PHOTONICS Research

FourierCam: a camera for video spectrum acquisition in a single shot

Chengyang Hu,[1,2,†] Honghao Huang,[1,2,†] Minghua Chen,[1,2] Sigang Yang,[1,2] and Hongwei Chen[1,2,*]

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

Received 13 October 2020; revised 21 February 2021; accepted 21 February 2021; posted 22 February 2021 (Doc. ID 412491); published 21 April 2021

The novel camera architecture facilitates the development of machine vision. Instead of capturing frame sequences in the temporal domain as traditional video cameras, FourierCam directly measures the pixel-wise temporal spectrum of the video in a single shot through optical coding. Compared to the classic video cameras and time-frequency transformation pipeline, this programmable frequency-domain sampling strategy has an attractive combination of characteristics for low detection bandwidth, low computational burden, and low data volume. Based on the various temporal filter kernel designed by FourierCam, we demonstrated a series of exciting machine vision functions, such as video compression, background subtraction, object extraction, and trajectory tracking.

https://doi.org/10.1364/PRJ.412491

1. INTRODUCTION

Humans observe the world in the space–time coordinate system, and traditional video cameras are also based on the same principle. The video data format in the unit of a time serial image frame is well understood for eyes and is the basis for many years of research in machine vision. With the development of optics, focal plane optoelectronics, and a post-detection algorithm, some novel video camera architectures have gradually emerged [1]. The single-shot ultrafast optical imaging system observes the transient events in physics and chemistry at an incredible rate of one billion frames per second (fps) [2]. An event camera with high dynamic range, high temporal resolution, and low power consumption asynchronously measures the brightness change, position, and symbol of each pixel to generate event streams and is widely used in autonomous driving, robotics, security, and industrial automation [3]. A privacy-preserving camera based on coded aperture has also been applied in action recognition [4]. Although the functions of these cameras are impressive, the essential sampling strategy is still to measure the reflected or transmitted light intensity of a scene in the temporal domain. In the lens system, pixels can be regarded as independent time channels, and the acquired signal is the temporal variation of light intensity at the corresponding position in the scene. It is well-known that the frequency domain feature of a visual temporal signal is more significant. For example, in general, a natural scene video has high temporal redundancies, so most information of a temporal signal concentrates on low-frequency components, which is a premise in video compression [5]. The static background of the scene appears as a DC component in the frequency domain, which provides insights for background subtraction [6–8]. In deep learning, performing high-level vision tasks based on spatial frequency domain data brings a better result [9]. By taking into account space–time duality, this strategy has the potential to be used for temporal frequency domain data. All of the above frequency characteristics imply that capturing video in the temporal frequency domain instead of the temporal domain will initiate a sampling revolution.

In this paper, we propose a temporal frequency sampling video camera: FourierCam, which is a novel architecture that innovates the basic sampling strategy. The concept of FourierCam is to perform pixel-wise optical coding on the scene video and directly obtain the temporal spectrum in a single shot. In contrast with the traditional cameras, the framework of single-shot temporal spectrum acquisition has a lower detection bandwidth. Furthermore, the data volume can be reduced by analyzing the temporal spectrum features for efficient sampling. Since the temporal Fourier transform is done in the optical system, its computational burden is lower compared to that of the time-frequency transformation pipeline (sampling–storing–transforming). In addition to the basic advantages, according to the clear physical meaning of the spectrum, a variety of temporal filter kernels can be designed to accomplish typical machine vision tasks. To demonstrate the capability of FourierCam, we present a series of applications, which cover video compression, background subtraction, object extraction, and trajectory tracking. These applications

can be easily switched only by adjusting the temporal filter kernels without changing the system structure. As a flexible framework, FourierCam can be easily integrated with existing imaging systems and is suitable for microimaging to macroimaging.

2. PRINCIPLE OF FOURIERCAM

FourierCam is suitably designed for acquiring a pixel-wise temporal spectrum of dynamic scenarios through optical coding. To optically acquire multiple Fourier coefficients, the input signal needs to be multiplied by sinusoids with different frequencies and phases and temporally integrated. However, ordinary natural signals are often nonreproducible for repeated operations. Therefore, a single-shot scheme is designed for parallel coding. The principle illustration and experimental optical setup of FourierCam are shown in Fig. 1(a). The dynamic scene is projected to a spatial light modulator (a digital micromirror device, DMD) by a camera lens and pixel-wise encoded. Then, the encoded light from the spatial light modulator is focused onto an image sensor (a charge-coupled device, CCD) and temporally integrated during exposure time. Figure 1(b) illustrates the coding strategy of FourierCam. The modulation units on the DMD are spatially divided into $m \times n$ coding groups (marked as CGs) for acquiring the temporal spectra of $m \times n$ pixels in the scene. A pixel at position j can be regarded as a temporal waveform (pixel temporal vector). The CG corresponding to the pixel temporal vector consists of $p \times q$ coding elements (marked as CEs) to obtain the Fourier coefficients of $p \times q$ frequencies. Each CE includes four DMD modulation

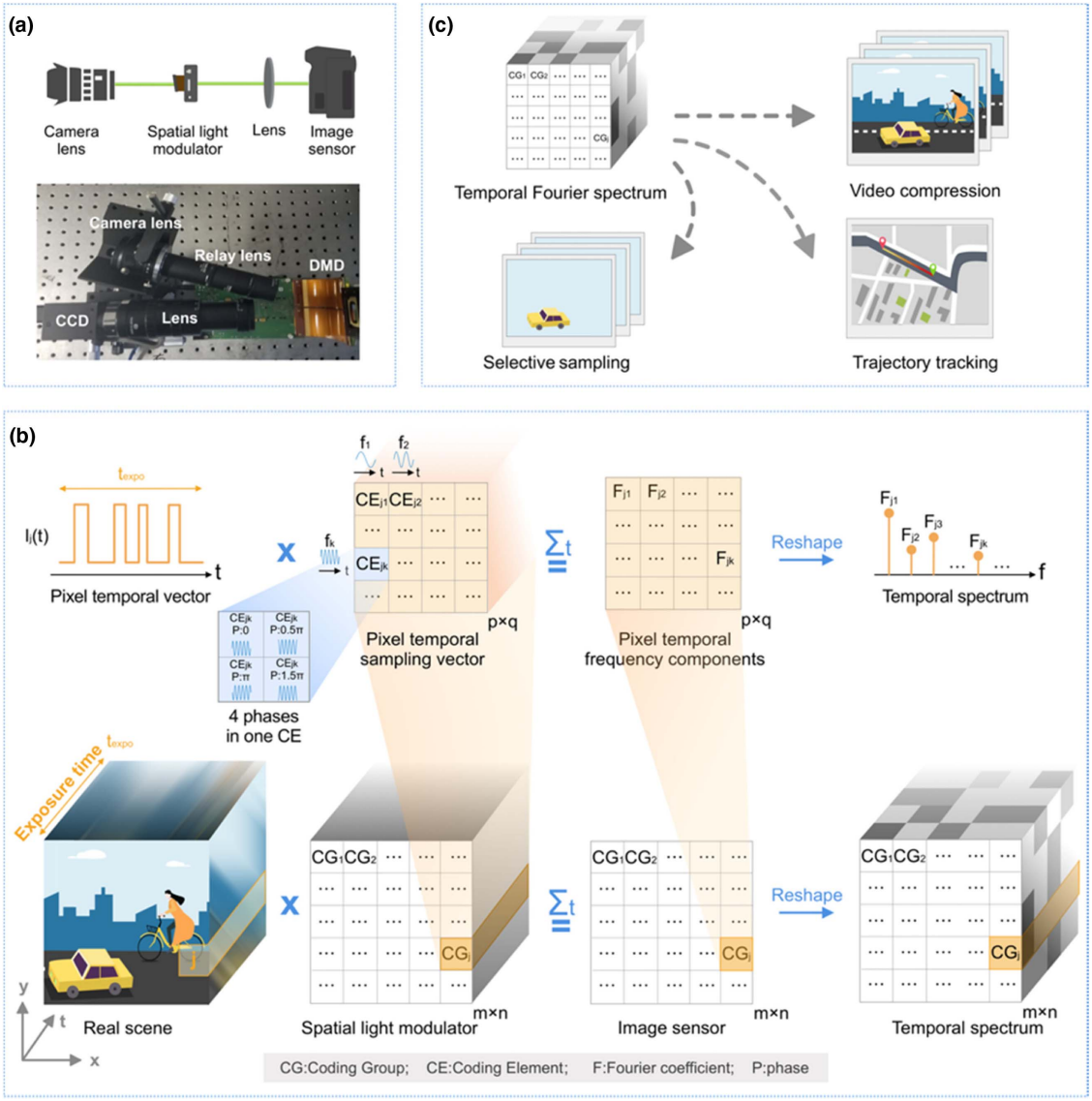

Fig. 1. Overview of FourierCam. (a) Schematic and prototype of FourierCam. (b) Coding strategy of FourierCam. The real scene is coded by a spatial light modulator (DMD) and integrated during a single exposure of the image sensor. The DMD is spatially divided into coding groups (5 × 5 coding groups are shown here, marked as CG), and each CG contains multiple coding elements (4 × 4 coding elements are shown here, marked as CE) to extract the Fourier coefficients of the pixel temporal vector. The Fourier coefficients of different pixel temporal vectors form the temporal spectrum of the scene. (c) Three demonstrative applications of FourierCam: video compression, selective sampling, and trajectory tracking.

