition matrix.

3. Retrospective nonlinear update and adaptive markov update at t_m : From the predicted state of each model at t_m^- , the sigma points are projected through the nonlinear measurement function (26) at the image-capture time. For model j , the innovation vector is defined as:

$$r_m^{(j)} = Z_m - \hat{Z}_m^{(j)} \quad (35)$$

and the innovation covariance matrix is:

$$S_m^{(j)} = P_{zz,m}^{(j)} + R_m \quad (36)$$

where, $P_{zz,m}^{(j)}$ is the predicted covariance in measurement space and R_m is the measurement noise covariance.

The Kalman gain of UKF branch j is:

$$K_m^{(j)} = P_{xz,m}^{(j)} (S_m^{(j)})^{-1} \quad (37)$$

The posterior state and covariance are updated as:

$$\begin{aligned} \hat{X}_{T,m}^{(j)} &= \hat{X}_{T,m-1}^{(j)} + K_m^{(j)} r_m^{(j)}, \\ P_{m,m}^{(j)} &= P_{m,m-1}^{(j)} - K_m^{(j)} S_m^{(j)} K_m^{(j)T}. \end{aligned} \quad (38)$$

The likelihood of each model is calculated from the Gaussian distribution:

$$\Lambda_m^{(j)} = \frac{\exp\left[-1/2 \left(r_m^{(j)}\right)^T \left(S_m^{(j)}\right)^{-1} r_m^{(j)}\right]}{\sqrt{(2\pi)^d |S_m^{(j)}|}} \quad (39)$$

Let $\bar{\mu}_m^{(j)}$ be the prior model probability obtained from the IMM interaction and prediction step. The posterior model probability after the measurement is updated as:

$$\mu_m^{(j)} = \frac{\Lambda_m^{(j)} \bar{\mu}_m^{(j)}}{\sum_{\ell=1}^M \Lambda_m^{(\ell)} \bar{\mu}_m^{(\ell)}}. \quad (40)$$

subject to the normalization condition: $\sum_{j=1}^M \mu_m^{(j)} = 1$.

After the UKF state update and model-probability update have been completed, the NIS of the CV branch is calculated according to (13). The adaptive state s_m and Markov transition matrix Π_m are then updated according to (14)-(17) at the measurement timestamp t_m itself:

$$(s_m, \Pi_m) = A\left(s_{m^-}, \Pi_{m^-}, \varepsilon_m^{\text{CV}}, t_m - t_{z,m^-}^{\text{last}}\right) \quad (41)$$

after which $t_{z,m}^{\text{last}}$ is assigned t_m . Here operator $A(\cdot)$ denotes the adaptation law defined in (14)-(17). The matrix Π_m is formed after the measurement at t_m has been processed and only affects subsequent IMM interaction steps. At the same time, $q_z^{\text{last}} \leftarrow q_m$, $\kappa_z^{\text{last}} \leftarrow \kappa_m$.

4. Chronological processing up to the current time: After inserting the OOSM E_m , the algorithm does not merely predict from t_m to t_k . All measurement events already processed within the retrospective interval, together with E_m , are sorted in ascending order by the key pair (t_ℓ, q_ℓ) . Let the resulting sorted sequence be:

$$E_{b:k}^{\text{RR}} = E_1, \dots, E_{N_R} \quad (42)$$

On the same time grid stored in the buffer, the Adaptive IMM state is reconstructed sequentially. For each interval $\Delta t_\ell = t_\ell - t_{\ell-1}$, an IMM prediction is carried out:

If a measurement exists at time t_ℓ the complete UKF update and model-probability update are repeated; NIS is then calculated and the adaptive state s_ℓ together with the Markov transition matrix Π_ℓ is updated according to (13)-(17). On each interval Δt_ℓ :

$$S_\ell^{\text{pre}} = P_\ell S_{\ell-1}^{\text{post}} \quad (43)$$

If a measurement event E_ℓ exists at t_ℓ with $U_\ell(\cdot)$ denoting the Adaptive IMM update operator defined in (35)-(41), the post-update state is:

$$S_\ell^{\text{post}} = U_\ell(S_\ell^{\text{pre}}, E_\ell) \quad (44)$$

If no measurement exists at t_ℓ , set $S_\ell^{\text{post}} = S_\ell^{\text{pre}}$. When there is no measurement, no new NIS exists and therefore no

new adaptation update is generated for s and Π . Every reconstructed state overwrites the corresponding entry in the historical buffer; consequently, the entire sequence $\hat{X}^{(j)}$, $P^{(j)}$, μ , s and Π after t_m correctly reflects the occurrence of the OOSM.

After the final event has been replayed, the filter is propagated to the current control time t_k yielding the state S_k^{RR} . Only the estimated state is rolled back; past NMPC commands and the past physical state of the camera are not rolled back. Frames that have already occurred are used as recorded data at their corresponding measurement events, while the corrected target state at t_k is supplied to the NMPC in the next control cycle.

Algorithm 1 Adaptive IMM-UKF with Timestamp-Consistent Rollback-Replay:

1: Input: $\mathcal{M}=\{\text{CV, CT-Singer}\}$, committed state S_z , buffers BS and BZ, retrospective window TR, current time t_k , new measurement events $\mathcal{E}_{\text{knew}}$, horizon N, sampling time T_s .
 2: Output: Fused target state $\hat{X}_{T,k}$ and prediction sequence X_{Tref} .

// Validate and classify new measurements

```

3: for each  $E_m \in \mathcal{E}_{\text{knew}}$  do
4:    $\kappa_m \leftarrow (\text{tm}, \text{qm})$ 
5:   if  $\text{tm} > t_k + \epsilon t$  then hold  $E_m$ 
6:   else if  $t_k - \text{tm} > \text{TR}$  then discard  $E_m$ 
7:   else insert  $E_m$  into BZ
8:   end if
9: end for
10: Remove duplicates and sort valid events by  $\kappa_m$ .
11:  $\mathcal{EOOSM} \leftarrow \{E_m : \kappa_m < \kappa_{\text{zlast}}\}$ 
    // Case 1: no out-of-sequence measurement
12: if  $\mathcal{EOOSM} = \emptyset$  then
13:   for each valid  $E_m$  in increasing  $\kappa_m$  do
14:      $S-m \leftarrow \mathcal{PIMM}(S_z, \text{tm} - t_{\text{zlast}})$ 
15:      $S+m \leftarrow \mathcal{UIMM}(S-m, E_m)$ 
16:      $\varepsilon\text{mCV} \leftarrow \text{NISCv}(S+m)$ 
17:      $[\text{sm}, \Pi\text{m}] \leftarrow \mathcal{A}(\text{sm}-1, \Pi\text{m}-1, \varepsilon\text{mCV}, \text{tm} - t_{\text{zlast}})$ 
18:      $S_z \leftarrow \text{COMMIT}(S+m, \text{sm}, \Pi\text{m}, \kappa_m)$ 
19:     Store  $S_z$  in BS
20:   end for

