Proceeding of the IEEE
International Conference on Robotics and Biomimetics
Dali, China, December 2019

A Dual-Camera-Based Ultrafast Tracking System for Simultaneous Multi-target Zooming

Shaopeng Hu and Kohei Shimasaki
Digital Monozukuri (Manufacturing) Education Research Center
Hiroshima University
3-10-32 Kagamiyama, Higashihiroshima,
Hiroshima 739-0046, Japan
{*hu, simasaki*}*@robotics.hiroshima-u.ac.jp*

Mingjun Jiang, Takeshi Takaki, and Idaku Ishii
Department of System Cybernetics
Hiroshima University
1-4-1 Kagamiyama, Higashihiroshima,
Hiroshima 739-8527, Japan
{*m-jiang, takaki, iishii*}*@robotics.hiroshima-u.ac.jp*

***Abstract*— In this paper, we develop a novel dual-camera system that can simultaneously capture zooming-in targets by combining an ultrafast pan-tilt camera and a fixed wide-view camera. According to the positions of all the targets recognized with deep learning in the wide-view camera, the pan and tilt angles of multiple pan-tilt cameras are controlled virtually on the ultra-fast pan-tilt camera through multi-thread viewpoint control to simultaneously capture the zooming-in images of all the targets. Our system can generate five hundred virtual cameras with different zooming views in a second, and the effectiveness of our system is demonstrated by showing experimental results in simultaneous zoom shooting for multiple running persons and cars in the range of 70 m or more in a natural outdoor scene.**

***Index Terms*— multi-target zooming, high-speed vision, ultrafast viewpoint control, dual camera system**

I. Introduction

Zooming-in on targets of interest, is an important technology to provide detailed visual information in fields of surveillance and monitoring for such as crime prevention, transportation safety, facial recognition, and human-computer interaction [1], [2], [3], [4]. Generally, the purpose of zooming is to achieve high-resolution target images with high quality. Optical zooming is an operation within a camera system with a high-magnification lens to observe a wider view of field without decreasing resolution. PTZ cameras that can mechanically change their views in the pan and tilt directions are well-used optical-zooming devices for capturing moving scenes in video surveillance applications. Many types of multi-camera systems [5] have been developed for multi-target zooming, and these are usually made up of multiple stationary cameras or PTZ cameras. Multiple PTZ cameras assisted by a wide-angle fixed camera are becoming a popular choice and have been rapidly developed for various applications. A fixed wide-angle camera can cover a large area to detect or track moving objects in a scene; then the PTZ camera will be controlled to zoom in on the object. The main challenge is how to schedule and control the PTZ cameras, especially when the number of targets changes or exceeds the number of active cameras [6], [7]. Moreover, the physical size of PTZ cameras and complexities in the control software should be considered as more serious problems as the number of PTZ cameras increases in a multi-camera network.

Recently, an ultrafast gaze control system that uses accelerated galvanomirrors [8], [9] was developed, and it was extended to function as multiple virtual pan-tilt cameras to observe simultaneously different views by switching their viewpoints hundreds of times in a second [10]. Its effectiveness was demonstrated in large-structure vibration sensing [10], [11] and monocular stereo sensing [12], [13]. If a single ultrafast pan-tilt camera could simultaneously observe different scenes distributed in a wider field of view instead of multiple PTZ cameras, the installation and management cost of the video surveillance system could be remarkably reduced for precise wide-area monitoring.

In this paper, we develop a concept of dual-camera-based simultaneous multi-target zooming system including a wide-view camera and an ultrafast mirror-drive pan-tilt camera. A wide-view camera with object recognition roughly detects the locations of multiple objects distributed in a wide area with CNN-based recognition at 25 fps, and the ultrafast pan-tilt camera with a telephoto lens functions as dozens of virtual tracking cameras to zoom in simultaneously on the detected multiple objects by accelerating its sensing, computation, and actuation with ultrafast gaze switching at 500 Hz. Its effectiveness is demonstrated by showing experimental results for multiple people and cars in a wide-area outdoor scene.

II. Concept

Multi-object tracking is an important methodology for localizing multiple objects in video sequences in computer vision applications. Most of these methods are based on target features captured by executing feature extraction and classification for sub-image regions selected when scanning the whole image with a sliding window.

