PHOTONICS Research

Single-shot 3D tracking based on polarization multiplexed Fourier-phase camera

JIAJIE TENG,[1,2] CHENGYANG HU,[1,2] HONGHAO HUANG,[1,2] MINGHUA CHEN,[1,2] SIGANG YANG,[1,2] AND HONGWEI CHEN[1,2,*]

[1]*Beijing National Research Center for Information Science and Technology (BNRist), Beijing 100084, China*
[2]*Department of Electronic Engineering, Tsinghua University, Beijing 100084, China*
**Corresponding author: chenhw@tsinghua.edu.cn*

Received 25 May 2021; revised 27 July 2021; accepted 27 July 2021; posted 29 July 2021 (Doc. ID 432292); published 13 September 2021

For moving objects, 3D mapping and tracking has found important applications in the 3D reconstruction for vision odometry or simultaneous localization and mapping. This paper presents a novel camera architecture to locate the fast-moving objects in four-dimensional (4D) space (x, y, z, t) through a single-shot image. Our 3D tracking system records two orthogonal fields-of-view (FoVs) with different polarization states on one polarization sensor. An optical spatial modulator is applied to build up temporal Fourier-phase coding channels, and the integration is performed in the corresponding CMOS pixels during the exposure time. With the 8 bit grayscale modulation, each coding channel can achieve 256 times temporal resolution improvement. A fast single-shot 3D tracking system with 0.78 ms temporal resolution in 200 ms exposure is experimentally demonstrated. Furthermore, it provides a new image format, Fourier-phase map, which has a compact data volume. The latent spatio-temporal information in one 2D image can be efficiently reconstructed at relatively low computation cost through the straightforward phase matching algorithm. Cooperated with scene-driven exposure as well as reasonable Fourier-phase prediction, one could acquire 4D data (x, y, z, t) of the moving objects, segment 3D motion based on temporal cues, and track targets in a complicated environment.

https://doi.org/10.1364/PRJ.432292

1. INTRODUCTION

Moving targets tracking in 3D space has found many applications in various fields, such as vehicle navigation, 3D reconstruction [1], and 3D motion estimation [2]. The most widely known image-free 3D sensor is LiDAR [3,4], which generally utilizes a laser as the active lighting source and high-bandwidth detector with complex data processing to achieve long-distance and high-speed 3D detection. However, due to the costly and complicated system architecture, it is hard to be widely utilized in normal 3D vision tasks. In contrast, image-based sensors or systems are with relatively low cost and common for a wider range of 3D visual applications, such as stereo vision [5,6] and monocular 3D systems [7], which allow the moving objects to track in different scales. Nevertheless, image-based systems often require a series of operations including camera pose estimation, multi-view calibration, feature extraction, and similarity matching, which increase the computational burden of the real-time 3D tracking.

With the advance of novel visional sensors, optoelectronics hybrid devices, and post-reconstruction algorithm, a serious of new camera architectures have emerged to tackle the challenging scenarios that are inaccessible to the traditional sensors. An event camera with high temporal resolution and high dynamic range provides a new data format with pixel-wise intensity changes asynchronously. This time-sparse event data reduces the power and bandwidth requirements and allows it to work in real-time 3D reconstruction for various vision applications [8,9]. For releasing the hardware requirement, a single-pixel detector is also applied to achieve a real-time image-free 3D tracking system [10]. Although this scheme faces trouble in the multi-objects tracking scenarios, it is still a low-cost and computation-efficient system. On the other hand, with the popularity of multi-dimensional encoded optoelectronic modulation devices, computational photography shows great potential in 3D reconstruction with single-shot stereo images [11–13], which is capable of cooperating with compressive sensing and adaptive reconstruction algorithms. Moreover, the lensless imaging system with a diffuser placed in front of the traditional sensor is demonstrated to achieve a single-shot 3D reconstruction in Refs. [14,15]. However, the effective resolution and computational overhead vary significantly with scene content and limit its practical application.

In this paper, we propose a four-dimensional (4D) information recording camera with multiplexed orthogonal polarization field-of-view (FoV), named polarization multiplexed Fourier-phase camera (PM-FPC). It is a novel camera framework, which is capable of reconstructing 4D data of moving objects in a single shot. The principle of PM-FPC is to perform pixel-wise optical coding on polarization multiplexed scenes to acquire the Fourier-phase maps of two orthogonal perspectives in one exposure. With the 8 bit grayscale quantized sinusoid modulation, the temporal resolution of the camera is increased to 256 times. Compared to the traditional image-based 3D stereo systems, it processes Fourier-phase transforming in the optical domain and has lower computational burden owing to the straightforward matching algorithm. Meanwhile, the image data volume and detection bandwidth get decreased due to the designed coding scheme with polarization multiplexing. Besides, it is able to plug in a standard camera system and adapt to various lighting environments with tunable exposure time. The experiment results with different 3D trajectories show its potential in real-time 3D motion estimation and recognition.

