Advances in
Bridge Engineering

ORIGINAL INNOVATION

Open Access

Bridge vibration measurements using different camera placements and techniques of computer vision and deep learning

Yongsheng Bai[1,2], Halil Sezen[1*], Alper Yilmaz[1] and Rongjun Qin[1]

*Correspondence:
sezen.1@osu.edu

[1] Department of Civil, Environmental and Geodetic Engineering, Ohio State University, 2070 Neil Ave, Columbus 43210, OH, USA
[2] Neural Image Corporation, 1 Lake Bellevue Dr, Bellevue 98005, WA, USA

Abstract

In this paper, a new framework is proposed for monitoring the dynamic performance of bridges using three different camera placements and a few visual data processing techniques at low cost and high efficiency. A deep learning method validated by an optical flow approach for motion tracking is included in the framework. To verify it, videos taken by stationary cameras of two shaking table tests were processed at first. Then, the vibrations of six pedestrian bridges were measured using structure-mounted, remote, and drone-mounted cameras, respectively. Two techniques, displacement and frequency subtractions, are applied to remove systematic motions of cameras and to capture the natural frequencies of the tested structures. Measurements on these bridges were compared with the data from wireless accelerometers and structural analysis. Influences of critical parameters for camera setting and data processing, such as video frame rates, data window size, and data sampling rates, were also studied carefully. The research results show that the vibrations and frequencies of structures on the shaking tables and existing bridges can be captured accurately with the proposed framework. These camera placements and data processing techniques can be successfully used for monitoring their dynamic performance.

Keywords: Vibration measurement, Deep learning, Camera placements, Mask R-CNN, Displacement and frequency subtractions

1 Introduction

Most bridges and buildings are large in size and volume. It is difficult and challenging for measuring their response to external excitations and assessing their performance at full scale for Structural Health Monitoring (SHM) missions. In practice, SHM can be implemented at a local scale when a limited number of sensors are placed only at critical locations of important infrastructure. Vibration- and vision-based sensors are commonly used for deformation and vibration measurements of existing structures. The former can record the acceleration or dynamic response of the structures, thus, inherent dynamic characteristics such as structural frequencies and vibration modes can be inferred from the measurements and observations, and utilized to assess the performance of the structures. The vision-based systems in civil engineering have been used not only for

detecting visual damage (Bai et al. 2023), but also for measuring deformations and vibrations (Dong and Catbas 2021). In recent years, the cost of cameras and Unmanned Aerial Vehicles (UAVs) has decreased dramatically to make them an attractive option for SHM.

Computer vision has been used for monitoring the movement of objects accurately, which our vision system can notice but is unable to quantify. The objective of this research is to apply vision-based technologies to quantitatively capture and analyze the motion of structures. We were motivated by studies using computer vision methods to measure the movement of structures (Feng et al. 2015; Chen et al. 2017). Also, some practices (Bai et al. 2021b, 2023) provide insights on how current knowledge of computer vision and deep learning can be facilitated to monitor and measure structural response to excitations in laboratory and field experiments.

In this research, three practical ways of camera placements are illustrated in Figs. 4, 5 and 6 are utilized to capture the motions of structures: 1) stationary cameras placed remotely and focused on a bridge for tracking one or multiple targets simultaneously, 2) structure-mounted cameras fixed on a bridge as contact sensors (e.g., accelerometers, see Section 2) to measure the bridge's vibrations, and 3) UAVs deployed to record both motions of the bridge and drones when nearby stationary objects are utilized as the reference. The first camera placement was tested with laboratory experiments and achieved subpixel accuracy (Bai et al. 2021a). Then, a more comprehensive study is implemented further using data from field experiments and other laboratory tests. A stationary camera itself can be the reference for the moving bridge. Otherwise, the nearby buildings or other motionless surroundings can be treated as a stationary reference to eliminate the movement of cameras caused by ambient motions, including the wind, ground motion induced by the surrounding traffic, and the UAV's movement during its flight. Also, it is strongly believed that the combined experience and expertise of structural engineering and computer vision are helpful to achieve good performance with the proposed framework when the appropriate targets and references are selected. For example, in the experiments of bridge vibration measurements, it is the mid-spans where the most significant motion happens and it is their joint regions where there are more textures that can be considered as the tracking targets. The cameras used in the field experiments cost 60 to 500 US dollars each and were operated at low speeds (i.e., 30 to 84 frames/second). Markers were not needed for the tests.

We have three main objectives for using camera-based technologies easily and effectively to monitor and assess the dynamic performance of in-service bridges: 1) integrating three possible camera placements mounted on different platforms to measure the vibrations of bridges using a deep learning framework, 2) validating the proposed framework of processing visual data with two shaking table tests and field experiments of pedestrian bridges, and 3) applying both displacement and frequency subtractions to remove camera motion in these measurements. These camera placements and techniques are addressed in Section 3, and their limitations are discussed in Sections 4, 5, and 6. Because pedestrian bridges are easily excited and the experiments can be conveniently repeated in the field, six of them were selected as the demonstration for the experimental studies in this paper. With the same framework, tests on traffic and railway bridges and a building in the progressive collapse study were also conducted and showed promising results for deflection and vibration measurements (Bai 2022). In all these

field experiments, accelerations measured by the wireless accelerometers and displacements measured by various cameras were not compared directly. Rather, the structural frequencies obtained from the measured accelerations and displacements (e.g., Figs. 15, 16 and 17 and Table 1) were used to validate the method proposed in this paper.

2 Related work

In order to better understand our work, research, and applications on displacement and vibration measurements with computer vision and deep learning methods, which inspired us to have a new framework in this paper, are addressed in this section. Also, studies about UAVs and wireless accelerometers are reviewed since we used both in our research.

Conventional computer vision techniques for measuring displacement or vibration can achieve high accuracy in practice. Feng et al. (2015) proposed an upsampled cross-correlation as a template matching algorithm to measure the displacement and vibration in a shaking table test and performed two field tests, including the tests on a railway and a pedestrian bridge. Chen et al. (2021) proposed a method with Digital Image Correlation to track the displacement or vibration of individual points on a model bridge. Dong et al. (2019) utilized optical flow estimation to track non-target objects on grandstand structures and implemented modal identification. Some researchers also employed various template-matching algorithms to track and measure the displacement or vibrations of buildings (Lee et al. 2017; Yin et al. 2014; Liu et al. 2016; Rajaram et al. 2017; Chen et al. 2017). Guo and Zhu (2016) applied the Lucas-Kanade template tracking algorithm on displacement measurement in an experiment, which is used as another method in this research. In these studies, how to obtain the subpixel accuracy for the measurements was addressed. An accuracy of 0.016 to 0.25 mm and 0.64 to 3.5 mm is reported, respectively, for the displacement measurements with cameras in these laboratory and field experiments.

Image-based deep learning methods for vibration measurements have achieved better performance since they can extract more useful features in the following research. Dong et al. (2020) implemented FlowNet2 on displacement and vibration measurements of various structures. Their approach to eliminating the movement of cameras during field tests is instructive to us about how to use displacement subtraction for true vibrations.

Table 1 Fundamental frequency (Hz) of six pedestrian bridges measured and extracted by three different camera placements and accelerometers (Accs), and calculated by structural analysis with SAP2000

	Remote	Accs	Structure-mounted	Accs	Drone-mounted	Accs	Structural analysis
CPB-1	M: 3.99 (L: 3.99)	4.03	M: 4.02 (L: 4.02)	3.99	D-: 4.04 (F-: 4.02)	3.99	4.04
CPB-2	M: 5.44 (L: 5.44)	5.44	M: 5.49 (L: 5.47)	5.47	D-: 5.45 (F-: 4.94)	5.45	5.44
CPB-3	M: 5.43 (L: 5.43)	5.43	M: 5.47 (L: 5.45)	5.45			5.44
CPB-4	M: 3.84 (L: 3.84)	3.85	M: 3.84 (L: 3.84)	3.85	D-: 3.91 (F-: 3.87)	3.84	3.88
CPB-5	M: 4.68 (L: 4.68)	4.66	M: 4.69 (L: 4.62)	4.64	D-: 4.70 (F-: 4.68)	4.59	4.75
CPB-6	M: 4.61 (L: 4.61)	4.62	M: 4.61 (L: 4.62)	4.61	D-: 4.68 (F-: 4.68)	4.60	4.75

Remote, Structure- and Drone- stand for remote, structure-mounted and drone-mounted cameras. M and L stand for the Mask R-CNN + SIFT and LK tracker methods, respectively. D- and F- refer to displacement subtraction and frequency subtraction

Xiao et al. (2020) investigated a proposed SHM system using deep learning algorithms to evaluate the structural responses with visual data fused with data from conventional sensors. Dong and Catbas (2019) applied the Visual Graph Visual Geometry Group network to extract features on the target and performed a field test on a two-span bridge. From these papers, an accuracy from 0.0087 to 0.08 mm was reported in the laboratory tests. Bai et al. (2021a) proposed a High-resolution Mask Regional Convolutional Neural Network (HR Mask R-CNN) to track and accurately measure the deflection of three concrete beams, and the vibrations of three masses on a shaking table in the laboratory tests. The deep learning method was trained by following standard data annotation, loss regulation, and parameter settings. Moreover, a measurement-smoothing technique referred to as the Scale-Invariant Feature Transformation (SIFT) was also introduced for high-accuracy measurements. Thus, the average error of deflection measurements from HR Mask R-CNN + SIFT for three test beams is 0.13 mm, and the difference between the extracted and input frequencies is less than 9% by identifying all the intended frequencies. This paper is based on our previous study but is more comprehensive.

Recently, UAVs are largely deployed for vibration measurements for their high mobility and efficiency. These studies (Yoon et al. 2018; Chen et al. 2021; Hoskere et al. 2019; Ribeiro et al. 2021; Perry and Guo 2021; Khuc et al. 2020) provided good examples for researchers to follow, and their methods to process the visual data from the drones are helpful to this research. For example, Khuc et al. (2020) utilized the UAVs to measure the swaying displacement of small-scale structures. Between two consecutive frames in a video, the keypoints on a target were located and matched so that their average movement represents the displacement. This is a method similar to our research on how to eliminate camera motion. But, in our study, all the frames in a video are aligned to the first frame by the affine transformation (Szeliski 2010).