```

// Case 2: one or more OOSMs

```

21: else
22:    $\kappa_{\text{min}} \leftarrow \min\{\kappa_m : E_m \in \mathcal{EOOSM}\}$ 
23:    $S_b \leftarrow$  latest checkpoint satisfying  $\kappa_b \leq \kappa_{\text{min}}$ 
24:    $S_z \leftarrow S_b$ 
25:    $\mathcal{ERR} \leftarrow$  historical events after  $\kappa_b \cup$  valid new events
26:   Remove duplicates and sort  $\mathcal{ERR}$  by  $\kappa_r$ 
27:   for each  $E_r \in \mathcal{ERR}$  do
28:      $S-r \leftarrow \mathcal{PIMM}(S_z, \text{tr} - t_{\text{zlast}})$ 
29:      $S+r \leftarrow \mathcal{UIMM}(S-r, E_r)$ 
30:      $[\text{sr}, \Pi\text{r}] \leftarrow \mathcal{A}(\text{sr}-1, \Pi\text{r}-1, \text{NISCv}(S+r), \text{tr} - t_{\text{zlast}})$ 
31:      $S_z \leftarrow \text{COMMIT}(S+r, \text{sr}, \Pi\text{r}, \kappa_r)$ 
32:     Rewrite corresponding checkpoint in BS
33:   end for
34: end if

```

// Prediction to control time and NMPC horizon

```

35:  $S-k|z \leftarrow \mathcal{PIMM}(S_z, t_k - t_{\text{zlast}})$ 
36:  $\hat{X}_{T,k} \leftarrow \text{IMM\_FUSION}(S-k|z)$ 

```

```

37: for  $i \leftarrow 1$  to N do
38:    $S-k+i|k \leftarrow \mathcal{PIMM}(S-k+i-1|k, T_s)$ 
39:    $\hat{X}_{T,k+i|k} \leftarrow \text{IMM\_FUSION}(S-k+i|k)$ 
40: end for
41:  $X_{Tref} \leftarrow \{\hat{X}_{T,k+i|k}\}_{i=0}^N$ 
42: return  $\hat{X}_{T,k}, X_{Tref}$ 

```

Mapping the Target State into a PTZ Dynamic Reference for NMPC

The predicted target-state sequence at the output of Algorithm 1 is $X_{T,k}^{ref} = \{\hat{X}_{T,k+i|k}\}_{i=0}^N$. At each node $k+i$ of the prediction horizon, the fused target state has the form:

$$\hat{X}_{T,k+i|k} = [\hat{\theta}_{T,k+i|k} \quad \hat{\phi}_{T,k+i|k} \quad \hat{\omega}_{T,k+i|k} \quad \hat{\psi}_{T,k+i|k} \quad \hat{\omega}_{T,k+i|k} \quad \hat{a}_{T,k+i|k}]^T$$

The PTZ reference vector is defined as:

$$r_{k+i|k} = [\theta_{r,k+i|k} \quad \phi_{r,k+i|k} \quad \omega_{p,r,k+i|k} \quad \omega_{t,r,k+i|k} \quad u_{p,r,k+i|k} \quad u_{t,r,k+i|k}]^T$$

where the components are defined as follows:

$$\begin{aligned} \theta_{r,k+i|k} &= \hat{\theta}_{T,k+i|k}, \quad \phi_{r,k+i|k} = \hat{\phi}_{T,k+i|k}, \\ \omega_{p,r,k+i|k} &= \hat{\omega}_{T,k+i|k} \cos(\hat{\psi}_{T,k+i|k}), \\ \omega_{t,r,k+i|k} &= \hat{\omega}_{T,k+i|k} \sin(\hat{\psi}_{T,k+i|k}), \\ \dot{\omega}_{p,r,k+i|k} &= \hat{a}_{T,k+i|k} \cos(\hat{\psi}_{T,k+i|k}) - \hat{\omega}_{T,k+i|k} \hat{\omega}_{T,k+i|k} \sin(\hat{\psi}_{T,k+i|k}), \\ \dot{\omega}_{t,r,k+i|k} &= \hat{a}_{T,k+i|k} \sin(\hat{\psi}_{T,k+i|k}) + \hat{\omega}_{T,k+i|k} \hat{\omega}_{T,k+i|k} \cos(\hat{\psi}_{T,k+i|k}), \\ u_{p,r,k+i|k} &= \omega_{p,r,k+i|k} + T_p \dot{\omega}_{p,r,k+i|k}, \\ u_{t,r,k+i|k} &= \omega_{t,r,k+i|k} + T_t \dot{\omega}_{t,r,k+i|k}. \end{aligned}$$

The complete PTZ dynamic-reference sequence over the prediction horizon is written as:

$$R_k = \{r_{k+i|k}\}_{i=0}^N. \quad (45)$$

IV. FORMULATION OF THE TARGET TRACKING CONTROLLER

This study utilizes a Nonlinear Model Predictive Control (NMPC) architecture rather than other feedback control configurations because NMPC is capable of directly handling state constraints, input constraints, and predicting trajectories over a finite horizon. This characteristic is highly suitable for target tracking problems featuring nonlinear dynamics and actuator limitations [19]–[24].

A. State-Space modeling of the camera

Including the velocity command in the augmented state allows the controller to model changes in the control command directly, similarly to velocity-form MPC structures or MPC with input-rate constraints [19], [20]:

$$X_c = [\theta, \phi, \omega_p, \omega_t, u_p, u_t]^T, \quad (46)$$

where the optimized MPC input vector $a_c \in \mathbb{R}^2$ represents the control acceleration $a_c = [a_p, a_t]^T = [\dot{u}_p, \dot{u}_t]^T$.

The continuous-time camera dynamics in (3) can be rewritten as:

$$\begin{cases} \dot{\theta} = \omega_p, \quad \dot{\phi} = \omega_t \\ \dot{\omega}_p = (u_p - \omega_p)/T_p, \quad \dot{\omega}_t = (u_t - \omega_t)/T_t \\ \dot{u}_p = a_p, \quad \dot{u}_t = a_t \end{cases} \quad (47)$$

where $T_p > 0$ and $T_t > 0$. The compact state form of (47) is

denoted by $\dot{X}_c = f_c(X_c, a_c; T_p, T_t)$ or in discrete time with sampling period T_s :

$$X_{c,k+1} = F_d(X_{c,k}, a_{c,k}; T_p, T_t, T_s), \quad (48)$$

The mapping (48) is used to predict the actuator state over the entire NMPC horizon.

B. Tracking errors and nonlinear observation constraints of the PTZ camera

At each control cycle k , the augmented Pan/Tilt state model in (46)-(48) is predicted over the horizon $i = 0, \dots, N$ using the dynamic-reference sequence R_k . Unlike approaches that optimize only angular error in joint space, this study directly incorporates the PTZ observation geometry into the NMPC problem. The predictive controller estimates not only the actuator state but also the target position on the image plane, thereby directly coupling tracking error, the current Zoom level, and the ability to keep the target within the field of view.

1. *Dynamic reference tracking error*: With the predicted state

$$X_{c,ik} = [\theta_{ik}, \phi_{ik}, \omega_{p,ik}, \omega_{t,ik}, u_{p,ik}, u_{t,ik}]^T$$

and the reference vector r_{ik} constructed in Section 3.3, three groups of dynamic tracking errors are defined:

$$\begin{cases} e_{\alpha,ik} = \begin{bmatrix} \text{wrap}_{\pi}(\theta_{ik} - \theta_{r,ik}) \\ \phi_{ik} - \phi_{r,ik} \end{bmatrix}, \\ e_{\omega,ik} = \begin{bmatrix} \omega_{p,ik} - \omega_{p,r,ik} \\ \omega_{t,ik} - \omega_{t,r,ik} \end{bmatrix}, \\ e_{u,ik} = \begin{bmatrix} u_{p,ik} - u_{p,r,ik} \\ u_{t,ik} - u_{t,r,ik} \end{bmatrix}. \end{cases} \quad (49)$$

where, e_{α} reflects the Pan/Tilt orientation error; e_{ω} reflects the error in the actual angular velocities of the actuator; and e_u evaluates the ability to track the inertia-compensated reference velocity command generated by the dynamic mapping in Section 3.3.