2. PRINCIPLES

A. Polarization Multiplexing and Demultiplexing

Polarization is a basic property of light, expressed as the vibration direction of the light-field. Here, we employ two orthogonal polarization states, polar-0° and polar-90°, to carry the duplet perspectives of the moving objects. A polarization beam splitter (PBS) is placed in the reverse direction to work as a combiner to generate the overlapping scene after the polarization multiplexing, which is shown in Fig. 1(b). Then, through the zooming lens, the overlapped scene is focused on the digital micromirror device (DMD), which temporally modulates each coding channel with four-phase-shifting sinusoid coding patterns to acquire the Fourier phase. Every coding channel is detected by the 4×4 binning polarization pixels, which are mounted in front of the polarization charge-coupled device (CCD). Thus, the phase information $(0, \pi/2, \pi, 3\pi/2)$ and polarization states (0°, 45°, 90°, 135°) can be acquired at the same time. Similar to the polarization extraction scheme in Ref. [13], with the known polarization array, it is able to reassemble the multiplexed scenes by simply extracting the polar-0° and polar-90° values and constructing two perspectives. The whole polarization multiplexing and demultiplexing process is depicted in Fig. 1. The extinction ratio of the PBS and the polarizer array in the camera will determine the orthogonal FoV's crosstalk level in the measurements.

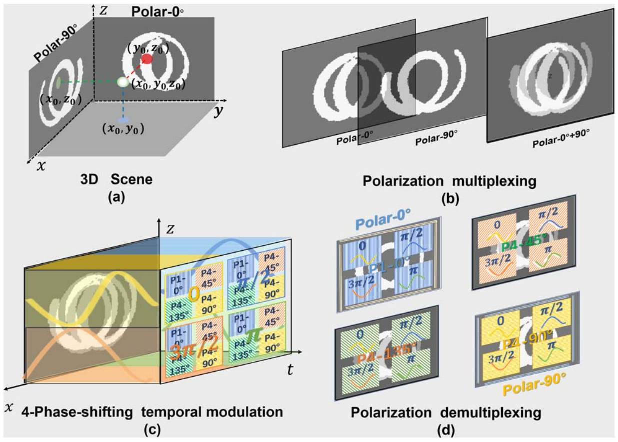

Fig. 1. Polarization multiplexing and demultiplexing process. (a) The 3D coordinate synthesized by the two orthogonal 2D plane. (b) Overlapping scene after the polarization multiplexing. (c) The detection image with four-phase-shifting temporal modulation. (d) Polarization demultiplexing process.

B. Fourier-Phase Transforming

When an object is moving through the scene, each detection channel in the sensor will get a similar temporal pulse with a different rising edge. According to the brightness constancy of the objects, $I(x, y, t) = I(x + \Delta x, y + \Delta y, t + \Delta t)$, the intensity of the voxel remains the same despite small changes of position and time period [16]. Thus, these temporal signals can be simply expressed as an impulse with spatial-variant time shifting. However, this information cannot be resolved in one exposure with a traditional camera. Herein, PM-FPC is designed to record this time-shift information through the optical coding method. With the principle of the discrete Fourier transform (DFT), a temporal signal can be represented by a series of discrete Fourier coefficients with different sampling frequencies. A shift in the time corresponds to the Fourier-phase shift in each encoded channel. To avoid phase unwrapping errors [17], we choose one period sinusoid pattern as the sampling frequency, which means only the first-order Fourier coefficients (1st DFT) in the optical coding process are used. The detailed time-encoded process in each channel is displayed in Fig. 2(a). On the sensor, 4×4 binning pixels consist of one temporal coding channel named channel i. With the pulse width modulation (PWM) mode of the DMD, four-phase-shifting sinusoid patterns $(0, \pi/2, \pi, 3\pi/2)$ have temporally modulated on each coding channel. In a single-shot image of the sensor, a series of Fourier-phase number F_{in} $(n = 1, 2, 3, 4)$ is detected. As illustrated in Eq. (1), the Fourier-phase number is the integration of the Hadamard product between the sinusoid pattern $\text{Wave}_n(t)$ and temporal signal s_i of coding channel i:

$$F_{\text{in}} = \int_0^{t_{\text{expo}}} s_i(t) \times \text{Wave}_n(t). \tag{1}$$

Thence, with the four-step phase-shifting method [18], the shifted Fourier phase of one coding channel i can be extracted as the following equation:

$$P_i = \arg\{(F_{i1} - F_{i4}) + j(F_{i2} - F_{i3})\}. \tag{2}$$

The time shifting on the 1st DFT can be summarized as

$$\begin{aligned} \mathfrak{F}_1\{s_i(t - T_i)\} &= S_i(f_1) \cdot \exp(-j2\pi f_1 T_i) \\ &= S_i(f_1) \cdot \exp(-jP_i), \end{aligned} \tag{3}$$

where $\mathfrak{F}_1$ is denoted as the 1st DFT operation, which only calculates the 1st DFT coefficient of the temporal signal in channel i, S_i is the amplitude of the Fourier coefficient, and f_1 is the frequency of the sinusoid coding pattern, which is inversely proportional to the exposure time $t_{\text{expo}} = \frac{1}{f_1}$. With the known exposure time and the detected Fourier-phase P_i, the time-shifting information T_i can be calculated out as the following equation:

$$T_i = \frac{P_i}{2\pi} \times t_{\text{expo}}. \tag{4}$$

The whole Fourier-phase transforming process is depicted in Fig. 2. After all of the Fourier-phase is extracted, the whole