Wireless accelerometers have been used as contact sensors to detect and capture the dynamic response of various bridges (Gheitasi et al. 2016; Gibbs et al. 2019; Baisthakur and Chakraborty 2020; White et al. 2020), and their data can validate the vibration measurements from cameras. In our study, several G-Link-200-8G wireless accelerometers (LORD 2022) were employed as ground-truth sensors. These accelerometers have high sensitivity for three axes (i.e., the input range is $\pm 2/2/4/8g$), and their bandwidth can reach $1\,kHz$. Its noise will be lowered to $25\,\mu g\sqrt{Hz}$, but its wireless range and sampling rate can be up to one kilometer and $4\,kHz$. Also, programmable high- and low-pass digital filters are utilized in the built-in program. This sensor can work continuously until the batteries lose their power. In addition, multiple accelerometers can be used simultaneously for lossless data collections, scalable network sizes, and node synchronizations of $\pm 50\,\mu s$. Two or three of them were placed on testing bridges as another vibration data source in our field experiments.

3 Methodologies

In this section, how a camera can perform the displacement or vibration measurements is introduced at first, then, three different camera placements, filters to suppress the noise, and two techniques to remove the camera motion are discussed. The experiments conducted on the pedestrian bridges belong to free-damped vibration, in which a bridge is subjected to an initial excitation (e.g., a jump on the bridge's deck)

and can vibrate at one or more frequencies. This oscillation diminishes from the peak to the standstill because of damping (Chopra 2019). In contrast, simulated structures on shaking tables are forced to vibrate by the tables and reach their resonant states when their frequencies are equal to the input ones from the tables.

3.1 Two-dimensional displacement measurements

Cameras can be used to capture the movement of an object or a target in the real world and save their projection on films (conventional cameras) or electronic storage devices (digital cameras). The pin-hole model is a typical model to interpret the relationship between a 2D image and the real world. The motion of this target can be recorded with a camera as shown in Fig. 1: the target represented by a point moves from position A to position B in real-world coordinates, but its motion can be projected to the image plane so that its trajectory ab is captured by the camera. This motion can be decomposed into two principal directions (i.e., dx and dy) defined by the image plane, in which the left corner is the origin while x and y axes refer to the direction from left to right and from top to bottom, respectively. The coordinates of each point on the image plane are in pixels, where one pixel is the smallest 2D square to divide the image plane evenly in two directions. Pixels can be used as the length unit of the image plane for measurement. For example, the convention to describe the pixel resolution is to use the set of two integer numbers, width and height. Width is the number of pixel columns in x direction and height is the number of pixel rows in y direction.

If each tracking target is assumed to be a rigid body, its motion can be represented by any point on it or by a bounding box to scope it. As shown in Fig. 2, the translation of a target between the first and ith frame, dx^i and dy^i, can be calculated as the position change of the bounding box or the average change of tracking points in the image plane of a stationary camera. The measured motions of a target can be directly used as its displacement since the camera is standstill and not affected by the target's movements. Otherwise, the motionless surroundings have to be utilized to remove the camera's motion in the measurements (see Eqs. 4 and 5).

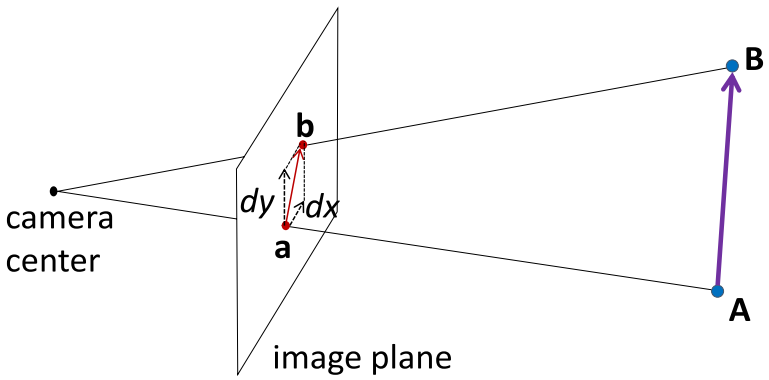


Fig. 1 Projection of displacement from A to B in the real world on an image plane

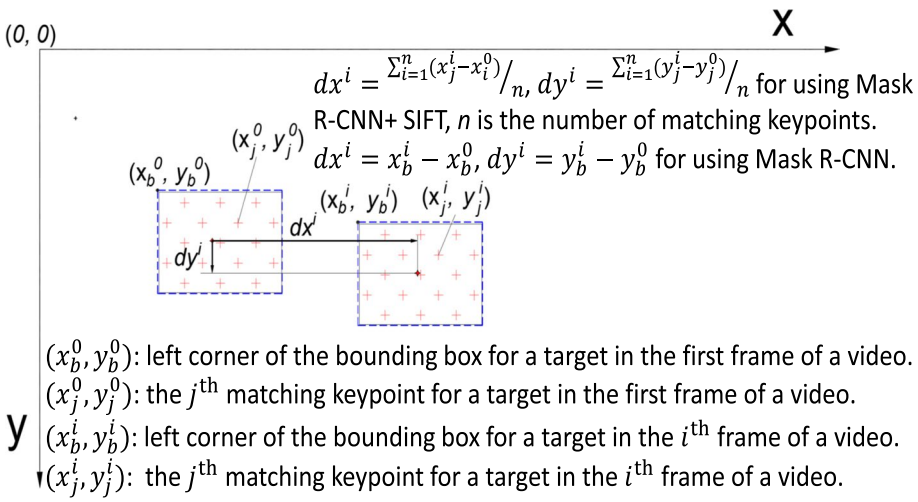


Fig. 2 Translation of a target measured by a bounding box or by matching keypoints

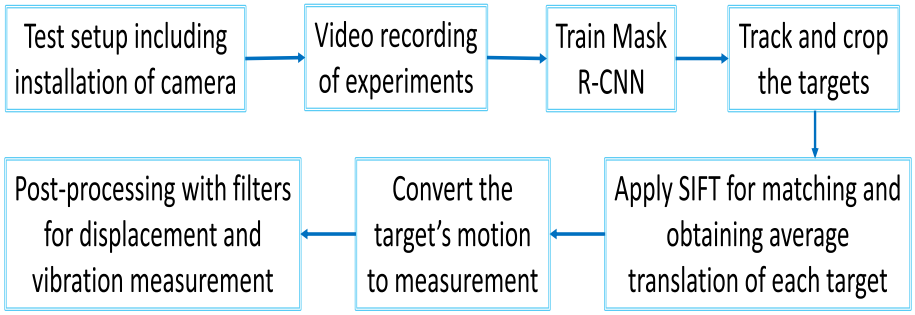


Fig. 3 Flowchart of Mask R-CNN + SIFT for automated displacement and vibration measurements with a stationary camera

3.2 Measuring the motions of targets with a deep learning method and an optical flow approach

To track the moving targets on a structure, template- and optical-flow-based methods have been proven to be effective in literature review. In this paper, a deep learning method using the High-resolution Network as the backbone, which is a template-based method and referred to as Mask R-CNN, is proposed to track and crop the targets with visual data (Bai et al. 2021b; Bai 2022). In addition, SIFT (Scale-Invariant Feature Transformation) is utilized to obtain more accurate measurements on the cropped images of the targets. This pipeline was verified with a static laboratory test and a shaking table test, whereas an optical flow approach named Lucas-Kanade (LK) tracker is employed for verification and comparison (Bai et al. 2021a). Figure 3 is the flowchart of Mask R-CNN + SIFT. The LK tracker has a similar framework, but directly tracks the target and computes optical flow between neighboring frames.

The displacement or vibration measurements can be converted from pixels to length units (inches or millimeters), which is also called a scalar, s. The horizontal and vertical displacements Δx^i and Δy^i of the target can be obtained as follows:

$$\Delta x^i = s \times dx^i \tag{1}$$

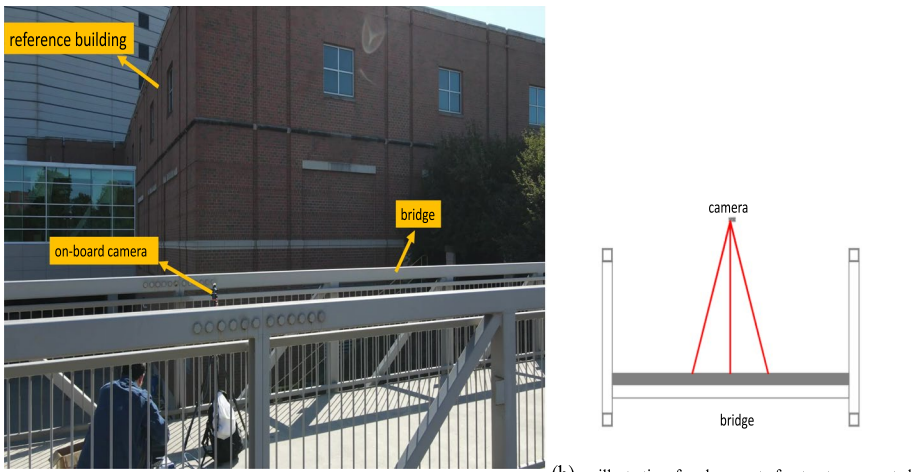
$$\Delta y^i = s \times dy^i \tag{2}$$

With this conversion, the measurements were converted into millimeters (mm). The scalar is in a range between 1.10 and 3.40 mm/pixel in our field experiments. Also, the SIFT helps us to obtain subpixel accuracy, thus, the error of displacement measurements for three concrete beams in a laboratory test can be 0.13 mm (Bai et al. 2021a). In field experiments, the same subpixel precision can be achieved but accuracy may be lower than tests in the laboratory because of the longer distance with our current cameras.

3.3 Three camera placements

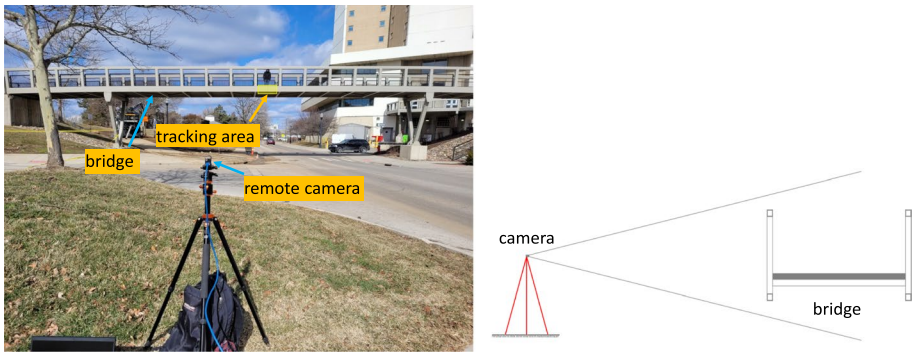
Three following different ways of camera placements are studied for potential real applications. Based on our research, structure-mounted cameras may be the first application used in field experiments by us.

