2025 IEEE/SICE International Symposium on System Integration (SII)
January 21-24, 2025, Munich, Germany

Saccade Argos: Hierarchical Robust Tracking System for High Spatio-temporal Resolution Vision

Leo Miyashita[1] and Masatoshi Ishikawa[1]

Abstract— **Target tracking is one of the most important tasks in computer vision to keep capturing an object with high spatio-temporal resolution. However, when the field of view is narrowed to capture a target in detail, it is easy to lose sight of the target due to the fast movement or occlusion. In this paper, we propose a system integration method that hierarchically handles multiple visions with different FOV from wide-angle to telephoto, and each vision independently tracks the target while sharing the target position to achieve robust tracking. The proposed method was applied to Saccade Argos, a three-tier tracking system that combines a fixed stereo camera and active tracking systems using Galvano scanners, and achieved robust target tracking against occlusion with high temporal resolution (1,000 fps) and high spatial resolution (248.3 px/deg.).**

I. INTRODUCTION

Tracking is an important task in the field of machine vision. Tracking targets range from industrial products to humans and animals, and its applications are diverse, including product management, security, motion and vibration analysis, sports, entertainment, and human-computer interface. Vision tracking is now an indispensable tool to grab the behavior of objects, especially when their movement is unpredictable. In vision tracking, it is expected to capture a target in high spatio-temporal resolution while maintaining robustness against the target's behavior and disturbances. However, there is a dilemma between field of view (FOV) and spatial resolution. With wide-angle vision, even when the actual velocity or acceleration of the target is high, the apparent movement in the image can be small, so the target is difficult to be lost. On the other hand, the effective resolution is reduced because the area of the target in the image would be smaller. In contrast, with telephoto vision, a larger number of pixels can be allocated to the target, but even small movements cause large changes in the image, making it easy to lose sight of the target. With a single vision, it is not easy to solve this dilemma and achieve high-resolution and robust tracking simultaneously. However, multiple visions with different FOV can be a solution. Furthermore, by treating these visions hierarchically, combining suitable tracking algorithms according to the FOV, and coordinating them with each other, robust tracking even against occlusion can be achieved.

For example, consider tracking the human eye from a distance to measure the gaze of a pedestrian or a runner as shown in Fig. 1. With wide-angle vision, it is not easy to capture fine changes of the gaze, and with telephoto vision, it is not easy to find the target itself or to continuously measure the head and eyes as they move quickly in the narrow FOV. Furthermore, simple tracking algorithms may lose sight of the target when the target blinks, turns away, or is occluded by some factor. In contrast, it is possible to achieve high spatio-temporal resolution and robust tracking of the target by constructing a system in which the wide-angle vision detects the human body, which is a macroscopic structure containing the target, rather than the target itself, and tracks the body while sharing the head position with the telephoto vision tracking the target eye. As long as the wide-angle vision can recognize the human body, the approximate location of the eye can be determined regardless of whether it is occluded or not, enabling seamless tracking while it is occluded and when it reappears.

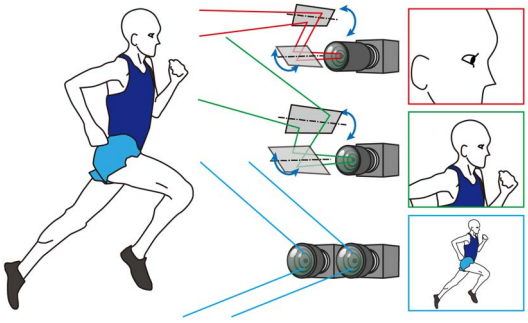

Fig. 1. Hierarchical tracking of a moving target.

In this paper, we propose a framework that integrates multiple vision tracking systems with different FOV and introduce the configuration, control, and tracking methods in a hierarchical manner. We also evaluate a novel three-tier tracking system, Saccade Argos, by three-dimensionally tracking a fast-moving object in high spatio-temporal resolution even with occlusion and discuss the performance and limitations.

II. RELATED WORKS

In this paper, we describe a tracking system that does not have actuators and has a fixed FOV as passive, and a system that can dynamically change the gaze direction by actuators as active. In addition, the unit that independently tracks the target is represented as a layer in a hierarchy for each FOV.