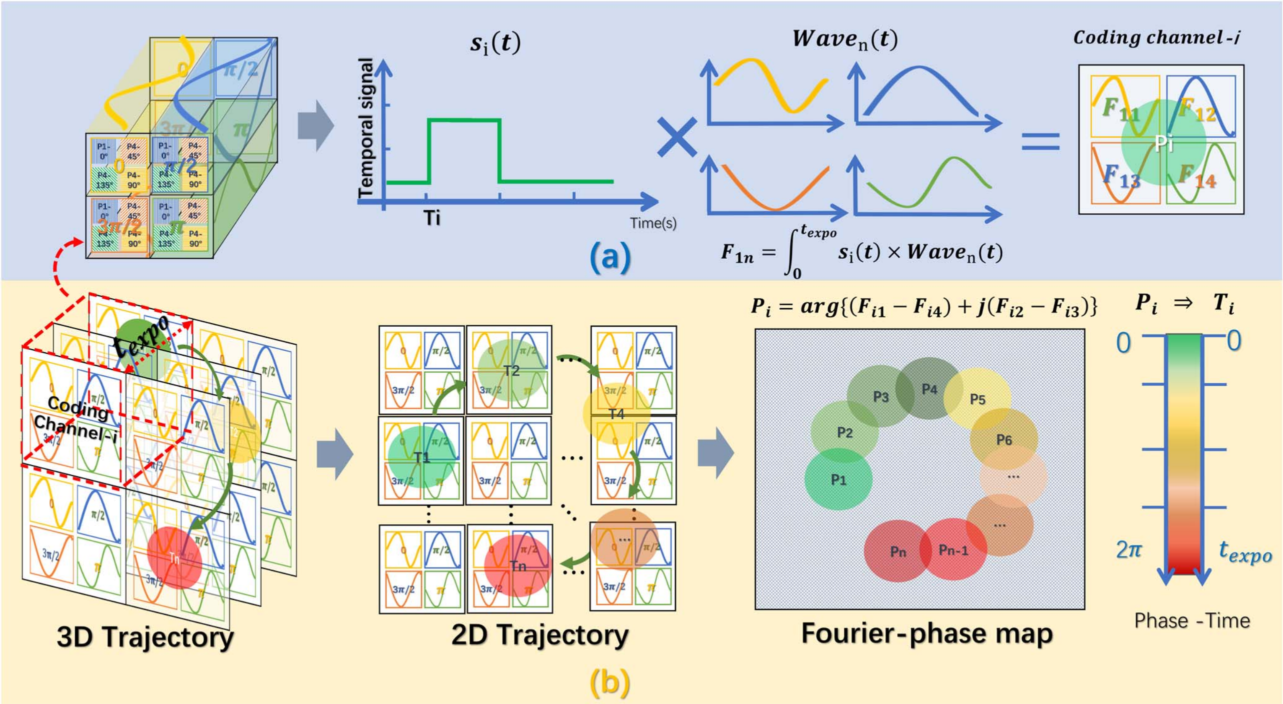

Fig. 2. Fourier-phase transforming process. (a) Time-encoded process in coding channel *i*. (b) Fourier-phase map with 3D trajectory.

image becomes a Fourier-phase map, and it is able to imply the appearance time of objects in each channel during the exposure.

C. 3D Mapping and Tracking

The proposed system follows a parallel 3D mapping and tracking philosophy, where the main modules operate in the one-way fashion estimating the final 4D data (x, y, z, t). A detailed overview of the flowchart is given in Fig. 3. Core modules of the system including Fourier-phase measuring, matching, and mapping processes are marked with dashed rectangles. The only input to the system is a single-shot 2D image of the PM-FPC. Through the measuring process, the Fourier-phase map of two orthogonal views is generated. For removing the phase interference caused by the environmental noise, it implements the mask extraction by setting the amplitude threshold, which is generally set to 1/10 of the average pixel intensity in the image. Besides, the initial phase of the exposure needs to be calibrated before the time-phase mapping process. The matching operation is applied between two orthogonal planes, the *XOZ* and *YOZ* planes. With the height-consistency calibration in the experiment, the moving target appears at the same height $(z_1 = z_2)$ of two orthogonal planes with its unique phase. Based on this, it can simply take any non-zero phase point $P_1(x_1, z_1)$ on the *XOZ* plane as the reference, to traverse all the pixels with the same height in the *YOZ* plane to find the correspondence point $P_2(y_2, z_1)$, which has the smallest phase difference with P_1. After this straightforward matching process, the 3D coordinates $(x = x_1, y = y_2, z = z_1)$ of the corresponding point are determined. Then, with the time-phase mapping relationship shown in Eq. (4), the time information of the object moving through the scenes can be calculated out through the precise phase measurement. Relying on the time-spatial consistency in the *XOZ* and *YOZ* plane, the 3D coordinates and time information of the object are synthesized out, which consist of the 4D datasets (x, y, z, t) as the final output.

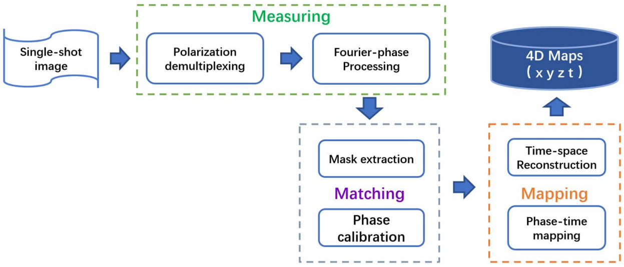

Fig. 3. Proposed system flowchart.