2. *Predicted image geometry and treatment of the zoom state*: To incorporate the visual-servo objective directly into the NMPC, at each prediction step the LOS deviation between the predicted target and the camera direction is determined as:

$$\begin{cases} \Delta\theta_{ik} = \text{wrap}_{\pi}(\theta_{r,ik} - \theta_{ik}), \\ \Delta\phi_{ik} = \phi_{r,ik} - \phi_{ik}. \end{cases} \quad (50)$$

Using the same nonlinear pinhole projection model developed in Section 3.2, the predicted target pixel coordinates are:

$$\begin{cases} \hat{u}_{ik} = c_x + f_{x,ik} \tan(\Delta\theta_{ik}), \\ \hat{v}_{ik} = c_y - f_{y,ik} \tan(\Delta\phi_{ik}). \end{cases} \quad (51)$$

where, f_x and f_y are treated as online optical parameters of the optimization problem:

$$f_{x,ik} = f_{x,k}, \quad f_{y,ik} = f_{y,k}, \quad i = 0, \dots, N$$

The direct image-plane error is defined as the displacement

from the predicted pixel position to the optical center:

$$e_{I,ik} = [\hat{u}_{ik} - c_x \quad \hat{v}_{ik} - c_y]^T \quad (52)$$

3. *Hard mechanical constraints*: The physical limits of the mechanism are treated as hard constraints and must be satisfied at every step of the prediction horizon

$$\begin{cases} \phi_{\min} \leq \phi_{ik} \leq \phi_{\max}, \\ |\omega_{p,ik}| \leq \omega_{p,\max}, \quad |\omega_{t,ik}| \leq \omega_{t,\max}, \\ |u_{p,ik}| \leq u_{p,\max}, \quad |u_{t,ik}| \leq u_{t,\max}, \\ |a_{p,ik}| \leq a_{p,\max}, \quad |a_{t,ik}| \leq a_{t,\max}. \end{cases} \quad (53)$$

where, $a_{p,\max}, a_{t,\max}$ are the limits on the rate of change of the velocity command as defined in Section 4.1.

4. *Soft field-of-view constraint*: In this study, the field of view is incorporated directly into the optimization problem through the predicted pixel coordinates in (51). Let $W_{\text{img}} \times H_{\text{img}}$ denote the image size and $m_u, m_v > 0$ the safety margins in the two directions. The desired observation region is defined as:

$$F_s = \{(u, v) : m_u \leq u \leq W_{\text{img}} - m_u, m_v \leq v \leq H_{\text{img}} - m_v\} \quad (54)$$

Nonnegative slack variables $m_{u,i}, m_{v,i}$ are used to soften the observation constraint (54):

$$\begin{cases} m_u - m_{u,i} \leq \hat{u}_{ik} \leq W_{\text{img}} - m_u + m_{u,i}, \\ m_v - m_{v,i} \leq \hat{v}_{ik} \leq H_{\text{img}} - m_v + m_{v,i}, \\ m_{u,i} \geq 0, \quad m_{v,i} \geq 0. \end{cases} \quad (55)$$

This formulation also highlights a characteristic specific to the PTZ problem: the predictive controller not only tracks the angular trajectory corrected by the Adaptive IMM-UKF and Rollback-Replay, but also directly predicts the image position and preserves target observability under the effects of Zoom and actuator limits.

C. Objective function and constrained nmpc optimization problem

Based on the augmented state model in Section 4.1 and the nonlinear observation geometry in Section 4.2, the Pan/Tilt control problem is formulated as an NMPC over a finite prediction horizon of N steps. Unlike an objective function that tracks only angular error, the proposed objective simultaneously optimizes the target position on the image plane, the actuator dynamic errors, command smoothness, and the level of FOV-constraint violation.

1. *Normalization of image-plane error*: The image error used in the objective function is normalized by half of the image dimensions:

$$\tilde{e}_{I,ik} = \begin{bmatrix} \hat{u}_{ik} - c_x & \hat{v}_{ik} - c_y \\ W_{\text{img}}/2 & H_{\text{img}}/2 \end{bmatrix}^T \quad (56)$$

Stage cost: At prediction step i , let the FOV-constraint slack vector be:

$$m_{ik} = [m_{u,ik} \quad m_{v,ik}]^T, \quad m_{ik} \geq 0.$$

The stage cost is defined as:

$$\begin{aligned} \ell_{i|k} = & \tilde{e}_{I,i|k}^T Q_I \tilde{e}_{I,i|k} + e_{\alpha,i|k}^T Q_\alpha e_{\alpha,i|k} + e_{\omega,i|k}^T Q_\omega e_{\omega,i|k} + \\ & + e_{u,i|k}^T Q_u e_{u,i|k} + a_{c,i|k}^T R_a a_{c,i|k} + \Delta a_{c,i|k}^T R_{\Delta a} \Delta a_{c,i|k} + \\ & + \rho_{F,1} \mathbf{1}^T \int_{i|k} + \int_{i|k}^T R_F \int_{i|k}. \end{aligned} \quad (57)$$

where, Q_I prioritizes bringing the target toward the optical center; Q_α maintains a small orientation error in LOS space; Q_ω forces the actual Pan/Tilt angular velocities to track the target-predicted angular velocities; and Q_u forces the velocity commands to track the dynamic references $u_{p,r}, u_{t,r}$ generated in Section 3.3. The terms R_a and $R_{\Delta a}$ respectively penalize the rate of change of the velocity command and the change of this quantity between successive steps, thereby reducing abrupt commands and high-frequency oscillations.

The input increment is defined as:

$$\Delta a_{c,i|k} = a_{c,i|k} - a_{c,i-1|k}, \quad a_{c,-1|k} = a_{c,k-1}^e \quad (58)$$

where $a_{c,k-1}^e$ is the command applied in the previous control cycle. The last two terms of (57) penalize only the observation slacks. The weights $\rho_{F,1} > 0$ and $R_F > 0$ are selected sufficiently large so that violation of the safe FOV region occurs only when the dynamic/mechanical limits prevent the observation requirement from being satisfied instantaneously.

2. Terminal cost of the prediction horizon: The terminal cost is constructed from the image error, orientation error, velocity error, and FOV violation at step N :

$$\begin{aligned} V_{N|k} = & \tilde{e}_{I,N|k}^T P_I \tilde{e}_{I,N|k} + e_{\alpha,N|k}^T P_\alpha e_{\alpha,N|k} + \\ & + e_{\omega,N|k}^T P_\omega e_{\omega,N|k} + \rho_{F,N} \mathbf{1}^T \int_{N|k} + \int_{N|k}^T P_F \int_{N|k}. \end{aligned} \quad (59)$$

The terminal weights are selected so that the target remains within the safe observation region at the end of the horizon. The total cost is:

$$J_k = \sum_{i=0}^{N-1} \ell_{i|k} + V_{N|k} \quad (60)$$

3. Constrained nonlinear optimization problem: At each cycle k , the NMPC determines the optimal sequences as:

$$\left\{ X_c^e, a_c^e, f^e \right\} = \arg \min_{\{X_c, a_c, f\}} J_k \quad (61)$$

The initial condition and discrete state equation are:

$$X_{c,0|k} = \hat{X}_{c,k}, \quad X_{c,i+1|k} = F_d(X_{c,i|k}, a_{c,i|k}, T_s), \quad (62)$$

with $i = 0, \dots, N-1$. Mechanical limits are imposed as the hard constraint set:

$$X_{c,i|k} \in X_{\text{mech}}, \quad a_{c,i|k} \in A_{\text{mech}}. \quad (63)$$

The observation constraint is imposed as a soft set with slack variables:

$$(X_{c,i|k}, f_{i|k}) \in C_{\text{FOV}}(r_{i|k}, f_{x,k}, f_{y,k}), \quad f_{i|k} \geq 0 \quad (64)$$

In (61)–(64), $\hat{X}_{c,k}$ is the current Pan/Tilt state used to initialize the optimization problem, and $F_d(\cdot)$ is the discrete model in (48). The sets X_{mech} and A_{mech} represent the hard mechanical limits of the state and input in (53). The set C_{FOV} represents the soft FOV inequalities in (55), parameterized by the target reference and current focal

length. Consequently, travel limits, actual velocity, velocity command, and command-rate limits always remain hard; only the observation constraints may be violated through slack variables when the target maneuvers rapidly or the actuator saturates.

