Cross-drone Binocular Coordination for Ground Moving Target Tracking in Occlusion-rich Scenarios

Yuan Chang, Han Zhou, Xiangke Wang, *Senior Member, IEEE*, Lincheng Shen, *Member, IEEE*, and Tianjiang Hu*, *Member, IEEE*

***Abstract*—How to work effectively under occlusion-rich environments remains a challenge for airborne vision-based ground target tracking, due to the natural limitation of monocular vision. Given this, a novel cross-drone binocular coordination approach, inspired by the efficient coordination of human eyes, is proposed and developed. The idea, derived from neural models of the human visual system, is to utilize distributed target measurements to overcome occlusion effects. Eventually, a binocular coordination controller is developed. It enables two distributed pan-tilt cameras to execute synergistic movements similar to human eyes. The proposed approach is able to work based on binocular or monocular vision, and hence it is practically appropriate for various environments. Both testbed experiments and field experiments are conducted for performance evaluation. Testbed experiments highlight its advantages over independent tracking in terms of accuracy while being robust to a partial perception ratio of up to 43%. Field experiments with a pair of drones further demonstrate its effectiveness in the real-world scenarios.**

***Index Terms*—Aerial Systems: Perception and Autonomy, Multi-Robot Systems, Visual Tracking, Cross-drone Binocular Vision.**

I. Introduction

SERVING as a real-time, low-cost and effective tool to monitor a large-scale area, unmanned aerial vehicles (UAVs) have been widely used in precision agriculture [1], disaster monitoring [2] and wildlife protection [3]–[5]. Ground target tracking, as a typical application of UAVs, enables continuous observation on a valuable target, such as a car [6], a human [7], a boat [8] or a wildlife [9]. However, it is challenging to conduct tracking tasks within occlusion-rich scenarios, e.g. urban areas, where massive buildings may lead to frequent and long-term target loss. To solve this, this paper proposes an aerial multi-view scheme for complementary perspectives, as shown in Fig. 1. Multiple drones are considered to form a collective swarm for such scenarios.

The challenges of target tracking in occlusion-rich scenarios mainly cover what to do when it is not feasible to achieve target-involved images. This leads to advances in fields like target detection, state estimation, and control strategies. State-of-the-art occlusion handling solutions in the target detection field are mainly concerned with partial occlusions [10]–[12]. However, there are barely effective methods for full occlusions. State estimation is an important complement to target tracking, but cannot guarantee the target remaining in the field of view of the camera [7]. In contrast, in this paper, we focus on the control strategy of the distributed active cameras, which plays a key role in the overall target tracking tasks.

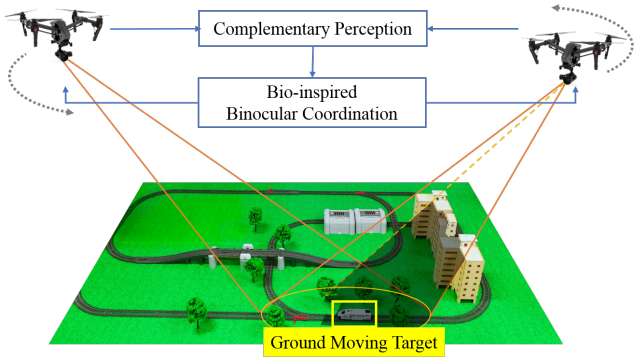

Fig. 1. The proposed aerial binocular system tracks the ground moving target in the context of partial occlusion. The drones share concurrent and complementary target measurements and generate commands for active cameras through binocular coordination inspired by human eye movements.

Airborne target tracking is typically divided into monocular, binocular and multi-vision modes. As for the monocular mode, either "zero-input" or "hold-input" control strategy is adopted as a common supplement [13]. The "zero-input" strategy, in which the camera rotation command is set to zero if the target is lost, and the "hold-input" strategy, in which the last control input is maintained once the target is missing, are only valid for a short period of time.

As for the multi-vision mode, the vehicles share self-states and target measurements through wireless communication. Their perspectives complement each other in time and space, thus reducing the probability of target loss. Despite the numerous studies of coordinated target tracking [14]–[22], only a few have considered the camera dynamics under swarm conditions [19]–[22], and even fewer have considered the impacts of occlusions, or out-of-frame events [22]. A typical solution to occlusion handling is to share target location within swarm nodes, which is estimated by monocular localization. However, visual-based monocular localization is so far a challenging

Manuscript received September 10, 2019; revised December 19, 2019; accepted February 1, 2020.

This paper was recommended for publication by Editor Jonathan Roberts upon evaluation of the Associate Editor and Reviewers' comments. This work was supported by National Natural Science Foundation of China under Grant 61973327. *(Corresponding author: Tianjiang Hu.)*

Yuan Chang, Han Zhou, Xiangke Wang and Lincheng Shen are all with College of Intelligence Science and Technology, National University of Defense Technology, Changsha 410073, China. Emails: {changyuan10, zhouhan, xkwang, lcshen}@nudt.edu.cn

Tianjiang Hu is with Sun Yat-sen University, Guangzhou 510725, China. Email: hutj3@mail.sysu.edu.cn

Digital Object Identifier (DOI): see top of this page.

issue, which limits its feasibility.

The binocular mode is the meta paradigm and a special case of multi-vision mode. In addition to the above-mentioned methods, bio-inspired methods have received a lot of attention. The human visual system provides a perfect model for binocular systems for its precise and robust target tracking performance. Human eyes always move in sync and focus on the same target. When one eye is covered, it still moves in accordance with the other eye, so that the target can be recaptured in time once the occlusion is removed. This feature is especially inspiring for occlusion-rich situations where the target is regularly covered by obstructions. As to the neural mechanisms of binocular coordination, the dominant hypothesis is proposed by Hering [23]. *Hering's Law* indicates that there are separate neural controllers for conjugate and vergence and that each eye receives an identical neural command from each controller [23]. For conjugate movements, the two eyes rotate in the same direction through the same angle. For vergence movements, the two eyes rotate conversely but also through the same angle. Since then, numerous neurophysiological models have been proposed for human visual systems. Zee et al. [24] developed a saccade-related vergence burst neurons (SVBNs) model to formalize premotor mechanisms of combined saccade-vergence eye movements. They also introduced the inhibition function. Another saccade generation model was presented by King and Zhou [25]. They suggest there are separate burst neurons for each eye, while the motoneurons receive binocular inputs. Recent developments of the binocular models are mostly derived from Zee and King's models [26]. Meanwhile, various robot eyes were also developed to imitate human eye movements [27], [28].