units that can be controlled independently. The four units in one CE modulated the light intensity in a predetermined sinusoid fashion with the same frequency and four different phases (0, 0.5π, π, 1.5π). Since a single exposure of the image sensor temporally integrates the encoded scene, one can extract the Fourier coefficient for a specific frequency by means of four-step phase-shifting in one CE. Therefore, different Fourier coefficients of the pixel temporal vector are acquired by CEs in one CG to form the temporal spectrum of the pixel temporal vector. With the same operation applied to all CGs, the temporal spectrum of the whole scene can be recovered. In the optical prototype, since the pitch size of the DMD is larger than the pitch size of the image sensor, we adjust the paraxial magnification of the zoom lens to match one DMD mirror with 3 × 3 image sensor pixels (i.e., larger effective image sensor pixel size) to ensure accurate DMD and image sensor alignment (see Appendix A for details). Moreover, although the DMD only modulates the light in a binary form, the pulse width modulation (PWM) technique can be utilized for the DMD [10] to produce sinusoidal coding. As the image sensor works as an integration detection, one just needs to keep the summation of the light intensity equivalent to an analog sinusoidal modulation.

In the experimental setup, the scene is imaged on a virtual plane through a camera lens (CHIOPT HC3505A). A relay lens (Thorlabs MAP10100100-A) transfers the image to the DMD (ViALUX V-9001, 2560 × 1600 resolution, 7.6 μm pitch size) for light amplitude distribution modulation. The reflected light from the DMD is then focused onto an image sensor (FLIR GS3-U3-120S6M-C, 4242 × 2830 resolution, 3.1 μm pitch size) by a zoom lens (Utron VTL0714V). Due to one DMD mirror being matched with 3 × 3 image sensor pixels, the effective resolution is one-third of the resolution of the image sensor in both the horizontal and the vertical directions (i.e., 1414 × 943).

The principle of the proposed FourierCam system is spatially splitting the scene into independent temporal channels and acquiring the temporal spectrum by the corresponding CG for each channel. Every CG contains some CEs to obtain Fourier coefficients for frequencies of interest. During one exposure time t_{expo}, the detected value $D_{jk\varphi}$ in CE k, CG j is equivalent to an inner product of pixel temporal vector $I_j(t)$ and pixel temporal sampling vector $S_{jk\varphi}(t)$:

$$\begin{aligned} D_{jk\varphi} &= \langle I_j(t), S_{jk\varphi}(t)\rangle \\ &= \int_{t_{\text{expo}}} I(t)[A + B\cos(2\pi f_k t + \varphi)]\mathrm{d}t, \end{aligned} \tag{1}$$

where $S_{jk\varphi}(t)$ is the sinusoidal pixel temporal sampling vector with frequency f_k and phase φ in CE k, CG j. A and B denote the average intensity and the contrast of $S_{jk\varphi}(t)$, respectively. The Fourier coefficient F_{jk} of f_k can be extracted by four-step phase-shifting as

$$2BC \times F_{jk} = (D_{jk0} - D_{jk\pi}) + i(D_{jk\frac{\pi}{2}} - D_{jk\frac{3\pi}{2}}), \tag{2}$$

where C depends on the response of the image sensor. The DC term A can be canceled out simultaneously by the four-step phase-shifting.

Based on the aforementioned principle of FourierCam, the temporal spectrum of the scene can be easily obtained. As a novel camera architecture with a special data format, FourierCam is of the following three advantages (see Appendix B for details).

Low detection bandwidth: Since the image sensor only needs to detect the integration of the coded scene for obtaining the temporal spectrum during the entire exposure time, the required detector bandwidth is much lower than the bandwidth of scene variation.

Low data volume: Natural scene is of high temporal redundancies; i.e., most information of it concentrates on low-frequency components. Besides, some special scenes, like periodic motions, have a narrow bandwidth in the temporal spectrum. FourierCam enables flexibly designing the sampling frequencies of interest to cut down the temporal redundancies and reduce data volume.

Low computational burden: The multiplication and summation operations of Fourier transform are realized by optical coding and long exposure in FourierCam; thus, the temporal spectrum can be acquired with low computational burden.

Here, we introduce three applications to demonstrate these advantages of FourierCam [illustrated in Fig. 1(c)]. The first application is video compression. We verify the temporal spectrum acquisition of FourierCam and demonstrate the video compression by using the low-frequency-concentration property of the natural scene. The second application is selective sampling. We show the FourierCam is able to subtract the static background, as well as extract the objects with a specific texture, motion period, or speed by applying designed temporal filter kernels to process the signals during sensing. The last application is trajectory tracking. The temporal phase reveals the time order of events so the FourierCam can be used to analyze the presence and trajectory of the moving objects. These applications show that the temporal spectrum acquired by FourierCam, as a new format of visual information, is able to provide physical features to assist and complete vision tasks.

3. TEMPORAL SPECTRUM ACQUISITION: BASIC FUNCTION AND VIDEO COMPRESSION

The basic spectrum acquisition function of FourierCam is demonstrated. For ordinary aperiodic moving objects or natural varying scenes, the energy in the temporal spectrum is mainly concentrated at low frequencies. This observation is exploited to record compressive video in the temporal domain by only acquiring the Fourier coefficients of low frequencies using FourierCam.

By using the above method, we assemble the Fourier coefficient F_{jk} of F_k in CG j. We can combine all Fourier coefficients in CG j to form its temporal spectrum as

$$F_j = \{F^*_{jh}, F^*_{jh-1}, \ldots, F_{jh-1}, F_{jh}\}, h = p \times q, \tag{3}$$

where h ($p \times q$) is the number of CEs in a CG, and F^*_{jh} denotes the complex conjugate of F_{jh}. The pixel temporal vector $I_j(t)$ can be reconstructed by applying inverse Fourier transform:

$$2BC \times R_j = \mathcal{F}^{-1}\{F_j\}, \tag{4}$$

where $\mathcal{F}^{-1}$ denotes the inverse Fourier transform operator. The result of the inverse transform R_j is proportional to the pixel

temporal vector $I_j(t)$ in CG j. By applying the same operation to all CGs, we can reconstruct the video of the scene.

The experiment setup and the corresponding coding signals of DMD are illustrated in Fig. 2(a). Nine frequencies ranging from 0 Hz (DC component) to 80 Hz are applied to encode the scene within 0.1 s exposure (corresponding to 10 fps). With the temporal spectrum acquired by FourierCam, via inverse Fourier transform, a video can be reconstructed at an equivalent 160 Hz frame rate with 16 times speedup compared to the original frame rate. To acquire nine frequency components, 3×3 CEs are set in each CG, resulting in a resolution of 235×157 in the reconstructed video. The frequency interval of the encoded signal satisfies the frequency domain sampling theorem and is determined by the exposure time (see Appendix C for details).

The first demonstrative scene in this application includes a toy car running in the field of view. A capture of the static toy car is shown in Fig. 2(b) (top left) as ground truth. The coded data acquired by FourierCam is shown in Fig. 2(b) (top right) in which the scene is blurred and features of the toy car cannot be visually distinguished. After decoding, the complex temporal spectrum of the scene can be extracted. The corresponding amplitude and phase are shown in Fig. 2(b) (middle row) with their zoom-in view (bottom row). In addition to the toy car with a translating motion, a rotating object is also used for demonstration. This scene is a panda pattern on a rotating disk with an angular velocity of ~20 rad/s. In Fig. 2(c), the static capture of the object (top left), coded data (top right), amplitude, and phase (middle row) are shown respectively.

To visually evaluate the correctness of the acquired temporal spectra, the videos of the two scenes are reconstructed using the inverse Fourier transform. Figure 2(d) displays three frames from the video of the toy car (left column) and the rotating panda (right column). These results clearly show the statuses of the dynamic scenes at different times and indicate that FourierCam is able to correctly acquire the temporal spectrum.