Problems occurred in recognition accuracy because handcrafted low-level feature descriptors and discriminatively trained shallow models cannot sufficiently describe the semantic relationship in images when observing complex scenes. Recently, deep learning technologies have brought about remarkable advances in multi-object detection. CNN is the most representative model of deep learning and CNN-based

978-1-7281-6321-5/19//$31.00 ©2019 IEEE

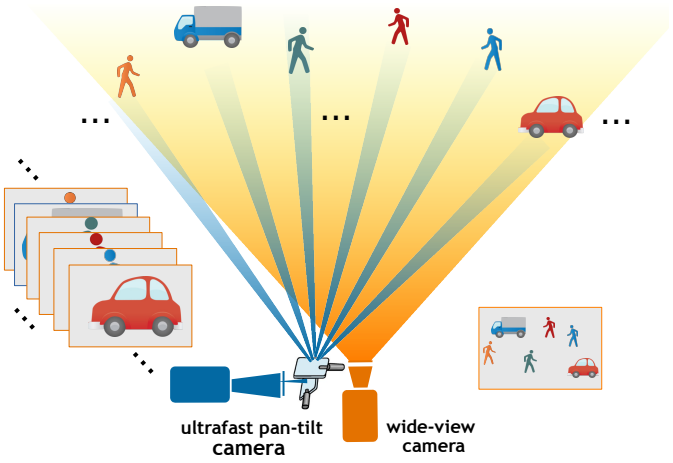

Fig. 1. Simultaneous multi-object zooming with multithread gaze control.

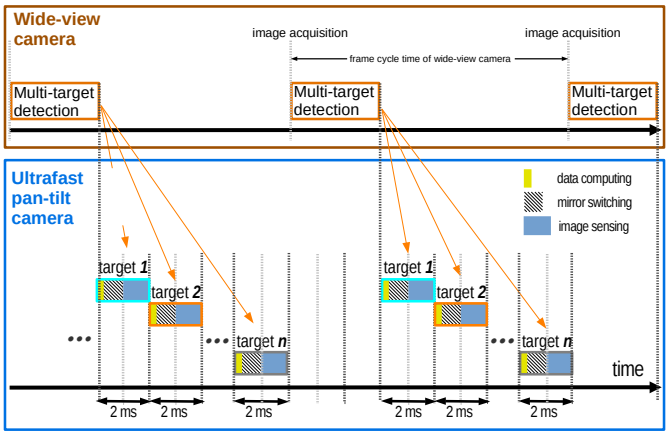

Fig. 2. Time-division thread processes for virtual pan-tilt cameras in multi-thread gaze control for simultaneous multi-object zooming.

methods such as YOLO [14] and SSD [15] can detect multiple objects in an image even when they appear in the low-resolution regions dozens-of-pixels wide. However, they cannot always obtain high-resolution images at the same time when a single-camera system is used for precise wide-area monitoring due to insufficient spatial resolution of the image sensor and camera lens.

In order to solve this problem for single-camera-based wide-area monitoring, we propose a concept of simultaneous multi-object zooming by introducing ultrafast active vision with a telephoto lens. Fig. 1 shows the concept of the proposed dual-camera system including an ultrafast mirror-drive pan-tilt camera (consisting of a high-speed camera and an ultrafast pan-tilt mirror device) with multithread gaze control to generate virtual cameras at hundreds of fps and a wide-view camera with multi-object detection at dozens of fps. These virtual cameras can be unified to behave as integrated real cameras and they can be controlled with great rapidity compared with standard real pan-tilt cameras.

Multithread gaze control enables an ultrafast pan-tilt camera to observe different views at the same time by parallelizing a series of operations with video shooting, processing, and gaze-control into time-division-thread processes with a fine temporal granularity. Fig. 2 shows the time-division thread processes for virtual pan-tilt cameras in multi-thread gaze control for simultaneous multi-object zooming. After the wide-view camera detects multiple targets in real time, the ultrafast pan-tilt camera can generate multiple dozens-of-fps pan-tilt cameras that will zoom in on the multiple objects distributed in a wide area according to these time-division thread processes.

