2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)
October 14-18, 2024. Abu Dhabi, UAE

An Ultrafast Multi-object Zooming System Based on Low-latency Stereo Correspondence

Qing Li[1,2], Shaopeng Hu[2], Kohei Shimasaki[2], Idaku Ishii[2,*]

Abstract— **In this paper, we develop a multiple-object zooming system which can capture clear images of different objects at an ultrafast speed. The system consists of a panoramic HFR stereo camera and a galvanometer-based reflective pan-tilt-zoom (PTZ) camera. In order to alleviate the impact of brightness, noise, and viewing angle in the image, we use the high speed motion information of the object for stereo correspondence. According to the spatial positions of all objects obtained from HFR stereo correspondence, we can obtain the control voltage of the pan and tilt mirror of the galvanometer-based reflective PTZ camera through the mapping relationship. Then, PTZ camera captures clear images of multiple objects in a time-division multiplexed manner at an extremely fast speed. Experimental results show that we can distinguish multiple fast-moving people indoors in the HFR stereo camera and capture their high-definition facial images simultaneously.**

I. INTRODUCTION

Zooming in on objects of interest in images is a critical technique in the field of image processing. This technique enhances image resolution and enables the acquisition of high-definition images of targeted objects. This improvement facilitates better observation and analysis of the objects of interest. It finds extensive application in various domains such as video surveillance [1], object tracking [2], medical image processing [3], and map and satellite remote sensing image processing [4].

At the same time, wide observation range can provide more information, which is also very important. However, the acquisition of high-definition images and the observation of wide fields of view are contradictory due to the limitations of camera imaging principles. Currently, there are two popular schemes for image zooming: 1) Image super-resolution reconstruction method [5], and 2) Optical zooming method based on PTZ cameras [6].

Image super-resolution refers to the process of reconstructing low-resolution images into high-resolution images. Early methods were mainly based on traditional image processing techniques such as bicubic interpolation [7] and convolution kernel filters [8]. With the development of machine learning technology, more and more researchers have begun to use the deep neural network (DNN) to solve the problem of super-resolution reconstruction. The current state-of-the-art approach uses generative adversarial networks (GANs) to solve the super-resolution reconstruction problem [9]. These methods work by training a generator network to generate high-resolution images and using a discriminator network to evaluate the quality of the generated images. High-resolution reconstruction methods have shown good performance in various applications, but there are still some shortcomings [10], such as loss of details, lack of global consistency, high computational load, high computational delay, etc. It is difficult to apply to real-time systems.

Different from software-based super-resolution reconstruction, the optical zoom method assists the panoramic camera to acquire high-resolution images by adding a camera equipped with a telephoto lens [11]. The PTZ camera system drives the motor to make the camera rotate in the horizontal and vertical directions to obtain high-definition images from different viewing angles. Due to the large volume of the camera and the telephoto lens, the auxiliary telephoto camera of this active mechanical movement often moves slowly and is not suitable for multi-target observation. To obtain high-definition images of multiple objects, many combination systems of multi-telephoto PTZ cameras have been proposed [12]. With the increase of PTZ cameras, the volume and size of the entire optical zooming system are also increasing exponentially, making it difficult to install and control.

Compared to the active motion PTZ cameras mentioned above, the galvanometer-based reflective PTZ camera system offers higher movement speeds and can switch between multiple objects extremely rapidly, making it highly suitable for high-definition image acquisition of multiple objects. Our laboratory has long been committed to researching reflective PTZ camera systems and has achieved significant research outcomes. In a recent paper, we propose a simultaneous multi-objective zooming system using a galvanometer-based PTZ camera [13]. This novel dual-camera system can simultaneously capture multiple zoomed-in images. However, due to the significant delay in AI detection, this zooming system exhibits considerable offset when observing multiple moving objects.

In this paper, we aim to improve the system proposed in [13] by introducing a HFR stereo camera. Correspond to multiple moving objects based on high-synchronous motion information. Then, the galvanometer-based reflective PTZ camera can capture high-definition images at the spatial position obtained through stereo correspondence. The mapping relationship between the control voltage and the spatial

*This work was supported by the Postdoctoral Fellowship Program (Grade C) of China Postdoctoral Science Foundation under Grant Number GZC20240879. (Corresponding author: Idaku Ishii)

[1]Qing Li is now with the Department of Electronic Engineering, Tsinghua University, Beijing 100084, China. [2]He was with Smart Robotics Lab., Hiroshima University, Hiroshima 739-8527, Japan. E-mail: soleilor@mail.tsinghua.edu.cn.