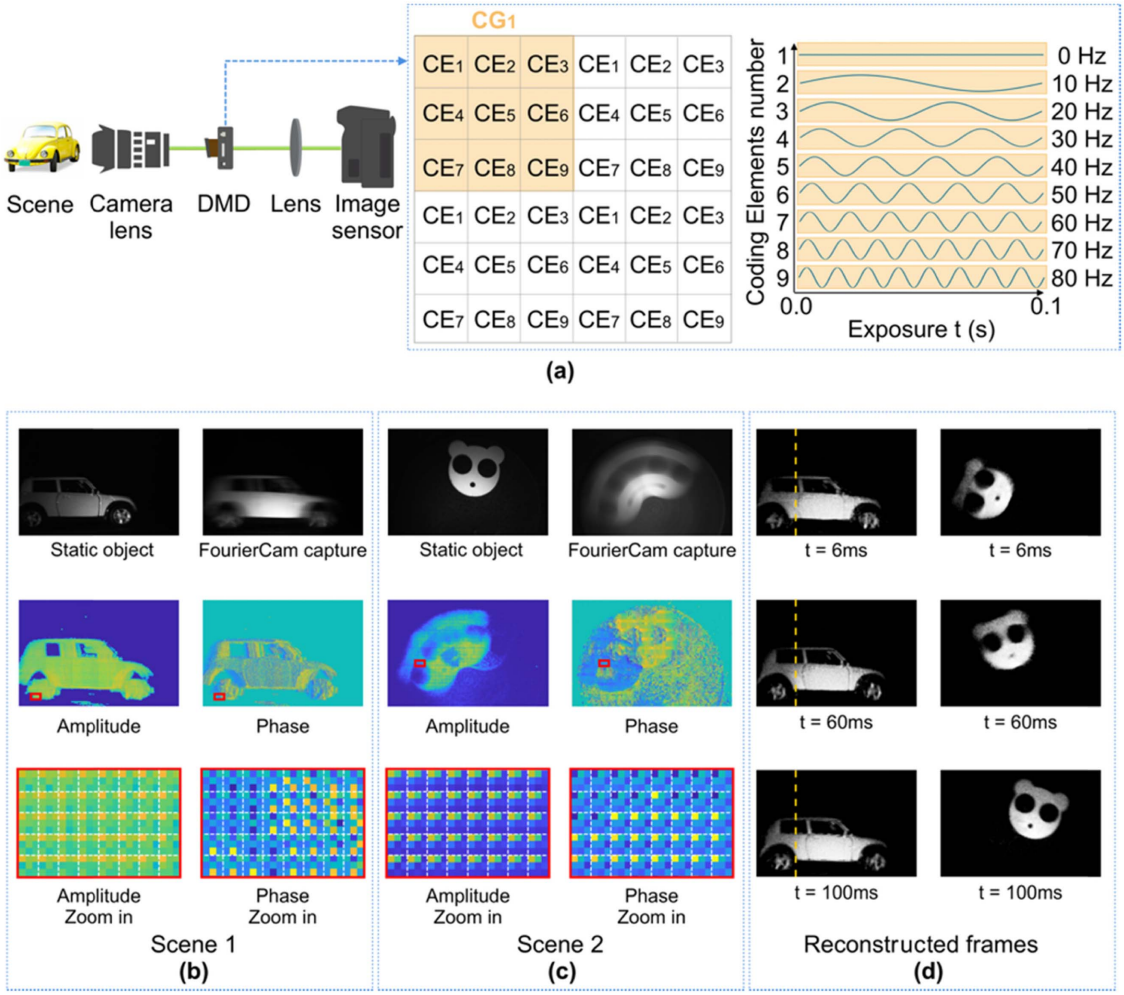

Fig. 2. Capturing aperiodic motion video using FourierCam. (a) Illustration of experiment setup and coding pattern on DMD. Each CG contains nine CEs (3×3, ranging from 0 Hz to 80 Hz) to encode the scene. (b) A toy car is used as a target. Top left: static object as ground truth. Top right: coded data captured by FourierCam. Middle left: amplitude of temporal spectrum. Middle right: phase of temporal spectrum. Bottom row: zoom in of middle row. A white-dotted mesh splits into different CGs. (c) A rotating disk with a panda pattern is used as a target. Top left: static object as ground truth. Top right: coded data captured by FourierCam. Middle left: amplitude of temporal spectrum. Middle right: phase of temporal spectrum. Bottom row: zoom in of middle row. A white-dotted mesh splits into different CGs. (d) Three frames from the reconstructed videos of the two scenes in (b) and (c). A yellow-dotted line is shown as reference.

As the single-shot detection data includes the information of multiple frames (16 frames for demonstration), FourierCam realizes the effect of (16×) video compression. (See Appendix D for the numerical analysis about the performance of video compression. The reconstructed toy car video is shown as an example in Visualization 1).

4. SELECTIVE SAMPLING: FLEXIBLE TEMPORAL FILTER KERNELS

FourierCam provides the flexibility for designing the combination of frequencies to be acquired, which is termed temporal filter kernels in this paper. By considering the prior of the scenes and objects, one can achieve selectively sampling the object of interest. In this part, three scenes are demonstrated: periodic motion video acquisition, static background subtraction, and object extraction based on speed and texture.

Periodic motions widely exist in medical, industry, and scientific research, such as heartbeat, rotating tool bit, and vibration. Since a periodic signal contains energy only in the direct current, fundamental frequency, and harmonics, it has a very sparse representation in the Fourier domain (see Appendix E for details). By taking the temporal spectrum characteristics into account as prior information, we use FourierCam to selectively acquire several principal frequencies in the temporal spectrum.

As shown in Fig. 3(b) (top left), a rotating disk with periodic patterns is designed as the target. The disk rotates at a speed as high as 5460 r/min. The disk contains four rings with 3, 5, 7, and 11 spatial periods from inner to outer; thus, the temporal frequencies of these four rings are 273, 455, 637, and 1001 Hz, respectively. We apply these frequencies to DMD to encode the scene [Fig. 3(a)] during a 0.5 s exposure (2 Hz frame rate) and further reconstruct a video of the rotating disk. Here, the equivalent maximum frame rate is 2002 Hz, so the frame rate improvement is 1001 times (corresponding compression ratio is 0.1%). The acquiring of four frequencies needs 2 × 2 CEs in each CG; thus, the resolution of the reconstructed video is 353 × 235. Four frames from the video are shown in Fig. 3(b) (bottom). The reconstructed video is provided as Visualization 2.

Subtracting the background and extracting moving objects are significant techniques for video surveillance and other video processing applications. In the frequency domain, the background is concentrated on the DC component. By filtering the DC component, one can subtract the background and extract moving objects. Some moving object extraction approaches performed in the frequency domain [6–8] have been proposed, which need to acquire the video first and then perform Fourier transform, and thus suffer from relatively high computational cost and low efficiency. Thanks to the capability