1) Structure-mounted cameras: A camera can be fixed on the bridge deck and have the same motion as the standing point. First, the camera will focus on nearby motionless objects such as buildings and the ground, and then the excitation (e.g., jumps on a pedestrian bridge) is applied to the structure. In this case, the camera plays a role of an accelerometer to reflect the vibration where the camera is seated. Figure 4 illustrates how a camera and its tripod are placed on the deck of a pedestrian bridge and the nearby building is used as the reference for bridge motion. The camera has the same motion as the bridge since heavy bags are hung on the hook of the camera and partially supported by the bridge's deck. The locations for these cameras must be far from the bridge supports and undergo motion that is signified due to excitations during the experiments, for example, all cameras in field experiments of this study were placed near or on the mid-span of pedestrian bridges.



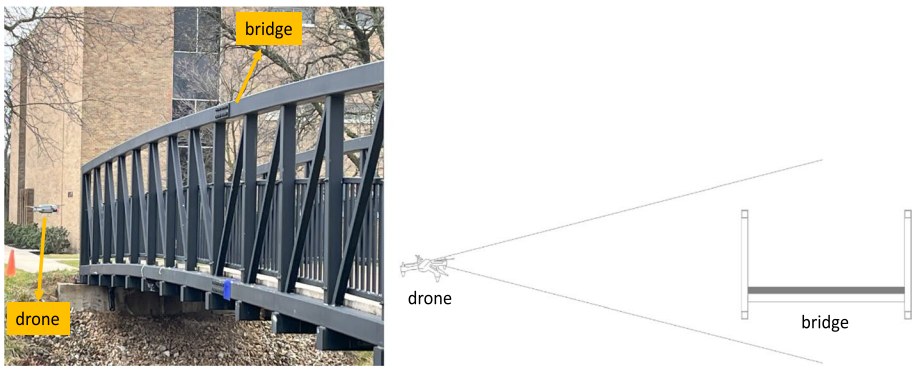
(a) an example for on-board camera placement. (b) an illustration for placement of a structure-mounted camera.

Fig. 4 An example of a structure-mounted camera to measure the vibration of a pedestrian bridge



(a) an example for remote camera placement. (b) an illustration for a remote camera measuring a bridge.

Fig. 5 An example of a remote camera to measure the vibration of a railway bridge



(a) an example for drone placement. (b) an illustration for a drone measuring a bridge.

Fig. 6 An example of a drone for vibration measurement of a pedestrian bridge

2) Remote cameras: In this case, a camera is placed remotely (e.g., far or close away) from the monitored bridges. The camera should be kept stationary during the testing process as much as possible, but it is not necessary to prevent slight motion of the camera since that motion can be canceled with displacement and frequency subtractions (see next subsection). In addition, the focal length of a camera is decisive for long-distance measurements. The longer the focal length is, the farther the camera can be placed from the bridges. A typical illustration of this camera placement is shown in Fig. 5. There are several advantages for remote cameras in measuring the vibration or displacements of the targets: First, a camera can capture the motion of multiple locations for the bridge simultaneously while a structure-mounted camera is limited to one target. Second, there are more available observation locations to fix a remote camera so that there are fewer risks for structural engineers who perform SHM missions. Finally, the targets can be mounted with markers or not, so it is convenient and less time-consuming for structural experts.

3) Drone-mounted cameras: A drone has high mobility and can be quickly deployed in SHM tasks. In Fig. 6, a drone flies to a certain height and keeps stable when focusing on a part of the pedestrian bridges and nearby reference objects

during the experiment. The motion of the drone can be obtained by aligning all the frames of the video to the first frame with the affine transformation (Szeliski 2010): First, the Lucas-Kanade tracker is used to focus on the reference (i.e., stationary objects) and obtain the affine transformation matrix between the first and current frame in the drone video. Second, the affine matrix is utilized to align the current frame to have the same perspective or viewpoint as the first frame. We can use Eq. 3 to map points and parallel lines with this affine transformation and not to change their geometrical relationships.

$$X' = AX + T \tag{3}$$

where $X' = (x_i', y_i')^T$ and $X = (x_i, y_i)^T$ represent intensity values of the i^{th} pixel located at position (x_i', y_i') on the aligned image and (x_i, y_i) on the original image, $A = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$ and $T = \begin{pmatrix} t_1 \\ t_2 \end{pmatrix}$ are linear transforms (i.e., rotation and scale) and translations in the affine transformation. a, b, c, and d are the rotation coefficients whereas t_1 and t_2 are translation in x and y directions. Third, the LK tracker and Mask R-CNN + SIFT can be employed to measure the bridge motion and drone motion simultaneously. Finally, the pure motion of the observed bridges can be obtained from Eq. 4.

$$U = U_0 - U_r \tag{4}$$

where U, U_0, and U_r are the pure bridge motion, measured bridge motion in the drone and drone movement from aligned images, respectively. This is referred to as displacement subtraction. On the other hand, if F, F_0 and F_r, respectively, are defined as the frequencies extracted from the pure bridge motion, measured bridge motion in the drone and drone movement itself. We can obtain the distribution of the frequencies of the observed bridges with drone-mounted cameras as this:

$$F = F_0 - F_r \tag{5}$$

This is a direct way of using the subtraction of the frequency domain. Therefore, it is called frequency subtraction. Both techniques utilized to eliminate the camera or drone motion are addressed in Section 4.

3.4 Filters for vibration signals and frequency extraction

At the beginning of this study, we found that the trend of vibration measured by the cameras is not flat when a free-damped vibration occurred by an initial excitation (see Fig. 7a). Also, noise caused by very low or high frequency affects the accuracy of frequency extraction methods. In addition, we noted that some filters are utilized to suppress the low frequencies for the data from the wireless accelerometers used in our experiments. Therefore, several filters had been compared so that the robust one can be used to filter the original vibration signals and obtain low-noise data from cameras. Figure 7 shows an example in which various filters are applied to the measured vibration for the free-damped vibration of a pedestrian bridge. But the trends for the filtered vibration by both Convolution and Median filters are not retrieved to be flat. The other three filters, including FIR (finite impulse response), IIR (infinite impulse response), and

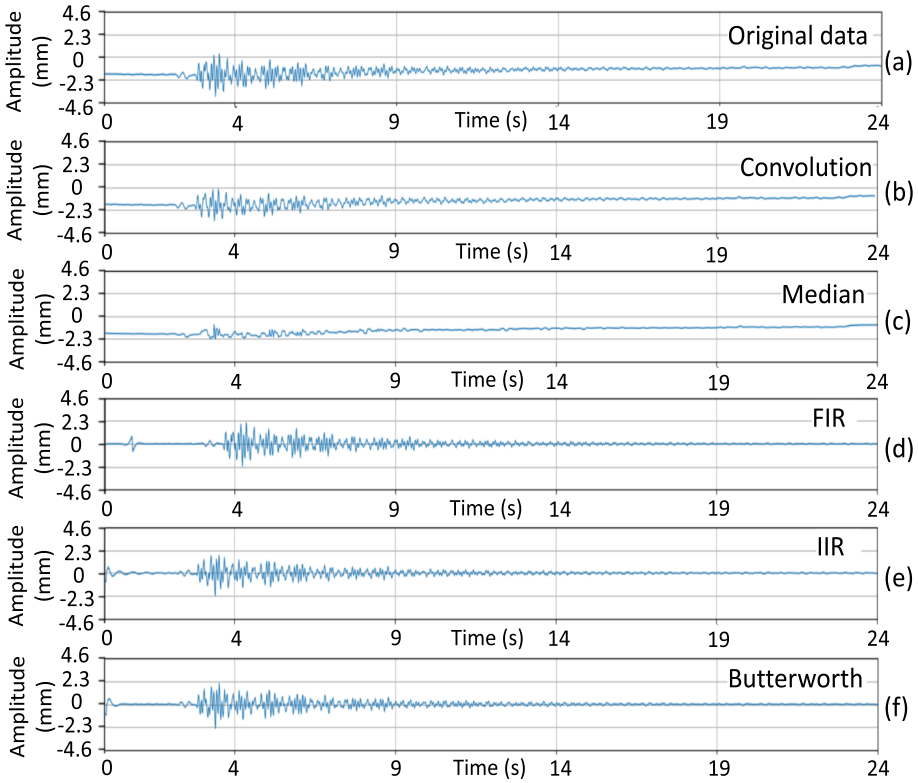


Fig. 7 Filtered vertical vibration of a pedestrian bridge (i.e., CPB-4 in Fig. 14d) with different filters. **a** to **f** are results from Convolution, Median, FIR (finite impulse response), IIR (infinite impulse response), and Butterworth band filters, respectively

Butterworth band filters (Press and Teukolsky 1990), can detrend well and retrieve back the similar shape of free damped vibration. It can be noted that, however, the IIR and Butterworth band filters can have a good agreement with the shape and the phases of the original signal. In Fig. 8, the frequencies extracted from the filtered vibration by these three filters have the same patterns and values. The natural frequencies of the bridges are detectable and consistent with each other. Since the very low and high frequencies can be suppressed significantly by a band passing technique, the Butterworth band passing filter is finally chosen to detrend and eliminate very low and high frequencies in the measured vibration signals with cameras.

The procedure for frequency extraction of the vibration is as follows: The vibration signals obtained by applying the Mask R-CNN + SIFT or LK tracker for visual data are filtered by the Butterworth band filter at first, then, frequencies of these filtered signals are extracted by Fast Fourier Transform (FFT) (Cooley and Tukey 1965) separately. In this paper, the highest frequency that our method can detect is the Nyquist frequency, which is half of the frame rate.