[2]Shaopeng Hu, Kohei Shimasaki, and Idaku Ishii are with Smart Robotics Lab., Hiroshima University, Hiroshima 739-8527, Japan (e-mail: hsp@hiroshima-u.ac.jp; simasaki@hiroshima-u.ac.jp; iishii@robotics.hiroshima-u.ac.jp).

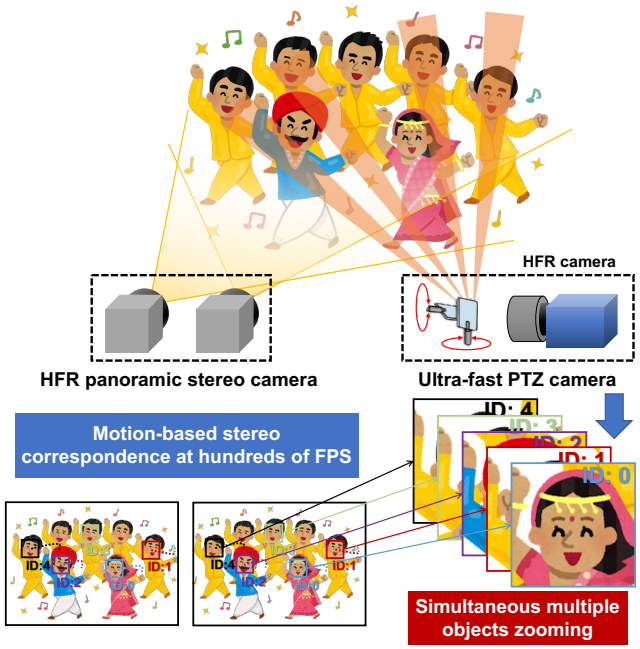


Fig. 1. Concept of ultrafast multi-object zooming system based on low-latency stereo correspondence.

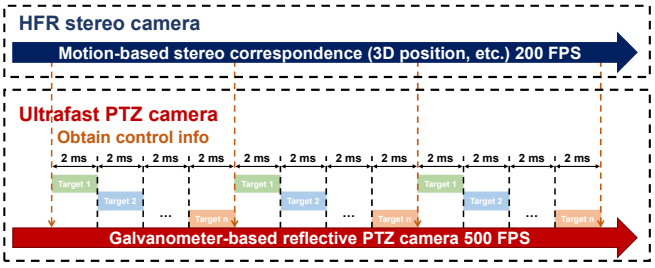


Fig. 2. Asynchronous architecture between multi-threaded gaze control of simultaneous multi-object zooming and motion-based stereo correspondence.

position of the galvanometer-based reflective PTZ camera was proposed in our previous paper [14]. The remainder of the paper is organized as follows. A detail algorithm analysis and concept illustration will be presented in Section II. Section III contains the fully experimental test platform description, and specific experimental design and process, followed by the discusssion of the experiments. Finally, the conclusion will be offered in Section IV.

II. ULTRAFAST MULTI-OBJECT ZOOMING SYSTEM BASED ON LOW-LATENCY STEREO CORRESPONDENCE

A. Concept

In this section, we introduce the framework of ultrafast multi-object zooming system based on low-latency stereo correspondence. Figure 1 shows the concept of the proposed hybird camera system. Unlike close-range, parallel-fixed dual-camera systems, in this paper we consider a more general scenario. There is a certain distance between the galvanometer-based reflective PTZ camera and

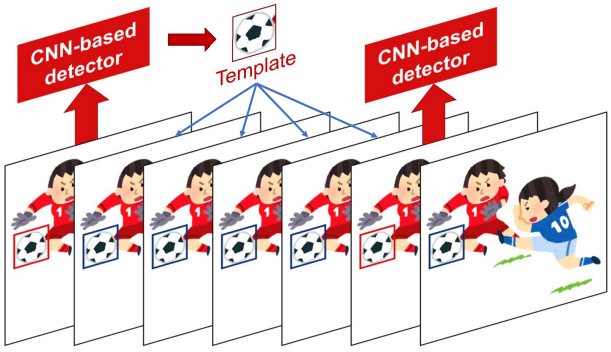


Fig. 3. CNN-based hybird template matching method.