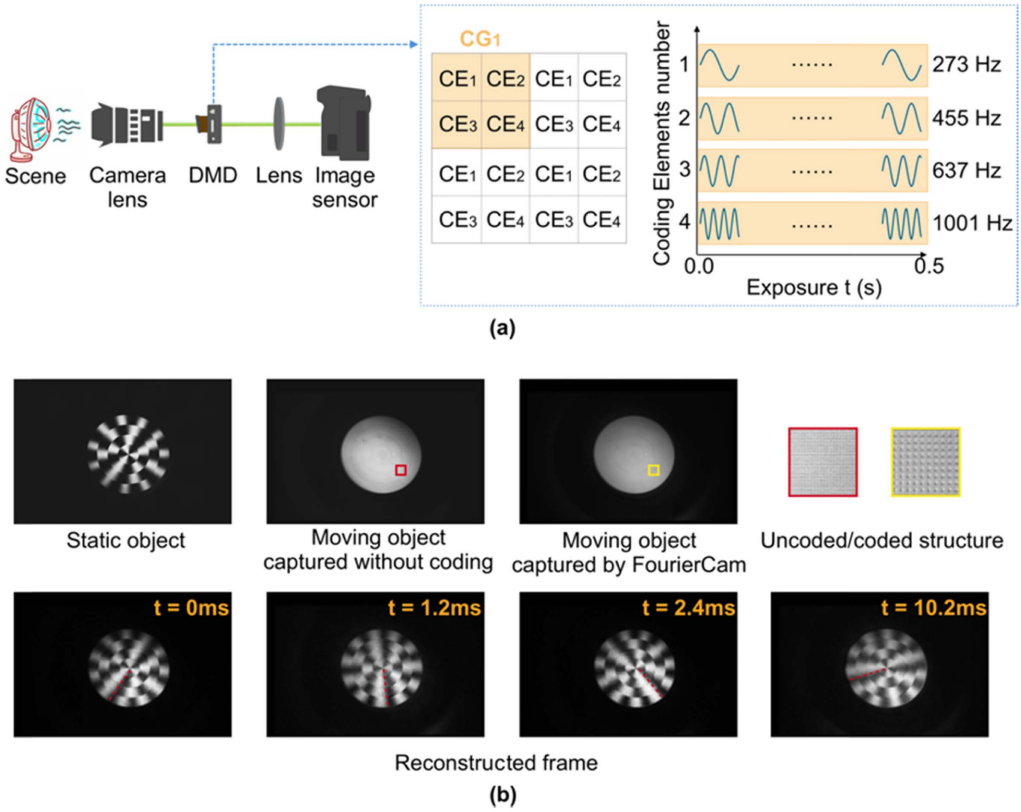

Fig. 3. Capturing periodic motion video using FourierCam. (a) To capture a periodic motion with four frequencies, each CG contains four CEs (2 × 2) to encode the scene. (b) A rotating disk is used as target. Top left: static object as ground truth. Top right: the zoom-in view of the captured data with and without coding, corresponding to normal slow cameras and FourierCam, respectively. Ordinary slow cameras blur out the details of moving objects while coded structure in FourierCam capture provides sufficient information to reconstruct the video. Bottom: four frames from the reconstructed video. Red-dotted lines are shown in each frame to indicate the direction of the disk.

of FourierCam to directly acquire specific temporal spectral components in the optical domain, it can overcome the drawbacks of the aforementioned methods. In addition to subtracting the background, preanalysis on the temporal spectrum profile of the objects of interest gives the prior for one to design coding patterns for FourierCam to realize specific object extraction.

To demonstrate the background subtraction capability of FourierCam, we capture a scene that has a rotating disk as a target object and a static poker card as background [Fig. 4(a), left]. The exposure time is 0.5 s, and the temporal frequencies of these four rings are 273, 455, 637, and 1001 Hz, respectively. Only the frequency that corresponds to one ring is applied for coding [Fig. 4(a)] in this case. In this way, each ring can be separately extracted without the background static poker card [Fig. 4(b)]. The results also indicate that one can distinguish objects with the same rotating speed but different textures. In comparison, objects with the same texture but different speeds can also be extracted separately. In Fig. 4(c) (left), two identical disks, both with six stripes, are present in the scene. They rotate at 1980 r/min (200 Hz in the temporal spectrum) and 800 r/min (80 Hz in the temporal

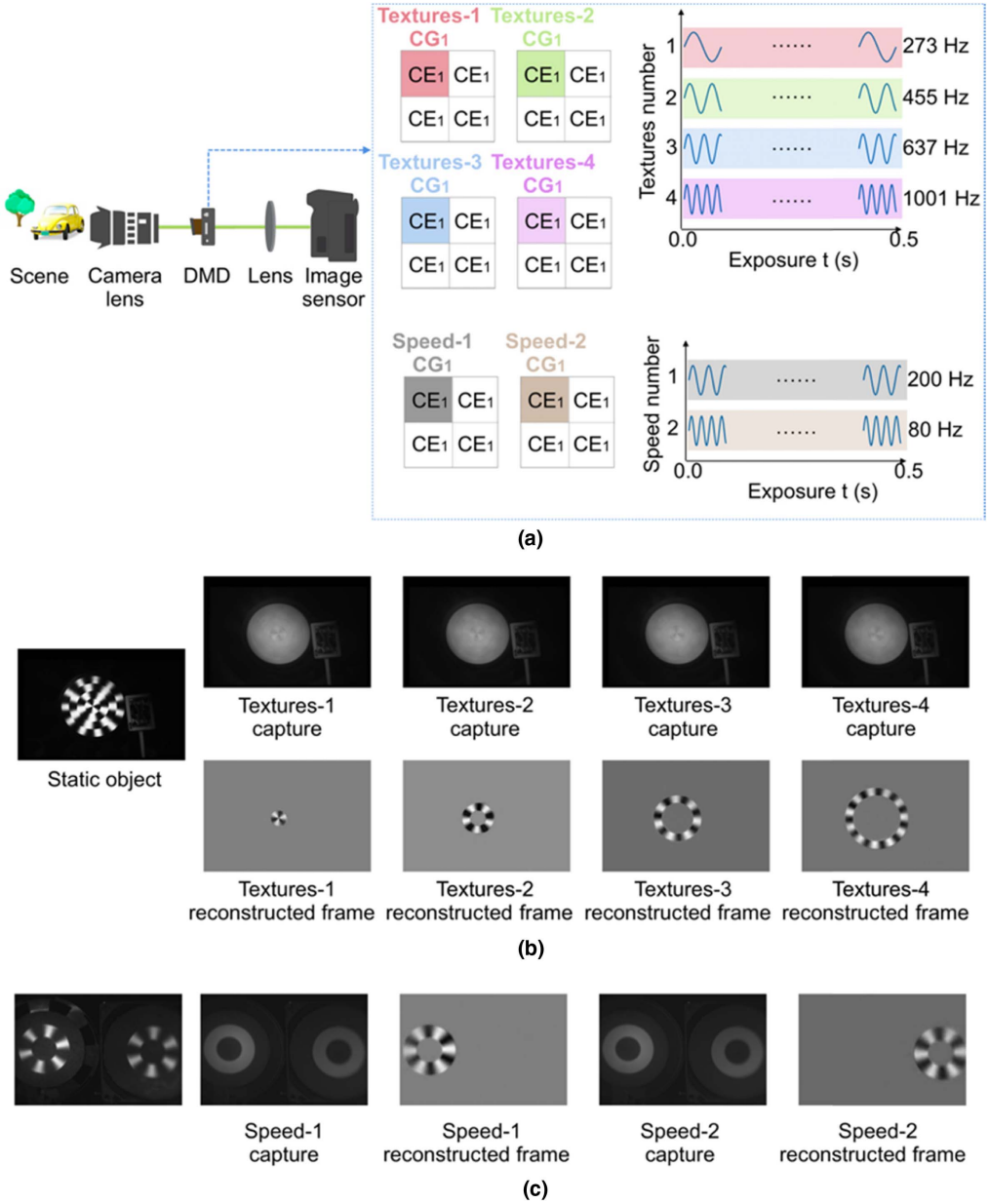

Fig. 4. Object extraction by FourierCam. (a) Illustration of object extraction. The coding frequencies are based on the spectrum of the objects of interest. In this demonstration, the four rings on the disk are regarded as four objects of interest. Each ring only contains one frequency so that one CE is used in one CG. (b) Left: reference static scene with a disk and a poker card. The disk is rotating when capturing, and the four rings share the same rotating speed. Four right columns: FourierCam captured data for four rings extraction and corresponding results. For each extracted ring, other rings and static poker card are neglected. (c) Results for two identical rings rotating at different speed (1980 and 800 r/min, respectively). FourierCam enables extraction of a specific one out of these two rings.

spectrum) relatively, and they appear the same in the capture of the ordinary slow camera. With FourierCam one can see the difference in the coding data and can extract a specific one out of them [Fig. 4(c)]. For simplicity, the above results are all one frame in the reconstructed video.

The results show that FourierCam enables background subtraction and object extraction based on the temporal spectrum difference. Although only one frequency was used in the experiment, in principle it allowed using multiple frequencies to reconstruct more complex scenes, as long as the spectral difference is sufficiently obvious. It is worth noting that in some special cases objects with different textures and speeds may have the same spectral features, making FourierCam fail to distinguish them (see Appendix F for details).

5. TEMPORAL PHASE: TRAJECTORY TRACKING

Object detection and trajectory tracking for a fast-moving object have found important applications in various fields. In general, object detection is to determine the presence of an object, and object tracking is to acquire the spatial–temporal coordinates of a moving object. For the temporal waveform of a pixel where the object would pass by, the moving object takes the form of a pulse at a specific time. As the object is moving, the temporal waveforms at different spatial positions are of different temporal pulse positions, resulting in a phase shift in their temporal spectra. Since Fourier transform is a global-to-point transformation, one can extract the information of the presence and position of the pulse in the temporal domain from the amplitude and phase of a single Fourier coefficient. From this perspective, one can use FourierCam to determine the presence or/and simultaneously acquire the spatial trajectory and temporal position of a moving object.