3.5 Displacement or frequency subtraction for systematic motion removal

The systematic motion of a camera or a drone occurs due to the wind, ground motion, or the drone's movement in the air. If camera motion is small (e.g., less than one pixel), the pure vibration of the excited structures can be obtained via subtraction of

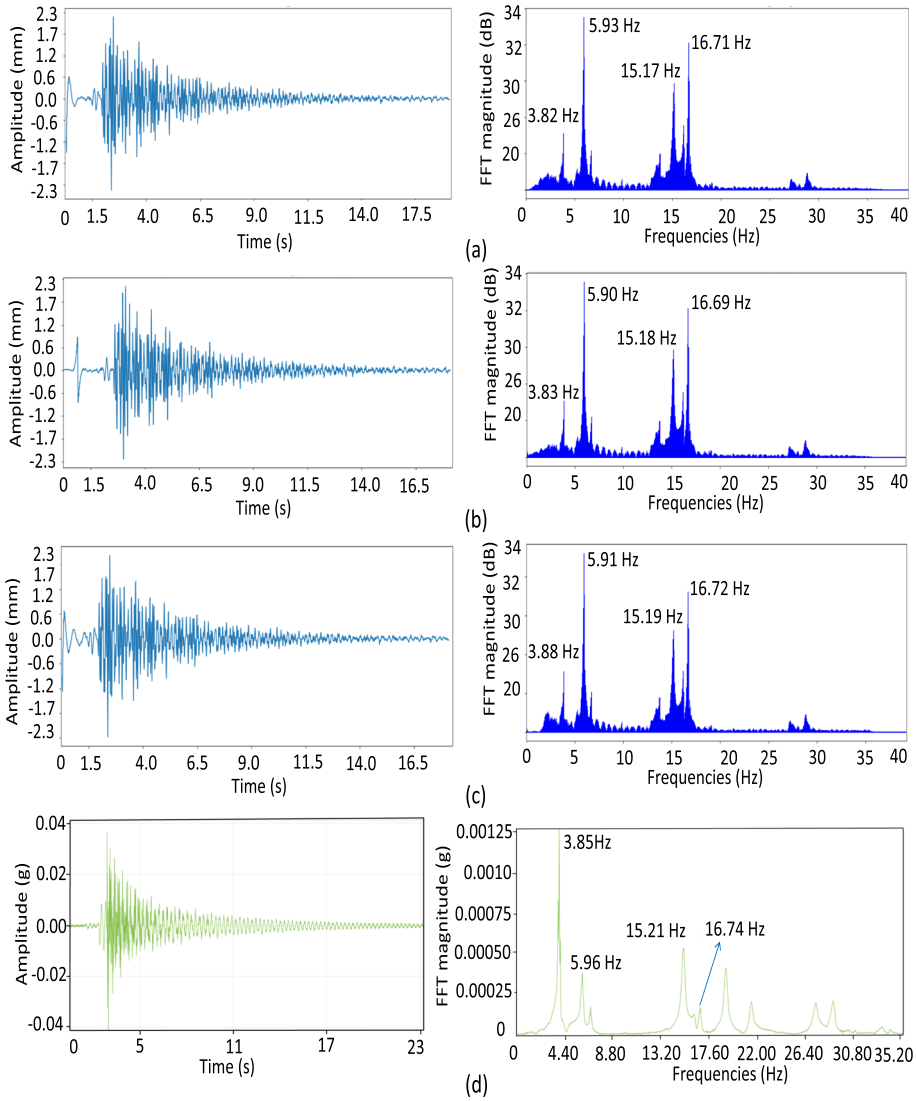


Fig. 8 Filtered vertical vibration of camera data and extracted frequencies by three different filters and acceleration data for CPB-4 pedestrian bridge. **a**, **b**, and **c** are results from band filters of FIR, IIR, and Butterworth, respectively. **d** is the measurement by an accelerometer

the motion of the reference and targets on the structure, which is called displacement subtraction (Chen et al. 2021; Dong et al. 2020; Nishi and Matsuda 2017). Then, the dynamic characteristics of structures like natural frequencies can be further extracted since the contribution of the camera motion is directly removed in the time domain (Eq. 4). Displacement subtraction is an intuitive way to assess the actual vibration of structures. Also, frequency subtraction is employed in this paper to extract the frequencies of vibration for infrastructure from visual data, including 1) extracting the frequencies of camera motion and the vibration of the excited structure in the camera separately, and 2) subtracting them directly to obtain the frequencies of the vibrated structures. (see Eq. 5 and Fig. 17).

4 Experiments

Shaking table tests and field experiments on pedestrian bridges are used to validate our proposed framework. Other field experiments on traffic and railway bridges were discussed in the dissertation of the first author (Bai 2022).

4.1 Shaking table tests in the laboratory

Shaking table tests can be experiments to simulate and assess the dynamic performance of structures in a controlled environment. Two online videos are utilized for this purpose, and our framework is applied directly to capture the vibrations and frequencies of the simulated structures. Since there is no geometric information in the shaking table tests, we used the pixel unit to indicate our results. All the measured vibrations are in the horizontal direction.

4.1.1 The shaking table test 1 with multiple single-DOF masses

Our proposed method was applied on a shaking table test with three separate masses (Mstkwon 2008b) to check its applicability in monitoring the dynamic movement of objects. In this test, there are three rectangles (masses) fixed on the shaking table at different heights (see Fig. 9a). Each rectangle, which is supported by two sticks and has its unique resonant frequency in the horizontal direction. This is due to differences in the lateral stiffness of each pair of sticks. The frequencies of the applied shaking are increased from 4.0 to 13.65 Hz to excite these masses and cause their harmonic vibrations. From the recorded video, 150 frames are randomly selected from a total of 6,674 frames and labeled for training the Mask R-CNN. The video has an image size of 640× 480 and a frame rate of 30 per second. SIFT is not applied to smooth the measurements since the goal of this test is to detect the frequencies instead of accurate amplitudes of the vibration. Thus, the motion of the bounding box represents the translation of each object. The LK tracker is utilized to verify our method by tracking the same vibration of the shaking table. On one hand, all the raw data are processed by the Butterworth filter, and FFT is applied to extract the frequencies for each tracking target. The filtered vibrations of the shaking table with the LK tracker and Mask R-CNN are shown in the left figures in Fig. 10. There are three major frequencies of vibration at approximately

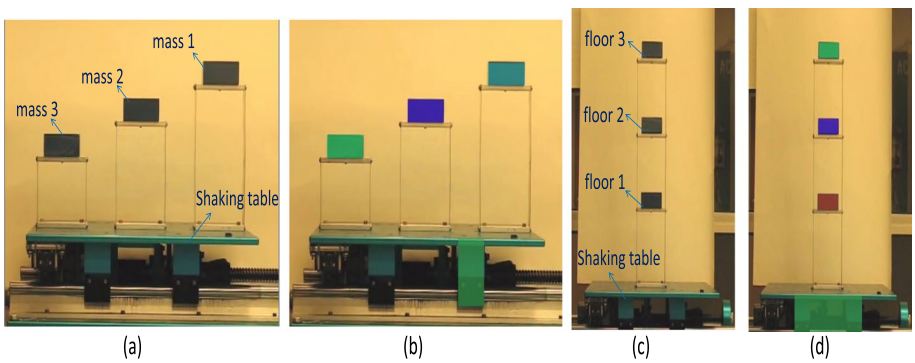


Fig. 9 Examples of training data for two shaking table tests (Mstkwon 2008b, a). The original images are shown in (**a**) and (**c**), and the tracking targets for the masses, floors and shaking tables are in different colors in (**b**) and (**d**)

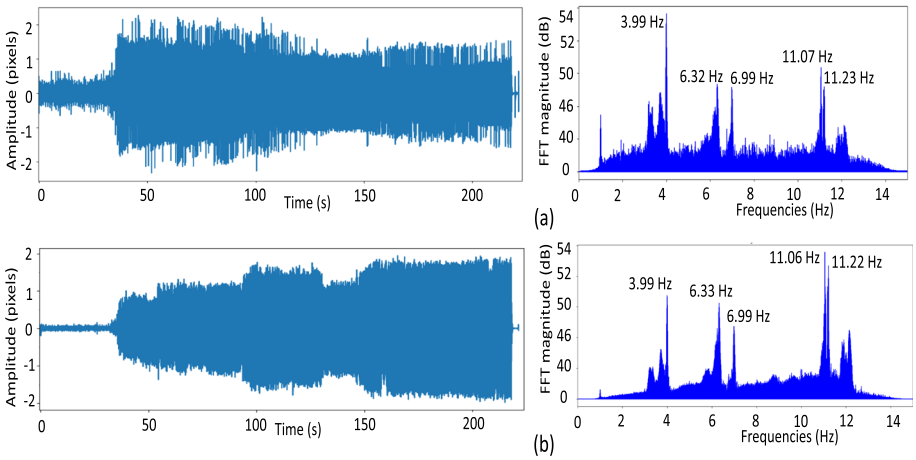


Fig. 10 Vibrations (left) and frequencies (right) of the shaking table in Fig. 9a measured by two methods. **a** and **b** are the results from Mask R-CNN and the LK tracker, respectively

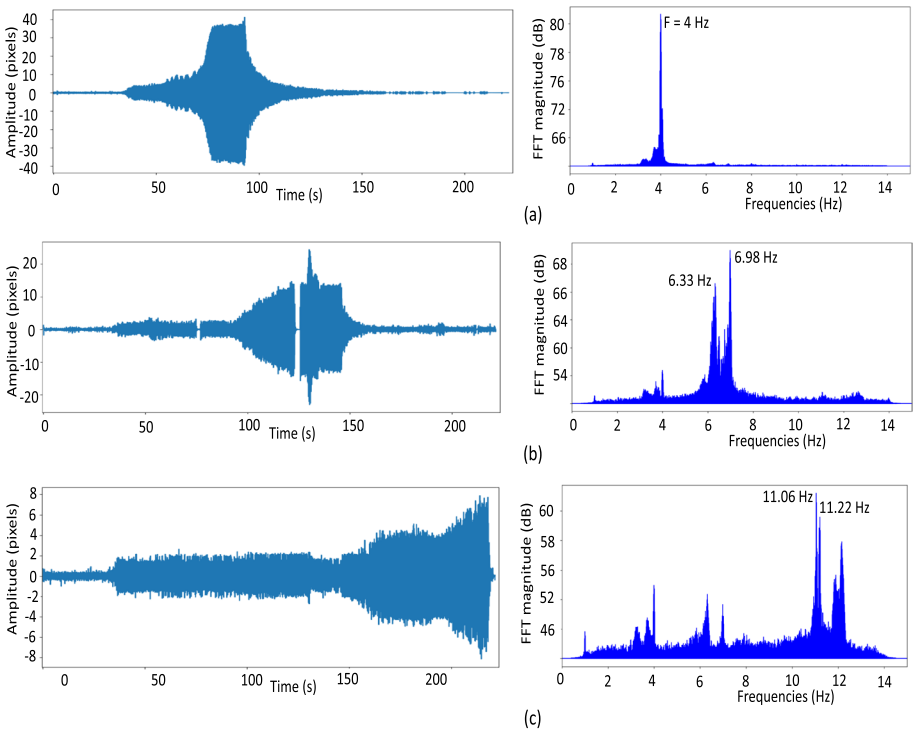


Fig. 11 Vibrations (left) of three masses measured by Mask R-CNN and the corresponding frequencies (right) calculated by FFT in shaking table test 1. **a**, **b** and **c** are the results for mass 1 to 3 in Fig. 9a

4.00, 6.35, and 11.35 Hz excited by the table. Both methods capture these frequencies with a less than 9.9% error. On the other hand, the vibrations of three rectangles are measured by the Mask R-CNN too. As shown in Fig. 11, their resonant frequencies are very close to the frequencies identified in Fig. 10 (i.e., 4.0, 6.35, and 11.35 Hz). The error rates for resonant frequencies of three masses are 0% ((4.0-4.0)/4.0=0%), 9.9% ((6.98-6.35)/6.35=9.9%) and −2.6%((11.06-11.35)/11.35=−2.6%), respectively.