3. EXPERIMENT AND RESULTS

The schematic diagram of the experiment is shown in Fig. 4. Light from two orthogonal views is reflected by mirrors M1 and M2 and filtered by the PBS. Respectively, the duplet views get tagged with two polarization states (polar-0° and polar-90°) and combined to be a view-overlapped scene, which is further imaged and projected on a DMD (ViALUX V-9500) through the imaging lens. The DMD is employed to implement the pixel-wise temporal modulation on each coding channel [19], which is modulated by the parallel four-phase-shifting sinusoidal pattern $(0, \pi/2, \pi, 3\pi/2)$. Each phase-encoded channel is composed of 2×2 sub-pixels corresponding to four polarizations (0°, 45°, 90°, 135°). The polarization camera (FLIR BPS-PGE-51S5P-C) applied in the system has 2448×2048 pixels

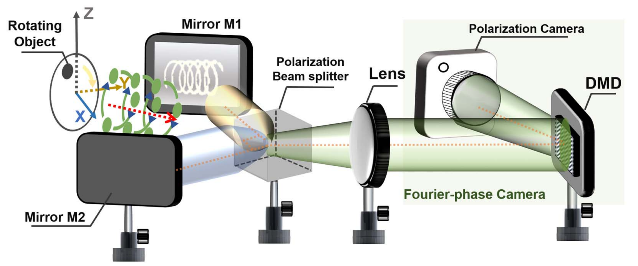

Fig. 4. Schematic diagram of the polarization multiplexed Fourier-phase camera.

mounted with different polarizers, and the pixel size is 3.45 μm. The DMD has a 1920 × 1080 micro mirror, whose pixel size is 7.6 μm. Through a strict optical calibration (Appendix B), which creates a suitable FoV (50 mm × 50 mm) with pixel-by-pixel correspondence (one micro-mirror corresponding to 2 × 2 polar pixels), the Fourier-phase map of the duplet view is measured, respectively. After the mask extraction and phase calibration steps, the matching process is applied between the *XOZ* and *YOZ* phase map. Then, the 3D position and time information of the object are derived by the time-phase mapping results. In the experiment, a motorized stage (Zolix LA100-60-ST) is utilized to build the horizontal linear movement of the target. When it comes to the circle motion scenes, an optical chopper (Thorlabs MC200B) with tunable frequency is implemented to work as a rotating stage. For a more complex motion scene, the vertical rotation and horizontal linear movement are combined to produce a spiral motion. The exposure time with each trajectory is different depending on the illumination and lighting conditions. Based on the Fourier-phase coding scheme, one coding channel corresponds to 4 × 4 pixels on the polarization camera, so the maximum spatial resolution of the reconstructed 3D space is 1/4 of camera resolution $(2048/4 = 512)$. However, for better phase measurement results, we utilize 12 × 12 binning pixels on the camera to perform as one coding channel and choose an illuminous LED as the moving object. In the first place, we test a dynamic scene with one object, including linear and circular motion, which is depicted in Fig. 5(a). The diameter of the objects is 5 mm, and the horizontal movement speed is 20 mm/s. In the one-line test, the exposure time is set at 2 s with 100 pixel length trajectory in the picture, which is consistent with the actual FoV and zoom ratio (0.645). In the circular motion, the period of the object motion is 320 ms, which is accordant to the rotation frequency (3.1 Hz) of the chopper. The reconstructed 3D position of these scenes is marked with the solid sphere with the changing color indicating the time information. The effective temporal resolution of the reconstruction is dependent on the exposure time and gray quantization bit number of the DMD, which is discussed in the Appendix C. Here, with the 8 bit