The above-mentioned vision modes provide inspiration and design guidelines for an efficient target tracking system. In this paper, we develop a cross-drone binocular coordination approach embodying a pair of flying vehicles (see Fig. 1). The impetus behind our work is to adopt the adaptive features of human eyes to promote the tracking robustness against occlusions. It is derived from the traditional binocular system, but in particular, the relative position and posture of the two cameras are varied. Therefore, we decouple the 3-D target tracking problem using humanoid representation. Then we design a binocular coordination controller based on the SVBNs model [24]. In case of occlusions, the tracking system adaptively chooses the dominate vision to generate outputs for both cameras. The camera that detects the target acts as the dominant eye, and the other camera moves in consistence with it in a conjugated manner. In this manner, the target still remains within the field of view of the occluded camera once the occlusion has passed, such that the tracking robustness is enhanced.

The main contributions of this paper are summarized as follows. First, a novel binocular coordination model is proposed and developed for cooperative target tracking, with inspirations from the human visual system. Then, the cross-drone target tracking framework is generalized for various scenarios with different occlusion levels, attached with an analysis of stability and convergence rate. Finally, the tracking performance is evaluated through a batch of testbed experiments and field experiments. It is verified that the proposed binocular target tracking approach is more robust in occlusion-rich areas compared with independent tracking. It also outstands in terms of tracking accuracy while being localization-free, which is essential for practical applications.

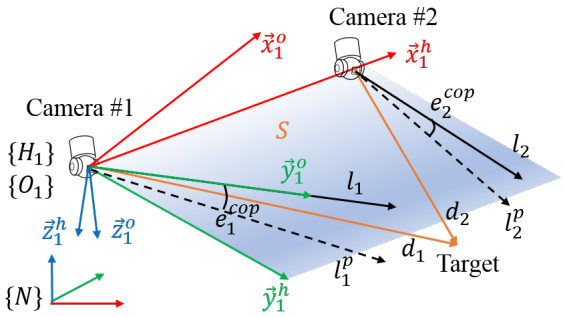

Fig. 2. Binocular target tracking geometry. The observation plane (blue part) coincides with the X-Y plane of the head frame. The dashed vectors represent the projections of the lines of sight on the observation plane.

II. Cross-drone Binocular Coordination

A. Problem Formulation

Consider the target tracking system with a pair of drones, a subscript $i \in \{1, 2\}$ is used in the following text to distinguish between them (see Fig. 2). The position of the flying vehicle $p_i \in \mathbb{R}^3$ is expressed in the local ENU frame $\{N\}$. It is assumed that a pan-tilt gimbal is installed at the center of the drone. The gimbal posture $x_i^{ptu} = [\theta_i, \psi_i]^T$ is also attached to $\{N\}$. Both vehicle position and gimbal posture can be measured by onboard sensors. Therefore, the current line of sight vector of the camera $l_i \in \mathbb{R}^3$ can be derived as

$$l_i = \begin{bmatrix} cos\theta_i sin\psi_i \\ cos\theta_i cos\psi_i \\ sin\theta_i \end{bmatrix}. \tag{1}$$

Each drone is equipped with an onboard processor to independently detect the target. Particularly, YOLOv3 network [29] is adopted and modified in our system for fast speed and high precision. Once the pre-defined target is detected, a pinhole model is then established between the target position $p_{tar} \in \mathbb{R}^3$ and the pixel coordinates $m_i = [u_i, v_i]^T$. Although p_{tar} cannot be directly obtained by monocular vision, the direction vector of the target relative to the camera $d_i \in \mathbb{R}^3$ is derived as

$$d_i = \frac{p_{tar} - p_i}{\|p_{tar} - p_i\|} = \frac{{}^N_C R_i}{\sqrt{(\Delta u_i)^2 + (\Delta v_i)^2 + f^2}} \begin{bmatrix} \Delta u_i \\ \Delta v_i \\ f \end{bmatrix}, \tag{2}$$

where f is the camera focal length. $[\Delta u_i, \Delta v_i]^T$ is the pixel difference between m_i and the image center $[u_0, v_0]^T$. ${}^N_C R_i \in \mathbb{R}^3$ represents the rotation matrix from the camera frame $\{C_i\}$ to the local frame $\{N\}$, which can be obtained from the pan-tilt angles of the camera.

Our objective is to design a control law for the angular velocity of the camera $\omega_i = \left[\dot{\theta}_i, \dot{\psi}_i\right]^T$, such that the line of sight l_i is directed to d_i. The vehicle position p_i, gimbal

posture x_i^{ptu} and target measurements m_i of both drones are shared through wireless communication.

B. Fusion Sensing and Intersection Controller

As human eyes always focus on the same point, an observation plane S is potentially constructed by binocular lines of sight. In this paper, S is defined by two drones and the target, as illustrated in Fig. 2. The three-point collinearity of plane S can be avoided by setting the same flight altitude for two drones. For clarity, we define two frames, the head-fixed frame $\{H_i\}$ and the orbit-fixed frame $\{O_i\}$, on each vehicle. For $\{H_i\}$, $\vec{x}_i^h$ points to the other drone from itself, $\vec{y}_i^h$ lies on the observation plane, perpendicular to the x-axis and points to the target side. $\vec{z}_i^h$ is determined by the right-hand rule. $\{O_i\}$ is obtained by rotating $\{H_i\}$ with its Y-axis $\vec{y}_i^o$ coinciding with $\vec{l}_i$. Such a definition does not distinguish between the two vehicles, which is very necessary for a distributed algorithm.

Remark 1. Without loss of generality, inconsistencies in the z-axis direction of $\{H_1\}$ and $\{H_2\}$ can be eliminated by coordinate transformation. For potential swarm applications, the swarm can be decomposed into a couple of binocular systems, and the outputs of different binocular systems can be synthesized correspondingly.

Thus, the normal vector of S is available for a node if the target is successfully detected, which is expressed by

$$n_i = norm(\vec{x}_i^h \times d_i). \tag{3}$$

Considering missed detections and false detections, a complementary filter is adopted for higher precision. Its discrete form at the time step k is

$$\tilde{n}_i(k) = \frac{n_i(k-1) + (-1)^{i+1}\sum_{j=1}^{2}(-1)^{j+1}\delta_j(k)n_j(k)}{1+\sum_{j=1}^{2}\delta_j(k)}, \tag{4}$$

where the boolean $\delta_i(k)$ indicates whether the target is successfully detected. $\delta_i(k) = 1$ if the target is detected and confirmed by motion continuity. Note that the denominator of (4) is nonzero. The first term of its numerator is used for smoothing, while the second term integrates the measurements of two distributed drones.