the panoramic camera system and the their line of sight is not parallel. In this case, it is difficult to determine the rotation angle of the PTZ camera only by the 2D position of the object in the image. Owing to our work in [14], we can establish a mapping between spatial position and galvanometer-based reflective PTZ camera's rotation angle. As shown in Figure 1, in order to detect multiple objects and obtain their spatial positions, we add a HFR stereo camera to the ultrafast multi-object zooming system. However, efficiently establishing correspondence for the same moving object in stereo images poses a significant challenge. Traditional appearance-based stereo correspondence algorithms are easily affected by lighting, noise, and viewing angles, etc. The motion characteristics of different objects vary, making it feasible to utilize object-specific motion information for accurate correspondence. Figure 2 shows the asynchronous architecture between multi-threaded gaze control of simultaneous multi-object zooming and motion-based stereo correspondence. Among them, the high-speed tracking and stereo correspondence of the objects are completed at a speed of 200 FPS in the stereo camera, and the spatial position of each object is updated with a delay of 5 ms. And the galvanometer-based PTZ camera deflects the mirrors at a speed of 500 FPS to observe different moving objects from different angles. At the end of each observation period, the PTZ camera will obtain new position information from the HFR stereo camera.

B. CNN-based HFR tracking

While HFR stereo camera provides high frame rate and low-latency images, it also brings a huge computational load. In 2012, with the great success of AlexNet in the image recognition competition [15]. Various deep learning models based on convolutional neural network (CNN) have been developed for target detection, such as YOLO [16], SSD [17], RCNN [18], etc. These CNN-based algorithms have achieved good performance in the detection of various objects, and are very suitable for the detection of various objects. However, due to the multi-level neuron structure of CNN, it can only run at a speed of dozens of frames per second, which is difficult to meet our high-speed needs.

As depicted in Figure 3, our laboratory has introduced a CNN-based template matching method. Following each

AI detection cycle, images of detected objects are captured. These updated screenshots serve as templates for object template matching. The hybrid template matching algorithm, combining CNN and traditional methods, operates efficiently at hundreds of frames per second.

C. Motion-based stereo correspondence

In this study, we employ motion characteristics of objects rather than appearance-based features to achieve stereo correspondence. This approach reduces sensitivity to lighting and noise, and enhances the ability to distinguish between multiple objects with similar appearances. From the previous section, we can track different objects $O(t) = \{o_1(t), o_2(t), \ldots, o_J(t)\}$ in the panoramic stereo camera at 5 ms intervals. For the j-th detected object $(j = 1, \ldots, J)$, each detection result $o_j(t)$ is composed of the following parameters:

$$o_j(t) = \{x_j(t), y_j(t), w_j(t), h_j(t)\}, \tag{1}$$

where $x_j(t), y_j(t), w_j(t), h_j(t)$ represent the upper left corner coordinates, width, and height of the object detection box at time t, respectively. We calculate the velocity $v_j(t)$ of the j-th object by taking the center $c_j(t) = \{cx_j(t), cy_j(t)\}$ of the detection bounding box as the object's position. We sample N frames of images backward from the current moment. The set of object motion velocities between frames during this period is referred to as Short-Term Velocities (STVs). The STVs $V_j(t)$ of j-th object at time t is,

$$V_j(t) = \{v_j(t-N+1), \ldots, v_j(t-1), v_j(t)\}\,(n = 0, 1, \ldots, N-1), \tag{2}$$

where $v_j(t-n) = [dx_j(t-n), dy_j(t-n)]$ is the velocity in x and y direction at the pixel sacle before n frames.