Compared with multi-PTZ-camera-based surveillance systems, simultaneous multi-object zooming with multi-thread gaze control on such a dual-camera system has the following advantages: (1) cost and space saving because only two cameras are used, (2) easy software-expandability that enables multi-target zooming without considering the time-varying target number and camera schedule, and (3) easy and precise calibration that can reduce the complexity in camera calibration and enables precise multi-target zooming because the camera internal parameters such as focal length, gain, and exposure time, are the same in multiple virtual pan-tilt cameras.

III. Dual-Camera-Based Simultaneous Multi-Target Zooming System

A. System Configuration

We developed a dual-camera-based ultrafast tracking system for simultaneous multi-target zooming system. Our developed system was divided into two parts: wide-view multi-target detection and ultrafast pan-tilt zooming. The two parts used different personal computers (PCs) coordinated through user datagram protocol (UDP) based communication. Fig. 3 shows an overview of the hardware system.

For wide-view multi-target detection, a USB 3.1 camera (DMK37BUX287, Imaging Source, Germany) functioned as a wide-view camera to capture the whole scene, and the focal distance of its lens was set to 8.5 mm. The PC for the wide-view camera had installed GPGPU for deep learning acceleration, an Intel Core i7-6850K CPU, GPGPU (GeForce GTX Titan Xp, NVIDIA, US), OS (Windows 10 Pro 64 bit), and RAM (32 GB DDR4).

The ultrafast pan-tilt camera consisted of an ultrafast pan-tilt mirror and a high-speed camera with a telephoto lens. The pan-tilt mirror device including two galvano-mirrors (6210H, Cambridge Technology, US) could control two degrees-of-freedom gazes and switch to any pan and tilt direction within 1 ms. A USB 3.1 camera (DFK37BUX287, Imaging Source, Germany) functioned as a high-speed camera operating at 500 fps, which means that 500 virtual cameras could be generated in one second. A $f = 200$ mm telephoto lens was attached to the camera head for zooming in. The specifications of the PC controlling the pan-tilt mirror and high-speed camera were CPU (Intel Xeon E5-1650v4), OS (Windows 7 Enterprise 64 bit), and RAM (16 GB DDR4). The pan-tilt mirror device was controlled through A/D and D/A boards. The sizes of the pan and tilt mirrors were 10.2 mm^2 and 17.5 mm^2, respectively,

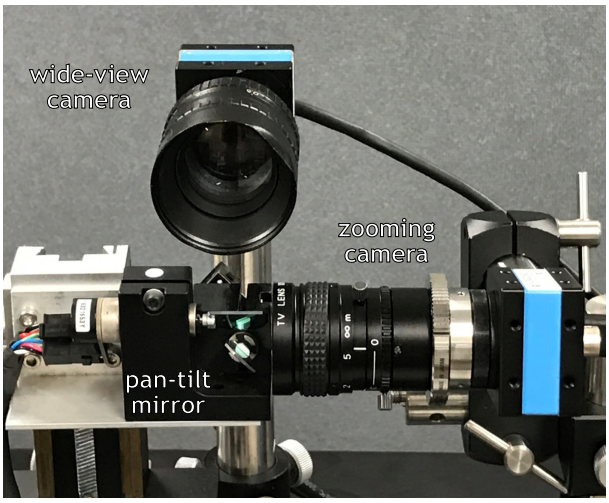

Fig. 3. Dual-camera-based multi-target zooming system.

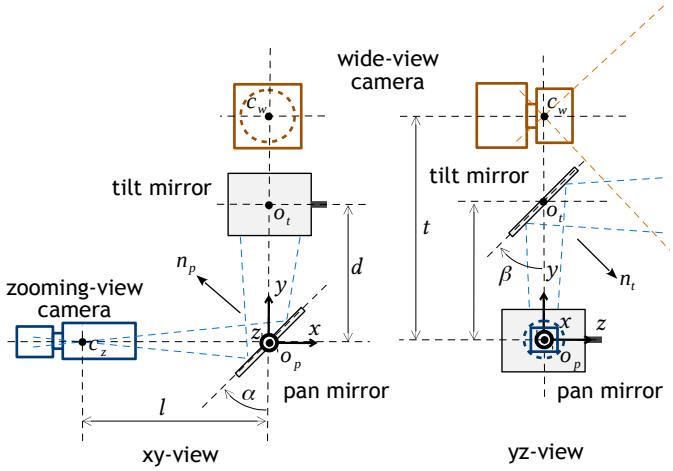

Fig. 4. Geometry of dual-camera-based multi-target zooming system.

and the view angles of the pan and tilt mirrors were movable and could cover the area within the range of -20 to 20 degrees.