To detect and track the moving object, only one frequency is needed to encode the scene. In this case, we let $p = q = 1$ and $f = f_0 = 1/t_{\text{expo}}$. Thus, f_0 is the lowest resolvable frequency, and its Fourier coefficient F_{j0} provides sufficient knowledge of presence or/and motion of object. The amplitude A_{j0} of F_{j0} is $A_{j0} = \text{abs}(F_{j0})$, where abs$(*)$ denotes the absolute operation. As a static scene does not contain the f_0 component in the temporal spectrum, moving object detection can be achieved by applying a threshold on A_{j0} that an A_{j0} larger than the threshold indicates the presence of moving objects.

For moving object tracking, since the long exposure has already given the trace of the object, the phase P_j of F_{j0} is utilized to further extract the temporal information: $P_j = \arg(F_{j0})$, where arg$(*)$ denotes the argument operation. A temporal waveform with a displacement of t_j in the temporal domain results in a linear phase shift of $-2\pi f_0 t_j$ in the temporal spectrum:

$$I_j(t - t_j) = \mathcal{F}^{-1}\{F_{j0} \times \exp(-i2\pi f_0 t_j)\}. \tag{5}$$

Therefore, the temporal displacement can be derived through

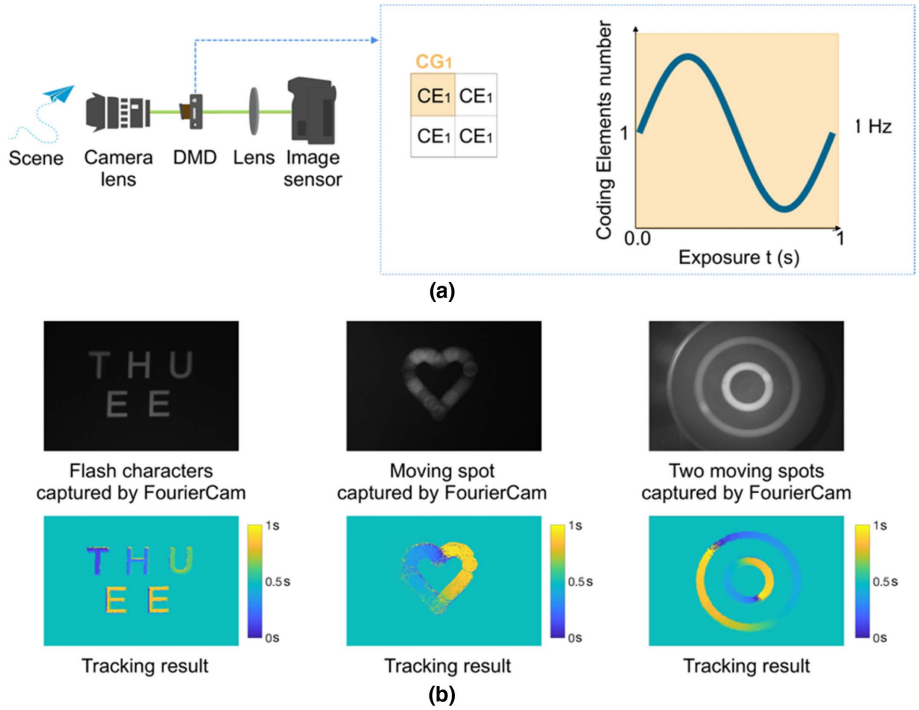

Fig. 5. Moving object detection and tracking by FourierCam. (a) Only one frequency is needed to encode the scene for moving object detection and tracking. The period of sinusoidal coding signal is equal to the exposure time. Thus, only one CE is contained in each CG. (b) Coded data captured by FourierCam and tracking results. Left column: characters 'T', 'H', 'U', 'EE' sequentially displayed by a screen with a 0.25 s duration for each. The color indicates the distribution of appearing time. Middle column: results for a displayed spot moving along a heart-shaped trajectory. Right column: results for two spots moving in circular trajectories with different radii. The spots are printed on a rotating disk driven by a motor.

$$t_j = t_{\text{expo}} \times \frac{P_j}{2\pi}. \tag{6}$$

By applying the same operation to all CGs, we can extract the temporal information for all CGs and acquire the spatial–temporal coordinates of a moving object in the scene.

To test this capability of FourierCam, we capture several targets ranging from flash characters, a single moving object, and multiple objects. The image sensor exposure time is 1 s, and the corresponding coding signal on DMD is 1 Hz [Fig. 5(a)] to ensure one period is contained by a single exposure to avoid 2π phase ambiguity due to the periodicity of the Fourier basis coding. First, a screen displays 'T', 'H', 'U', 'EE' sequentially with a 0.25 s duration for each (see Visualization 3). The raw capture and the extracted temporal position are shown in Fig. 5(b) (the left column), which indicates that FourierCam is able to detect the objects via amplitude and distinguish different temporal positions of objects via phases. Then a spot moving along a heart-shaped trajectory, displayed on the screen, is used as a target to test the tracking capability of FourierCam (see Visualization 4). This result [Fig. 5(b), the middle column] shows FourierCam can resolve the spatial and temporal position of the object. We also test FourierCam on actual multiple objects, which are two spots moving in circular trajectories with different radii [Fig. 5(b), the right column]. The spots are printed on a rotating disk driven by a motor at a speed of 60 r/min. The scene is also recorded by a relatively high-speed camera for reference (see Visualization 5). The temporal resolution is determined by both the exposure time and coding frequency (see Appendix G for details); that is, the higher the coding frequency is, the higher temporal resolution will be, but the temporal range also narrows at the same time. For the current setup, the temporal resolution is 3.9 ms. By applying phase unwrapping algorithms, the trade-off between temporal resolution and temporal range can be overcome to further improve the tracking performance.

6. DISCUSSION AND CONCLUSION

The main achievement of this work is the implementation of a high-quality temporal spectrum vision sensor that represents a concrete step toward the low detection bandwidth, low computational burden, and low data volume novel video camera architecture. In the experiment, we demonstrate the advantages of FourierCam in machine vision applications such as video compression, background subtraction, object extraction, and trajectory tracking. Among these applications, prior knowledge is not required for aperiodic video compression, background subtraction, and trajectory tracking (see Table 1 in Appendix H for details). These applications cover the most common scenarios and can be integrated with existing machine vision systems, especially autonomous driving and security [11]. The emergence of prior knowledge makes FourierCam lose some flexibility but gain better performance. Applications that require prior knowledge (periodic video compression and specific object extraction) have special scenarios (e.g., modal analysis of vibrations). Several engineering disciplines rely on modal analysis of vibrations to learn about the physical properties of structures. Relevant areas include structural health monitoring [12] and nondestructive testing [13,14]. These special scenarios are usually stable (i.e., require less flexibility) and allow better performance at a higher cost.

It is worth mentioning that the FourierCam is built to enhance the flexibility of information utilization with the given limited data throughput. First, by taking the low-frequency properties of a natural scene, one can only sample the most significant low-frequency components to perform data compression during data acquisition with the frequency sampling flexibility of FourierCam. This compression based on frequency is similar to the JPEG [15] compression based on spatial frequency, that is, to store more significant information within limited data capability. In general, this is a kind of lossy compression, and it can also be lossless for some sparse scenes (such as periodic motion). Second, the FourierCam directly obtains the temporal spectrum as a special data type with abundant physical information of the dynamic scenes. Although the process that uses multiple DMD pixels and camera pixels to decode one frequency component brings data cost, the phase-shift operation of the multiple pixels can also reduce the background noise so that the quality of the data can be increased.

The temporal and spatial resolutions are the key parameters of the FourierCam. The temporal resolution (the highest frequency component that can be acquired) is determined by the bandwidth of the modulator. In the present optical system, the PWM mode reduces the DMD refresh rate. Zhang *et al.* [16] used error diffusion dithering techniques to binarize the Fourier basis patterns in space, which can be referenced in the temporal domain to maintain the refresh rate of DMD. In terms of spatial resolution, each Fourier coefficient is in need of 4 pixels for four-step phase-shifting. Although the four-step phase-shifting offers better measurement performance, one can also utilize three-step phase-shifting [16] or two-step phase-shifting [17] for a higher spatial resolution. Furthermore, taking a closer look at the process, one can notice that the principle of FourierCam is similar to the color camera based on the Bayer color filter array (CFA) [18]. CFA and FourierCam use different pixels to collect different wavelengths and temporal Fourier coefficients in parallel, respectively. Therefore, the demosaicing algorithm in CFA can be introduced into FourierCam to improve the spatial resolution [19,20]. Although a monochrome image detector is used in the experiments, the possibility of combining FourierCam with a color image detector is obvious, as long as the coding structure of FourierCam needs to be adjusted according to the distribution of CFA. It is worth mentioning that in machine vision based on deep learning, training and inference on the temporal spectrum is feasible through complex-valued neural networks, without the need for image restoration as an intermediate step [21,22]. We believe that the data format of the temporal spectrum provided by FourierCam has the potential to be used in multimodal learning for high-level vision tasks like optical flow or event flow [23]. In addition, proposing a more compact and lightweight design will help develop a commercial FourierCam. One can borrow the compact optical design from miniaturized DMD-based projectors, or one can integrate the modulator on the sensor chip, which is still challenging with current technology. And in some applications with

loose frame rate requirements, a commercial liquid crystal modulator can be used instead of DMD to reduce costs. Beyond machine vision, we believe that the flexible temporal filter kernel design properties of FourierCam can play a role in other fields, for example, using FourierCam to perform frequency division multiplexing demodulation in space optical communication or to extract specific signals in voice signal detection.