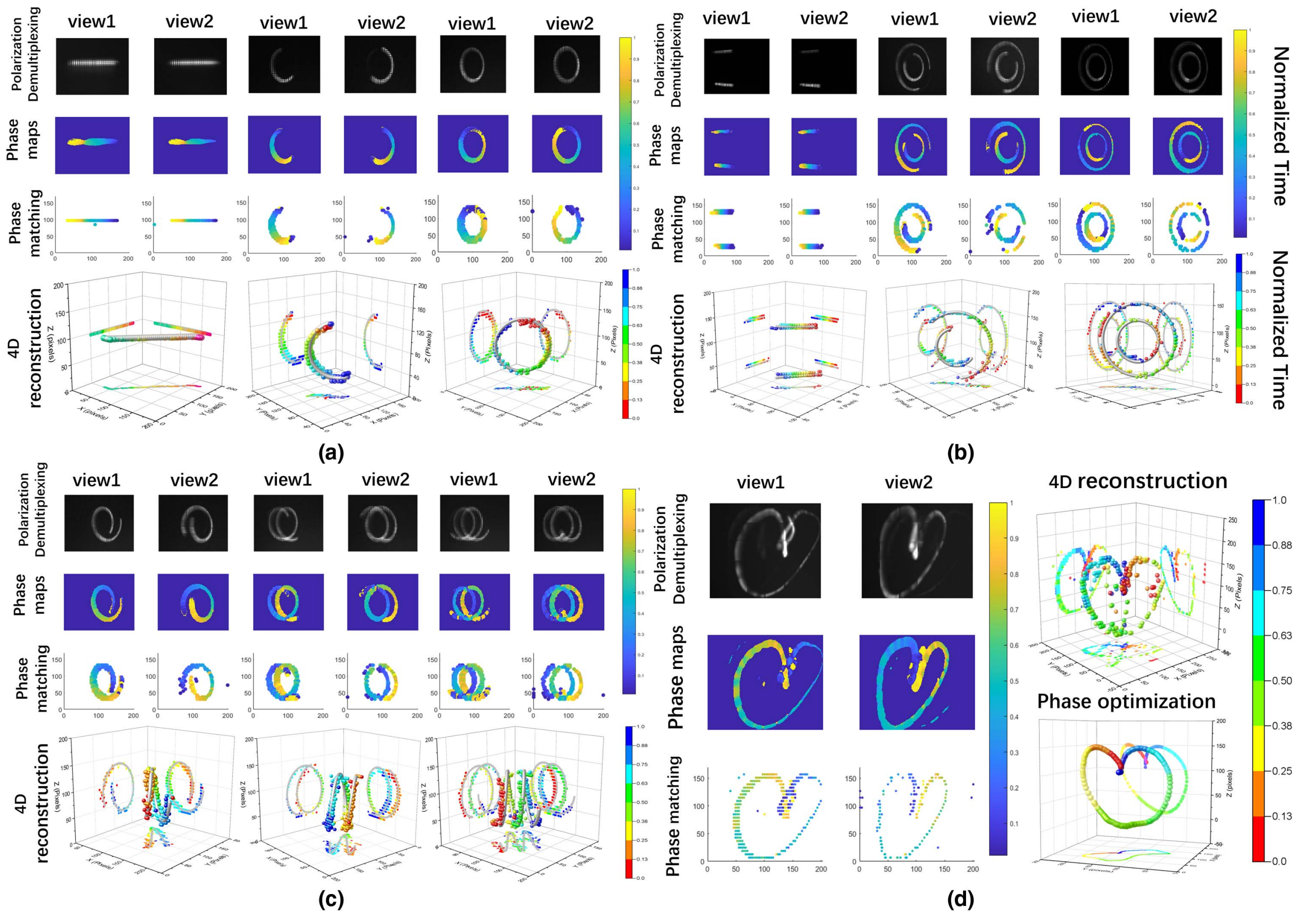

Fig. 5. (a) One object motion. (b) Two objects motion. (c) Rotation. (d) Handwriting heart (see Visualization 1).

grayscale DMD and 200 ms exposure time, the equivalent temporal resolution is 1280 fps (256 × 5 fps; fps, frames per second). Then, an object is added in the scene, and the results of two objects 3D tracking are shown in Fig. 5(b). In the multi-target tracking, the spatial-temporal restriction is utilized, which is the Euclidean distance [20] between the previous phase point and the current one. Based on the distance restriction, the multiple targets with different motion can also be distinguished in a single shot.

Furthermore, the object spiral motion with different exposure time (200 ms, 400 ms, 600 ms) is recorded and reconstructed, where the spiral forward speed is also 20 mm/s, and the rotation period is 160 ms. As shown in Fig. 5(c), these 3D tracking results all fit well to the real 3D trajectory marked with the gray sphere in the time 3D maps. To further verify the system potential in the 3D motion recognition, we also tested one handwriting trajectory with the heart shape, the results are displayed in Fig. 5(d). It is noted that in the complicated motion scene, the trajectory crossing area is unpreventable with the increasing exposure, which leads to inaccurate 4D reconstruction and poor matching results. To solve this problem, phase loss function and motion estimation are added to mitigate the effect of phase error. Some phase optimization in the crossing area is discussed in Appendix F, such as the *K*-means clustering [19], cubic spline interpolation [21], and Kalman filter [22]. Owing to these optimizations, a dynamic 3D handwriting trajectory of the heart and its matching process are displayed in Visualization 1.

4. CONCLUSION

In conclusion, we proposed a single-shot 3D tracking system based on a novel camera architecture and a new image format: Fourier-phase map. The principle of the system is to acquire the time-phase shifting of the target in two orthogonal views with different polarization states. Only one 2D image with motion trajectory is needed to reconstruct the 4D data (x, y, z, t) of the target. Owing to the polarization multiplexing and optical coding method, the detection bandwidth gets significantly decreased, which makes it work well in a low-cost polarization camera with efficient reconstruction algorithms. Meanwhile, the Fourier-phase transforming in the optical domain reduces the computational overhead from the data acquisition and quantization in the calculation. Compared with the traditional tracking system based on the frame difference method, simple algorithms, such as the phase-matching and time-phase mapping processes, relevantly mitigate the cost of the computation. In the experiment, the effective frame rate of the camera is achieved at 1280 fps under the exposure time of 200 ms, which breaks the exposure constraint of a traditional camera with the preset temporal resolution. With the long-exposure detection scheme, PM-FPC has higher SNR than normal high-speed cameras, providing it with the ability to capture the dynamic objects under low-light 3D scenes. In addition, owing to the pixel-mounted polarizer array, PM-FPC can filter the inevitable glare and specular interference, which is a great challenge to a traditional camera. Besides, to achieve more efficient and accurate phase-to-depth mapping, the universal stereo vision system is considered to replace the orthogonal view system. For the wider application in the 3D motion estimation, the phase prediction process is added on the trajectory crossing area. Furthermore, it is expected to realize a real-time 3D motion prediction and multi-target detection system with the development of the neural network under new data types [23,24].