The lines of sight of the cameras need to remain within plane S to intersect on the target. The intersection error is defined by the deviation between the line of sight and its projection $l_i^p(k) \in \mathbb{R}^3$ on S. $l_i^p(k)$ is given by

$$l_i^p(k) = l_i(k) - \tilde{n}_i(k)\frac{l_i(k)\cdot\tilde{n}_i(k)}{\tilde{n}_i(k)\cdot\tilde{n}_i(k)}. \tag{5}$$

Thus, the intersection error $e_i^{cop}(k)$ can be obtained by

$$e_i^{cop}(k) = sgn((l_i(k)\times l_i^p(k))\cdot\vec{x}_i^h)acos(\frac{l_i(k)\cdot l_i^p(k)}{|l_i(k)|\cdot|l_i^p(k)|}). \tag{6}$$

In this paper, an intersection controller is applied as

$$\omega_i^{cop}(k) = C_{cop}e_i^{cop}(k), \tag{7}$$

where C_{cop} is a positive coefficient. $\omega_i^{cop}(k)$ is the desired angular velocity of the camera expressed in the frame $\{O_i\}$.

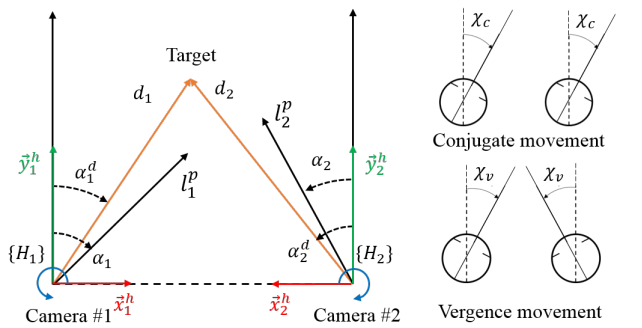

Fig. 3. Target tracking geometry within the observation plane. The conjugate and vergence movements are also illustrated.

C. Binocular Coordination Controller

When the intersection constraint is satisfied, only the motion within the observation plane needs to be considered. According to *Hering's Law*, eye movements can be represented as a linear summation of conjugate and vergence [23]. As depicted in Fig. 3, in this paper, the conjugate angle $\chi_c(k)$ and vergence angle $\chi_v(k)$ are defined as

$$\begin{bmatrix}\chi_c(k)\\ \chi_v(k)\end{bmatrix} = \frac{1}{2}\begin{bmatrix}1 & -1\\ 1 & 1\end{bmatrix}\begin{bmatrix}\alpha_1(k)\\ \alpha_2(k)\end{bmatrix}. \tag{8}$$

Similarly, the desired conjugate angle $\chi_c^d(k)$ and desired vergence angle $\chi_v^d(k)$ are given by

$$\begin{bmatrix}\chi_c^d(k)\\ \chi_v^d(k)\end{bmatrix} = \frac{1}{2}\begin{bmatrix}1 & -1\\ 1 & 1\end{bmatrix}\begin{bmatrix}\alpha_1^d(k)\\ \alpha_2^d(k)\end{bmatrix}. \tag{9}$$

The SVBNs model [24] describes the oculomotor path of humanoid target tracking through binocular coordination, as shown in Fig. 4(a). It suggests that the vergence channel and conjugate channel are coupled. In this paper, a bilateral core (Fig. 4(b)) is derived from the SVBNs model. Some assumptions are adopted in the derivation of the control model as follows: 1) The transfer functions of the human visual system are considered linear [30]; 2) The omni-directional pause neurons (OPN) are neglected because the physiological characteristics of human eyes are different from the aerial target tracking system.

Based on the bilateral core, the desired angular velocity expressed in the orbit-fixed frame is obtained by

$$\begin{bmatrix}\omega_1^{bio}(k)\\ \omega_2^{bio}(k)\end{bmatrix} = \begin{bmatrix}C_c+C_n & C_v\\ C_n-C_c & C_v\end{bmatrix}\begin{bmatrix}\chi_c^d(k)-\chi_c(k)\\ \chi_v^d(k)-\chi_v(k)\end{bmatrix}, \tag{10}$$

where C_c and C_v are control gains for conjugate and vergence movements, respectively, and $C_c > C_v > 0$. The coefficient $C_n \geq 0$ controls the coupling portion.

D. Outputs Calculation

The desired angular velocities within the observation plane and perpendicular to the observation plane have been obtained with (7) and (10). After then, we take the vectorial sum of the two terms and calculate the desired outputs through coordinate transformations.

$$\omega_i = \begin{bmatrix}1 & 0 & 0\\ 0 & 0 & 1\end{bmatrix}{}_H^N R_i {}_O^H R_i \begin{bmatrix}\omega_i^{cop}(k)\\ 0\\ \omega_i^{bio}(k)\end{bmatrix}, \tag{11}$$

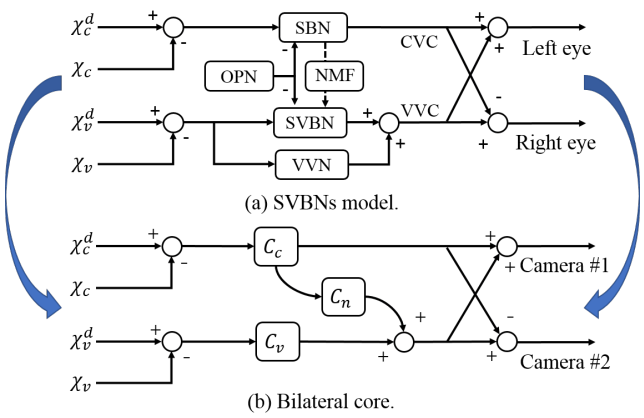

Fig. 4. Derivation of the bilateral core from the saccade-related vergence burst neurons (SVBNs) model [24] through migration and simplification. (a) In the SVBNs model, the saccadic burst neurons (SBN) are used for pure saccades while vergence velocity neurons (VVN) are used for pure vergence. (b) In the bilateral core, the omni-directional pause neurons (OPN) are neglected, the rest nuclei are combined and linearized. Note that the relation between SBN and SVBN is described by a nonlinear modulation function (NMF), while both of them are gated by OPN.

where ${}^H_O R_i \in \mathbb{R}^3$ denotes the rotation matrix from the orbit-fixed frame $\{O_i\}$ to the head-fixed frame $\{H_i\}$, and ${}^N_H R_i \in \mathbb{R}^3$ is the rotation matrix from $\{H_i\}$ to the local frame $\{N\}$.