Passive tracking systems often use wide-angle vision to provide a large enough FOV to track a target, which reduces the spatial resolution available to the target, so high-resolution image sensors are often used. However, the passive system cannot feedback the information of a target, so the

[1]Research Institute for Science and Technology, Tokyo University of Science, Tokyo, Japan miyashita@ishikawa-vision.org

performance for detailed target observation is not comparable with an active system.

For an active system, a motorized platform is often used to change the gaze direction, pan and tilt [1], [2], [3]. However, it is not easy to track an object moving agilely with large acceleration, because a motorized platform directly moves a camera and lens having large inertia.

In contrast, a vision system that performs pan and tilt using a motorized rotating mirror such as a Galvano scanner [4], [5], [6], [7] or fast steering mirror (FSM) [8] has been proposed. Since mirrors can be lightweight compared to cameras and lenses, high-speed pan and tilt can be achieved by rotating the mirror at high speed in front of the lens to virtually change the gaze direction. In particular, since the rotation angle of the optical axis will be twice of the mirror, an active tracking system using mirrors can achieve high tracking performance when image capturing and processing are as fast as the dynamics of the mirror to feedback the target position.

However, there are not many studies of tracking systems that combine and handle multiple tracking layers hierarchically. Zhao proposed an active tracking system that handles four different FOV [9], but the system uses the same tracking algorithm and a single actuator unit for all visions, and so the system does not make the most of the hierarchical structure. There are also some proposed vision systems in which two visions with different FOV perform different actions [10], [11], [12], but they are master-slave systems in which the telephoto vision determines the gaze direction based only on the instruction of the wide-angle vision. While these systems have simultaneously realized imaging with different characteristics, they are essentially one-tier tracking systems without mutual coordination. Qing proposed a two-tier tracking system [13] but it does not take into account the three-dimensional information of a target obtained thanks to multiple layers, nor discuss extensions to a hierarchy having three or more layers.

In this paper, we discuss effective calibration methods and generalized configurations that can be applied to more than three layers, in which each layer performs tracking in parallel while cooperating together to be robust against occlusion and loss of sight.

III. MODELING OF A LAYER

A. Layer model

In this paper, we treat both the passive and active tracking units mentioned in the previous section as a single layer for tracking. For the passive layer, which has a fixed FOV, the following perspective projection is used to model a camera in this paper.

$$s\tilde{\boldsymbol{u}} = I[R|\boldsymbol{t}]\tilde{\boldsymbol{x}} \tag{1}$$

$$\tilde{\boldsymbol{u}} = \begin{bmatrix} u & v & 1 \end{bmatrix}^{\mathrm{T}}, \quad \tilde{\boldsymbol{x}} = \begin{bmatrix} x & y & z & 1 \end{bmatrix}^{\mathrm{T}}$$

Note that $\boldsymbol{u}$ are 2D coordinates on an image coordinate system and $\boldsymbol{x}$ are 3D coordinates in world coordinate system, and $\tilde{\cdot}$ means homogeneous coordinates. $I, R, \boldsymbol{t}$ are intrinsic and extrinsic parameters of a camera. If there are multiple cameras in a layer, 3D position of a target can be obtained by the stereo vision using the equation in conjunction.

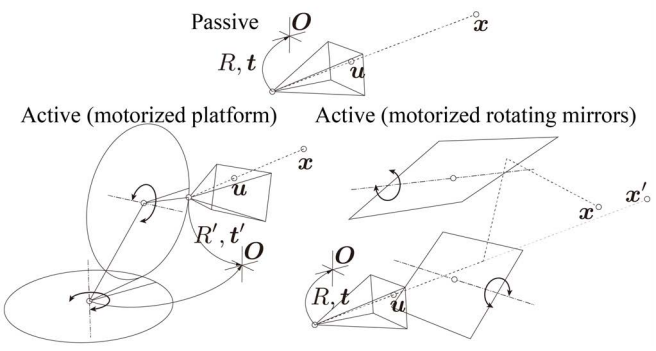

Fig. 2. Model of a passive layer and active layers. In the case of motorized platform, the pose and position of the camera will be complicated. In the case of motorized rotating mirrors, the target position has to be converted with complicated matrices. However they can be handled with the same method described in the section IV .