Therefore, we can utilize the cosine distance between objects' STVs for stereo correspondence. Identical objects exhibit higher cosine similarity. Since the cosine distance only considers the directionality between vectors, in order to increase the influence of the motion amplitude on the cosine distance, we introduce the scale cosine distance s,

$$s = \frac{A \cdot B}{\max(|A|, |B|)^2}, \tag{3}$$

where $|A|$ and $|B|$ are the modulo lengths of the vector A and B. Therefore, the scale cosine distance S between N-dimensional STVs ${}^lV_i(t)$ and ${}^rV_j(t)$ is,

$$S(i,j) = \frac{1}{N}\sum_{m=0}^{N-1} \frac{{}^lv_i(t-m) \cdot {}^rv_j(t-m)}{\max\left(|{}^lv_i(t-m)|, |{}^rv_i(t-m)|\right)^2}. \tag{4}$$

The higher the motion similarity of objects, the larger the scale cosine distance between STVs, which is close to 1. After completing the correspondence of moving objects, the spatial positions $P_w = \{x_w, y_w, z_w\}$ of different objects can be calculated by triangulation. Finally, based on the mixed similarity of short-term velocities, a bipartite graph S can be reconstructed for I targets in the left camera and J targets in the right camera,

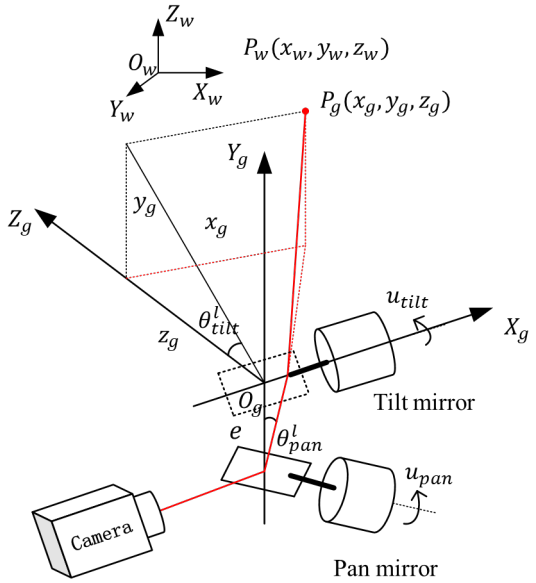


Fig. 4. Structure of galvanometer-based reflective PTZ camera.

$$S = \begin{bmatrix} S(1,1) & S(1,2) & \cdots & S(1,J) \\ S(2,1) & S(1,2) & \cdots & S(2,J) \\ \cdots & \cdots & \cdots & \cdots \\ S(I,1) & S(I,2) & \cdots & S(I,J) \end{bmatrix}. \tag{5}$$

Using the Hungarian matching algorithm, identical targets in stereo images can be quickly matched. Finally, spatial positions of different objects can be obtained through disparity.

D. Relationship between spatial position and PTZ camera's rotation angle

From the previous section, the spatial position P_w of the object has been determined. However, the critical task remains: how to convert P_w into the control voltages $U = \{u_{pan}, u_{tilt}\}$ for the galvanometer-based reflective PTZ camera.

Figure 4 shows the structure of the galvanometer-based reflective PTZ camera, in which the rotation axes of the pan mirror and tilt mirror of the galvanometer are perpendicular to each other. The camera was installed in a way parallel to the tilt mirror rotation axis, while perpendicular to the rotating axis of the pan mirror. When the galvanometer is in the initial state, the galvanometer coordinate system $O_gX_gY_gZ_g$ is established. The intersection of the optical axis and the rotation axis of the tilt mirror is the origin O_g, the rotation axis of the tilt mirror is the X_g axis, and the rotation axis of the pan mirror is the Z_g axis.

First, we need to convert point P_w in the world coordinate system to point P_g in the galvanometer coordinate system, and P_g is,

$$P_g = R \cdot P_w + T. \tag{6}$$

R and T is the rotation and translation matrix connecting the world coordinate system and the galvanometer coordinate

system. Next, the relationship between the point P_g in the galvanometer coordinate system and the control voltage $U = \{u_{pan}, u_{tilt}\}$ is,

$$\begin{cases} u_{tilt} = \frac{1}{2k_t} \arctan \frac{y_g}{z_g} \\ u_{pan} = \frac{1}{2k_p} \arctan \frac{x_g}{z_g \sec 2\theta_t + e}. \end{cases} \quad (7)$$

Among them, k_p and k_t are the linear coefficients between the control voltage of the pan mirror and the tilt mirror and the deflection angle, respectively. In [14], we proposed a flexible calibration method to solve the mapping relationship between spatial points and control voltage. So we can use the galvanometer-based reflective PTZ camera to capture the moving objects at the specified position in real time.