D. Design of the Gain-Scheduled PID Zoom Controller

Target size in the image and maneuver level: Let $w_k = x_{\max,k} - x_{\min,k}$ be the width of the target bounding box at cycle k , and W_{img} be the image width. The target size is normalized as:

$$\gamma_k = \frac{w_k}{W_{\text{img}}}, \quad 0 < \gamma_k \leq 1 \quad (65)$$

To construct a Zoom adjustment law according to the tracking condition, the LOS velocity estimated by the IMM-UKF and the normalized image error from Section 4.2 are converted to bounded quantities:

$$\bar{v}_k = \text{sat}_{[0,1]} \left(\frac{\hat{v}_{T,k}}{v_{\text{ref}}} \right), \quad \eta_k = \|\tilde{\mathbf{e}}_{I,k}\|_2. \quad (66)$$

where $v_{\text{ref}} > 0$ is a reference LOS velocity used for normalization. The desired target size is determined as:

$$\gamma_k^* = \text{sat}_{[\gamma_{\min}, \gamma_{\max}]} \left[\frac{\gamma_0}{1 + K_v \bar{v}_k + K_e \eta_k} \right], \quad (67)$$

where γ_0 is the desired image ratio when the target moves steadily and remains near the image center; $K_v, K_e \geq 0$ are tuning coefficients; $\gamma_{\min}$ and $\gamma_{\max}$ bound the allowable target-size range. According to (67), an increase in maneuver level or image error reduces γ_k^* corresponding to a tendency to Zoom out and widen the FOV; conversely, under good tracking conditions, the target may be Zoomed in to increase observation resolution.

1. Conversion of the desired image size to reference focal length and fov: With the pinhole model, at the same observation time the apparent size of the same target is proportional to the focal length. Therefore, the reference focal length can be determined from the image-size ratio without requiring direct knowledge of the target distance D or actual size W_{real} :

$$f_{x,k}^* = \text{sat}_{[f_{x,\min}, f_{x,\max}]} \left[f_{x,k} \frac{\gamma_k^*}{\max(\gamma_k, \gamma_\epsilon)} \right], \quad (68)$$

where $\gamma_\epsilon > 0$ prevents degeneracy when the bounding box is very small. From the reference focal length, the desired horizontal FOV is determined and constrained within the physical range of the Zoom mechanism:

$$\text{FOV}_k^* = \text{sat}_{[\text{FOV}_{\min}, \text{FOV}_{\max}]} \left[2 \arctan \left(\frac{W_{\text{img}}}{2 f_{x,k}^*} \right) \right]. \quad (69)$$

In the feedback-control system, the deviation caused by changes in target distance is corrected in subsequent frames rather than by adding a separate distance estimator.

2. Zoom error and pid gain-scheduling variable: The Zoom-control error in the FOV domain is:

$$e_{z,k} = \text{FOV}_k^* - \text{FOV}_k \quad (70)$$

To vary PID sensitivity according to the tracking condition while ensuring that all gains remain bounded, a normalized scheduling variable is defined as:

$$\chi_k = \text{sat}_{[0,1]}(w_v \bar{v}_k + w_e \eta_k), w_v + w_e = 1, w_v, w_e \geq 0 \quad (71)$$

The PID gains are interpolated within bounded intervals:

$$\begin{aligned} K_{P,k} &= K_{P,\min} + (K_{P,\max} - K_{P,\min}) \chi_k, \\ K_{I,k} &= K_{I,\max} - (K_{I,\max} - K_{I,\min}) \chi_k, \\ K_{D,k} &= K_{D,\min} + (K_{D,\max} - K_{D,\min}) \chi_k. \end{aligned} \quad (72)$$

When the target maneuvers strongly or approaches the image boundary, K_P and K_D increase to improve response speed and reduce oscillation, whereas K_I decreases to limit error accumulation when the Zoom command is prone to saturation [25], [26].

3. *Discrete PID law, derivative filtering, and anti-windup*: The derivative component is filtered by a first-order filter to limit amplification of bounding-box size noise. Let:

$$D_{z,k} = \alpha_D D_{z,k-1} + (1 - \alpha_D) \frac{e_{z,k} - e_{z,k-1}}{T_s}, \quad \alpha_D = \frac{\tau_D}{\tau_D + T_s} \quad (73)$$

The unsaturated Zoom command is determined as:

$$v_{z,k}^{\text{raw}} = K_{P,k} e_{z,k} + I_{z,k} + K_{D,k} D_{z,k} \quad (74)$$

where,

$$I_{z,k} = \text{sat}_{[I_{\min}, I_{\max}]} \left[I_{z,k-1} + K_{I,k} T_s e_{z,k} + K_{aw} T_s (v_{z,k-1}^{\text{sat}} - v_{z,k-1}^{\text{raw}}) \right].$$

The final command sent to the Zoom mechanism is subject to the hard velocity limit:

$$v_{z,k} = v_{z,k}^{\text{sat}} = \text{sat}_{[-V_{z,\max}, V_{z,\max}]} (v_{z,k}^{\text{raw}}) \quad (75)$$

The limits $FOV_{\min} \leq FOV_k \leq FOV_{\max}$ and $|v_{z,k}| \leq V_{z,\max}$ are the physical limits of the Zoom mechanism.

4. *Handling loss of visual measurements*: The control law (68)-(78) is updated only when a valid target bounding box is available. When the detector temporarily provides no measurement, ρ_k is undefined and is not replaced by an artificial size. In this case, the integral component is frozen and the reference FOV is moved to a safe target-reacquisition value

$$\text{if } Z_k = \emptyset, \text{ then } FOV_k^* = FOV_{\text{reacq}},$$

$$e_{z,k}^{\text{reacq}} = FOV_{\text{reacq}} - FOV_k,$$

$$I_{z,k} = I_{z,k-1}, \quad D_{z,k} = 0,$$

$$v_{z,k}^{\text{raw}} = K_{\text{reacq}} e_{z,k}^{\text{reacq}}, \quad v_{z,k} = \text{sat}_{[-V_{z,\max}, V_{z,\max}]} (v_{z,k}^{\text{raw}}).$$

When the bounding box reappears, set $e_{z,k-1} \leftarrow e_{z,k}$ before the first derivative calculation to avoid derivative kick.