B. Geometry and Algorithms

Here, we describe the geometry of our system as illustrated in Fig. 1. It includes the locations and orientations of virtual cameras formed by the ultrafast pan-tilt camera, and the relationships between each virtual camera and the wide-view camera. To describe these relationships, we defined the xyz-coordinate of the whole system as shown in Figure 4. We present the xy-view and yz-view of the geometrical configuration of the dual-camera pan-tilt mirror system. The x-axis is defined as the optical axis of the camera crossing the central area of the pan mirror, and the y-axis is defined as the line connecting the center points of the pan mirror and tilt mirror. The target zooming direction corresponds to the z-axis.

The origin of the xyz-coordinate system was set to the center of the pan mirror $o_p = (0,0,0)^T$. The pan mirror could switch along the z-axis, and its normal vector was defined as $n_p = (-\cos\alpha, \sin\alpha, 0)^T$. The center of the tilt mirror was located at $o_t = (0,d,0)^T$, where d was the distance between the centers of the pan and tilt mirrors. The tilt mirror could switch around a straight line parallel to the x-axis at a distance d, and its normal vector was defined as $n_t = (0, -\sin\beta, \cos\beta)^T$. The angle α and β mean the pan and tilt angles (respectively) used to control the position and orientation of each virtual camera. The optical center of the high-speed camera with zooming lens was set as $c_z = (-l,0,0)^T$, where l is the distance from the center of the pan mirror along the x-direction. The optical center of wide-view camera was set as $c_w = (0,t,0)$, where t presents the distance from the center of the pan mirror along the y-direction.

The virtual camera can be generated by pan and tilt mirror reflection [12], which can be described as cascaded planar mirror reflections using the above geometrical relationship. The final position $\boldsymbol{c}_v$ and orientation $\boldsymbol{q}_v$ of virtual camera $\boldsymbol{v}(\boldsymbol{c}_v, \boldsymbol{q}_v)$ can be described as follows:

$$\boldsymbol{c}_v(\alpha,\beta) = \begin{pmatrix} l\cos 2\alpha \\ -(l\sin 2\alpha + d)\cos 2\beta + d \\ -(l\sin 2\alpha + d)\sin 2\beta \end{pmatrix}, \tag{1}$$

$$\boldsymbol{q}_v(\alpha,\beta) = \begin{pmatrix} -\cos 2\alpha \\ \sin 2\alpha \cos 2\beta \\ \sin 2\alpha \sin 2\beta \end{pmatrix}. \tag{2}$$

Multiple virtual pan-tilt zooming cameras were controlled according to the positions of multiple targets in the image captured by the wide-view camera. The virtual pan-tilt camera and fixed wide-view camera should be calibrated for the purpose of zooming up specified targets accurately. The virtual camera c_v and wide-view camera c_w are in the same world coordination system O-XYZ. The wide-view camera is always fixed in the same place and its three axes are parallel to the axes of the world coordinate, while the position and orientation of the virtual camera were changing according to different pan and tilt angles.

The target $P(X_p, Y_p, Z_p)$ was detected using a CNN-based method, and we can obtained its position in the pixel coordinate uv-system as $u(u_p, v_p)$. According to the intrinsic parameters of the wide-view camera under the image coordinate xy-system, the point $p_w(x_w, y_w)$ projected on the imaging plane was calculated. Then, pan and tilt mirrors were controlled to form a virtual camera so that the target was projected at the center of the virtual camera view according to the geometric relationship. For both virtual cameras and the actual wide-view camera, the intrinsic parameters such as focal length and optic center were constant. We assumed that the targets to be zoomed were sufficiently distant from our system; this distance was much longer than the distance from the optical center of the virtual camera to that of the wide-view camera.