The active layer that performs the pan and tilt can be modeled by replacing extrinsic parameters $R, \boldsymbol{t}$ with $R', \boldsymbol{t}'$ in the case of motorized platform, and by replacing $\boldsymbol{x}$ with $\boldsymbol{x}'$ in the case of motorized rotating mirrors as shown in Fig. 2. Note that as a model of motorized rotating mirrors, the figure shows combination of single-axis motors like a Galvano scanner but dual-axis motor like FSM can be modeled with the same model by setting the rotation centers to the same point. However, when a layer tracks the target independently with a single camera, the absolute gaze direction is not important because tracking is performed based on the relative positioning of a target in the image. Therefore, there is no need to accurately calibrate $R', \boldsymbol{t}'$, or conversion matrices between $\boldsymbol{x}$ and $\boldsymbol{x}'$. Then, we introduce a simple calibration method for active layers in the next section.

B. Tracking and reference

In an active layer, the actuator has to be driven appropriately to track a target. As shown in Fig. 3, we define the center of the FOV as the center point, detected landmarks on the target as keypoints, a specified keypoint to be captured at the center point as the tracking point, and a specified keypoint to be given for complement of tracking points in the other layers as the reference point. The tracking strategy in this paper is to control the actuator based on the position of the obtained tracking point, always keeping the distance between the tracking point and the center point small. However, the tracking point may be lost when the object is moving at high speed or due to occlusion. In such cases, the tracking point can be complemented by using reference points on other layers appropriately as mentioned in the next section.

IV. POSITION COMPLEMENT AMONG LAYERS

When there are multiple layers that perform independent tracking, the target that each layer tracks does not necessarily have to be the same. If the target in a layer has inclusion

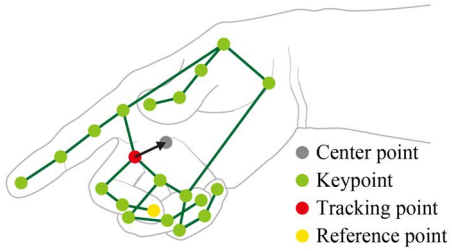

Fig. 3. Point definitions. Tracking point is a selected keypoint to minimize the distance to the center point. Reference point is a selected keypoint for position complement in other layers.

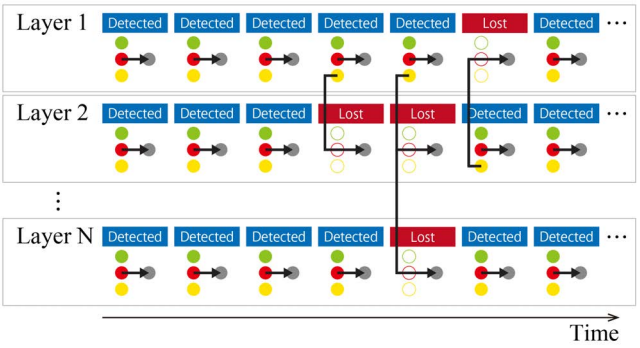

Fig. 4. Example timeline including complements of the lost tracking point in a layer with a detected reference point in another layer.

relationship with a target in another layer, and the FOV has an overlap, they can be linked by sharing the target position with each other. If the targets are the same, this link enables tracking with high positional accuracy, and if the targets are in inclusion relationship, robust tracking that is resistant to occlusion and disturbance can be realized. In this paper, we focus on the case of inclusion relationship and complement a lost tracking point of a layer based on a detected reference point of another layer as shown in Fig. 4. Although the reference points of multiple layers can be involved in this process, the information from the layer with narrower FOV is generally considered to be more accurate. Therefore, the reference point of the layer detecting a target with the narrowest FOV is used for the complement. On the other hand, since each layer does not necessarily have the same system configuration, some layers are supposed to output the 2D position in the image coordinate system or the 3D position in the world coordinate system as the reference point. Then we have to organize how to use the reference point in each case.

A. Complement from 3D to 3D position

When a layer has two or more cameras and can observe the same target as a stereo vision, or when a TOF camera is used, the parameter s, which determines the z-coordinate in Equation (1) is derived and the 3D coordinates of the reference point can be measured. In this case, a lost 3D tracking point can be substituted by the measured 3D reference point.