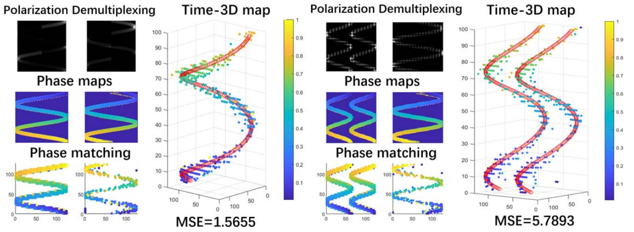

Fig. 6. Simulation results.

APPENDIX A: QUANTITATIVE ANALYSIS ON THE PERFORMANCE OF PM-FPC

To validate the proposed scheme of the PM-FPC, a numerical simulation is designed before the experiment. There are two simulated 3D scenes: the one is a sphere with a radius of 10 pixels moving along a spiral path, and the other is a pair of parallel spheres moving through the scenes, which verified its ability of multi-objects 3D tracking. As shown in Fig. 6, the designed trajectory is marked with a red circle, and the reconstructed 3D map is marked with a solid ball, which also indicates the time information by color. Mean square error (MSE) of the 4D data (x, y, z, t) is calculated as Eq. (A1) to quantitatively evaluate the effect of the reconstruction:

$$\mathrm{MSE} = \frac{1}{N}\sum_{i}^{N}\frac{(x_i - \hat{x}_i)^2 + (y_i - \hat{y}_i)^2 + (z_i - \hat{z}_i)^2 + (t_i - \hat{t}_i)^2}{4}, \tag{A1}$$

where (x_i, y_i, z_i, t_i) is the reconstructed point location in the time 3D map, and the real point coordinate is expressed as $(\hat{x}_i, \hat{y}_i, \hat{z}_i, \hat{t}_i)$. Here, the MSE of the two scenes is shown in the bottom of the time 3D maps. The reconstructed 3D trajectory is nicely fit to the real one. We also noted that with more objects appearing in the scene, the worse the matching effect and the larger MSE of the reconstruction. In addition, when the trajectory of the object coincides within the exposure time, the phase measurement and match of this area are inaccurate, which causes remarkable reconstruction error.

APPENDIX B: OPTICAL CALIBRATION

Similar to an optical coding camera system, the plug-in optical devices need to be carefully calibrated before the Fourier-phase measurement. First, the imaging lens and the relay lens in the system are both designed for clear imaging effects with suitable FoV. It is important to record the whole object trajectory during the exposure time. Another key parameter in PM-FPC is the pixel-to-pixel correspondence between the DMD and the polarization sensor. There are two common errors in the pixel-to-pixel calibration: the mismatch and the misalignment, which are difficult to observe through the imaging. Fortunately, the sensor and DMD can be regarded as two spatial gratings, so the Moire fringe operation illustrated in Ref. [25] can be applied in the calibration. In the experiment, we adjust the zoom ratio of the zooming lens and the rotation angle of the sensor, which will contribute to the accurate pixel-to-pixel correspondence of the system and the precise phase measurement for 3D tracking.

APPENDIX C: EFFECTIVE TEMPORAL RESOLUTION

As the object moves in the scene, the temporal signals at different pixel channels are of different time-pulse positions, leading to coincidence phase shift in the 1st DFT coefficient. Here, the accuracy of the Fourier-phase measurement is the temporal resolution of the 3D tracking in PM-FPC. The phase measurement accuracy is determined by the DMD quantization bits for temporal grayscale coding and the exposure time (t_{expo}) of the image sensor. The temporal resolution T is illustrated as the following equation:

$$T = \frac{t_{\text{expo}}}{\text{DMD}_{\text{grayscale}}} = \frac{t_{\text{expo}}}{2^N}, \tag{C1}$$

where $\text{DMD}_{\text{grayscale}}$ is the practical DMD grayscale level, and it can be calculated by the quantization bits (N bit). Since we use the DMD with the PWM mode as the temporal-spatial coding device, the light is digitally modulated by 8 bit quantized grayscale. Therefore, the temporal resolution of the PM-FPC is $t_{\text{expo}}/256$, which improves the camera tracking frame by 256 times.

APPENDIX D: SNR IN THE PM-FPC

SNR is utilized to describe the quality of the measurement, which is particularly important in applications requiring a precise object tracking system. More specifically, it is the ratio of the measured signal to the overall measured noise (frame-to-frame) during the CCD's exposure time. Here, we only consider the overall camera SNR under global exposure, not the pixel-level SNR, which is usually variable with the objects' motion. There are three primary sources of noise in a camera system: photo noise, dark noise, and read noise [26]. Photo noise is the inherent natural variation of the incident photo flux. It has a square root relationship with the signal and cannot be reduced via camera design. Dark noise arises from the statistical variation of the thermally generated electrons and is the square root of the number of thermal electrons generated with a given exposure time. Read noise comes from the inherent electronic uncertainty of the on-chip preamplifier and spurious charge of the camera. With the on-chip binning method, the M adjacent pixels on the sensor array can consist of one super-pixel, which allows the system to reach a better photo signal at the cost of the spatial resolution. Therefore, the final SNR of the system can be illustrated as