The ultrafast pan-tilt camera of our system could function as N number of virtual pan-tilt cameras for zooming in on multiple targets. The number of targets N and each target's

pixel position $\boldsymbol{p}_n(u_n, v_n)$, which are detected in the wide-view camera, were expressed as

$$\boldsymbol{D}_w = \{N, \boldsymbol{p}_1(u_1, v_1), \boldsymbol{p}_2(u_2, v_2), \cdots, \boldsymbol{p}_n(u_n, v_n)\}, \quad (3)$$

and $\boldsymbol{D}_w$ was simultaneously transferred to the ultrafast pan-tilt camera for multithread gaze control. After receiving the data, N number of virtual cameras functioned with frame-by-frame gaze control to zoom in on each target simultaneously. For each virtual camera $\boldsymbol{v}_n$, computing with calibration data was required to obtain the pan and tilt angles $\boldsymbol{\theta}_n(\alpha_n, \beta_n)$ using $\boldsymbol{p}_n(u_n, v_n)$:

$$\boldsymbol{\theta_n}(\alpha_n, \beta_n) = \boldsymbol{f}(\boldsymbol{p_n}(u_n, v_n), \boldsymbol{g}_{vw}), \quad (4)$$

where $\boldsymbol{g}_{vw}$ represents the geometric relationship between the virtual camera and wide-view camera, including the extrinsic parameters such as locations in world coordinates and intrinsic parameters such as focal lengths. Note that for the wide-view camera, we have to correct the lens distortion which is easy because of fixed focal length; and for the zooming-view it can zoom the area within the depth of field because the focal length is also fixed, thus we will consider using zoom lens or combining optical and digital zoom together in the future.

IV. Experimental Results

To verify the performance of our system, we conducted an experiment to zoom in on persons and cars running in an outdoor scene. In the experiment, we used YOLOv3 [16] as a CNN-based object detector for multi-target detection and localization in the image. YOLOv3 has obvious advantages for small object detection, and was suitable for the detection of targets in our wide-view camera images.

Fig. 5 shows the experimental environment. We detected and zoomed in on cars and persons 70 m or more in front of our system, observing them as different types of targets. Five cars remained in place, while three persons were running at different speeds in an outdoor scene. Fig. 6 shows examples of the experimental results of multi-object detection and simultaneous zooming for the targets. The image resolutions in both the wide-view camera and the simultaneous zooming camera were set to 640×480 pixels. As shown in Fig. 6, the resolutions of detected targets in the wide-view camera were very low and the targets were too vague to be manually authenticated. In spite of the low resolution, YOLOv3 could still detect multiple targets. The positions of all the targets detected in the wide-view camera were transmitted to the zooming PC, and multiple virtual cameras were used in order to simultaneously zoom in on all the targets. Thus, the five cars and three persons could be clearly observed in the zoomed-in images, as shown in Fig. 6.

Fig. 7 shows multiple target detection and zooming results from 0.1 s to 6.1 s at intervals of 1 s. Three persons were running in different directions, and the ultrafast pan-tilt mirror switched to generate multiple virtual cameras on the basis of multi-thread control for all the targets, zooming according to their positions in the image. Fig. 7 shows the target detection results in the wide-view camera and the zooming results of

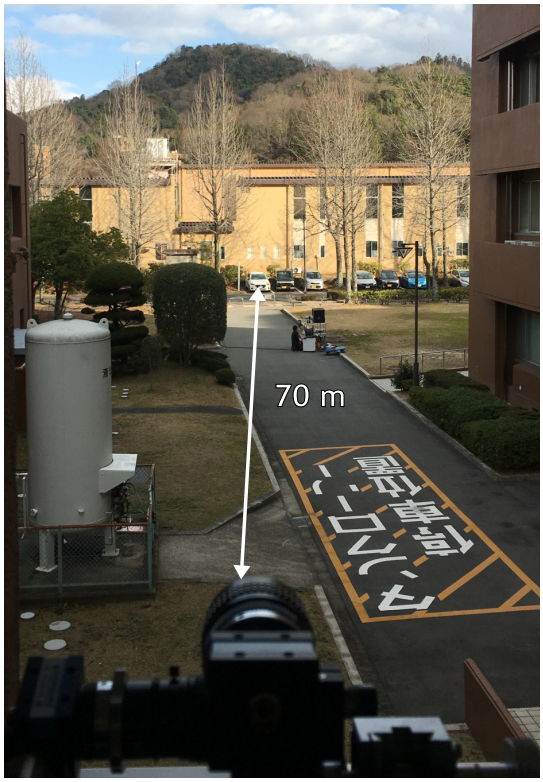

Fig. 5. Experimental environment for wide-area multi-target zooming.

three running persons and a static car. Person 1 and Person 2 were in the centers of the zoomed-in images, and Person 3 was far from the image center. This is because Person 3 ran so quickly that there was displacement during the delay time in the CNN-based detection and unsynchronized viewpoint switching became non-negligible. The zooming images indicate that static and moving targets can be simultaneously tracked and zoomed using multiple virtual cameras.