B. Complement from 3D to 2D position

Complementing a lost 2D tracking point in a passive layer with the detected 3D reference point can be easily accomplished by substituting the reference point into Equation (1). For an active layer, it is also possible to use models with $R', \boldsymbol{t}'$, or $\boldsymbol{x}'$ in the same way, but the calibration for these parameters can be complicated due to the nonlinearity with respect to rotation angles[14]. In particular, when a system consists of many layers and forms a large hierarchy, calibration of each layer is expected to be as simple as possible. Therefore, assuming that the nonlinearity caused by the actuators can be approximated by a polynomial of at most the second order, we use the detected 3D reference point to directly determine the rotation angle of the actuator using the following Lagrange interpolation polynomial.

$$\boldsymbol{\theta}' = K_3 \boldsymbol{\lambda}_3 \tag{2}$$

$$\boldsymbol{\theta}' = \begin{bmatrix} \theta'_x & \theta'_y \end{bmatrix}^{\mathrm{T}}, \quad \boldsymbol{\lambda}_3 = \begin{bmatrix} 1 & x & y & z & x^2 & y^2 & z^2 & xy & yz & zx \end{bmatrix}^{\mathrm{T}}$$

Here, $\boldsymbol{\theta}$ is the rotation angle for each axis in the case that the target is successfully tracked and $\boldsymbol{\theta}'$ represents the complemented rotation angle. They can be replaced with an input to the actuator such as voltage if the actuator has linearity. In system calibration, samples corresponding to $\boldsymbol{\theta}$ and $\boldsymbol{\lambda}_3$ are easily obtained by independently tracking the same point in each layer. Since Equation (2) is linear, the coefficients of the polynomial K_3, can be derived from a pseudo-inverse matrix composed of a sufficient number of samples. The validity of this approximation is examined in Section V.

C. Complement from 2D position

A single 2D reference point is insufficient for complement of target position in another layer, and so highly accurate complement cannot be performed in principle. However, the smaller the distance between layers and the further away the target is, the smaller the disparity becomes, and thus the smaller the complement error becomes. In this paper, we consider only the complement to 2D position as follows.

$$\boldsymbol{\theta}' = K_2 \boldsymbol{\lambda}_2 \tag{3}$$

$$\boldsymbol{\lambda}_2 = \begin{bmatrix} 1 & \theta_x & \theta_y & \theta_x^2 & \theta_y^2 & \theta_x \theta_y & u & v \end{bmatrix}^{\mathrm{T}}$$

Note that $\boldsymbol{\lambda}_2$ is composed of parameters of a layer which is successfully tracking a target, and $\boldsymbol{\theta}'$ is a variable of another layer which is losing sight of the target. In the system calibration, K_2 can be derived in the same way with K_3 simultaneously.

V. SACCADE ARGOS

A. System setup

We actually constructed a hierarchical tracking system, Saccade Argos shown in Fig. 5, which has three layers with different FOV and tracking methods. Table I shows the specifications of the Saccade Argos in this paper. The ratio of the FOV between the adjacent layers is approximately $4:1$. The most telephoto layer has a focal length of about

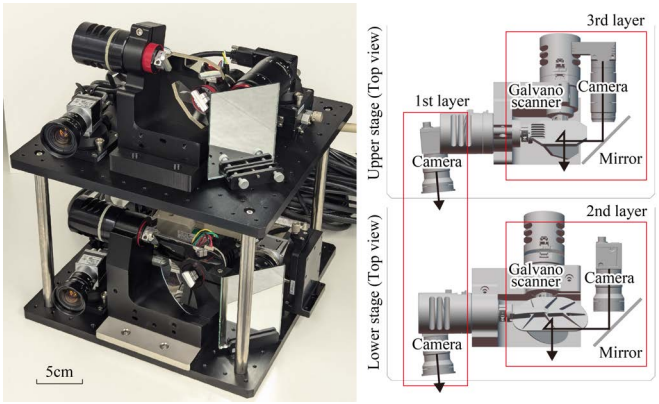

Fig. 5. Saccade Argos and top view of each stage. Three tracking layers are constituted on upper and lower stages. The cameras in the 1st layer are fixed on each stage and form a stereo vision.