4.1.2 A simulated multi-story structure in the shaking table test 2

This is another laboratory experiment from a video including 3,573 frames in a two-minute video (Mstkwon 2008a), and the proposed framework of Mask R-CNN + SIFT is used to track and measure the dynamic performance of this simulated structure. The frame rate is 30 frames per second and the image size is 640×480 for this video. 208 frames are randomly selected for labeling (see Fig. 9c). The Mask R-CNN has been trained to track three rectangles and the shaking table simultaneously. In the test, there are two designated frequencies, 4.35 and 12.55 Hz, for this three-story structure. The measured vibrations and frequencies of three rectangles are shown in Fig. 12, in which the extracted frequencies are almost identical to the frequencies of the shaking table as a resonant response. This is because the amplitude of each rectangle is more than one pixel, so there is no need to use the SIFT to achieve the subpixel accuracy of the vibrations of these rectangles. In addition, the modes of this structure are captured in video and found by our model. Figure 13 shows two examples of its modes in the video and modal extraction with our method. These two modes corresponding to the resonance excitation are captured. The third mode, however, is not detected since there is no excitation. It can be inferred that the modes of the structure can be measured by the proposed method and have good agreement with the law of structural dynamics.

4.2 Field experiments on pedestrian bridges

Six pedestrian bridges on the main campus of The Ohio State University (OSU) were tested in this study (see Fig. 14). All of them are steel truss structures with steel tubes and concrete slabs since the steel tube trusses for pedestrian bridges are lightweight

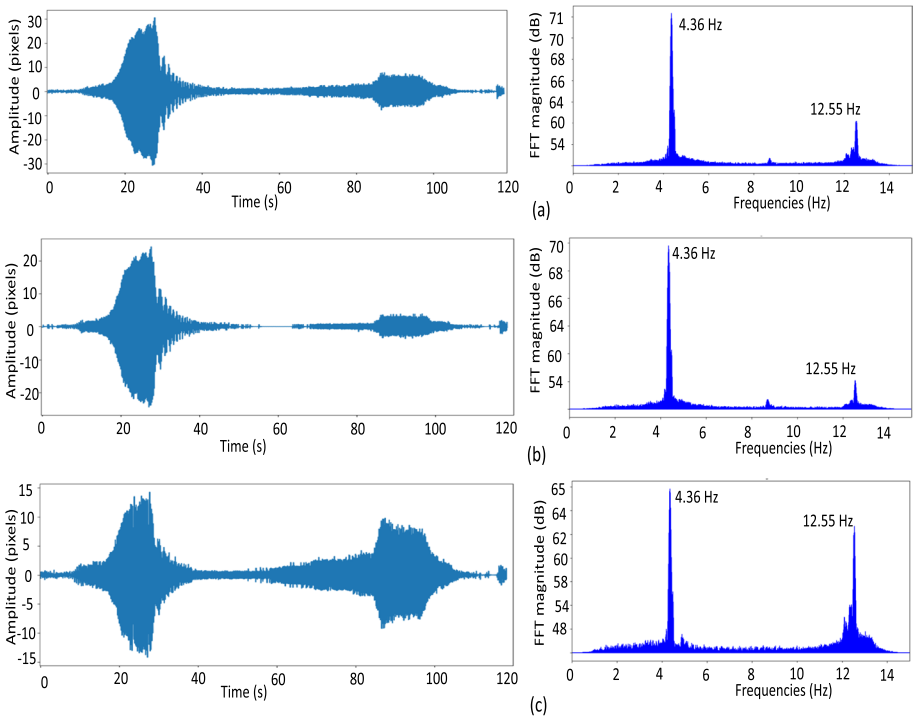


Fig. 12 Vibrations (left) of three floors measured by Mask R-CNN and the corresponding frequencies (right) calculated by FFT. **a**, **b** and **c** are the results for floors 3 to 1 in Fig. 9c

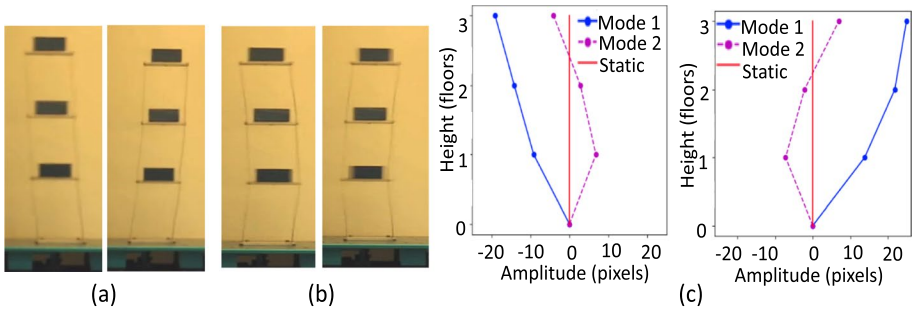


Fig. 13 Dynamic modes of the multi-story structure in Fig. 9c captured by our method. **a** and **b** are two modes in the video, and **c** is the measured modes by our method

Fig. 14 Six pedestrian bridges tested in this study

and the components can be manufactured in a factory with low cost and high quality. CPB-1 is a three-span bridge and its middle span is 78 ft long. The other five bridges have a single span and simple supports. CPB-2 and CPB-3 are almost the same with a 30-ft-long span, Also, CPB-5 and CPB-6 have an identical configuration of an 80.75-ft-long span. CPB-4 is a 72-ft-long pedestrian bridge. All of the vibrations measured by cameras, drones, and accelerometers are in the vertical direction for these pedestrian bridges.

It is common that the stiffness of these bridges usually is not high, which means the vibration caused by walking and running is acceptable. Therefore, some bridges on the OSU main campus are chosen as the way to validate our methods. In these field tests, jumps from a person are the source to cause the vibrations of these bridges. Since this excitation leads to the free-damped vibration of each pedestrian bridge, the three proposed camera placements are applied and our proposed framework for visual data processing is tested. To monitor the motion of a target on the bridge, the Mask R-CNN was trained at first with 50 frames randomly selected from the videos, thus, the target can be recognized and tracked with the proposed pipeline of or displacement vibration measurements.

4.2.1 Tests on CPB-1

This is a three-span bridge. The vibrations of its midspan in the second span were monitored. Three ways of camera placements, including remote, structure-mounted, and drone-mounted cameras as shown in Figs. 4, 5 and 6, are used to measure the vibrations of this bridge. In data processing steps, both Mask R-CNN + SIFT and LK tracker are used. The frame rate of cameras placed on the deck and nearby the bridge is set as 45 and 60, while the drone has a video with 48 frames per second. Their resolutions (image size) are 1,920×1,200 and 1,920×1,080, respectively. The remote and structure-mounted cameras have a 25-mm long lens each. The long lenses enable the distant reference buildings to be clearly captured by the cameras, thus, high-accuracy measurement can be achieved. Since the drone can fly very close to the bridge, the targets can have a good definition in these videos. The filtered vibrations and extracted frequencies with different camera placements are shown in Figs. 15, 16 and 17. For a free-damped vibration of the bridge, structure-mounted and remote cameras can be used to measure the vibration shape and magnitudes of the bridge. In Figs. 15a and 16a, compared to the fundamental frequency (the largest peak on each FFT plot) of this bridge captured by the accelerometers, the Mask R-CNN + SIFT method has a difference of −1.0% ((3.99-4.03)/4.03=−1.0%) and 0.8% ((4.02-3.99)/3.99=0.8%) using remote and structure-mounted cameras.

Figure 17 shows that the drone-mounted camera does not capture the exact dynamic response of the bridge accurately and the frequency plot has more noise than the other

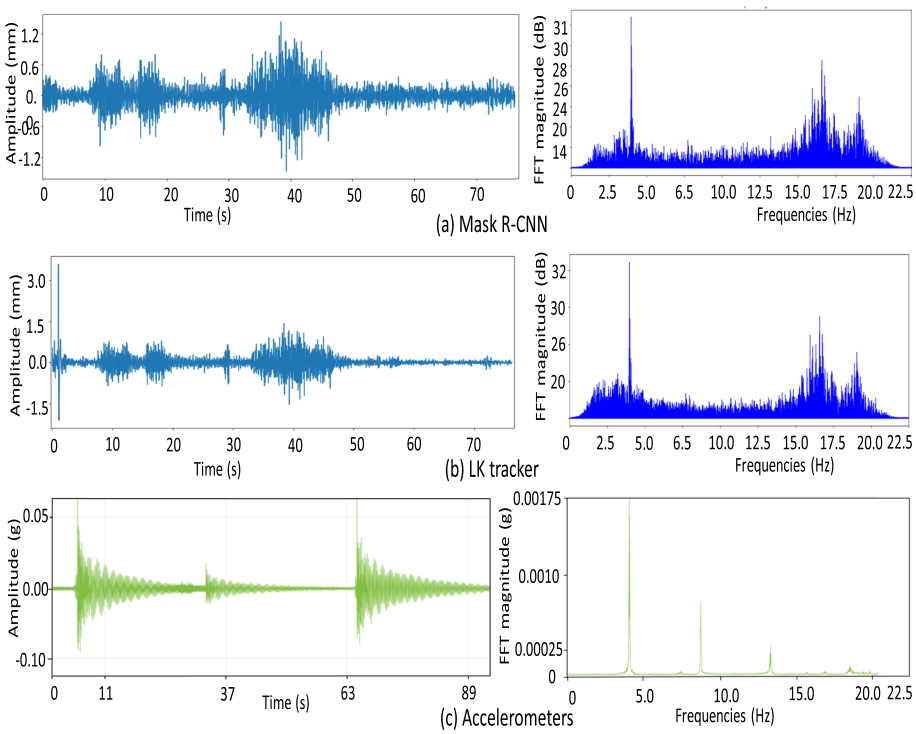


Fig. 15 Processed visual data measured by a remote camera (top two recordings) and accelerometers (bottom recording) on the CPB-1. **a**, **b** and **c** are measured vibrations and extracted frequencies from Mask R-CNN, LK tracker, and accelerometers, respectively