III. Biological Inspirations and Advantages

A. Analysis of Stability and Convergence Rate

First, we prove that the proposed controller is capable of guaranteeing stable tracking on the target.

Theorem 1. Consider the binocular tracking problem formulated in subsection II-A, where the drones and the target are stationary and the tracking movement is limited within the observation plane, the proposed control law (10) with coefficients $C_c > 0$, $C_v > 0$ and $C_n \geq 0$ ensures that the tracking error defined by

$$e = \begin{bmatrix} e_1 \\ e_2 \end{bmatrix} = \begin{bmatrix} \alpha_1^d - \alpha_1 \\ \alpha_2^d - \alpha_2 \end{bmatrix} \tag{12}$$

asymptotically converges to 0.

Proof. By taking time derivative of (12) and combining with (8,9,10), the error dynamics is derived as

$$\dot{e} = Ae, \tag{13}$$

where

$$A = \frac{1}{2}\begin{bmatrix} -C_n - C_c - C_v & C_n + C_c - C_v \\ -C_n + C_c - C_v & C_n - C_c - C_v \end{bmatrix}. \tag{14}$$

Consider the following Lyapunov function candidate.

$$V(e) = e^T P e, \tag{15}$$

where P is a pending symmetric matrix. By differentiating (15) with respect to time and combining with (14), the following equation can be obtained.

$$\dot{V}(e) = e^T (A^T P + PA) e. \tag{16}$$

For any given symmetric positive definite matrix Q, if there exists a symmetric positive definite matrix P, so that $A^T P + PA = -Q$, then the linear system described by (13) is asymptotically stable.

Let $Q = I$, we have

$$A^T P + PA = -I, \tag{17}$$

where

$$P = \begin{bmatrix} p_{11} & p_{12} \\ p_{21} & p_{22} \end{bmatrix}, A = \frac{1}{2}\begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix}. \tag{18}$$

Note that $p_{12} = p_{21}$ since P is a symmetric matrix.

By solving linear equations with (14), (17) and (18), it is derived that

$$\begin{cases} p_{11} = -(a_{21}^2 + a_{22}^2 + a_{11}a_{22} - a_{12}a_{21})/den \\ p_{12} = p_{21} = (a_{11}a_{21} + a_{12}a_{22})/den \\ p_{22} = -(a_{11}^2 + a_{12}^2 + a_{11}a_{22} - a_{12}a_{21})/den \end{cases}, \tag{19}$$

where the denominator $den = (a_{11}a_{22} - a_{12}a_{21})(a_{11} + a_{22}) = -8C_cC_v(C_c + C_v)$. Since $C_c > 0$ and $C_v > 0$, the denominator is non-zero.

Then, by variable substitution, we have the determinant

$$|p_{11}| = \frac{a_{21}^2 + a_{22}^2 + 4C_cC_v}{8C_cC_v(C_c + C_v)} > 0, \tag{20}$$

and

$$\begin{vmatrix} p_{11} & p_{12} \\ p_{21} & p_{22} \end{vmatrix} = \frac{C_n^2 + C_c^2 + C_v^2 + 2C_cC_v}{4C_cC_v(C_c + C_v)^2} > 0. \tag{21}$$

According to Sylvester's criterion, P is a positive definite matrix. Thus, the stability of the closed-loop system can be guaranteed.

Remark 2. The coupling term C_n improves the convergence rate of $V(e)$, which is defined as

$$\eta := \left\{ -\frac{\dot{V}(e)}{V(e)}, e \neq 0 \right\}. \tag{22}$$

By combining (14)-(19) and (22), we have $\eta \leq 1/\mu_{min}$, where μ_{min} is the minimum eigenvalues of P. It is derived that $\mu_{min} \in R+$ is a decreasing function of C_n. Thus, η increases with the coupling term C_n. It indicates that a positive C_n improves the dynamic response performance of the system.

B. Dealing with Occlusions

In case of occlusions, the binocular controller can be further extended to deal with partial perception cases. According to the binocular rivalry principle, the human visual system adaptively selects the dominate vision based on high-level cognition [31]. The occluded eye follows the dominate eye and they move in sync. Therefore, instead of (9), a pure conjugate movement is introduced to by freezing vergence angle as

$$\tilde{\chi}_v^d(k) = \chi_v^d(k-1), \tag{23}$$

where $\tilde{\chi}_v^d(k)$ is an estimate of the desired vergence angle, which is the input of (10).

Theorem 2. Consider the binocular tracking problem with a constant flying altitude h and a fixed baseline l, which is the distance between two drones, the proposed occlusion handling strategy (23) ensures the target tracked if

$$\frac{l}{2h} < tan\frac{\delta}{2}, \tag{24}$$

where δ is the camera's field of view.

Proof. Assume that the target is moving on a flat ground, according to (8), the relationship between χ_c and χ_v during target movement is derived as

$$tan(\chi_v + \chi_c) + tan(\chi_v - \chi_c) = \frac{l}{h}. \tag{25}$$

Note that the following inequalities must be satisfied due to the intersection constraints.

$$\begin{cases} -\pi/2 < \chi_c < \pi/2 \\ \chi_v - \chi_c < \pi/2 \\ \chi_v + \chi_c < \pi/2 \end{cases}. \tag{26}$$

By applying (23), $\tilde{\chi}_v$ is a constant as the target moves. Then, the target can be successfully tracked if the deviation between $\tilde{\chi}_v$ and χ_v is less than $\delta/2$ during the pure conjugate movement. It can be proved that χ_v is minimized at $\chi_c = \pm\pi/2$ and maximized at $\chi_c = 0$. Therefore, by solving (25) with $\chi_c = \pm\pi/2$ and $\chi_c = 0$, respectively, (24) is derived.

Remark 3. Note that by applying (23), there is deviation between $\tilde{\chi}_v^d(k)$ and $\chi_v^d(k)$. This is inevitable since only monocular vision is available. Meanwhile, it provides important information about the trend of target movements. This makes it possible for the covered camera to recapture the target as soon as it emerges from the occlusions.

If the target is invisible for both eyes, a hold-input strategy is applied as

$$\begin{cases} \tilde{\chi}_c^d(k) = \chi_c^d(k-1) \\ \tilde{\chi}_v^d(k) = \chi_v^d(k-1) \end{cases}. \tag{27}$$

The hold-input strategy sometimes fails, for instance, when there is a long-time target occlusion, where three or more cameras are required to participate in the target tracking. Above all, the proposed approach is summarized as follow.