19.3 times, FOV of about 1/18.2, and a spatial resolution of about 18.4 times per unit compared to the most wide-angle layer. Since the apparent velocity of the object increases with telephoto, the telephoto layer was composed of faster devices.

The cameras in each layer were calibrated for the intrinsic parameters in Equation (1) using Zhang's method [15], and for the stereo camera in the first layer, stereo calibration was performed to calibrate the extrinsic parameters in Equation (1). The calibration between layers was performed based on Equation (2) and (3) by tracking a bright spot by all layers and sampling their positions and inputs to the actuators at appropriate intervals. As a result, 77 samples were used to optimize K_3 and K_2, and the average error between the first and the second layers was about $0.67\,\mathrm{deg.}$ in the pan angle, about $0.41\,\mathrm{deg.}$ in the tilt angle. The average error between the first and the third layers was about $0.57\,\mathrm{deg.}$ in the pan angle and about $0.47\,\mathrm{deg.}$ in the tilt angle. These errors are smaller than half of the FOV, and so the accuracy is sufficient when each layer performs tracking independently. Therefore, it is reasonable to approximate the nonlinearity in active layers with a function of at most the second order.

Since each layer performs tracking independently, the framerate depends on the algorithm of each layer. Therefore, we adopted a fast algorithm for the telephoto layer. All deep-learning-based algorithms are processed on nVidia GeForce RTX 4090 and the others are on AMD Ryzen Threadripper 7980X. In an active layer, actuators are driven with PD control method based on the position of the tracking point relative to the center point.

B. Tracking including occluded scenes

In this paper, we demonstrate Saccade Argos by performing two types of tracking; Demonstration A for the right eye and Demonstration B for the tip of left middle finger.

In Demonstration A, the human body pose was detected in the first layer using a deep-learning-based YOLOv8-pose, and the 3D position of each key point of the body was calculated from the disparity of the images acquired from each camera of the stereo vision. The first and second rows of Fig. 6 show the images captured by the stereo vision in the first layer and the detected 17 keypoints. Since the final tracking target of Demonstration A is the right eye, the 3D position of the right eye is set as the reference point for a lost tracking point in the other layer. In the second layer, five keypoints of the face were detected using a deep-learning-based YOLOv8-pose model that is trained for facial landmarks. The third row of Fig. 6 shows the captured images and detected keypoints in the second layer. In the second layer, the nose was set as the tracking point and the right eye as the reference point. In the third layer, eyes were detected using a cascade classifier of Haar-like features [16] trained for eyes. The fourth row of Fig. 6 shows the captured image of the third layer and the center of the bounding box of the detected eye. Figure 6 shows that the right eye, which is the final tracking target, was tracked continuously with high spatio-temporal resolution in the third layer. In addition, even when occlusion occurred, the reference points in the other layers were used for continuous tracking where the target was thought to be located, allowing seamless tracking when the target reappeared. The working distance for Demonstration A was about $4\mathrm{m}$, and the trackable velocity at this distance is calculated to be $23.3\mathrm{m/s} = 83.8\mathrm{km/h}$ based on the Galvano scanner's step response for $0.4\,\mathrm{deg.}$ rotation.

TABLE I

SPECIFICATIONS OF EACH LAYER. FOCAL LENGTH REPRESENTS 35 MM EQUIVALENT VALUE.

		1st layer	2nd layer	3rd layer
Tracking type		Passive stereo	Active mono	Active mono
Actuation type		–	Galvano scanner	Galvano scanner
Focal length [mm]		41	187	790
FOV (H×V) [deg.]		47.4×36.4	11.0×8.2	2.61×1.96
Camera		Basler acA640-750um	Basler acA640-750um	OMRON sentech HSV-MC1
Resolution [px/deg.]		13.5	58.4	248.3
Actuator		–	VantagePro QS45XY-AG	VantagePro QS30XY-AG
Optical angle [deg.]		–	±25.5	±25.5
Step response 0.4 deg. [ms]		–	1.2	1.2
A	Algorithm	YOLOv8 body pose	YOLOv8 facemark	Haar-like cascade eye
	Framerate [fps]	37	78	1,000
B	Algorithm	YOLOv8 body pose	OpenPose hand pose	Binarization
	Framerate [fps]	39	75	1,000