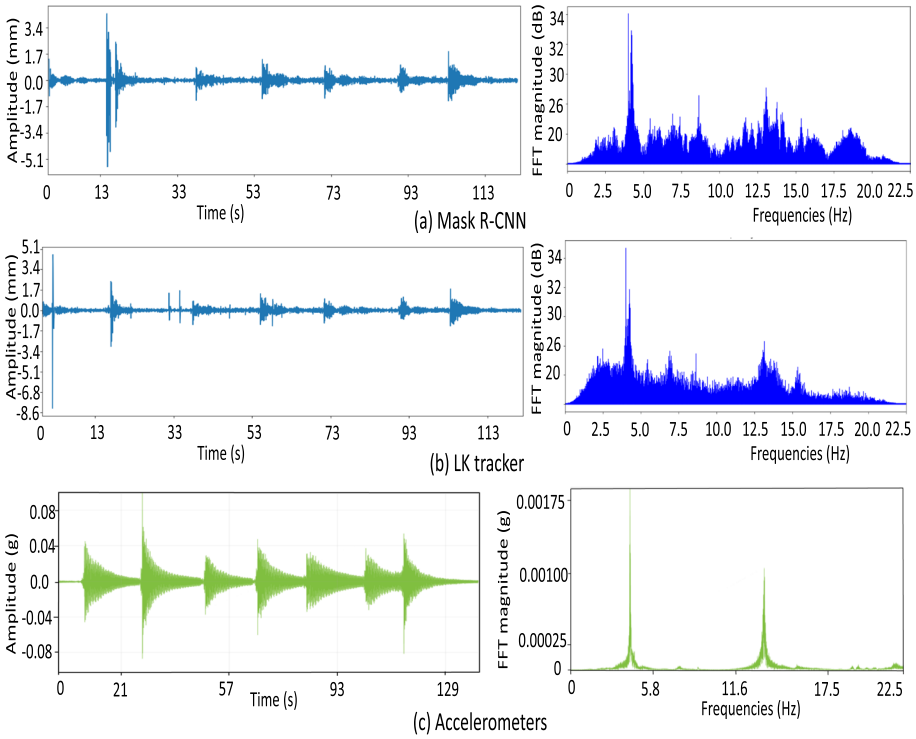


Fig. 16 Processed visual data measured by a structure-mounted camera (top two recordings) and accelerometers (bottom recording) on the CPB-1. **a**, **b** and **c** are the measured vibrations and extracted frequencies by Mask R-CNN, LK tracker, and accelerometers, respectively

two camera placements. If the largest peak in these FFT plots is the fundamental frequency of this bridge, both displacement and frequency subtractions can find it (around 4.0 Hz) precisely as we did in Figs. 15 and 16. These results show that the proposed framework and techniques for visual data processing work well for all these camera placements. But it also indicates that a drone-mounted camera may be more easily influenced by the drone itself and weather conditions in the field. In addition, we found that appropriate excitations on these bridges are the key for significant movements of targets to be captured by our proposed framework when other experiments on traffic and railway bridges were conducted (Bai 2022).

4.2.2 Experiments on six pedestrian bridges

The measurements for the fundamental frequency of six pedestrian bridges are shown in Table 1 when each bridge was excited with one or multiple jumps in the field experiments. It should be pointed out that the drone data for CPB-3 are missing because it has a similar structure to CPB-2 and the drone-mounted camera test was not performed. The results of the CPB-1 have already been shown in Figs. 15, 16 and 17. Table 1 indicates our proposed framework can accurately capture all the fundamental frequencies of these tested pedestrian bridges. In this table, SAP2000 (CSI 2022) is used as a structural analysis method to model and compute the frequencies of these bridges as described in the dissertation of the first author (Bai 2022). The calculated fundamental frequencies are very close to the measured ones with the proposed pipeline.

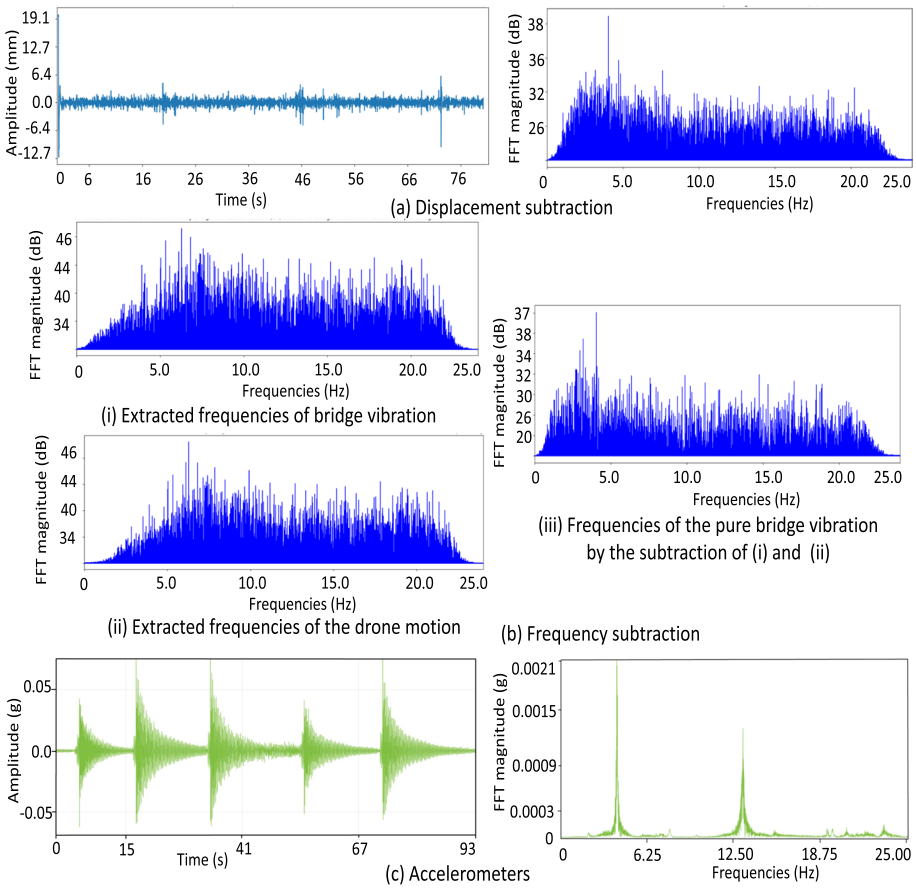


Fig. 17 Processed visual data measured by a drone-mounted camera (top two recordings) and accelerometers (bottom recording) on the CPB-1. **a** and **b** are displacement and frequency subtractions, respectively, and **c** is the accelerometer data

4.2.3 Sensitivity of structure-mounted cameras for measuring the vibrations of bridges

Figure 8 shows the measured vibrations and extracted frequencies (peaks on FFT plots) when the camera was placed on the deck of the CPB-4 bridge, which means that the camera is used as a contact sensor. It shows that our proposed methods can not only accurately detect the vibrations of the bridge due to a jump, but also capture the top four frequencies of this bridge as the wireless accelerometers did. The same observation was found in the field experiments of CPB-1 in Fig. 16.

5 Ablation study

5.1 Effect of different speeds for camera operation

We conducted several experiments on the CPB-4 bridge with the structure-mounted camera placement to evaluate the influence of different frame rates, window sizes, and sampling rates on visual data acquisition and processing. First, the camera was used with various frame rates when an excitation like a jump was applied, and the bridge's vibrations were precisely captured. The test result is shown in Fig. 18. In each test, the camera was operated from 30 to 84 frames per second when the bridge was excited by one jump. The reference is the building around the bridge (see Fig. 4), so the camera can capture almost the same shape of the free-damped vibration as the accelerometers