$$\text{SNR} = \frac{MPQ_E T}{\sqrt{(P+B)MQ_E D + MDT + N_r^2}}, \tag{D1}$$

where P is the photo flux incident on the sensor, B is the background photo flux, Q_E is the quantum efficiency of the camera, D is the dark current, N_r is the read noise, and T is the exposure time. Under low-light conditions, the camera system becomes photo-noise-limited at the longer exposure. Therefore, the SNR equation can be simplified to the equation as follows:

$$\text{SNR} = \frac{PQ_E}{\sqrt{(P+B)Q_E + D}} \cdot \sqrt{MT}. \tag{D2}$$

Meanwhile, the number of binned pixels allows the system to reach a photo-noise-limited mode more quickly, which is tunable with different light environments. When it comes to the PM-FPC, the M drops to 1/4 of the original (corresponding to only one of the 2×2 binned polarization pixels utilized in the measurement), and the equivalent exposure time is 256 times that of the traditional camera ($t_{\text{expo}} = 256T$). Thence, compared with the traditional camera system, the PM-FPC has improved the SNR by 64 times.

APPENDIX E: DATA VOLUME

Assuming this 3D tracking task is finished by the stereo cameras with F frames and N pixels in each frame, its data volume is $F \times N$ bytes (1 byte for one pixel channel). PM-FPC processes the same resolution 3D tracking through a polarization camera with a single-shot image. In the optical coding process, each phase-shifting sinusoidal pattern corresponds to N_{polar} polarization filters on the sensor. N_{phase}-step phase-shifting patterns are utilized for more precise phase measurement. It means that $N_{\text{polar}} \times N_{\text{phase}}$ binning pixels consist of one coding channel to measure the Fourier phase of the signal. With the ideal correspondence between the DMD and the image sensor, the minimum spatial-encoded pixel size is $N_{\text{polar}} \times N_{\text{phase}}$. Here, we propose a ratio value R to represent the data volume comparison between the stereo cameras and PM-FPC, and the equation is listed as follows:

$$R = \frac{2 \times F \times N}{1 \times 1 \times N_{\text{polar}} \times N_{\text{phase}} \times N}, \quad N_{\text{polar}} \geq 4, \; N_{\text{phase}} = 2,3,4 \,. \tag{E1}$$

In the simulation results mentioned above, the frame rate of a single-view video is 256 fps, and the resolution is 256 × 256 pixels; with the coding phase patch of 4 × 4 binning pixels, the volume ratio result is 32, which saves huge bandwidth resources for data transmission and storage.

APPENDIX F: PHASE OPTIMIZATION IN THE CROSSING AREA

As mentioned above, the principle of PM-FPC is the time-shift impulse leading to the pixel-wise phase shift in the Fourier domain during the exposure time. Therefore, when it has two impulses occurring in the same pixel, the phase measuring scheme will lose its precision in this area, which is also called phase entanglement. This effect directly leads to phase mismatching and time information loss. For wider applications, phase estimation and prediction are needed to solve this problem through the spatio-temporal continuity with the object motion. Assuming that the phase change of the moving target is continuous in the phase maps, the real phase information in the crossing area can be estimated among the neighborhood phase values. Here, we choose the K-means clustering [27] to acquire multiple centroid points with reliable phase-spatial

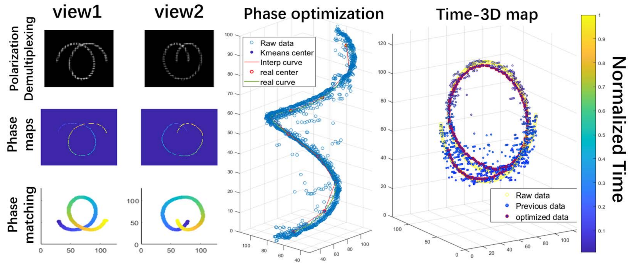

Fig. 7. Phase optimization in the crossing area.

information. According to these K-group centroid points, the cubic spline interpolation algorithm is applied to fit out the real phase changing curve in the phase map. With the phase optimization, the phase loss area gets a reasonable phase value, which has favorable effects on the farther phase-matching process. In the simulation, we design one typical crossing trajectory as the 3D scene; then, applying phase optimization in the 3D tracking process, the phase maps after the optimization, the fitting curve, and the comparison of the 3D reconstruction are displayed in the Fig. 7. In the 3D reconstruction, the actual 3D trajectory is marked with a yellow circle, the previous reconstruction result is marked with a blue circle, and the optimized reconstruction result is marked with the red circle. It can be found that with reasonable phase estimation, the reconstruction effect gets improved relevantly. However, this optimization is immature for solving more complicated 3D dynamic scenes, such as multiple targets crossing and objects moving periodically. On the other hand, it also can change different exposure times to avoid this phase crossing, like cutting complex motion trajectories into fragments of simple motion trajectories. Much more intelligent algorithms are needed to study for more precise 4D location of the objects, which is also an important area in our future research.

Funding. National Key Research and Development Program of China (2019YFB1803500); National Natural Science Foundation of China (61771284); Institute for Guo Qiang Tsinghua University.

Disclosures. The authors declare no conflicts of interest.