Algorithm 1 Aerial Binocular Coordination (ABC)

Input:
-The position of each drone p_i
-The posture of each pan-tilt unit x_i^{ptu}
-The image captured by each camera
Output: The servo command for each pan-tilt unit ω_i
Workflow:
Acquire m_i based on the target detection algorithm
Obtain l_i and d_i from (1) and (2)
Calculate the intersection term ω_i^{cop} by (3)-(7)
Prepare χ_c, χ_v according to (8)
switch target detection states **do**
  case full perception
    Calculate χ_c^d, χ_v^d using (9)
  case partial perception
    Apply (23) for χ_v^d and estimate χ_c^d by (9)
  case no perception
    Apply (27) for estimates of χ_c^d and χ_v^d
Calculate the coordination term ω_i^{bio} using (10)
Derive ω_i from coordinate transformations by (11)
return ω_i

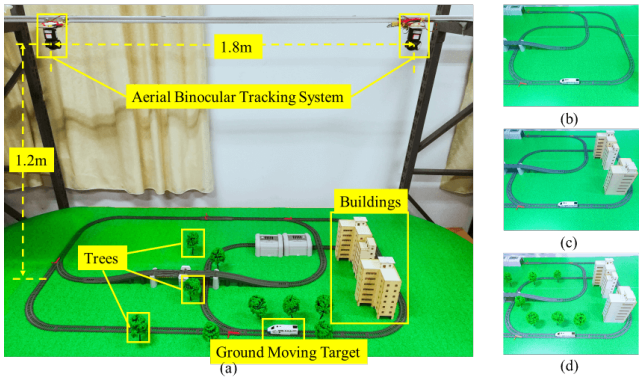

Fig. 5. The scaled testbed experiments: (a) the overall prototype and configurations; (b)-(d) three scenarios with different occlusion extents. The camera's field of view is 32 degrees. Its resolution is 640*480. With a scale ratio of 1:100, this testbed demonstrates a typical mission configuration with a flying height of 120 m and a fixed baseline of 180 m.

IV. Results and Discussions

A. Part A: Testbed Experiments

A testbed platform has been built to simulate real-world scenarios with a scale ratio of 1:100, as shown in Fig. 5. Two pan-tilt cameras are mounted above the scaled field to analog rotorcrafts. A model train serves as the moving target with a speed of 0.3 m/s (equivalent to 108 km/h in the real-world). Its movements contain linear motion in two directions and several turns. Some artificial buildings and tree models are placed around the rail as obstructions. Compared to field tests, our testbed maintains the authenticity of the camera response, target movement, and image processing, while offering the following advantages: 1) it costs less time and effort to conduct experiments; 2) it is more convenient to evaluate tracking performance at different occlusion levels since the occlusions can be adjusted by increasing or decreasing these modular obstructions.

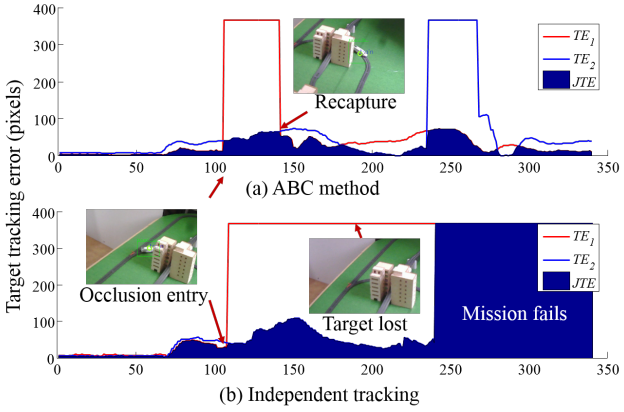

Fig. 6. Comparison of the tracking performance of different methods in testbed experiments. For both methods, the target is occluded for camera 1 at SEQ.108 and is occluded for camera 2 at SEQ.240. (a) For the ABC method, the target emerges in the camera's field of view at SEQ.142 and 269, with a low JTE throughout the mission. (b) For the independent tracking, the mission fails after both cameras lost the target.

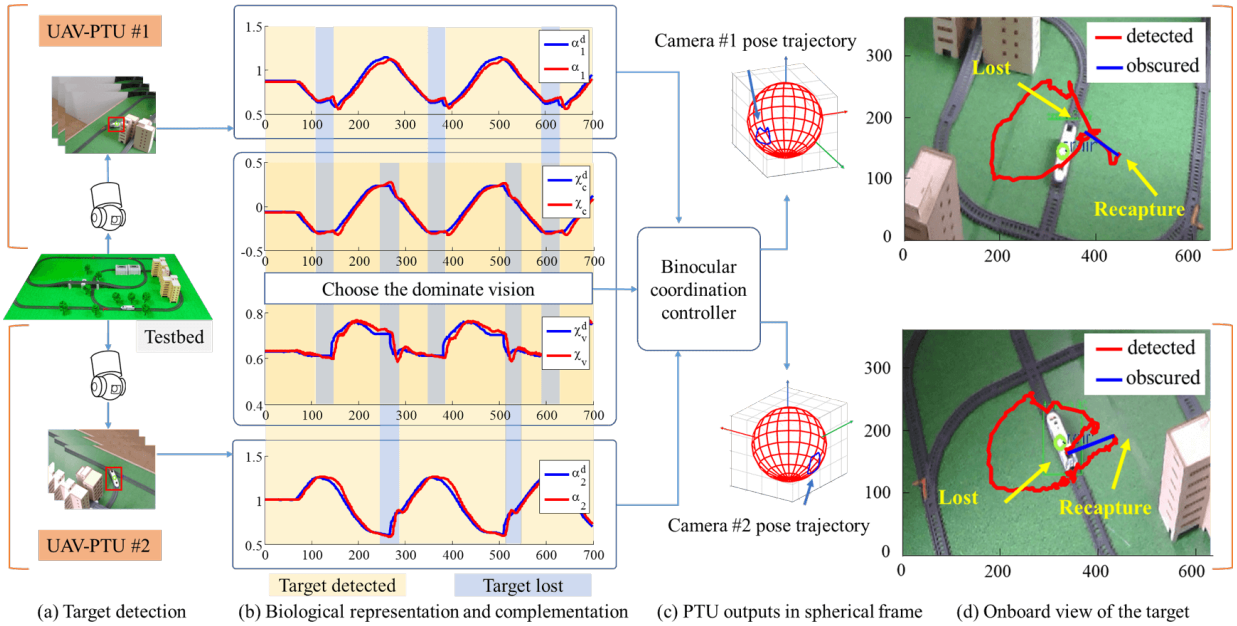

Fig. 7. Workflow of the aerial binocular coordination algorithm. (a) Each camera performs target detection independently. (b) They exchange target measurements and represent the tracking problem in a humanoid way. On considering the target occlusion, a switch control architecture is applied to calculate the desired conjugate and convergence angles based on reliable information. The yellow and blue parts indicate the complementation of the target measurements. (c) The vergence error and conjugate error are eliminated by generating desired angular velocities for cameras from a binocular coordination controller. (d) The red curve represents the image trajectory of the target, and the blue curve is obtained by connecting the vanishing point and the appearance point of the target. It suggests that the system keeps steady tracking of the moving target.