APPENDIX A: CORRESPONDENCE BETWEEN DMD AND IMAGE SENSOR IN FOURIERCAM

In the FourierCam the most important thing is to adjust each mirror of the DMD so as to correspond exactly to the pixel of the image sensor, such as CCD or CMOS. Under the premise of complete correspondence, FourierCam can achieve high-precision decoding. However, since the sizes of the CCD and DMD are very small, it is difficult to accurately align. Fortunately, CCD and DMD can be regarded as two gratings, so they can be aligned by observing the moiré fringes formed between them [24]. There are two kinds of errors: mismatch and misalignment. Mismatch means line spatial frequency disagreement, and misalignment means rotational disagreement. When each mirror of the DMD and each pixel of the CCD are not corresponding exactly, a diverse moiré fringe pattern according to the mismatch and misalignment conditions will appear. Figure 6 shows the experimental results when we adjust the pixel-to-pixel correspondence in the FourierCam. Figure 6(a) shows the moiré fringe patterns when the mismatch and misalignment occur between the CCD pixels and the DMD. Adjusting the rotation angle of the DMD can eliminate misalignment as shown in Fig. 6(b). Next, adjusting the magnification of lens, the moiré pattern does not appear in the FourierCam as shown in Fig. 6(c). In the statement of Fig. 6(c), the adjustment error is 0.02%, which means that for every 5000 pixels, a pixel offset will occur. Therefore, high-precision correspondence between DMD and CCD is realized in FourierCam.

APPENDIX B: DETAILED DISCUSSION ABOUT FEATURES OF FOURIERCAM

Detection bandwidth: To measure a temporal significance with max frequency $f_{\max}$, the required minimum detection bandwidth of traditional cameras equals $f_{\max}$. For FourierCam acquiring h Fourier components, the required minimum detection bandwidth is $\frac{f_{\max}}{2h}$ according to the frequency domain sampling theorem (see Appendix C). For example, in the natural scene demonstration (toy car and panda in the manuscript), $f_{\max}$ is 80 Hz and eight Fourier components except from the direct current are obtained; thus, the required detection bandwidth of FourierCam is 5 Hz, while for traditional cameras it is 80 Hz.

Assuming a video is captured by traditional cameras with M frames and N pixels in each frame, its data volume is $M \times N$ bytes (assuming 1 byte for one pixel). FourierCam obtains h Fourier components of the same video, and the data volume is $2h \times N$ bytes since a complex Fourier coefficient needs twice the capacity than a real number. Generally, M is larger than $2h$. For example, in the "running dog" video in Appendix D, $M = 100$, $h = 16$, and $N = 1080^2$; thus, the data volumes for a traditional camera and FourierCam are 116.64 and 18.66 megabytes, respectively. By considering the prior information of the object and applying selective sampling, the data volume can be further reduced.

Floating point operations (FLOPs) comparison between FFT and FourierCam: FLOPs include the standard floating-point operations of additions and multiplications to evaluate the computational burden. To calculate the temporal spectrum of a video with M frames and N pixels in each frame, the fast Fourier transform (FFT) needs $5MN \log_2 M$ FLOPs. In FourierCam, since the multiplication and summation operations of Fourier transform are realized by optical coding, only $3MN$ FLOPs are required for the four-phase-shifting operation. Therefore, the required FLOPs for the temporal spectrum acquisition can be reduced by $(5M \log_2 M - 3M) \times N$. For example, in the demonstration of the periodic motion in application II in the paper, 3.9 GFLOPs can be neglected by FourierCam.

Light throughput in FourierCam: In addition to the above advantages, light throughput plays an important role in high-speed photography and is worthy of discussion. Two types of high-speed cameras (including normal high-speed shutter and impulse coding cameras) are used for comparison. The impulse coding cameras turn on the pixels in a spatial block at a certain time to capture high-speed video [25,26]. Considering one coding group, the average light intensity at a coding group is L, the active area is A, the video has N frames, and the entire duration is T. So the frame rate requirement for the capture device is $\frac{N}{T}$. For high-speed shutter cameras, the whole area A will be an active area, and the light throughput of one frame is $L \times A \times \frac{T}{N}$. Therefore, the light throughput of N frames video is $L \times A \times T$. For impulse coding cameras, the whole area A will be divided into N exposure groups, with each group exposing sequentially. The light throughput per frame (exposure groups)

Fig. 6. Phase analysis of the moiré fringe pattern obtained by the phase-shifting moiré method. (a) There are two errors: mismatch and misalignment. (b) Only mismatch error. (c) FourierCam with high-precision correspondence.

is $L \times \frac{A}{N} \times \frac{T}{N}$, and the light throughput of N frames video is $L \times \frac{A}{N} \times T$. For FourierCam, each coding group will be divided into $p \times q \times N_{\text{phase}}$ smallest units (N_{phase} is the number of phases and in aperiodic scenes $p \times q = \frac{N}{2}$), and each unit is modulated by a sinusoidal signal during the whole exposure time of the image detector, so the light throughput of each unit is $\frac{L \times A \times T}{N \times N_{\text{phase}}}$. Similar to the abovementioned temporal domain sampling strategy, which superimposes all frames to calculate the light throughput, the FourierCam should also add all frequency components to calculate the video light throughput. Therefore, the light throughput of FourierCam is $\frac{L \times A \times T}{N \times N_{\text{phase}}} \times p \times q = \frac{L \times A \times T}{2 \times N_{\text{phase}}}$. In summary, the light throughput of the FourierCam is lower than that of high-speed shutter cameras (even when $N_{\text{phase}} = 1$); this is introduced by sinusoidal modulation. However, the light throughput of the FourierCam has nothing to do with the number of frames N, while the impulse coding cameras are related to it. When N increases, the light throughput advantage of FourierCam compared to impulse coding cameras becomes more obvious. In principle, FourierCam uses at least two phases which have a 180-deg shift. Fortunately, by using the light from both ON and OFF reflection angles of DMD and adding a second sensor, it is possible to complete the temporal spectrum acquisition with each sensor collecting only one phase. This means that $N_{\text{phase}} = 1$, which can realize the competitive light throughput as high-speed shutter cameras.

APPENDIX C: FRAME RATE AND FREQUENCY DOMAIN SAMPLING IN FOURIERCAM

Traditional cameras can be regarded as the temporal-domain sampling process when capturing video, and the frame rate is the temporal sampling rate. Considering each pixel temporal waveform, given the frame rate, the highest frequency component, $f_{\max}$, that can be acquired is $\frac{f_s}{2}$, where f_s is the frame rate. Unlike the temporal-domain sampling process of traditional cameras, FourierCam is based on frequency-domain sampling. FourierCam directly acquires frequency components. When the highest frequency component it collects is $f_{\max}$, the equivalent frame rate of FourierCam is $2f_{\max}$. In addition, the frequency domain sampling interval (Δf) of FourierCam needs to satisfy the frequency domain sampling theorem to ensure that the reconstructed video does not alias in the time domain. The frequency domain sampling interval is determined by the exposure time of the image detector (t_{expo}), $\Delta f \le \frac{1}{t_{\text{expo}}}$. For example, the exposure time of an image detector is 1 s, and the frame rate is 1 Hz. If the frame rate is increased to 10 Hz, the frequency components to be acquired are 1 Hz, 2 Hz, 3 Hz, 4 Hz, and 5 Hz. Its frequency interval is 1 Hz, which satisfies the frequency domain sampling theorem.