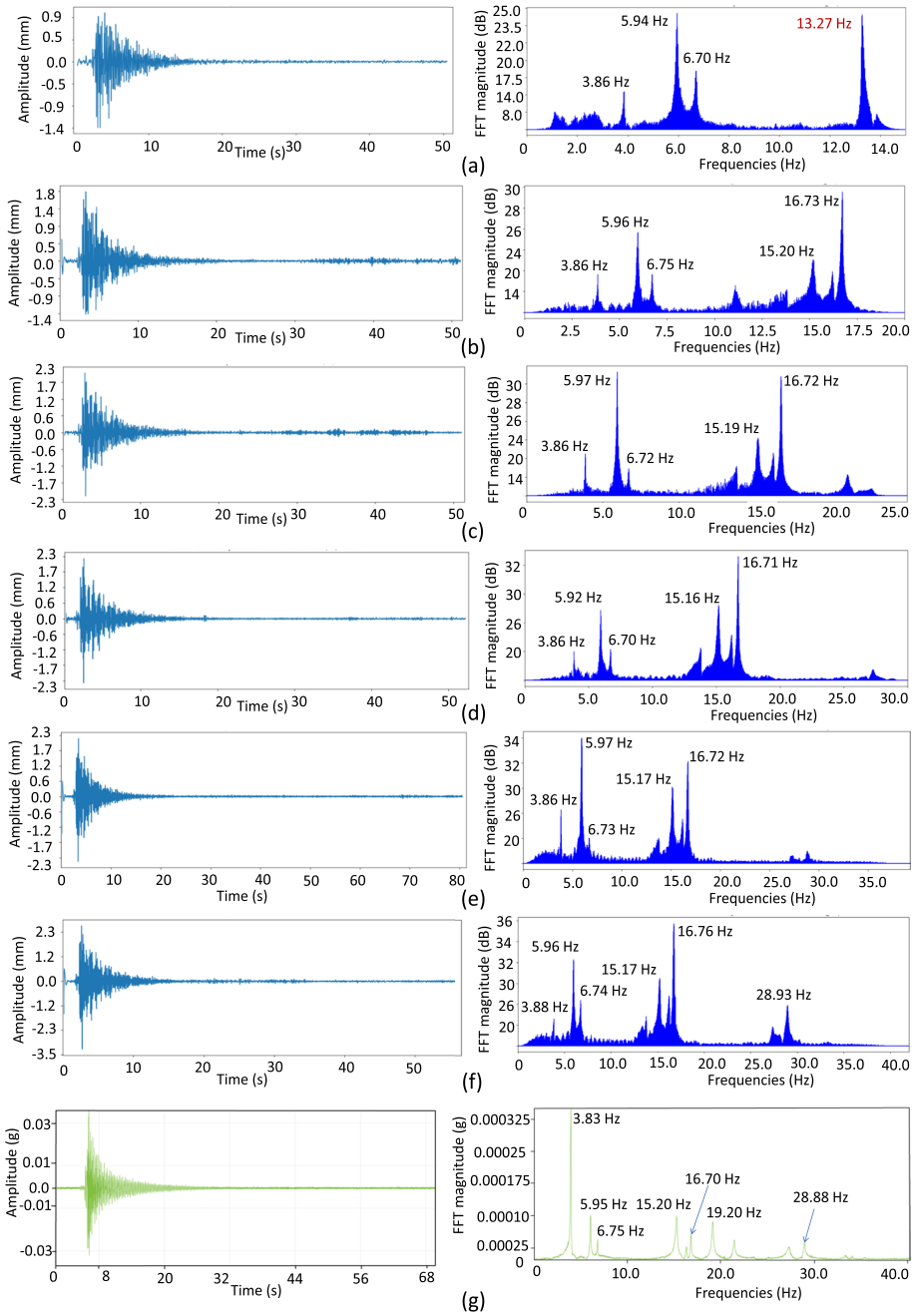


Fig. 18 Measured vibration and extracted frequencies from a structure-mounted camera with varied operation speeds for CPB-4. **a** to **f** are the results from our method to process the visual data when fps = 30, 40, 50, 60, 78.5, and 84, respectively, and **g** is accelerometer data (e.g., the maximum range of the frequency axis is half of the sample rate)

have done for this bridge. The five frequencies (peaks on FFT plots), which are detected by the accelerometers as 3.83, 5.95, 6.75, 15.20, and 16.70 Hz, are also captured by the structure-mounted camera accurately except for the case when the frame rate is 30. An unexpected frequency at 13.27 Hz, as shown in red, is provided by the latter. In all of these cases, the fundamental frequency can be well detected. This study shows that the

structure-mounted cameras can have the same accuracy as traditional accelerometers in SHM missions. Also, it can be concluded that the low-speed camera standing on the bridge can not identify higher frequencies of the pedestrian bridges. Based on our experiments and the studies (Bai 2022), the cameras to monitor the vibrations of pedestrian bridges should have a frame rate of no less than 30 frames per second, such that the fundamental frequency and other significant frequencies can be captured. The higher the camera speed is, the more frequencies including higher ones can be detected.

5.2 Effects of different window sizes and sampling rates on visual data

A study on the effect of different windows sizes and locations is conducted on processing the above visual data when a camera was placed on the CPB-4 bridge recording the bridge motion with 78.5 frames per second (see Fig. 18e). Figure 19 shows

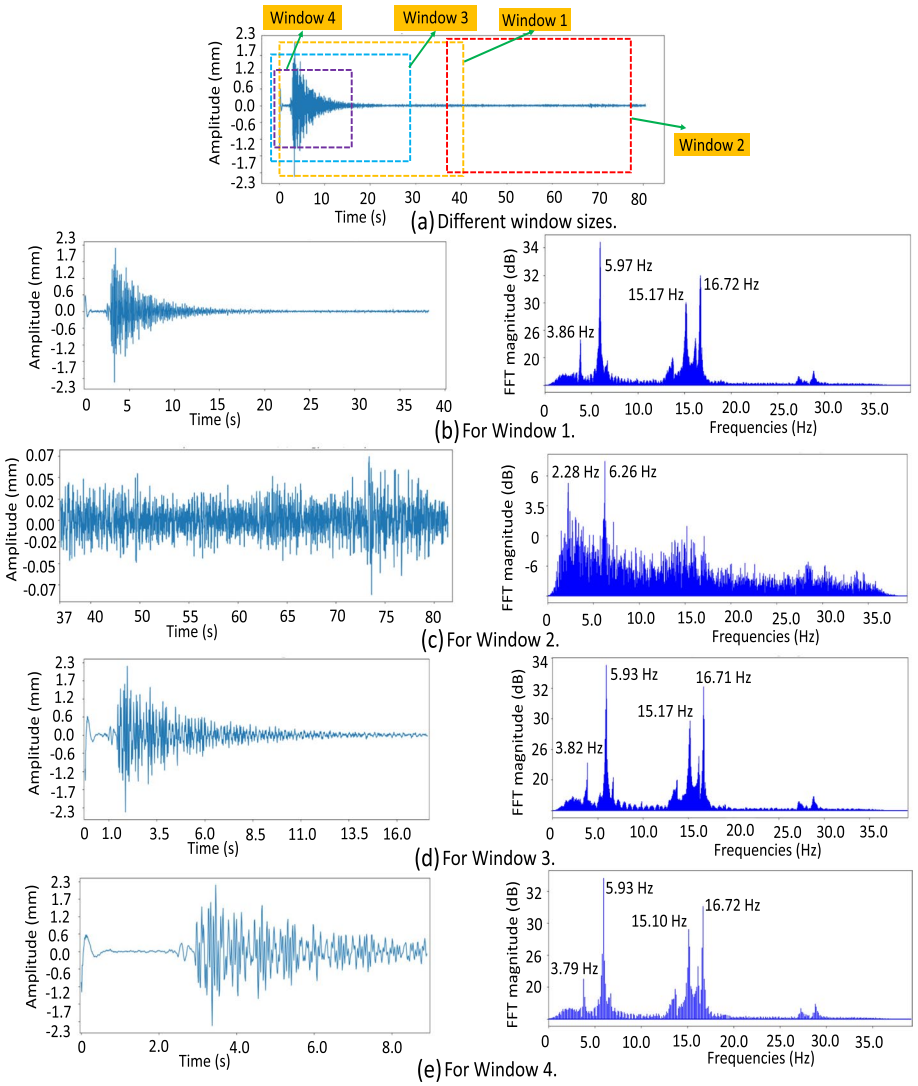


Fig. 19 Different window sizes and locations to sample datapoints in vibration signals measured on CPB-4 with a structure-mounted camera. The sizes and locations of windows to sample data are shown in (**a**), and **b** to **e** are the filtered vibration of corresponding windows on original data and extracted frequencies

that four correct frequencies of this bridge, which are around 3.86, 5.97, 15.17, and 16.72 Hz, can be captured accurately by the proposed framework with Mask R-CNN + SIFT only if the sampling windows (e.g., Windows 1, 3 and 4) include the whole free damped vibration of the bridge. But the noise becomes obvious in Window 2 since the structural vibration has already disappeared in this period. In addition, the same data are sampled at every one, two, three, and four datapoints (see Fig. 20) to extract the frequencies during the data processing. Some datapoints are discarded to investigate if the remaining ones are sufficient to represent the characteristics of these measured vibrations. It shows that the extracted frequencies lose the accuracy (e.g., the frequencies shown in Fig. 20c and d in red colors) when the sample rates decrease to every three and four datapoints, but the fundamental frequency (i.e., around 3.85 Hz) can be found.

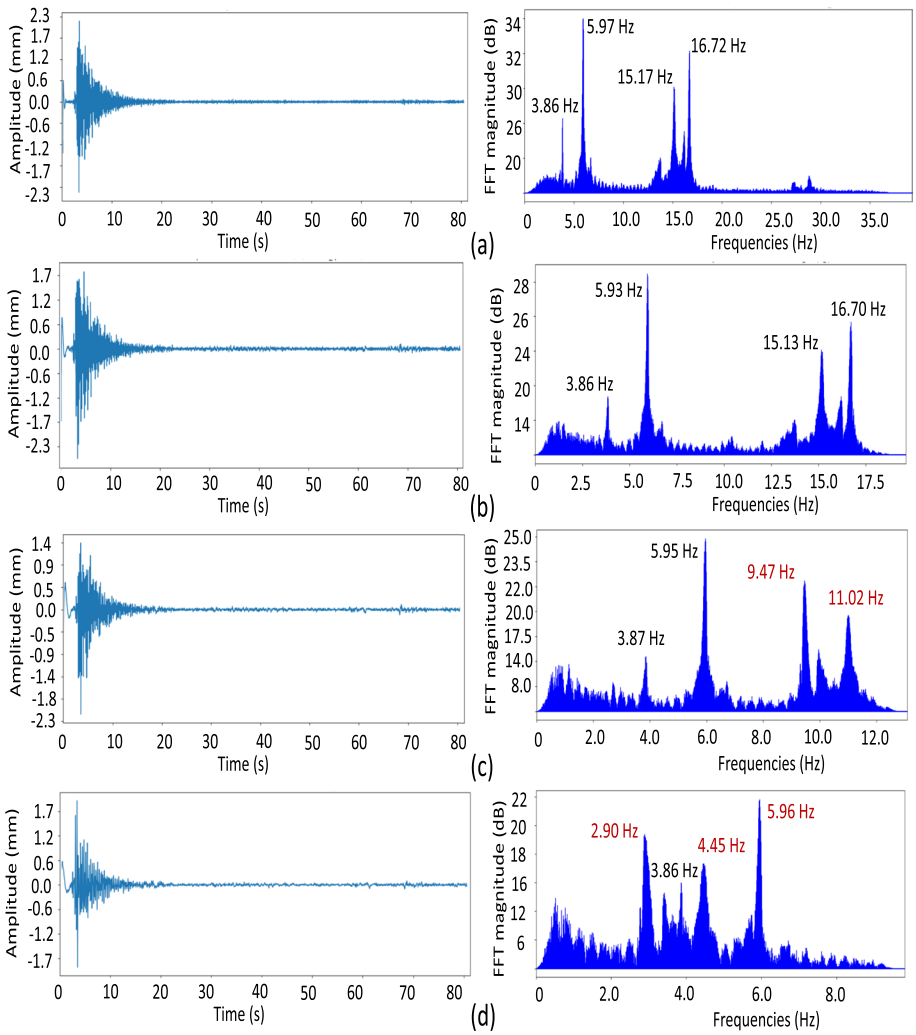


Fig. 20 Measured vibration and extracted frequencies from a structure-mounted camera on CPB-4 with different sampling rates. **a** to **d** are resulted by sampling every one, two, three, and four datapoints on the original data (e.g., the maximum range of the frequency axis is half of the sample rate)

6 Discussion

The proposed framework has shown its effectiveness in processing data from remote, structure-mounted, and drone-mounted cameras in laboratory and field experiments. These camera placements and visual data processing techniques may offer engineers more options using cameras to assess bridges in SHM missions. Its performance is perfect in the shaking table test since the vibration shapes and modes are all captured accurately. This is also consistent with our experimental results when our framework achieves an accuracy of 0.13 mm for measuring the mid-span deflection of concrete beams (Bai et al. 2021a). Our method will be automatic and real-time for repeated tests both in the field and laboratory after training. Also, other deep learning algorithms instead of the HR Mask R-CNN that we used in this paper can be employed to track targets, and then our data processing techniques are able to smooth the measurements, denoise the signals, and extract the frequencies with these visual data.