Fig. 8 shows the centroid positions of all the targets in the wide-view image with Yolov3. u and v refer to the image coordinates in horizontal and vertical directions, respectively, of the wide-view camera. The centroid positions of five static cars basically remained stable, while the centroid positions of the three running persons varied with time according to their running status. Fig. 9 shows the pan and tilt angles of each virtual camera operated using the ultrafast pan-tilt camera with multi-thread gaze control. These pan and tilt angles were controlled to correspond with the centroid positions u and v detected in the wide-view camera. Note that the pan and tilt angles for Person 3 were broken off as shown in the red boxes in Fig. 9; Person 3 was not detected at this moment because of occlusion. When Person 3 was not detected, the number of targets to be detected changed from 8 to 7, and the number of virtual cameras rapidly changed from 8 to 7. The gaze control thread for Person 3 kept the last image captured before occlusion until the person was detected again with Yolov3. These results indicate that our system can simultaneously zoom in on multiple static and moving persons and objects,

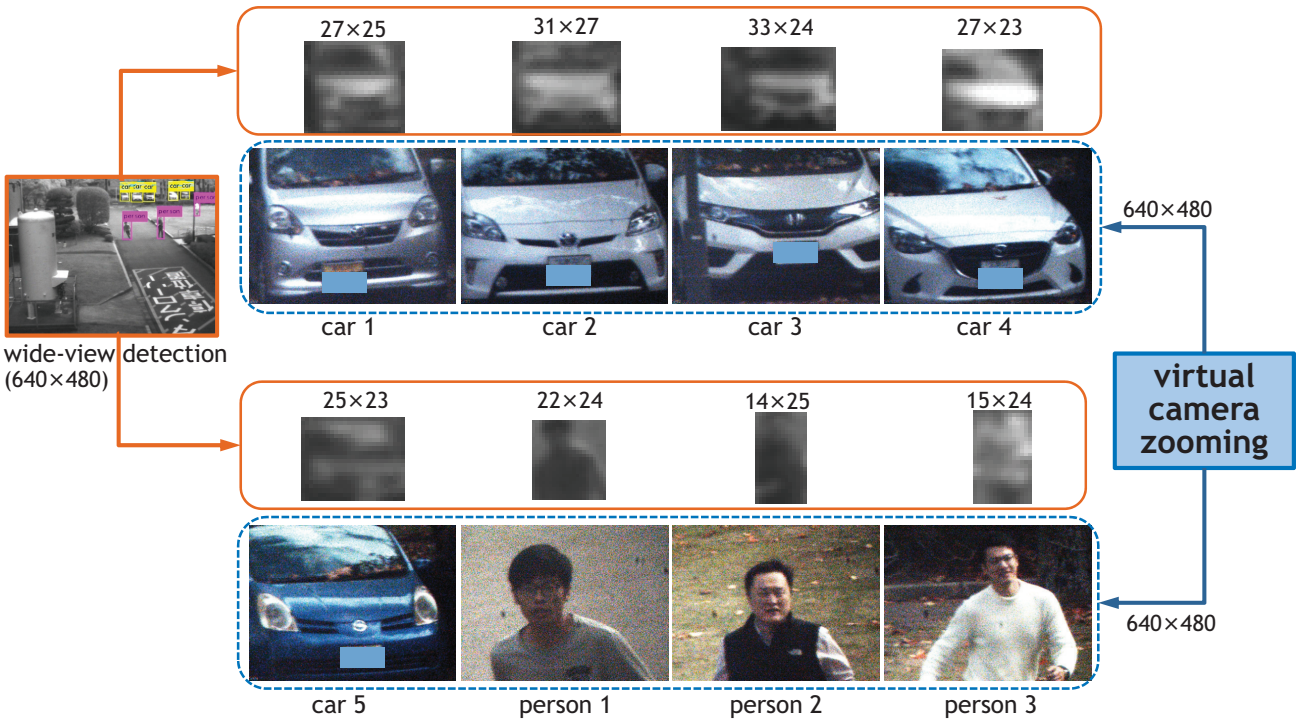

Fig. 6. Multiple persons and cars detected in wide-view camera and their zoomed-in images in the ultrafast pan-tilt camera.