Algorithm 2 NMPC Pan/Tilt and Gain-Scheduled PID Zoom Control

1: Input: Current camera state $X_{c,k}$, target prediction XT_{ref} , FOV_k , fx_k , fy_k , current bboxk, NMPC parameters $\mathcal{PN}$, Zoom parameters $\mathcal{PZ}$, previous move ac_{k-1} and PID states $I_{z,k-1}$, $D_{z,k-1}$.
2: Output: Pan/Tilt commands $up_{k,\text{cmd}}$, $ut_{k,\text{cmd}}$ and Zoom command $vz_{k,\text{cmd}}$.

```
// Generate PTZ dynamic reference
3:  $R_k \leftarrow \mathcal{GPTZ}(XT_{\text{ref}})$ 
// ----- NMPC Pan/Tilt -----
4: Set  $X_{c,0|k} \leftarrow X_{c,k}$ 
5: Freeze  $fx_k$ ,  $fy_k$  and  $FOV_k$  over the current prediction horizon
6: for  $i \leftarrow 0$  to  $N-1$  do
7:    $X_{c,i+1|k} \leftarrow F_d(X_{c,i|k}, ac_{i|k})$ 
8:   Compute dynamic tracking errors  $e\alpha_{i|k}$ ,  $e\omega_{i|k}$ ,  $e\theta_{i|k}$ 
9:   Compute LOS errors and predicted pixel position
10:  Compute normalized image error  $eI_{i|k}$ 
11: end for
12: Apply hard actuator and pinhole-validity constraints
13: Apply soft FOV constraints using slack  $\delta i|k \geq 0$ 
14: Set  $ac_{-1|k} \leftarrow ac_{k-1}$  appl
15: Construct stage cost  $\ell_i$  and terminal cost  $V_N$ 
16:  $\{a^*c, X^*c, \delta^*\} \leftarrow \arg \min J_k$ 
// Receding-horizon implementation
17:  $ac_{k,\text{appl}} \leftarrow a^*c_{0|k}$ 
18:  $X^*c_{1|k} \leftarrow F_d(X_{c,k}, ac_{k,\text{appl}})$ 
19:  $[up_{k,\text{cmd}}, ut_{k,\text{cmd}}]T \leftarrow Su X^*c_{1|k}$ 
// ----- Gain-Scheduled PID Zoom -----
20: if bboxk is valid then
21:    $\rho_k \leftarrow wk / W_{\text{img}}$ 
22:    $[\bar{v}_k, \eta_k] \leftarrow \text{TRACKING\_INDICATORS}(\hat{X}T_k, eI_k)$ 
23:    $\rho^*k \leftarrow \text{DESIRED\_TARGET\_SIZE}(\bar{v}_k, \eta_k)$ 
24:    $FOV^*k \leftarrow \text{IMAGE\_SIZE\_TO\_FOV}(\rho^*k, \rho_k, FOV_k)$ 
25:    $\chi_k \leftarrow \text{SCHEDULING\_VARIABLE}(\bar{v}_k, \eta_k)$ 
26:    $[K_{p,k}, K_{i,k}, K_{d,k}] \leftarrow \text{GAIN\_SCHEDULE}(\chi_k)$ 
27:    $ez_k \leftarrow FOV^*k - FOV_k$ 
28:    $Dz_k \leftarrow \text{FILTERED\_DERIVATIVE}(ez_k, Dz_{k-1})$ 
29:    $Iz_k \leftarrow \text{ANTI\_WINDUP\_INTEGRAL}(Iz_{k-1}, ez_k)$ 
30:    $vz_{k,\text{cmd}} \leftarrow \text{SAT}(K_{p,k} ez_k + K_{i,k} Iz_k + K_{d,k} Dz_k)$ 
31: else
32:    $Iz_k \leftarrow Iz_{k-1}$ 
33:    $FOV^*k \leftarrow FOV_{\text{reacq}}$ 
34:    $vz_{k,\text{cmd}} \leftarrow \text{REACQUISITION\_ZOOM}(FOV^*k - FOV_k)$ 
35:   Reset derivative state for the next valid bbox
36: end if
37: Enforce  $FOV_{\min}/FOV_{\max}$  directional Zoom limits
// Apply control and close the optical loop
38: Send  $up_{k,\text{cmd}}$ ,  $ut_{k,\text{cmd}}$  and  $vz_{k,\text{cmd}}$  to the PTZ system
39: Store  $ac_{k,\text{appl}}$ ,  $Iz_k$  and  $Dz_k$  for the next cycle
40: Update  $FOV_{k+1}$ ,  $fx_{k+1}$  and  $fy_{k+1}$  from the new Zoom state
41: return  $up_{k,\text{cmd}}$ ,  $ut_{k,\text{cmd}}$ ,  $vz_{k,\text{cmd}}$ .
```

V. SIMULATION

The objective of the simulation is to evaluate the influence of measurement latency in the closed-loop PTZ target-tracking system and, at the same time, to examine the ability of the retrospective update and state re-propagation (Rollback-Replay) mechanism in the IMM-UKF estimator to compensate for that latency.



Figure 2. Simulation environment

The study builds a simulation environment in which CoppeliaSim models the camera and the complete PTZ mechanism, while the CoppeliaSim scene provides a flexibly moving target. The controller is designed and implemented on an embedded Jetson computer. The simulation accounts for both image-processing latency and mechanical latency of the Pan/Tilt/Zoom axes.

The scenarios are constructed to compare three principal factors: an ideal condition without delay, a delayed condition without delay compensation, and a delayed condition combined with Rollback-Replay.

The target moves along a sinusoidal trajectory, with its velocity linearly increasing from 5 m/s to 10 m/s in the first 20 seconds, and then maintained at 10 m/s. The basic total simulation time is 60 seconds. Measurement latency is simulated by delaying the measurement image data before feeding it into the detection and estimation processing pipeline. In the delayed cases, the nominal latency is set to 40 ms; actual payload latency recorded fluctuates within the range of 40-80 ms. The PTZ controller operates at a frequency of 50 Hz, corresponding to a sampling period of

20 ms. The NMPC controller utilizes a prediction horizon of 20 steps, equivalent to a prediction time of 400 ms. The optimization problem is solved using acados/CasADi with the SQP-RTI configuration. The dynamic model is discretized using the 4th-order Runge-Kutta (RK4) integration method, and the optimal solution is initialized via a warm start mechanism to reduce the solving time in the subsequent cycle.

Four simulation cases are designed as follows:

Case 1: No delay, no rollback, no feedback. This is the ideal case, used to evaluate the target tracking capability of the controller and estimator when completely unaffected by measurement latency.

Case 2: With delay, no rollback, with feedback. This case evaluates the adverse impact of measurement latency when the Rollback and Replay mechanism is not utilized. Feedback from the UKF to the Vision module remains enabled to test whether the target re-locking support mechanism can compensate for the delay-induced deviations.

Case 3: With delay, with rollback, with feedback. This case is used to evaluate the effectiveness of the Rollback and Replay delay compensation mechanism in the presence of measurement latency.

Case 4: Testing the stress-test condition for the Rollback and Replay. This case increases the difficulty level of the problem by setting the target velocity to 50 m/s and increasing the nominal latency to 150 ms. Actual results recorded an average latency of about 160 ms and a maximum of about 200 ms.

Simulation results:

Table 1. Simulation results of the first 3 cases.