Two sets of experiments have been conducted. The purpose of the first experiment is to demonstrate the superiority of the aerial binocular coordination (ABC) method over independent tracking. Some buildings are placed around the rail to simulate a small-scale town. For each camera, the target is intermittently blocked, but at any time, at least one camera can detect the target. The independent tracking controller is described by $\omega_i^m = C_m e, i \in \{1, 2\}$, where e is the coordinate difference between the target and the image center, and C_m is the control gain. If the target is occluded, the camera simply applies a zero-input strategy. For a fair competition, both the ABC method and independent tracking adopt binocular cameras as inputs, and we set $C_c = 1.50, C_v = 1.22, C_n = 0.60, C_m = 2.00$ so that their control gains are approximately identical. The only difference is that there is coordinated movement between the two cameras of the ABC method, while for monocular tracking, the two cameras work independently.

The tracking error of each camera is defined by

$$TE_i = \sqrt{(\Delta u_i)^2 + (\Delta v_i)^2}, i \in \{1, 2\}, \tag{28}$$

where $(\Delta u_i, \Delta v_i)$ is the coordinate difference between the target and the image center. In case of target lost, the image coordinates are set to $(0, 0)$. In addition, we define the joint tracking error (JTE) as $JTE = \min\{TE_1, TE_2\}$ to evaluate its performance as a system.

The results are presented in Fig. 6. When there is no occlusion, both methods enable a stable tracking of the target. The mean values of JTE are 57.29 and 68.01 pixels for the ABC method and independent tracking, respectively. The ABC method has higher accuracy due to the coupling term C_n, which improves the convergence rate η as introduced in Subsection III-A. When the target is partially occluded, for the ABC method, the occluded camera follows the movement of its partner, and recapture the target after a short while. However, for independent tracking, the occluded camera just stays still and is unlikely to detect the target again. (In the considered scenario, the target moves periodically along the rails, so it may reappear in the camera's image, but this is rare in practice.) This means that for independent tracking, stable target tracking cannot be achieved by increasing the number of cameras. The multiple cameras will be "left behind" one by one in occlusion-rich scenarios. In contrast, the ABC method is better at task robustness due to the complementary features.

To gain insight into the collaboration between two cameras in the ABC method, the information flow of the first experiment has been illustrated in Fig. 7. Obviously, the two cameras complement each other in case of partial perception events. It is also noted that the amplitude of the conjugate angle is 0.6 rad (-0.33∼0.27 rad), whereas the variation of the vergence angle is 0.17 rad (0.59∼0.76 rad), which confirms that the conjugate movement plays a major role in ground target tracking, while the vergence movement assists in adjusting the target to the image center. This result further supports (18). In terms of tracking accuracy from the binocular geometry view, the root-mean-square error (RMSE) of conjugate movement is 0.035 rad, while the RMSE of vergence movement is 0.022 rad. Both channels are controlled with high accuracy.

Then, we continue with the second experiment to study the effect of different extents of occlusion on the tracking performance of the ABC method. Three scenarios are considered, as depicted in Fig. 5(b)-(d). A comparison of the JTE between three scenarios is presented in Fig. 8. It is regarded that the

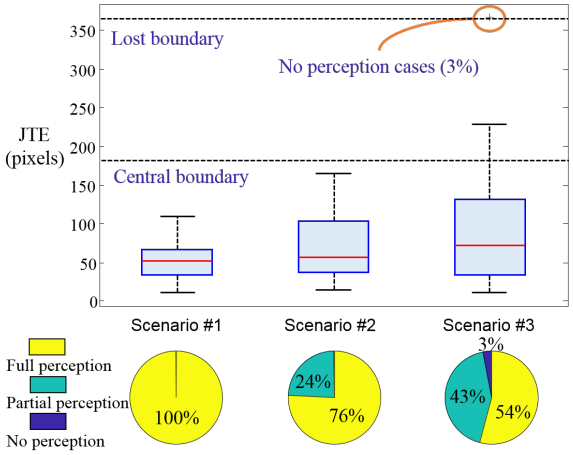

Fig. 8. Comparison of the joint tracking error (JTE) within three scenarios of different occlusion extents. The central boundary and lost boundary are half the diagonal length of the center horizon and the entire image, respectively. The pie charts at the bottom show the occlusion distributions.

tracking accuracy is good if the target remains in the central horizon, which is defined to by an area of 320*180 centered in the image. (Recall that the image is 640*480 pixels, the length and width of the central horizon are each half of the entire image.) The results suggests that with the increase of occlusion degree, the JTE increases gradually. Meanwhile, the target was kept within the central horizon in all the considered scenarios. It also proves that the proposed approach is practically feasible in occlusion-rich scenarios. For scenario #3, where the partial perception ratio is up to 43% and there are even no perception events, the mean of JTE is still less than 100 pixels.

B. Part B: Field experiments

The purpose of this subsection is to validate the proposed approach with rotorcrafts through field experiments. It is important because some assumptions are no longer satisfied in real-world applications. Several issues existing in practice such as a varying baseline, the presence of delay and loss in the cross-drone communication, may have an unknown impact on the tracking performance.

The tracking system consists of a pair of DJI M210 rotorcrafts labeled as UAV 1 and UAV 2, separately. Each drone is equipped with a pan-tilt camera for active sensing, an NVIDIA TX2 processor for onboard processing, and a data transmission module. The camera's resolution is 1280*720. Both drones are operated manually, whereas the cameras are autonomous. A car with unknown movements serves as the target.

The mission is divided into two phases, as depicted in Fig. 9. Some screenshots of the onboard views are shown in Fig. 10 and the results are drawn in Fig. 11. In the first stage, the drones circle and hover at an altitude of 40 m, whereas the target moves randomly at a speed of 5 m/s. For most of the time in this stage, both vehicles are capable of detecting the target. The mean value of the JTE is 57.01, which indicates that the proposed approach enables an effective tracking of a moving target with high accuracy.