Fig. 7. Wide-view images with object recognition and zoomed-in images for three running persons and a car for 6 s at intervals of 1 s.

and that the number of virtual zooming cameras can be flexibly adjusted according to the number of targets to be detected in the wide-view camera.

V. Conclusions

In this study, we developed a novel dual-camera-based simultaneous multi-target zooming system based on an ultrafast pan-tilt camera that can generate multiple virtual cameras and

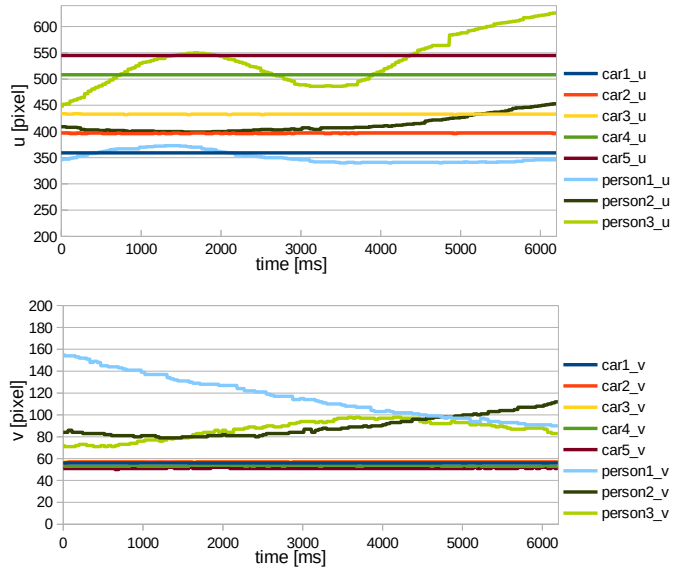

Fig. 8. Centroid positions of targets in the wide-view camera.

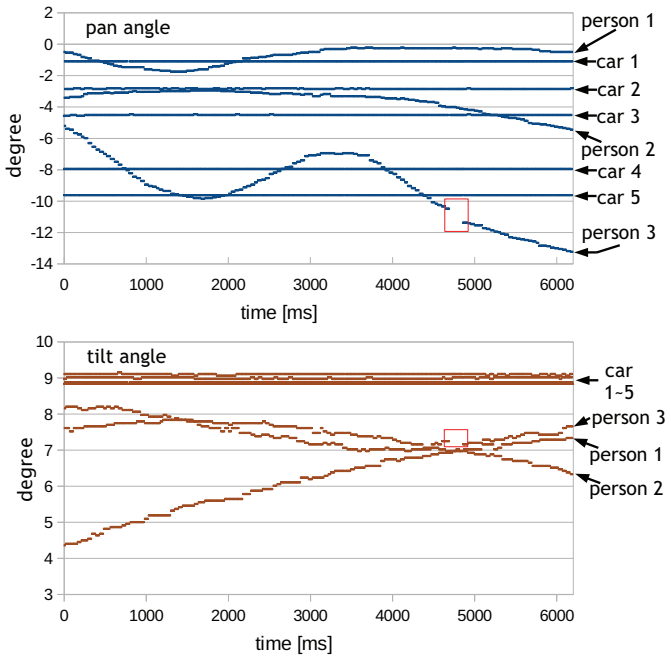

Fig. 9. Pan and tilt angles of virtual cameras operating in the ultrafast pan-tilt camera.

a wide-view camera with CNN-based multi-object detection at dozens of fps. It has the advantage that there is no need to physically add cameras even when the number of targets to be observed changes dynamically, including a case when the number of targets is much larger than the number of cameras. This is because our system can generate 500 virtual cameras in a second with ultrafast multithread gaze control. The experimental results verified the effectiveness of our system to simultaneously track and zoom multiple different types of static and moving targets using virtual cameras without considering the number of targets.