Evaluation parameters	Case 1	Case 2	Case 3	Case 3 vs. Case 2	Case 3 vs. Case 1
Time simulation (s)	59.98	26.42	59.98	-	-
Pan RMS (deg)	0.498	1.810	0.474	Reduced by 73.8%	Reduced by 4.8%
Tilt RMS (deg)	0.032	1.659	0.032	Reduced by 98.1%	0.0%
U pixel RMS (px)	25.43	84.97	16.86	Reduced by 80.2%	Reduced by 33.7%
V pixel RMS (px)	1.44	67.40	1.22	Reduced by 98.2%	Reduced by 15.3%
Estimator RMS (px)	2.98	345.73	5.30	Reduced by 98.5%	Increased by 77.9%
Innovation RMS (px)	4.37	283.51	7.81	Reduced by 97.2%	Increased by 78.7%
NIS consistency relative to d	0.12 (6.0% d)	4.91 (245.4% d)	1.02 (51.0% d)	Reduced by 79.2%	Increased by 750.0%
Tilt command RMS (deg/s)	0.118	6.067	0.110	Reduced by 98.2%	Reduced by 6.8%
Tilt acceleration RMS (deg/s ²)	0.851	102.34	0.846	Reduced by 99.2%	Reduced by 0.6%
UKF time mean (ms)	1.70	1.68	2.58	Increased by 53.6%	Increased by 51.8%
NMPC time mean (ms)	1.80	1.96	2.12	Increased by 8.1%	Increased by 17.8%

The results show that the Rollback and Replay mechanism has a decisive impact on the stability of the system under delayed measurement conditions.

In Case 1, when there is no delay, the system runs for the full 59.98 seconds with small tracking errors: Pan RMS reaches 0.498 deg, Tilt RMS reaches 0.032 deg, U-pixel RMS reaches 25.43 px, and V-pixel RMS reaches 1.44 px. This is the ideal reference level of the system when measurements are updated synchronously with the current state. In Case 2, when there is delay but Rollback and Replay is not used, the system becomes unstable after 26.42 seconds. The Pan RMS error increases to 1.810 deg, Tilt RMS increases to 1.659 deg, U-pixel RMS reaches 84.97 px, and V-pixel RMS reaches 67.40 px. Notably, the

projection error of the estimator increases sharply, with Estimator RMS reaching 345.73 px and Innovation RMS reaching 283.51 px. The mean NIS increases to 4.91, indicating that delayed measurements, when updated directly at the current time, cause significant geometric deviations, misleading the filter into interpreting them as strong maneuvers and ultimately leading to estimation instability. In Case 3, when Rollback and Replay and feedback are enabled, the system successfully sustains the full 59.98 seconds of simulation. Compared to case 2, Pan RMS Reduced bys from 1.810 deg to 0.474 deg, representing a 73.8% reduction; Tilt RMS Reduced bys from 1.659 deg to 0.032 deg, a 98.1% reduction. On the image plane, U-pixel RMS Reduced bys from 84.97 px to

16.86 px, while V-pixel RMS Reduced bys from 67.40 px to 1.22 px. Estimator metrics also improve significantly: Estimator RMS drops from 345.73 px to 5.30 px, Innovation RMS drops from 283.51 px to 7.81 px, and the mean NIS Reduced bys from 4.91 to 1.02.

Regarding control signals, case 2 generates large command oscillations, especially on the Tilt axis: Tilt command RMS reaches 6.067 deg/s and Tilt acceleration RMS reaches 102.34 deg/s². When Rollback and Replay is enabled in case 3, Tilt command RMS Reduced bys to 0.110 deg/s and Tilt acceleration RMS Reduced bys to 0.846 deg/s², returning nearly to the levels of case 1. This demonstrates that Rollback and Replay not only improves estimation quality but also significantly reduces aggressive control efforts caused by deviations from delayed measurements. The mean computation time of the IMM-UKF increases from

approximately 1.68–1.70 ms to 2.58 ms when Rollback-and-Replay is enabled. The mean NMPC computation time increases from approximately 1.80–1.96 ms to 2.12 ms. Nevertheless, these values remain much smaller than the 20 ms control period, indicating that the proposed delay-compensation mechanism satisfies the real-time requirement. These results indicate that Rollback and Replay helps synchronize the delayed measurement with the filter state at the exact time the image was captured. Consequently, the estimator avoids updating at the wrong time, reduces abnormal innovations, keeps the covariance stable, and provides a more reliable target state for the NMPC. As a result, the system maintains closed-loop stability and tracking quality that is close to the ideal non-delayed case

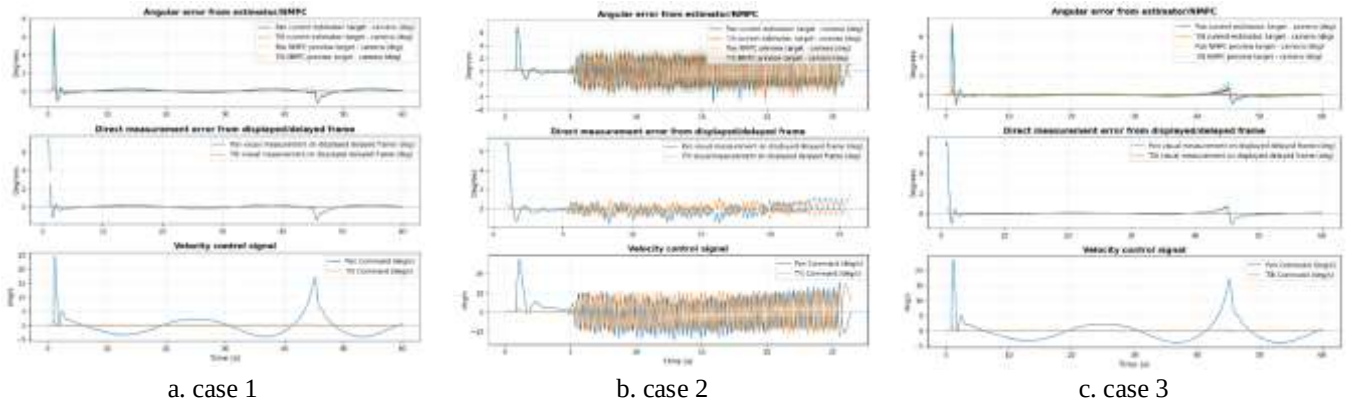


Figure 3. Pan/Tilt angle error and control signal.