APPENDIX D: QUANTITATIVE ANALYSIS ON THE PERFORMANCE OF FOURIERCAM

To quantitatively evaluate the reconstruction, we perform a simulation of FourierCam with a "running dog" video, which has 100 frames with a spatial resolution of 1080 × 1080 pixels. We obtain the temporal spectrum of the video with 16 frequencies (the number of acquired Fourier coefficients $h = 16$). Figure 7(a) compares the long exposure capture and the FourierCam encoded capture. The long exposure with low temporal resolution results in an obvious motion blur, and the details of the object are lost, whereas the temporal spectrum contains information of the motion to further reconstruct the dynamic scene. In the reconstructed video, the SSIM (structural similarity index) keeps stable with an average of 0.9126 and a standard deviation of 0.0107 [shown in Fig. 7(b)]. In Fig. 7(c), we also display a visual comparison of three exemplar frames from the ground truth video and the FourierCam reconstructed results, respectively. These results illustrate that FourierCam is able to reconstruct a clear video with only low-frequency coefficients.

The trade-off between temporal resolution and spatial resolution exists in FourierCam. By acquiring h Fourier coefficients, the frame rate can be improved by $2h$ times at the cost of L times reduction in spatial resolution. Each Fourier coefficient is in need of four pixels for a four-step phase-shifting operation; thus $L = 4h$. Therefore, the reconstructed spatial resolution is inversely proportional to h [shown in Fig. 8(a)]. To illustrate this relationship, we capture the "running dog" video with 4, 9, 16, 25 frequencies. Four corresponding frames

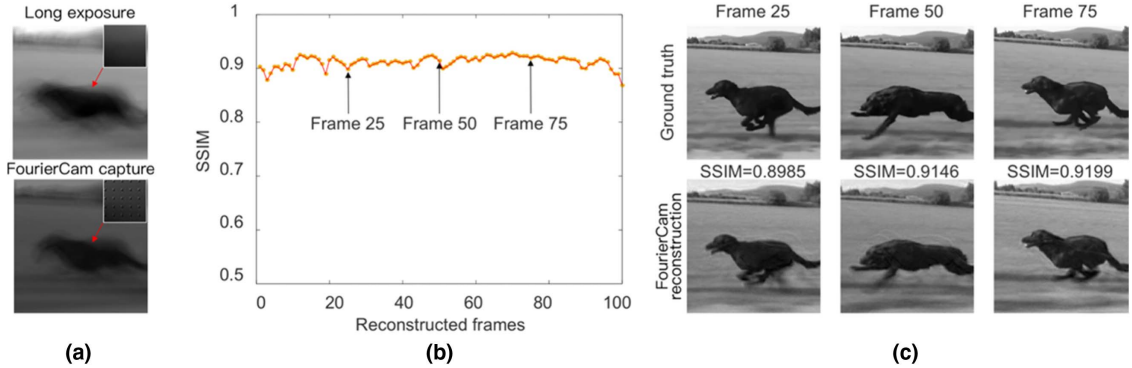

Fig. 7. Simulation of FourierCam video reconstruction. (a) Long exposure capture with all frames directly accumulating together, corresponding to a slow camera and the FourierCam encoded capture. The insets show the zoom-in view of the areas pointed by the arrows. (b) In the reconstructed video with 16 Fourier coefficients, the SSIM of each frame keeps stable with an average of 0.9126 and a standard deviation of 0.0107. (c) Three exemplar frames from the ground truth and reconstructed video.

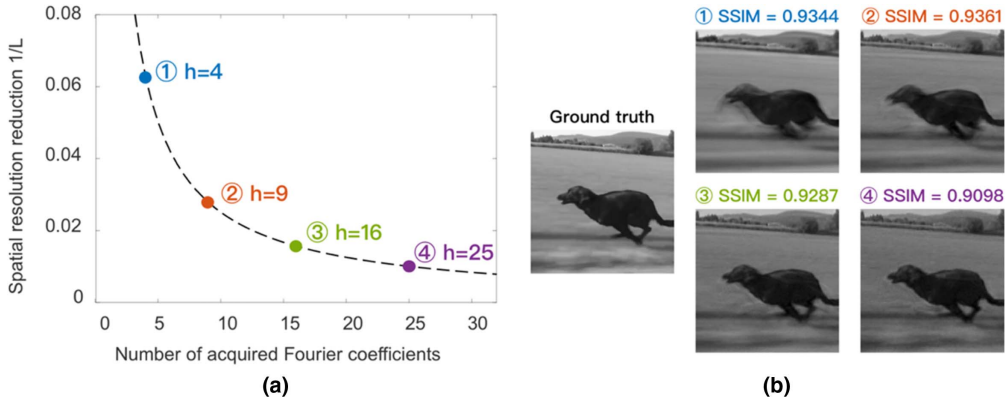

Fig. 8. Quantitative analysis on the performance of FourierCam. (a) Relation between number of acquired Fourier coefficients h and spatial resolution reduction L of FourierCam. (b) Comparison of reconstructed frames with different numbers of acquired Fourier coefficients, corresponding to point 1 to point 4 in (a).

from the four reconstructed videos with ground truth are shown in Fig. 8(b). With h increasing, the SSIM remains stable. If h becomes too large, the SSIM slightly decreases but still remains larger than 0.9. The reason is that the motion blur gets eased with the effective frame rate improved, but the increase of the number of frequencies causes the reduction of the spatial resolution. These results indicate that one may properly decide the number of frequencies in FourierCam based on the concrete need and scenario to balance the spatial and temporal resolutions.

APPENDIX E: FOURIER DOMAIN PROPERTIES OF PERIODIC AND APERIODIC MOTION

Consider the signal at a position where a periodic motion passes: it is in periodic form in time domain. Fourier transform of a periodic signal with period P contains energy only at the frequencies that are an integer multiple of repetition frequency $\frac{1}{P}$, and therefore the periodic signal has a sparse representation in the Fourier domain. When the period of the periodic signal becomes infinitely long, the periodic signal comes to an aperiodic signal with a single pulse and its spectrum becomes continuous. Figure 9 provides a graphical illustration of the spectrum of periodic and aperiodic signals.

APPENDIX F: TEMPORAL RESOLUTION OF OBJECT TRACKING IN FOURIERCAM

As the object is moving, the temporal waveforms at different spatial positions are of different temporal pulse positions, resulting in a phase shift in their temporal spectra. The phase-shift detection accuracy is the temporal resolution of object tracking in FourierCam. The phase-shift accuracy is determined with the DMD grayscale level and the exposure time of the image detector, so the temporal resolution is $\frac{t_{\text{expo}}}{\text{DMD}_{\text{grayscale}}}$. Since we use a DMD with PWM mode as the spatial light modulator in FourierCam, the light is digitally modulated by 8-bit grayscale. Therefore, during a single exposure t_{expo}, the temporal resolution of object tracking is $\frac{t_{\text{expo}}}{256}$.

APPENDIX G: FOURIER DOMAIN PROPERTIES OF MOVING OBJECT

Changes in both the texture and the speed of the moving object can cause a difference in the Fourier domain. As illustrated in Fig. 10(a), when a block with sinusoidal fringe texture is moving at a speed of v, the detected waveform at the red point is also in a sinusoidal form. In Fig. 10(b), a block with a higher

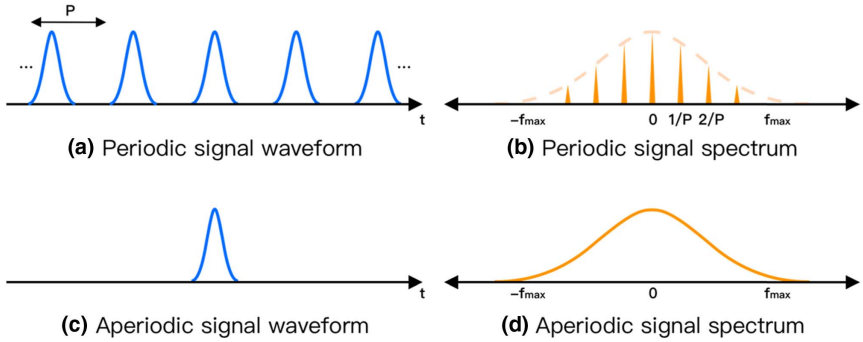

Fig. 9. Fourier domain properties of periodic and aperiodic signals. The (a) periodic signal has a (b) sparse spectrum while the (c) aperiodic signal has a (d) continuous spectrum.