In Demonstration B to track the tip of left middle finger, the tracking algorithm in the first layer was the same with that in Demonstration A, but the reference point is set to the left wrist. In the second layer, the pose of the left hand was detected using deep-learning-based OpenPose[17]. From the first to third rows of Fig. 7 show the results of the first and second layers. In the second layer, the root of the middle finger was set as the tracking point and the tip of the middle finger as the reference point. In the third layer, a bright area detection using binarization in a region of interest (ROI) was used to detect the tip of the finger and setting the centroid as the tracking point and reference point. Although this detection algorithm is fast enough to achieve $1,000\,\mathrm{fps}$, it is not robust to changes in shading. Moreover, keypoints in the first layer does not include a fingertip, and so a fingertip cannot be set as a reference point of the first layer. The second layer also cannot provide an accurate complement because only 2D keypoints are obtained, resulting in erroneous tracking of a tip of the index finger as shown in the right column of Fig. 7. This result of erroneous tracking is discussed in the next section. The working distance for Demonstration B was about $3\,\mathrm{m}$, and the trackable velocity at this distance is calculated to be $17.5\,\mathrm{m/s} = 62.8\,\mathrm{km/h}$ at least.

VI. DISCUSSION

A. High-speed image processing and accurate complement

The apparent movement of a target on the image increases in the telephoto layer, so tracking requires faster image capture, faster image processing, and faster actuators when the layer is active. While image capture and actuators are device-dependent and can perform independently of the target, image processing, which recognizes the target semantically, is target-dependent. The more telephoto the image is, the smaller area on the target is observed, but this does not mean that the object can be represented by a simple model. Rather, the requirement for detailed observation means that there is a structure to be recognized in that area, and its semantic recognition requires large computational cost. Furthermore, even if semantic recognition can be accelerated, it is not easy to distinguish whether the focused object is the tip of the middle finger or the tip of the index finger, as in Demonstration B, when the FOV is limited. Hence, it is important to use multiple FOV to grasp the larger structure containing a target.

In Demonstration B, semantic recognition was omitted to achieve $1,000\,\mathrm{fps}$ high-speed image processing, and so the layer lacked the ability to recognize the target. Moreover, even though multiple FOV were used, it was not possible for the other layers to set the exact location of the target as a reference point and the system caused erroneous tracking. From the above, it is expected for improvement to speed up semantic recognition and to three-dimensionally grasp the reference point in all layers.

B. Telephoto macro lens

In order to observe an object in as much detail as possible, it is effective to get close to the object and observe it with a telephoto vision. On the other hand, since the FOV becomes extremely narrow at such a telephoto and close-up situation, it is difficult to grasp the current observation area and to capture the target by manual operation. Because of this background, few lenses that enable telephoto and macro imaging have been manufactured, except for microscope-scale lenses. This limited the design freedom of the telephoto side of the system presented in this paper. On the other hand, if it becomes easier to capture an object by automation as shown in this paper, the applicability of telephoto macro imaging will expand greatly. In the future, the development of lens systems with unprecedented optical characteristics for automation is expected.

VII. CONCLUSIONS

In this paper, we propose a method for integrating and controlling a hierarchical tracking system with multiple FOV for the purpose of continuously observing a dynamic object with high spatio-temporal resolution. The method is generalized to be applicable to various tracking styles, and we have demonstrated in tracking a distant object with a high temporal resolution of $1,000\mathrm{fps}$ and a high spatial resolution of $248.3\,\mathrm{px/deg.}$ using Saccade Argos, which integrates three tracking systems with different FOV. Furthermore, the proposed method has realized robust tracking against occlusion by complementing the position of the target among tracking layers. We also discussed the effectiveness of the hierarchy and its issues based on the case of erroneous tracking. Future works will include constructing a system that can recognize various targets at high speed with three-dimensional keypoints and extension for high-speed focusing with a liquid lens.

REFERENCES

[1] Y. Xie, L. Lin and Y. Jia, "Tracking Objects with Adaptive Feature Patches for PTZ Camera Visual Surveillance," 20th International Conference on Pattern Recognition, pp. 1739-1742, 2010.

[2] P. D. Z. Varcheie and G. -A. Bilodeau, "Adaptive Fuzzy Particle Filter Tracker for a PTZ Camera in an IP Surveillance System," IEEE Transactions on Instrumentation and Measurement, vol. 60, no. 2, pp. 354-371, 2011.