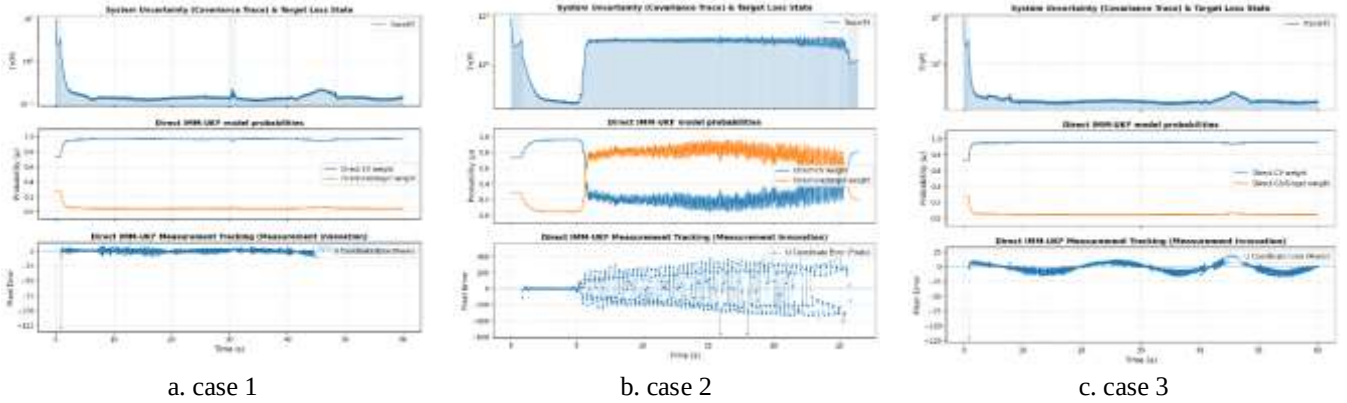


Figure 4. Signal from the IMM-UKF filter

Figure 3 shows the Pan/Tilt angular errors and control signals across the three cases 1, 2, and 3. In case 1, angular errors are small and control signals are relatively smooth, reflecting ideal zero-delay operating conditions. In case 2, after measurement latency appears, the error increases rapidly and the control signal becomes highly oscillatory. In particular, the Tilt axis exhibits large oscillations, which aligns with the sharply increased Tilt RMS and Tilt command RMS values in the data table. This indicates that the feedback from the UKF to the Vision module is insufficient to compensate for the deviation caused by updating the delayed measurement at the wrong time. In case 3, errors and control signals are restored close to those in case 1. The Pan/Tilt errors are kept within a significantly smaller region compared to case 2, and the control signals exhibit fewer oscillations. This confirms that Rollback and

Replay reduces the phase lag between the measurement and the current state, helping the NMPC receive a more consistent target state.

Figure 4 illustrates the signals related to the IMM-UKF filter. In case 2, NIS and the estimator's projection error increase sharply, showing that the filter is no longer consistent when measurement latency is introduced into the update at the current time. This increase also heavily skews the CT-Singer model probability upward, reflecting the filter's misinterpretation of latency as strong target maneuvering. In case 3, estimation errors, measurement errors, and NIS reduced markedly compared to case 2. Model probabilities are also more stable and are not biased toward the maneuvering model. This shows that Rollback and Replay restores the temporal consistency of

measurements and helps the IMM-UKF maintain a stable estimated state.

Results of case 4.

Case 4 is used to evaluate the stress-test condition for the Rollback and Replay mechanism under conditions where the target moves faster and the latency is larger. With a target velocity of 50 m/s and an average latency of about 160 ms, the system still successfully maintains the simulation for the full 120 seconds. This demonstrates that Rollback and Replay is capable of keeping the closed-loop system stable even under large latency conditions.

Table 2. Simulation results for Case 4.

Evaluation parameters	Value
Time simulation	120 s
Payload latency mean	0.1601 s
Payload latency max	0.2000 s
Pan current RMS	1.5156 deg
Tilt current RMS	0.0547 deg
U-pixel RMS	190.34 px
V-pixel RMS	8.76 px
Estimator projection RMS	150.32 px
Innovation RMS	37.37 px
NIS consistency relative to d	1.4038
Pan command RMS	19.54 deg/s
Tilt command RMS	2.86 deg/s
IMM-UKF mean time	3.72 ms
NMPC mean time	2.62 ms

Tracking performance degrades significantly compared with Case 3. The Pan current RMS increases to 1.5156 deg, the U-pixel RMS reaches 190.34 px, the estimator projection RMS increases to 150.32 px, and the innovation RMS reaches 37.37 px.

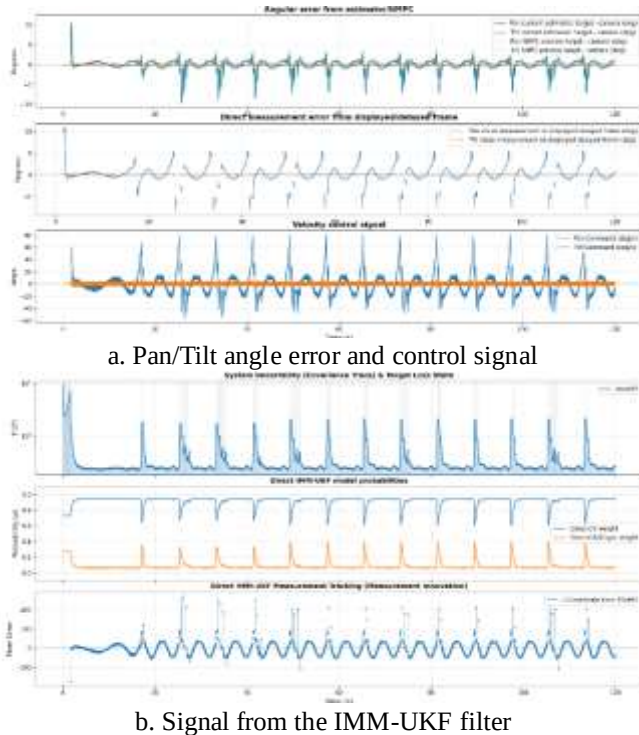


Figure 5. Case 4

At the same time, the control commands become more aggressive, with the Pan command RMS reaching 19.54 deg/s. These results indicate that, when the latency increases to approximately 160–200 ms, the Rollback-and-Replay

mechanism can still prevent closed-loop divergence; however, the replay accuracy becomes strongly dependent on the target motion model. Under fast-maneuvering target conditions, the prediction error from the past to the present increases, degrading tracking quality and forcing the NMPC to generate larger control commands. The simulation results show that measurement latency has a massive impact on the stability of the PTZ camera target tracking system. When there is latency but Rollback and Replay is not used, the measurement is updated at the wrong time, sharply increasing innovation, NIS, and the estimator's projection error, ultimately leading to closed-loop divergence. When Rollback and Replay is enabled, the system resynchronizes the measurement according to its timestamp, maintaining estimation stability and bringing tracking quality close to the ideal zero-delay case. However, under high-velocity conditions and large latencies of around 160–200 ms, Rollback and Replay only prevents the system from diverging, while tracking quality visibly degrades because replay error strongly depends on the target prediction model. Therefore, Rollback and Replay is a necessary mechanism to compensate for latency in real-time PTZ systems, but it must be combined with adaptive adjustments according to latency and the target's maneuvering level to improve accuracy under extreme conditions.

Comparative and Ablation Evaluation

To further isolate the contribution of the proposed architecture, two controlled comparisons were conducted under the same target-motion and PTZ dynamic conditions. First, the proposed NMPC was compared with a bounded PD controller to evaluate the benefit of predictive constrained control. Second, the estimator was evaluated using four configurations combining Rollback–Replay/no-Rollback processing with adaptive/fixed Markov transition matrices. These comparisons were designed as mechanism-level ablation tests rather than as a new simulation scenario.

Table 3. Comparison results between the PD and NMPC

Comparison	Metric	Bounded-PD	NMPC	%
Controller	Pan RMS	0.492°	0.394°	19.9%
Controller	Tilt RMS	0.280°	0.203°	27.5%
Controller	U-pixel RMS	5.32 px	4.27 px	19.7%
Controller	V-pixel RMS	3.03 px	2.19 px	27.7%
OOSM, 1.0-s delay	Pan RMS	No-RB: 2.068°	RB: 1.429°	30.9%
OOSM, 1.0-s delay	Tilt RMS	No-RB: 1.127°	RB: 0.737°	34.6%
Adaptive Markov, 1.0-s delay	Pan RMS	Fixed: 1.545°	Adaptive: 1.429°	7.5%
Adaptive Markov, 1.0-s delay	Tilt RMS	Fixed: 0.934°	Adaptive: 0.737°	21.1%

The controller comparison shows that NMPC consistently outperforms the bounded PD baseline under the same closed-loop conditions. The Pan and Tilt RMS errors are reduced by approximately 19.9% and 27.5%, respectively,

while the U- and V-pixel RMS errors decrease by approximately 19.7% and 27.7%. This improvement is attributed to the predictive use of the actuator dynamics, nonlinear LOS-to-pixel geometry, and explicit state/input/FOV constraints in the NMPC formulation. The result therefore indicates that the control benefit does not arise solely from the estimator.