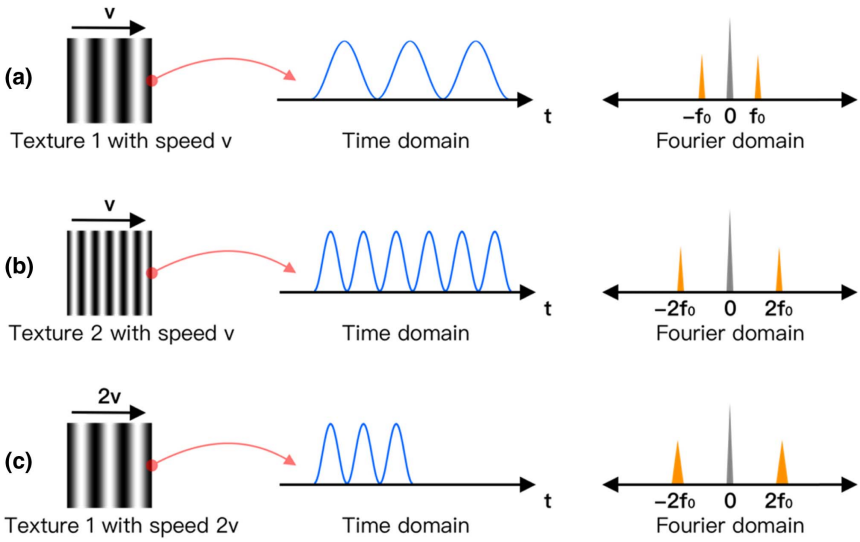

Fig. 10. Illustration of Fourier domain properties of moving objects with different texture and speed. (a) Block with sinusoidal fringe texture moving at a speed of v. The temporal waveform of the red point is shown with its Fourier spectrum. (b) Block with higher spatial frequency texture, also moving at the speed of v. (c) Block identical to (a) but moving at a higher speed $2v$.

Table 1. Comparison Between Different Application for FourierCam

Application	Prior Knowledge	Scenario	Coding Method
Video compression	×	Normal	Multifrequency coded signals depend on exposure time
Selective sampling (Periodic motion video acquisition)	Motion period	Periodic	Multifrequency coded signals depend on motion period
Selective sampling (Background subtraction)	×	Normal	Multifrequency DC components are not included
Selective sampling (Object extraction)	Temporal spectrum profile of the interest objects	Normal	Multifrequency coded signals depend on prior knowledge
Trajectory tracking		Normal	Single-frequency coded signals depend on exposure time

spatial frequency texture but also moving at the speed of v corresponds to a higher frequency in the Fourier domain compared to Fig. 10(a). By selectively acquiring a specific range of frequency (e.g., $2f_0$), we can extract a specific object [e.g., the one in Fig. 10(b)]. Also, the change in moving also causes a difference in the spectrum [Fig. 10(c)]; thus, we can also extract it from the one in Fig. 10(a). However, because of the joint effect of texture and speed, the spectrum in Figs. 10(b) and 10(c) is quite similar. To distinguish these two objects, we can add more constraints such as the length of the waveform, which is one of our future works.

APPENDIX H: COMPARISON BETWEEN DIFFERENT APPLICATION FOR FOURIERCAM

The comparison between different applications for FourierCam is shown in Table 1. In periodic compressive video reconstruction, *a priori* knowledge can be used to achieve higher compression ratios. It is also possible not to use prior knowledge, in which case the compression ratio is the same as the aperiodic.

Funding. National Key Research and Development Program of China (2019YFB1803500); National Natural Science Foundation of China (61771284).

Disclosures. The authors declare no conflicts of interest.

†These authors contributed equally to this paper.

REFERENCES

1. J. N. Mait, G. W. Euliss, and R. A. Athale, "Computational imaging," Adv. Opt. Photon. **10**, 409–483 (2018).
2. J. Liang and L. V. Wang, "Single-shot ultrafast optical imaging," Optica **5**, 1113–1127 (2018).
3. G. Gallego, T. Delbruck, G. M. Orchard, C. Bartolozzi, B. Taba, A. Censi, S. Leutenegger, A. Davison, J. Conradt, K. Daniilidis, and D. Scaramuzza, "Event-based vision: a survey," IEEE Trans. Pattern Anal. Mach. Intell. (2021).
4. Z. W. Wang, V. Vineet, F. Pittaluga, S. N. Sinha, O. Cossairt, and S. B. Kang, "Privacy-preserving action recognition using coded aperture videos," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)* (IEEE, 2019), pp. 1–10.
5. T. Ouni, W. Ayedi, and M. Abid, "New low complexity DCT based video compression method," in *International Conference on Telecommunications* (IEEE, 2009), pp. 202–207.

6. W. Wang, J. Yang, and W. Gao, "Modeling background and segmenting moving objects from compressed video," IEEE Trans. Circuits Syst. Video Technol. **18**, 670–681 (2008).
7. D.-M. Tsai and W.-Y. Chiu, "Motion detection using Fourier image reconstruction," Pattern Recogn. Lett. **29**, 2145–2155 (2008).
8. T.-H. Oh, J.-Y. Lee, and I. S. Kweon, "Real-time motion detection based on discrete cosine transform," in *19th IEEE International Conference on Image Processing* (IEEE, 2012), pp. 2381–2384.
9. K. Xu, M. Qin, F. Sun, Y. Wang, Y.-K. Chen, and F. Ren, "Learning in the frequency domain," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (IEEE, 2020), pp. 1740–1749.
10. D. Doherty and G. Hewlett, "10.4: phased reset timing for improved digital micromirror device (DMD) brightness," SID Symp. Dig. Tech. Papers **29**, 125–128 (1998).
11. S. Ojha and S. Sakhare, "Image processing techniques for object tracking in video surveillance: a survey," in *International Conference on Pervasive Computing (ICPC)* (IEEE, 2015), pp. 1–6.
12. I. Ishii, S. Takemoto, T. Takaki, M. Takamoto, K. Imon, and K. Hirakawa, "Real-time laryngoscopic measurements of vocal-fold vibrations," in *Annual International Conference of the IEEE Engineering in Medicine and Biology Society* (IEEE, 2011), pp. 6623–6626.
13. A. Davis, M. Rubinstein, N. Wadhwa, G. J. Mysore, F. Durand, and W. T. Freeman, "The visual microphone: passive recovery of sound from video," ACM Trans. Graph. **33**, 79 (2014).
14. A. Davis, K. L. Bouman, J. G. Chen, M. Rubinstein, F. Durand, and W. T. Freeman, "Visual vibrometry: estimating material properties from small motion in video," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (IEEE, 2015), pp. 5335–5343.
15. G. Wallace, "The JPEG still picture compression standard," IEEE Trans. Consum. Electron. **38**, xviii–xxxiv (1992).
16. Z. Zhang, X. Wang, G. Zheng, and J. Zhong, "Fast Fourier single-pixel imaging via binary illumination," Sci. Rep. **7**, 12029 (2017).
17. L. Bian, J. Suo, X. Hu, F. Chen, and Q. Dai, "Efficient single pixel imaging in Fourier space," J. Opt. **18**, 085704 (2016).
18. B. E. Bayer, "Color imaging array," U.S. patent 3,971,065 (July 20, 1976).
19. H. S. Malvar, L.-W. He, and R. Cutler, "High-quality linear interpolation for demosaicing of Bayer-patterned color images," in *IEEE International Conference on Acoustics, Speech, and Signal Processing* (IEEE, 2004), paper iii-485–8.
20. R. Ramanath, W. E. Snyder, G. L. Bilbro, and W. A. Sander, "Demosaicking methods for Bayer color arrays," J. Electron. Imaging **11**, 306–315 (2002).
21. W. Chen, J. Wilson, S. Tyree, K. Q. Weinberger, and Y. Chen, "Compressing convolutional neural networks in the frequency domain," in *22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (ACM, 2016), pp. 1475–1484.
22. H.-S. Choi, J.-H. Kim, J. Huh, A. Kim, J.-W. Ha, and K. Lee, "Phase-aware speech enhancement with deep complex U-Net," in *International Conference on Learning Representations* (2019), pp. 1–20.
23. T. Baltrušaitis, C. Ahuja, and L.-P. Morency, "Multimodal machine learning: a survey and taxonomy," IEEE Trans. Pattern Anal. Mach. Intell. **41**, 423–443 (2019).
24. 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).
25. G. Bub, M. Tecza, M. Helmes, P. Lee, and P. Kohl, "Temporal pixel multiplexing for simultaneous high-speed, high-resolution imaging," Nat Methods **7**, 209–211 (2010).
26. D. Hutchison, T. Kanade, J. Kittler, J. M. Kleinberg, F. Mattern, J. C. Mitchell, M. Naor, O. Nierstrasz, C. Pandu Rangan, B. Steffen, M. Sudan, D. Terzopoulos, D. Tygar, M. Y. Vardi, G. Weikum, M. Gupta, A. Agrawal, A. Veeraraghavan, and S. G. Narasimhan, "Flexible voxels for motion-aware videography," in *Computer Vision—ECCV*, K. Daniilidis, P. Maragos, and N. Paragios, eds. (Springer, 2010), Vol. **6311**, pp. 100–114.