[3] N. Liu, H. Wu and L. Lin, "Hierarchical Ensemble of Background Models for PTZ-Based Video Surveillance," IEEE Transactions on Cybernetics, vol. 45, no. 1, pp. 89-102, 2015.

[4] K. Okumura, H. Oku and M. Ishikawa, "High-speed gaze controller for millisecond-order pan/tilt camera," IEEE International Conference on Robotics and Automation, pp. 6186-6191, 2011.

[5] D. Wood, M. Bishop, "A Novel Approach to 3D Laser Scanning," Australasian Conference on Robotics and Automation, 2012.

[6] K. Okumura, K. Yokoyama, H. Oku, and M. Ishikawa, "1ms Auto PanTilt - video shooting technology for objects in motion based on Saccade Mirror with background subtraction," Advanced Robotics, vol. 29, no. 7, pp. 457–468, 2015.

[7] K. Iida, and O. Hiromasa, "Saccade Mirror 3: High-speed gaze controller with ultra wide gaze control range using triple rotational mirrors." IEEE International Conference on Robotics and Automation, pp. 624-629, 2016.

[8] J. M. Hilkert, G. Kanga, and K. Kinnear "Line-of-sight kinematics and corrections for fast-steering mirrors used in precision pointing and tracking systems", SPIE 9076, Airborne Intelligence, Surveillance, Reconnaissance (ISR) Systems and Applications XI, 90760F, 2014.

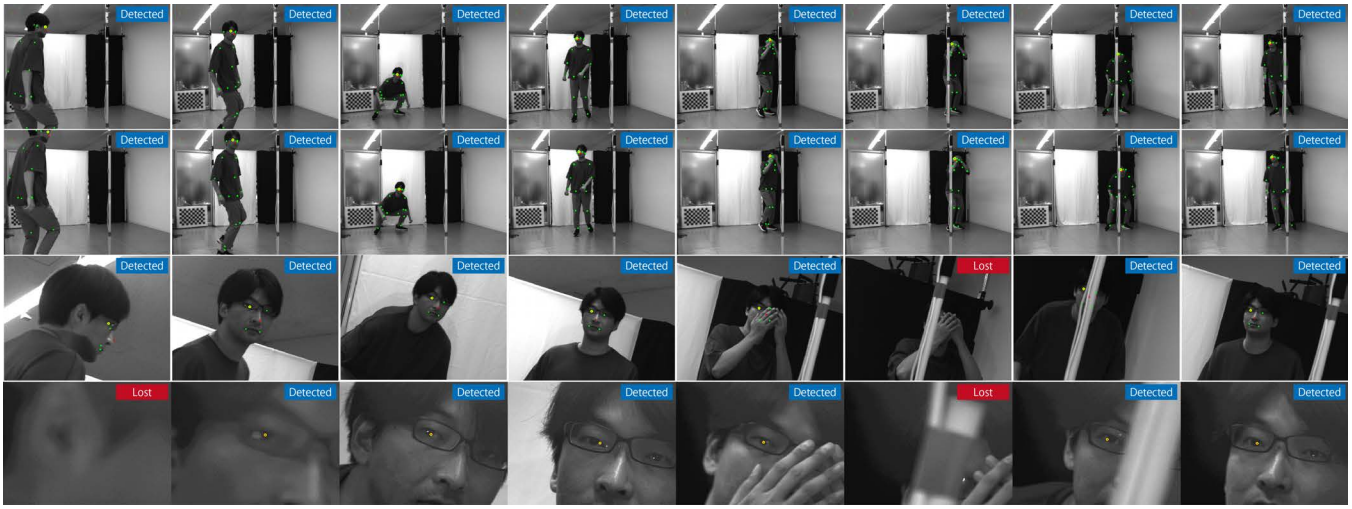

Fig. 6. Demonstration A. The target, right eye is available as a keypoint in all layers, and so complement of the tracking point was performed accurately.