The second stage begins at 160 seconds when UAV 2 starts moving from $P2$ to $P2'$ until the line of sight is blocked

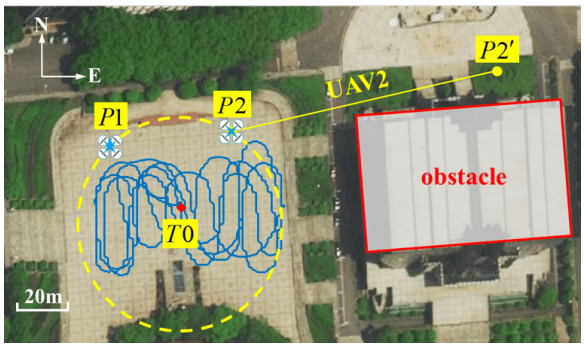

Fig. 9. A coherent mission that combines target motion, baseline changes, and partial perception. $T0$ is the target initial position, and the blue curve is the target trajectory recorded by GPS. The quadrotor icons denoted by $P1$ and $P2$ are the takeoff positions of the drones. As UAV 2 approaches $P2'$, the camera would be obstructed from the target by the building.

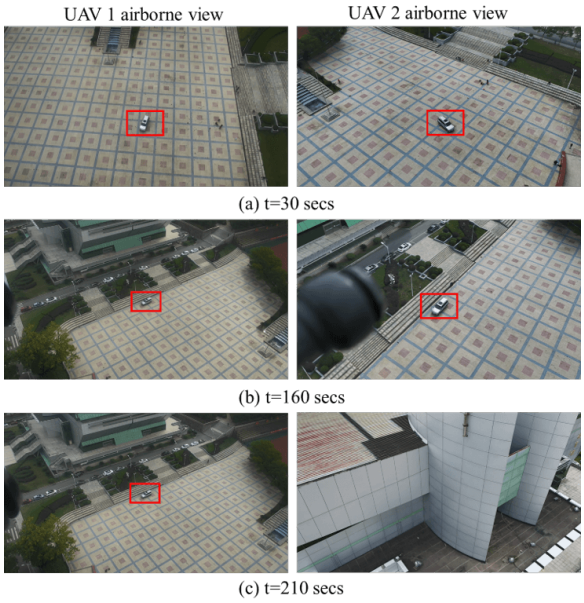

Fig. 10. The screenshots of conducted field experiments: (a) the initial condition of the mission; (b) the start of the second stage when UAV 2 starts moving towards point $P2'$; (c) the condition when UAV 2 is at $P2'$ and is obscured by the building.

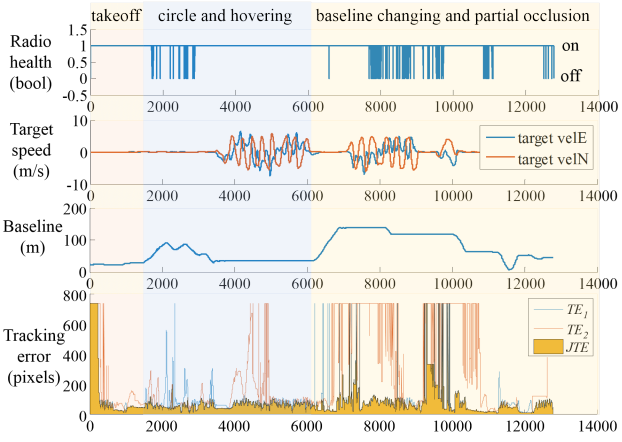

Fig. 11. Tracking performance of the field experiment accompanied by radio losses, target movement and baseline changes. Three phases of the mission are labeled using different colors.

by the building. UAV 2 stays at $P2'$ for 90 s, and returns back to $P2$. Meanwhile, UAV 1 holds its position. Then, as the car performs a random movement, the tracking system works in partial perception state for 120s. During this stage, the joint tracking error is 76.86 pixels, which is slightly larger than no-occlusion scenarios. During the movement of UAV 2, the length of the baseline varies from 30 m to 130 m, which does not have a significant impact on the tracking accuracy. Moreover, there are sporadic communication losses and detection failures during the experiment, but the tracking performance is not severely affected. This demonstrates the practicality of the proposed approach.

V. Concluding Remarks

This paper has proposed and developed a novel target tracking approach that is effective in occlusion-rich scenarios through cross-drone binocular coordination. Due to complementary perspectives of distributed cameras, it is significantly more robust against occlusions compared with independent tracking. Moreover, it has more practical potentials for diverse scenarios since it is localization-free. It has been demonstrated through testbed and field experiments that the JTE is less than 100 pixels in the specified situations. The target is still stably tracked even with a partial perception ratio of 43% in the testbed experiments.

Future studies will focus on enhancing its versatility by adding cognitive consistency constraints for dealing with multiple targets. Furthermore, multi-UAV applications with collaborate control would be considered as well.

VI. Acknowledgment

The authors would like to thank Xiaojia Xiang, Dengqing Tang and Yong Zhou from National University of Defense Technology and Minghui Li from Sun Yat-sen University for their assistance and efforts on onboard vision detection and tracking algorithms. Thanks are also extended to Bosen Lin, Ziye Chen and Tengxiang Li for their dedication in the field experiments.

References

[1] P. Tokekar, J. V. Hook, D. Mulla, and V. Isler, "Sensor planning for a symbiotic UAV and UGV system for precision agriculture," *IEEE Trans. Robot.*, vol. 32, no. 6, pp. 1498-1511, Dec. 2016.

[2] L. Merino, F. Caballero, J. R. Martínez-De-Dios, I. Maza, and A. Ollero, "An unmanned aircraft system for automatic forest fire monitoring and measurement," *J. Intell. Robotic Syst.*, vol. 65, nos. 1-4, pp. 533-548, Jan. 2012.

[3] J. C. Hodgson, R. Mott, S. M. Baylis, T. T. Pham, S. Wotherspoon, A. D. Kilpatrick, R. R. Segaran, I. Reid, A. Terauds, and L. P. Koh, "Drones count wildlife more accurately and precisely than humans," *Methods Ecol. Evol.*, vol. 9, no. 5, pp. 1160-1167, May 2018.

[4] K. S. Christie, S. L. Gilbert, C. L. Brown, M. Hatfield, and L. Hanson, "Unmanned aircraft systems in wildlife research: current and future applications of a transformative technology," *Frontiers Ecol. Environ.*, vol. 14.5, no. 2016, pp. 241-251.