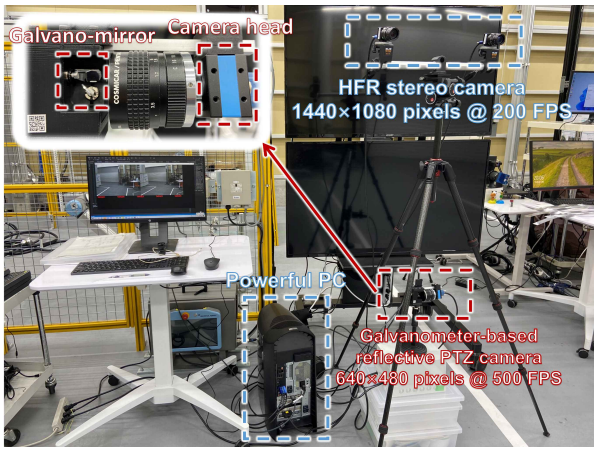
III. EXPERIMENTS

A. System Configuration

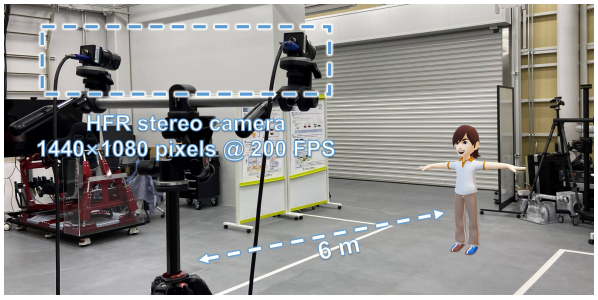
We extend the ultrafast multi-object zooming system by adding a panoramic HFR stereo camera. The system we proposed was divided into 2 parts: HFR motion-based object stereo correspondence and ultrafast pan-tilt zooming. The two parts work asynchronously at different frequencies. Motion-based object stereo correspondence is implemented online on the HFR stereo camera at 200 FPS and simultaneous multi-object zooming is implemented online at the galvanometer-based reflective PTZ camera at 500 FPS. The HFR stereo camera system consists of 2 high-speed camera USB 3.0 camera heads (DFK 37BUX273, Imaging Source Corp., Germany). The camera has a compact dimension of $36 \times 36 \times 25$ mm, weights 70 grams, without 6-mm lens, and can capture and transfer 10 bit color images of 1440×1080 pixels to RAM at 238 FPS via USB 3.0 interface. The galvanometer-based reflective PTZ camera consists of a high-speed USB 3.0 camera head (DFK37BUX287, Image Source, Germany), a 2-axis pan-tilt galvano-mirror (6210H, Cambridge Technology, US), and A D/A board (PEX-340416, Interface, Hiroshima, Japan) sent control signals to the galvano-mirror. The PTZ camera head attached with a 50 mm telephoto lens had a color CMOS sensor of 720×540 pixels, whose sensor and pixel sizes were 4.96×3.72 mm and 6.9×6.9 μm, respectively. It could capture and transfer 8-bit RGB 720×540 images at 539 fps to PC through a USB 3.1 interface. We used a powerful PC for driving the whole system, which have following hardware specifications: Intel Core i9-9900K @ 3.2 GHz CPU, 32 GB RAM, and a NVIDIA GeForce RTX 2080 Ti GPU. Figure 5(a) shows its overview. Figure 5(b) shows the next experimental environment. The distance between the person and the multi-target zooming system is about 6 meters, and they are doing different actions at the same time.

B. People in different actions indoors

We present the experimental results obtained when our system multi-zoomed multi persons who were doing different actions, such as shaking and jumping. Figure 6 shows the correspondence result in HFR stereo camera for $t = 20.50 - 21.00$ s. Among them, person 0 is shaking his body from left

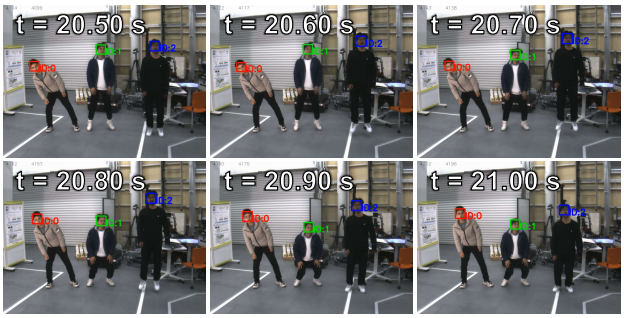


(a) Ultrafast multi-object zooming system based on low-latency stereo correspondence.