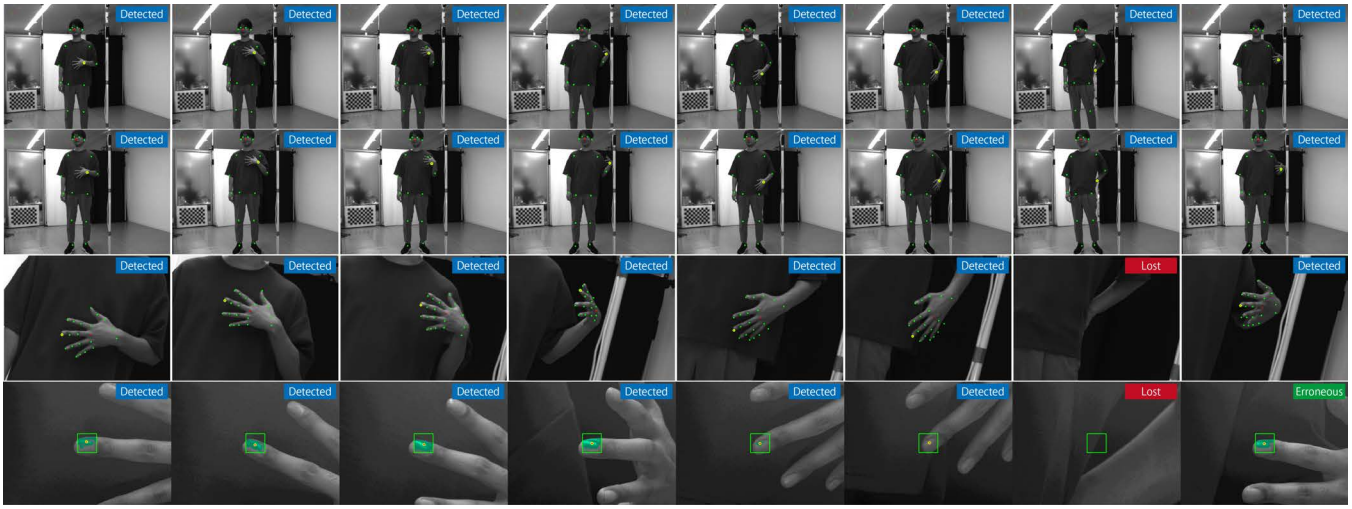

Fig. 7. Demonstration B. The target, tip of left middle finger is available as a keypoint only in the 2nd layer, which detects 2D positions of keypoints. As a result, erroneous tracking occurred after occlusion.

[9] X. Zhao, Q. Gu, T. Aoyama, T. Takaki and I. Ishii, "A fast multi-camera tracking system with heterogeneous lenses," IEEE/RSJ International Conference on Intelligent Robots and Systems, pp. 2671-2676, 2013.

[10] S. Hu, K. Shimasaki, M. Jiang, T. Senoo and I. Ishii, "A Simultaneous Multi-Object Zooming System Using an Ultrafast Pan-Tilt Camera," IEEE Sensors Journal, vol. 21, no. 7, pp. 9436-9448, 2021.

[11] K. Shimasaki, M. Ito, S. Hu, F. Wang and I. Ishii, "Smart Telescope System with Automatic Tracking," IEEE SENSORS, pp. 1-4, 2023.

[12] T. Zhang, Z. Li, Q. Wang, K. Shimasaki, I. Ishii and A. Namiki, "DoF-Extended Zoomed-In Monitoring System With High-Frame-Rate Focus Stacking and High-Speed Pan-Tilt Adjustment," IEEE Sensors Journal, vol. 24, no. 5, pp. 6765-6776, 2024.

[13] Li, Qing, Shaopeng Hu, Kohei Shimasaki, and Idaku Ishii, "An Active Multi-Object Ultrafast Tracking System with CNN-Based Hybrid Object Detection" Sensors, no. 8, 4150, 2023.

[14] T. Sueishi, H. Oku and M. Ishikawa, "Mirror-based high-speed gaze controller calibration with optics and illumination control," IEEE/RSJ International Conference on Intelligent Robots and Systems, pp. 3064-3070, 2015.

[15] 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.

[16] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. I-I, 2001.

[17] Z. Cao, G. Hidalgo, T. Simon, S.-E. Wei, and Y. Sheikh. "OpenPose: Realtime Multi-Person 2D Pose Estimation Using Part Affinity Fields," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 43, no. 1, pp. 172–186, 2021.