[5] D. Chabot, S. R. Craik, and D. M. Bird, "Population census of a large common tern colony with a small unmanned aircraft," *PloS ONE,* vol. 10, no. 4, 2015, Art. no. e0122588.

[6] W. Meng, Z. He, R. Su, P. K. Yadav, R. Teo, and L. Xie, "Decentralized multi-UAV flight autonomy for moving convoys search and track," *IEEE Trans. Control Syst. Technol.*, vol 25, no. 4, pp. 1480-1487, Jul. 2017.

[7] E. Price, G. Lawless, R. Ludwig, I.Martinovic, H. H. Bülthoff, M. J. Black, and A. Ahmad, "Deep neural network-based cooperative visual tracking through multiple micro aerial vehicles," *IEEE Robot. Automat. Lett.*, vol. 3, no. 4, pp. 3193-3200, Oct. 2018.

[8] H. H. Helgesen, F. S. Leira, T. I. Fossen, and T. A. Johansen, "Tracking of ocean surface objects from unmanned aerial vehicles with a pan/tilt unit using a thermal camera," *J. Intell. Robotic Syst.*, vol. 91, no. 3-4, pp. 775-793, 2018.

[9] O. M. Cliff, D. L. Saunders, and R. Fitch, "Robotic ecology: Tracking small dynamic animals with an autonomous aerial vehicle," *Sci. Robot.*, vol. 3, no. 23, 2018, Art. no. eaat8409.

[10] K. Meshgi and S. Ishii, "The state-of-the-art in handling occlusions for visual object tracking," *IEICE Trans. Inf. Syst.*, vol. 98, no. 7, 1260-1274, 2015.

[11] K. Meshgi, S. Maeda, S. Oba, H. Skibbe, Y. Li, and S. Ishii, "An occlusion-aware particle filter tracker to handle complex and persistent occlusions," *Comput. Vis. Image Und.*, vol. 150, pp. 81-94, 2016.

[12] Y. Isobe, G. Masuyama, and K. Umeda, "Occlusion handling for a target-tracking robot with a stereo camera," *Robomech J.*, vol. 5, no. 1, Apr. 2018.

[13] L. Schenato, "To zero or to hold control inputs with lossy links?," *IEEE Trans. Autom. Control*, vol. 54, no. 5, pp. 1093-1099, 2009.

[14] H. Oh, S. Kim, H. Shin and A. Tsourdos, "Coordinated standoff tracking of moving target groups using multiple UAVs," *IEEE Trans. Aerosp. Electron. Syst.*, vol. 51, no. 2, pp. 1501-1514, 2015.

[15] Z. Wang and D. Gu, "Cooperative target tracking control of multiple robots," *IEEE Trans. Ind. Electron.*, vol. 59, no. 8, pp. 3232-3240, Aug. 2012.

[16] W. Meng, Z. He, R. Su, P. K. Yadav, R. Teo, and L. Xie, "Decentralized multi-UAV flight autonomy for moving convoys search and track," *IEEE Trans. Control Syst. Technol.*, vol. 25, no. 4, pp. 1480-1487, Jul. 2017.

[17] A. Khan, B. Rinner, and A. Cavallaro, "Cooperative robots to observe moving targets," *IEEE Trans. Cybern.*, vol. 48, no. 1, pp. 187-198, Jan. 2018.

[18] C. Robin and S. Lacroix, "Multi-robot target detection and tracking: taxonomy and survey," *Auton. Robots*, vol. 40, no. 4, pp. 729-760, Apr. 2016.

[19] F. Morbidi and G. L. Mariottini, "Active target tracking and cooperative localization for teams of aerial vehicles," *IEEE trans. Control Syst. Technol.*, vol. 21, no. 5, pp. 1694-1707, Sep. 2013.

[20] L. W. Krakow, C. M. Eaton, and E. K. Chong, "Simultaneous Non-Myopic Optimization of UAV Guidance and Camera Gimbal Control for Target Tracking," *IEEE Conference on Control Technology and Applications (CCTA),* pp. 349-354, 2018.

[21] W. W. Whitacre and M. E. Campbell, "Decentralized geolocation and bias estimation for uninhabited aerial vehicles with articulating cameras," *J. Guid. Control Dynam,* vol. 34, no. 2, pp. 564-573, 2011.

[22] V. Dobrokhodov, I. Kaminer, K. Jones, I. Kitsios, C. Cao, M. Lili, N. Hovakimyan, and C. Woolsey, "Rapid motion estimation of a target moving with time-varying velocity," *AIAA Guidance, Navigation and Control Conference and Exhibit,* pp. 6746, 2007.

[23] W. M. King, "Binocular coordination of eye movements-Hering's Law of equal innervation or uniocular control?," *Eur. J. Neurosci.,* vol. 33, no. 11, pp. 2139-2146, 2011.

[24] D. S. Zee, E. J. Fitzgibbon, and L. M. Optican, "Saccade-vergence interactions in humans," *J. Neurophysiol.,* vol. 68, no. 5, pp. 1624-1641, 1992.

[25] W. M. King and W. U. Zhou, "Neural basis of disjunctive eye movements," *Ann. N. Y. Acad. Sci.*, vol. 956, no. 1, pp. 273-283, 2002.

[26] A. Gibaldi and M. S. Banks, "Binocular eye movements are adapted to the natural environment," *J. Neurosci.*, vol. 39, no. 15, pp. 2877-2888, 2019.

[27] Y. Song and X. Zhang, "An active binocular integrated system for intelligent robot vision," in *Proc. IEEE Int. Conf. Intell. and Secur. Inform.,* 2012, pp. 48-53.

[28] N. Pateromichelakis, A. Mazel, M. A. Hache, T. Koumpogiannis, R. Gelin, B. Maisonnier and A. Berthoz, "Head-eyes system and gaze analysis of the humanoid robot Romeo," in *Proc. IEEE Int. Conf. Intell. Robots Syst. (IROS)*, 2014, pp. 1374-1379.

[29] J. Redmon and A. Farhadi, "Yolov3: An incremental improvement," *arXiv preprint,* arXiv:1804.02767, 2018.

[30] X. Zhang and Y. Sato, "Cooperative movements of binocular motor system," in *Proc. IEEE Int. Conf. Auto. Sci. Eng.*, 2008, pp. 321-327.

[31] D. H. Arnold, P. Law and T. S. A. Wallis, "Binocular switch suppression: A new method for persistently rendering the visible 'invisible'," *Vision Res.*, vol. 48, no. 8, pp. 994-1001, 2008.