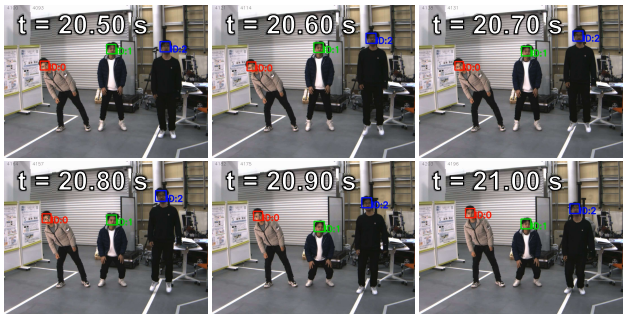


(b) Experiment environment for multi-object zooming.

Fig. 5. Overview of the experiment setup.

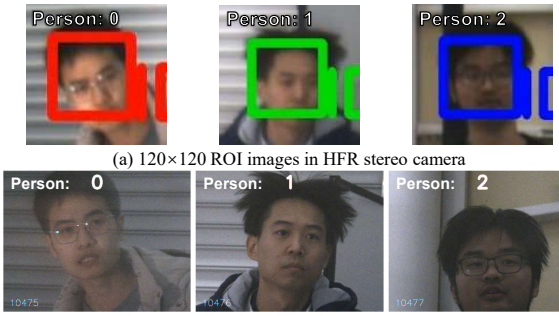


(a) Left image in HFR stereo camera.



(b) Right image in HFR stereo camera.

Fig. 6. The 1440 × 1080 input images and correspondence result.



(a) 120×120 ROI images in HFR stereo camera

(b) 640×480 input images of galvanometer-based PTZ camera

Fig. 7. Blurred 120×120 images of the faces in the panorama image and clear 640×480 images of the faces in the PTZ camera (t = 20.80 s).

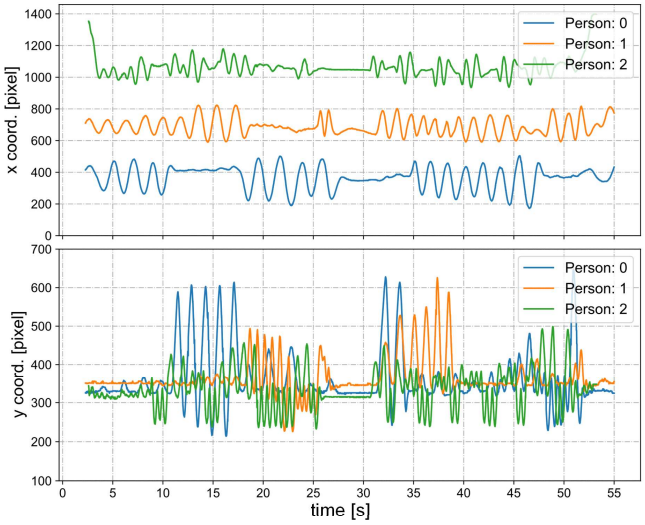
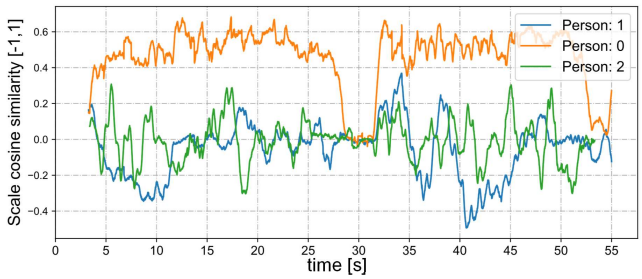
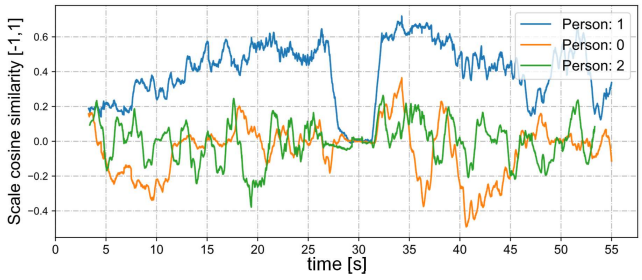