REFERENCES

1. D. A. Forsyth and J. Ponce, *Computer Vision: A Modern Approach* (Pearson, 2012).
2. D. Pathak, R. Girshick, P. Dollár, T. Darrell, and B. Hariharan, “Learning features by watching objects move,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2017), pp. 2701–2710.
3. S.-T. Park and J. G. Lee, “Improved Kalman filter design for three-dimensional radar tracking,” IEEE Trans. Aerosp. Electron. Syst. **37**, 727–739 (2001).
4. A. Chaikovsky, Y. O. Grudo, Y. A. Karol, A. Y. Lopatsin, L. Chaikovskaya, S. Denisov, F. Osipenko, A. Slesar, M. Korol, Y. S. Balin, and S. V. Samoilova, “Regularizing algorithm and processing software for Raman lidar-sensing data,” J. Appl. Spectrosc. **82**, 779–787 (2015).
5. E. Seemann, K. Nickel, and R. Stiefelhagen, “Head pose estimation using stereo vision for human-robot interaction,” in *6th IEEE International Conference on Automatic Face and Gesture Recognition* (2004), pp. 626–631.
6. R. Munoz-Salinas, E. Aguirre, and M. Garca-Silvente, “People detection and tracking using stereo vision and color,” Image Vis. Comput. **25**, 995–1007 (2007).
7. A. Mauri, R. Khemmar, B. Decoux, N. Ragot, R. Rossi, R. Trabelsi, R. Boutteau, J.-Y. Ertaud, and X. Savatier, “Deep learning for real-time 3D multi-object detection, localisation, and tracking: application to smart mobility,” Sensors **20**, 532 (2020).
8. Y. Zhou, G. Gallego, H. Rebecq, L. Kneip, H. Li, and D. Scaramuzza, “Semi-dense 3D reconstruction with a stereo event camera,” in *Proceedings of the European Conference on Computer Vision (ECCV)* (2018), pp. 235–251.
9. H. Rebecq, G. Gallego, E. Mueggler, and D. Scaramuzza, “EMVS: event-based multi-view stereo—3D reconstruction with an event camera in real-time,” Int. J. Comput. Vis. **126**, 1394–1414 (2018).
10. Q. Deng, Z. Zhang, and J. Zhong, “Image-free real-time 3-D tracking of a fast-moving object using dual-pixel detection,” Opt. Lett. **45**, 4734–4737 (2020).
11. Y. Sun, X. Yuan, and S. Pang, “Compressive high-speed stereo imaging,” Opt. Express **25**, 18182–18190 (2017).
12. Z. Zhang and S. Zhang, “One-shot 3D shape and color measurement using composite RGB fringe projection and optimum three-frequency selection,” Proc. SPIE **7511**, 751103 (2009).
13. M. Qiao, X. Liu, and X. Yuan, “Snapshot spatial–temporal compressive imaging,” Opt. Lett. **45**, 1659–1662 (2020).
14. N. Antipa, G. Kuo, R. Heckel, B. Mildenhall, E. Bostan, R. Ng, and L. Waller, “Diffusercam: lensless single-exposure 3D imaging,” Optica **5**, 1–9 (2018).
15. X. Feng and L. Gao, “Ultrafast light field tomography for snapshot transient and non-line-of-sight imaging,” Nat. Commun. **12**, 2179 (2021).
16. T. Yamazato, M. Kinoshita, S. Arai, E. Souke, T. Yendo, T. Fujii, K. Kamakura, and H. Okada, “Vehicle motion and pixel illumination modeling for image sensor based visible light communication,” IEEE J. Sel. Areas Commun. **33**, 1793–1805 (2015).
17. C. Zhang, H. Zhao, X. Gao, Z. Zhang, and J. Xi, “Phase unwrapping error correction based on phase edge detection and classification,” Opt. Lasers Eng. **137**, 106389 (2021).
18. H. Huang, C. Hu, S. Yang, M. Chen, and H. Chen, “Temporal ghost imaging by means of Fourier spectrum acquisition,” IEEE Photon. J. **12**, 6803012 (2020).
19. A. Likas, N. Vlassis, and J. J. Verbeek, “The global k-means clustering algorithm,” Pattern Recognit. **36**, 451–461 (2003).
20. T. Saito and J. I. Toriwaki, “New algorithms for euclidean distance transformation of an n-dimensional digitized picture with applications,” Pattern Recognit. **27**, 1551–1565 (1994).
21. S. McKinley and M. Levine, “Cubic spline interpolation,” Coll. Redwoods **45**, 1049–1060 (1998).
22. E. A. Wan and R. Van Der Merwe, “The unscented Kalman filter for nonlinear estimation,” in *Proceedings of the IEEE 2000 Adaptive Systems for Signal Processing, Communications, and Control Symposium* (2000), pp. 153–158.
23. C. Hu, H. Huang, M. Chen, S. Yang, and H. Chen, “Fouriercam: a camera for video spectrum acquisition in a single shot,” Photon. Res. **9**, 701–713 (2021).
24. C. Hu, H. Huang, M. Chen, S. Yang, and H. Chen, “Video object detection from one single image through opto-electronic neural network,” APL Photon. **6**, 046104 (2021).
25. S. Ri, M. Fujigaki, T. Matui, and Y. Morimoto, “Accurate pixel-to-pixel correspondence adjustment in a digital micromirror device camera by using the phase-shifting Moiré method,” Appl. Opt. **45**, 6940–6946 (2006).
26. O.-C. M. Gain, *On-Chip Multiplication Gain* (Roper Scientific, 2002).
27. K. Krishna and M. N. Murty, “Genetic k-means algorithm,” IEEE Trans. Syst. Man Cybernet. B **29**, 433–439 (1999).