Currently, the main challenge that remains with our system is to handle large deviations in zooming images when targets move faster, due to latency during the CNN-based recognition and gaze unsynchronized viewpoint switching. In the future, we intend to improve our system to track precisely and zoom in on fast-moving targets by implementing a fast local tracking algorithm with CNN-based recognition that can compensate for the latency in object recognition and viewpoint switching.

ACKNOWLEDGMENT

This work was supported in part by JST ACCEL Grant Number JPMJAC1601 and JSPS KAKENHI Grant Number 16H02348.

REFERENCES

[1] H. Proena and J. C. Neves, "Visible-wavelength iris/periocular imaging and recognition surveillance environments," *Image Vis. Comput.*, vol. 55, pp. 22–25, 2016.

[2] K. T. Song and J. C. Tai, "Dynamic calibration of pantiltzoom cameras for traffic monitoring," *IEEE Trans. Syst. Man Cybern. B Cybern.*, vol. 36, no. 5, pp. 1091–1103, 2006.

[3] U. Park, H. C. Choi, A. K. Jain and S. W. Lee, "Face tracking and recognition at a distance: A coaxial and concentric PTZ camera system," *IEEE Trans. Inf. Forens. Secur.*, vol. 8, no. 10, pp. 1665–1677, 2013.

[4] L. Fiore, D. Fehr, R. Bodor, A. Drenner, G. Somasundaram, and N. Pa-panikolopoulos, "Multi-camera human activity monitoring," *J. Intell. Robot. Syst.*, vol. 52, no.1, pp. 5–43, 2008.

[5] C. Ding, B. Song, A. Morye, J. A. Farrell, and A. K. Roy-Chowdhury, "Collaborative sensing in a distributed PTZ camera network, *IEEE Trans. Image Process.*, vol. 21, no. 7, pp. 3282–3295, 2012.

[6] Y. Xu and D. Song, "Systems and algorithms for autonomous andscalable crowd surveillance using robotic PTZ cameras assisted by a wide-angle camera, *Auton. Robot.*, vol. 29, no. 1, pp. 53–66, 2010.

[7] P. Natarajan, P. K. Atrey, and M. Kankanhalli, "Multi-camera coordination and control in surveillance systems : A survey, *ACM Trans. Multimed. Comput. Commun.*, vol. 11, no. 4, pp. 57, 2015.

[8] K. Okumura, H. Oku and M. Ishikawa,"High-speed gaze controller for millisecond-order pan/tilt camera," *Proc. IEEE Int. Conf. Robot. Automat.*, 2011, pp. 6186–6191.

[9] K. Kobayashi-Kirschvink and H. Oku, "Design principles of a high-speed omni-scannable gaze controller", *IEEE Robot. Automat. Lett.*, vol.1, vo. 2, pp. 836–843, 2016.

[10] T. Aoyama, L. Li, M. Jiang, K. Inoue, T. Takaki, J. Ishii, H. Yang, C. Umemoto, H. Matsuda, M. Chikaraishi and A. Fujiwara, "Vibration sensing of a bridge model using a multithread active vision system," *IEEE/ASME Trans. Mechatr.*, vol. 23, no. 1, pp. 179–189, 2018.

[11] T. Aoyama, L. Li, M. Jiang, T. Takaki, I. Idaku, H. Yang, C. Umemoto, H. Matsuda, M. Chikaraishi, and A. Fujiwara, "Vision-based modal analysis using multiple vibration distribution synthesis to inspect large-scale structures", *J. Dyna. Syst. Meas. Contr.*, vol. 141, no. 3, pp. 031007-1–031007-12, 2019.

[12] S. Hu, Y. Matsumoto, T. Takaki, and I. Ishii, "Monocular stereo measurement using high-speed catadioptric tracking," *Sensors*, vol. 17, no. 8, pp. 1839, 2017.

[13] S. Hu, J. Ming, T. Takaki, and I. Ishii, "Real-time monocular three-dimensional motion tracking using a multithread active vision system," *J. Robot. Mechatron.*, vol. 30, no. 3, pp. 453-466, 2018.

[14] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only lookonce: Unified, real-time object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2016, pp. 779–788.

[15] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, "SSD: single shot multibox detector," *Proc. Euro. Conf. Comput. Vis.*, 2016, pp. 21–37.

[16] J. Redmon and A. Farhadi, "Yolov3: An incremental improve-ment. arXiv preprint," *arXiv:1804.02767*, 2018.