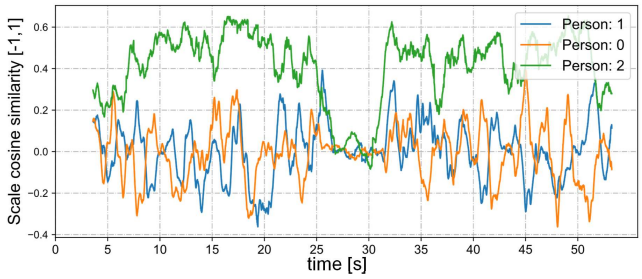
Fig. 8. Image centriods of faces in the left image of HFR stereo camera.



(a) Similarities between person 0 in the left image and all persons in the right image.



(b) Similarities between person 1 in the left image and all persons in the right image.



(c) Similarities between person 2 in the left image and all persons in the right image.

Fig. 9. Scale cosine similarities between persons in the HFR stereo camera.

to right, person 1 is squatting, and person 2 is jumping. Based on the stereo correspondence results, we draw rectangles of the same color for the same person in the HFR stereo images. Figure 7 shows the blurred 120×120 images of the faces in the panorama image and clear 640×480 images of the faces in the PTZ camera at time t = 20.80 s. We can clearly observe the facial features of different characters through the PTZ camera. The *xy* coordinate values of the image centroids of different faces in the left HFR stereo video are depicted in Figure 8. Figure 9 shows scale cosine similarities between different persons in the HFR stereo camera. We perform stereo correspondence by the scale cosine distance of STVs of different persons. Obviously, between 27 and 30 seconds, the similarity between different people is rapidly reduced, because different people stop moving at the same time. At other times, better performance was maintained. Figure 10 shows the 3D trajectories of different persons for $t = 20.50 - 21.10$ s. In the end, we show the control voltages of the galvo-based PTZ camera when tracking different people. The control voltage is completely driven by the spatial position obtained from the calculation in the previous step. The experimental results show that due to the low-latency detection results of the stereo camera, we can

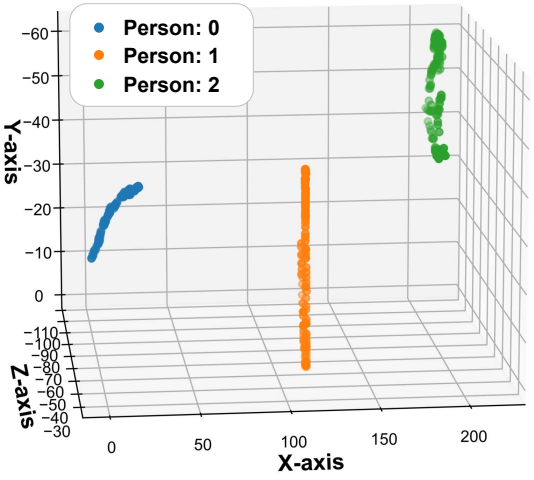


Fig. 10. The 3D trajectories of different persons (t = 20.5-21.1 s).

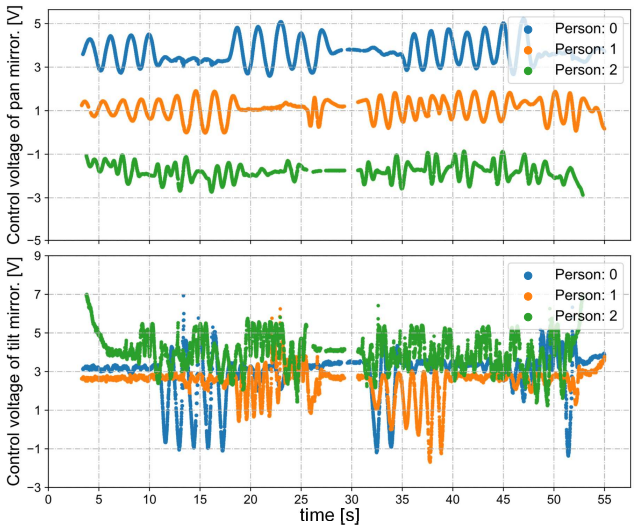


Fig. 11. Control voltages of pan and tilt mirror when tracking different persons.

complete the tracking and shooting of multiple objects in the center of the field of view of the PTZ camera at an extremely fast speed.