The estimator ablation confirms that the Rollback–Replay mechanism provides an additional benefit beyond the adaptive IMM itself. At a 1.0-s measurement delay, Rollback–Replay with Adaptive Markov reduces the Pan and Tilt RMS errors from 2.068° and 1.127° to 1.429° and 0.737° , corresponding to reductions of approximately 30.9% and 34.6%. Because the adaptive Markov configuration is kept unchanged in this comparison, the difference can be attributed primarily to timestamp-consistent retrospective reconstruction. This supports the role of Rollback–Replay as a delay-compensation mechanism rather than merely an implementation detail.

A further comparison was made between adaptive and fixed Markov transition matrices while keeping Rollback–Replay enabled. At a 1.0-s delay, the adaptive configuration further reduces the Pan RMS by about 7.5% and the Tilt RMS by about 21.1% compared with the fixed-transition baseline. This indicates that the proposed NIS-based Markov adaptation contributes independently to the tracking performance and is not merely a secondary consequence of Rollback–Replay.

The present study focuses on mechanism-level baselines within the same nonlinear PTZ formulation; direct comparison with standard Bar-Shalom one-step or multi-step OOSM implementations is left for future work.

REFERENCES

- [1] S. Hutchinson, G. D. Hager, and P. I. Corke, “A tutorial on visual servo control,” *IEEE Transactions on Robotics and Automation*, vol. 12, no. 5, pp. 651–670, 1996.
- [2] F. Chaumette and S. Hutchinson, “Visual servo control. I. Basic approaches,” *IEEE Robotics & Automation Magazine*, vol. 13, no. 4, pp. 82–90, 2006.
- [3] F. Chaumette and S. Hutchinson, “Visual servo control. II. Advanced approaches [Tutorial],” *IEEE Robotics & Automation Magazine*, vol. 14, no. 1, pp. 109–118, 2007.
- [4] R. Hartley and A. Zisserman, *Multiple View Geometry in Computer Vision*. Cambridge, U.K.: Cambridge University Press, 2003.
- [5] Z. Zhang, “A flexible new technique for camera calibration,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, no. 11, pp. 1330–1334, 2000.
- [6] S. Wang, D. Liu, Y. Yang, X. Chen, P. Cao, L. Li, L. Yan, Z. Cheng, and Y. Shen, “Distortion correction in surface defects evaluating system of large fine optics,” *Optics Communications*, vol. 312, pp. 260–268, 2014.
- [7] Y. Zhang et al., “ByteTrack: Multi-object tracking by associating every detection box,” in *Proceedings of the European Conference on Computer Vision*, Cham: Springer Nature Switzerland, 2022.
- [8] N. Aharon, R. Orfaig, and B.-Z. Bobrovsky, “BoT-SORT: Robust associations multi-pedestrian tracking,” *arXiv preprint arXiv:2206.14651*, 2022.
- [9] D. Reis et al., “Real-time flying object detection with YOLOv8,” *arXiv preprint arXiv:2305.09972*, 2023.
- [10] Y. Bar-Shalom, X. R. Li, and T. Kirubarajan, *Estimation with Applications to Tracking and Navigation: Theory, Algorithms and Software*. New York, NY, USA: John Wiley & Sons, 2001.
- [11] H. A. P. Blom and Y. Bar-Shalom, “The interacting multiple model algorithm for systems with Markovian switching coefficients,” *IEEE Transactions on Automatic Control*, vol. 33, no. 8, pp. 780–783, Aug. 1988.
- [12] R. A. Singer, “Estimating optimal tracking filter performance for manned maneuvering targets,” *IEEE Transactions on Aerospace and Electronic Systems*, vol. AES-6, no. 4, pp. 473–483, July 1970.
- [13] E. Mazor, A. Averbuch, Y. Bar-Shalom, and J. Dayan, “Interacting multiple model methods in target tracking: A survey,” *IEEE Transactions on Aerospace and Electronic Systems*, vol. 34, no. 1, pp. 103–123, Jan. 1998.
- [14] X. R. Li and V. P. Jilkov, “Survey of maneuvering target tracking. Part I: Dynamic models,” *IEEE Transactions on Aerospace and Electronic Systems*, vol. 39, no. 4, pp. 1333–1364, Oct. 2003.
- [15] S. J. Julier and J. K. Uhlmann, “Unscented filtering and nonlinear estimation,” *Proceedings of the IEEE*, vol. 92, no. 3, pp. 401–422, Mar. 2004.
- [16] E. A. Wan and R. Van Der Merwe, “The unscented Kalman filter for nonlinear estimation,” in *Proceedings of the IEEE 2000 Adaptive Systems for Signal Processing, Communications, and Control Symposium*, Lake Louise, AB, Canada, 2000, pp. 153–158.
- [17] Y. Bar-Shalom, “Update with out-of-sequence measurements in tracking: Exact solution,” *IEEE Transactions on Aerospace and Electronic Systems*, vol. 38, no. 3, pp. 769–777, 2002.
- [18] Y. Bar-Shalom, H. Chen, and M. Mallick, “One-step solution for the multistep out-of-sequence-measurement problem in tracking,” *IEEE Transactions on Aerospace and Electronic Systems*, vol. 40, no. 1, pp. 27–37, 2004.
- [19] G. Galuppini, I. Schimperna, and L. Magni, “Offset-free output feedback tracking MPC for δ -ISS nonlinear systems subject to input and input rate constraints,” *IEEE Transactions on Automatic Control*, 2025.
- [20] I. A. Jimoh and H. Yue, “A velocity form model predictive control of an autonomous underwater vehicle,” *IEEE Journal of Oceanic Engineering*, 2025.
- [21] Y. Du et al., “A visual-based target tracking framework for UAV using model predictive contouring control,” *IEEE Transactions on Industrial Electronics*, 2025.
- [22] R. Verschuere et al., “acados—A modular open-source framework for fast embedded optimal control,” *Mathematical Programming Computation*, vol. 14, no. 1, pp. 147–183, 2022.
- [23] J. Andersson et al., “CasADi—A software framework for nonlinear optimization and optimal control,” *Mathematical Programming Computation*, vol. 11, no. 1, pp. 1–36, 2019.
- [24] A. Amer et al., “Visual tracking nonlinear model predictive control method for autonomous wind turbine inspection,” in *Proceedings of the 21st International Conference on Advanced Robotics*, IEEE, 2023.
- [25] W. Zhan et al., “Target size adaptive PID control method for gimbal visual servo,” in *Proceedings of the 17th International Conference on Intelligent Human-Machine Systems and Cybernetics*, IEEE, 2025.
- [26] P. Shi et al., “Intelligent vehicle path tracking control based on improved MPC and hybrid PID,” *IEEE Access*, vol. 10, pp. 94133–94144, 2022.
- [27] E. Rohmer, S. P. N. Singh, and M. Freese, “V-REP: A versatile and scalable robot simulation framework,” in *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2013.

Le Van Khanh M.Sc. is with Hanoi, Vietnam (e-mail: levankhanhth@gmail.com, <https://orcid.org/0000-0002-6376-1769>)

Luong Thu Phong Ph.D.Eng. is with Hanoi, Vietnam (e-mail: luongthuphongmpei@gmail.com)