IV. CONCLUSIONS

In this study, we extend the simultaneous multi-object zooming system by adding a high-frame-rate (HFR) stereo camera. The newly proposed system can track and acquire the spatial positions of multiple objects at a speed of hundreds of frames per second, ultimately obtaining high-definition images of multiple objects simultaneously. The advantage of this system is that the galvanometer-based camera and the panoramic camera can be positioned arbitrarily, requiring only calibration prior to use. Furthermore, with the low-latency tracking provided by the high-frame-rate stereo camera, the deviation in images captured by the galvanometer-based camera is minimized.

However, due to the limitations of USB transfer speeds, high-resolution images can currently only be transmitted at approximately 200 FPS. We plan to consider a high-speed FHD camera capable of 500 FPS in future iterations. Additionally, to reduce computational load, the future system will operate in a heterogeneous mode where the server machine and the control machine are separated.

REFERENCES

[1] H. Proena and J. C. Neves, "Visible-wavelength iris/periocular imaging and recognition surveillance environments," *Image and Vision Computing*, vol. 55, pp. 22–25, 2016, recognizing future hot topics and hard problems in biometrics research. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0262885616300518

[2] S. Hu, K. Shimasaki, M. Jiang, T. Takaki, and I. Ishii, "A dual-camera-based ultrafast tracking system for simultaneous multi-target zooming," in *2019 IEEE International Conference on Robotics and Biomimetics (ROBIO)*, 2019, pp. 521–526.

[3] Y. Cha and S. Kim, "The error-amended sharp edge (ease) scheme for image zooming," *IEEE Transactions on Image Processing*, vol. 16, no. 6, pp. 1496–1505, 2007.

[4] R. O. Amoroso, A. M. Parma, J. Orensanz, and D. A. Gagliardini, "Zooming the macroscope: medium-resolution remote sensing as a framework for the assessment of a small-scale fishery," *ICES Journal of Marine Science*, vol. 68, no. 4, pp. 696–706, 2011.

[5] S. Farsiu, M. D. Robinson, M. Elad, and P. Milanfar, "Fast and robust multiframe super resolution," *IEEE transactions on image processing*, vol. 13, no. 10, pp. 1327–1344, 2004.

[6] C.-H. Chen, Y. Yao, D. Page, B. Abidi, A. Koschan, and M. Abidi, "Heterogeneous fusion of omnidirectional and ptz cameras for multiple object tracking," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 18, no. 8, pp. 1052–1063, 2008.

[7] R. E. Carlson and F. N. Fritsch, "Monotone piecewise bicubic interpolation," *SIAM journal on numerical analysis*, vol. 22, no. 2, pp. 386–400, 1985.

[8] S. Bell-Kligler, A. Shocher, and M. Irani, "Blind super-resolution kernel estimation using an internal-gan," *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[9] A. Bulat, J. Yang, and G. Tzimiropoulos, "To learn image super-resolution, use a gan to learn how to do image degradation first," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 185–200.

[10] W. Yang, X. Zhang, Y. Tian, W. Wang, J.-H. Xue, and Q. Liao, "Deep learning for single image super-resolution: A brief review," *IEEE Transactions on Multimedia*, vol. 21, no. 12, pp. 3106–3121, 2019.

[11] 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, 2014.

[12] P. Natarajan, P. K. Atrey, and M. Kankanhalli, "Multi-camera coordination and control in surveillance systems: A survey," *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, vol. 11, no. 4, pp. 1–30, 2015.

[13] 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.

[14] Q. Li, M. Chen, Q. Gu, and I. Ishii, "A flexible calibration algorithm for high-speed bionic vision system based on galvanometer," in *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2022, pp. 4222–4227.

[15] F. N. Iandola, S. Han, M. W. Moskewicz, K. Ashraf, W. J. Dally, and K. Keutzer, "Squeezenet: Alexnet-level accuracy with 50x fewer parameters and¡ 0.5 mb model size," *arXiv preprint arXiv:1602.07360*, 2016.

[16] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 779–788.

[17] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, "Ssd: Single shot multibox detector," in *European conference on computer vision*. Springer, 2016, pp. 21–37.

[18] R. Girshick, "Fast r-cnn," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 1440–1448.