In practice, factors including temperature, moving vehicles, wind, source of excitation of structures, and lenses of cameras must be taken into account for more accurate vibration measurements of existing bridges. Longer lenses and high-definition cameras are preferable so that a higher subpixel accuracy can be achieved, especially when the distance between cameras and targets is large. For example, we used 25-mm lenses on pedestrian bridges for remote and structure-mounted cameras in these tests, where the targets were 8 to 30 m from the cameras. The reference buildings and other motionless surroundings were in the view of cameras. Also, camera movements affect the accuracy of vibration measurements for bridges. Experiments addressed in Bai (2022) show that drone-mounted cameras do not work for these railway and traffic bridges because they are rigid and the corresponding excitation under normal load levels is insignificant to them. In addition, all these field experiments were conducted when the targets were in good visibility in cameras.

Each camera placement has its own merits and disadvantages. Drones can collect data quickly and can be close to the target bridges, but may not accurately capture frequencies of vibrations larger than the fundamental frequency of these pedestrian bridges. Structure-mounted cameras are able to overcome this shortcoming but need to be fixed on bridges. Otherwise, remote cameras can be placed far away from the bridges. Currently, limited by the cameras' quality, we only measured the vibrations of one target on a pedestrian bridge with these three camera placements, but multiple cameras can be synchronized to focus on different targets on the bridges as we did in the experiments on a railway bridge (Bai 2022). Also, it is possible to monitor multiple targets with high-quality cameras mounted on drones or placed remotely. However, whatever camera placements are used for data collection, the cameras must be standstill or have a stationary object in videos. We are still working on some guidelines for the consideration of all these factors.

7 Conclusions

In order to mimic human vision for measuring bridge vibrations, a framework using different camera placements and utilizing techniques of computer vision and deep learning is proposed to save time and cost in SHM missions. It can not only provide ways of

visual data acquisition, but also show how to use data processing techniques, including noise removal, data sampling, and camera motion removal, for achieving accurate and reliable results. There are some conclusions in this research:

1) Three camera placements were tested for the applications of the vibration measurements on existing bridges. Our experiments show that the proposed framework can help process the visual data effectively. Structure-mounted cameras, which were first tested by us, can be used as contact sensors for more accurately measuring vibration signals and extracting more natural frequencies. Remote cameras can perform long-distance measurements well with long lenses. Drone-mounted cameras can detect the fundamental frequencies of the structures even though they may not provide accurate magnitudes of the vibrations for pedestrian bridges. Both methods of frequency and displacement subtractions were successfully applied to eliminate the camera movement in field experiments.
2) Six pedestrian bridges, which are all in service and the normal traffic was not affected, were used to validate the proposed framework. Dynamic characteristics of these bridges under the free-damped vibration can be captured. Since the modal shapes of the simulated structures in shaking table tests can be found with our methods, it is possible to apply the framework to obtain the modal shapes and other dynamic characteristics of these bridges. It should be noted that this proposed framework can get better results when appropriate excitations on existing traffic and railway bridges are applied (Bai 2022). Also, the Mask R-CNN used for motion tracking in the proposed framework can be replaced with other deep learning algorithms to measure the vibrations of bridges.
3) The influences of frame rates of cameras, positions and sizes of sampling windows, and sampling rates (camera speed) are also studied for visual data processing in practice. Our experiments indicate that the camera speed should be larger than 30 frame per second, and the sampling window should include the vibrations from beginning to the end. Also, some visual datapoints can be discarded without affecting the accuracy if high-speed cameras (e.g., fps$\geq$78.5 in our experiments) are used for these in-service bridges.

Structural damage detection can be potentially applied to the visual data obtained from the proposed framework. Also, the applicability of cell phones with the same framework as discussed in this paper is worthy of testing for bridge vibration measurements in the future.

Acknowledgements
In this study, Dr. Aydin Demir, Dr. Bing Zha, and Jianli Wei helped us do the experiments in field experiments. We appreciate their time and participation.

Authors' contributions
Yonsgheng Bai did the tests and drafted this paper. Professors Halil Sezen, Alper Yilmaz, and Rongjun Qin supervised this study and revised the paper.

Authors' information
Yongsheng Bai was rewarded with his Ph.D. in Civil Engineering at The Ohio State University (OSU) in 2022. He is now a research scientist in computer vision and deep learning at Neural Image Corporation in Bellevue, WA, USA. Halil Sezen is a professor of Structural Engineering at OSU and a registered professional engineer (P.E., state of Ohio). His research focuses on structural analysis, design, and performance assessment of various structures under extreme loading conditions. He has published more than 150 peer-reviewed articles. Alper Yilmaz is a professor with appointments in the Civil, Environmental, and Geodetic Engineering (CEGE) and Computer Science and Engineering (courtesy) Departments at

OSU. He is a Fellow of the American Society for Photogrammetry and Remote Sensing (ASPRS) and a senior member of IEEE. Rongjun Qin is an associate professor in the Department of CEGE and Electrical and Computer Engineering in computer vision at OSU. He has strong expertise in using images/videos to perform accurate localization and detection of objects, multi- and stereo-view camera-based 3D vision, image-based monitoring, and disaster responses. He focuses on computational solutions for accurately measuring static and moving objects in an urban context using aerial/UAV imagery, LiDAR, and satellite images.

Funding
This material is based upon work partially supported by the National Science Foundation in the US under Grant No. 2036193.

Availability of data and materials
The data that support the findings of this study are available from the corresponding author, H.S., upon reasonable request.

Declarations

Competing interests
The authors declare that they have no competing interests.

Received: 5 August 2023 Accepted: 6 October 2023
Published online: 01 November 2023

References

Bai Y (2022) Deep learning with vision-based technologies for structural damage detection and health monitoring. PhD dissertation
Bai Y, Abduallah RM, Sezen H et al (2021a) Automatic displacement and vibration measurement in laboratory experiments with a deep learning method. In: 2021 IEEE Sensors. p 1–4
Bai Y, Sezen H, Yilmaz A (2021b) End-to-end deep learning methods for automated damage detection in extreme events at various scales. In: 25th International Conference on Pattern Recognition (ICPR). p 6640–6647
Bai Y, Zha B, Sezen H et al (2023) Engineering deep learning methods on automatic detection of damage in infrastructure due to extreme events. Struct Health Monit 22(1):338–352
Baisthakur S, Chakraborty A (2020) Modified hamiltonian monte carlo-based bayesian finite element model updating of steel truss bridge. Struct Control Health Monit 27(8):e2556
Chen JG, Davis A, Wadhwa N et al (2017) Video camera-based vibration measurement for civil infrastructure applications. Journal of Infrastructure Systems 23(3):B4016013
Chen G, Liang Q, Zhong W et al (2021) Homography-based measurement of bridge vibration using uav and dic method. Measurement 170:108683
Chopra A (2019) Dynamics of Structures: Theory and Applications to Earthquake Engineering. Pearson, London
Cooley JW, Tukey JW (1965) An algorithm for the machine calculation of complex fourier series. Math Comput 19(90):297–301
CSI (2022) SAP2000 Advanced Structural Analysis Program, Version 23.0.0. Computers and Structures, Inc., New York
Dong CZ, Catbas FN (2019) A non-target structural displacement measurement method using advanced feature matching strategy. Adv Struct Eng 22(16):3461–3472
Dong CZ, Catbas FN (2021) A review of computer vision–based structural health monitoring at local and global levels. Struct Health Monit 20(2):692–743
Dong CZ, Celik O, Catbas FN (2019) Marker-free monitoring of the grandstand structures and modal identification using computer vision methods. Struct Health Monit 18(5–6):1491–1509
Dong CZ, Celik O, Catbas FN et al (2020) Structural displacement monitoring using deep learning-based full field optical flow methods. Struct Infrastruct Eng 16(1):51–71
Feng D, Feng MQ, Ozer E et al (2015) A vision-based sensor for noncontact structural displacement measurement. Sensors 15(7):16557–16575
Gheitasi A, Ozbulut OE, Usmani S et al (2016) Experimental and analytical vibration serviceability assessment of an in-service footbridge. Case Stud Nondestruct Test Eval 6:79–88
Gibbs MM, Kwon DK, Kareem A (2019) Data-enabled prediction framework of dynamic characteristics of rural footbridges using novel citizen sensing approach. Front Built Environ 5:38
Guo J, Zhu C (2016) Dynamic displacement measurement of large-scale structures based on the lucas-kanade template tracking algorithm. Mech Syst Signal Process 66:425–436
Hoskere V, Park JW, Yoon H et al (2019) Vision-based modal survey of civil infrastructure using unmanned aerial vehicles. J Struct Eng 145(7):04019062
Khuc T, Nguyen TA, Dao H et al (2020) Swaying displacement measurement for structural monitoring using computer vision and an unmanned aerial vehicle. Measurement 159:107769
Lee J, Lee KC, Cho S et al (2017) Computer vision-based structural displacement measurement robust to light-induced image degradation for in-service bridges. Sensors 17(10):2317
Liu B, Zhang D, Guo J (2016) Vision-based displacement measurement sensor using modified taylor approximation approach. Opt Eng 55(11):114103
LORD (2022) G-link-200 rugged wireless accelerometer, 3 axis. https://www.microstrain.com/wireless-sensors/G-Link-20032. Accessed 15 Oct 2021

Mstkwon (2008a) Mdof system forced vibration. https://www.youtube.com/watch?v=OaXSmPgl1os. Accessed 20 May 2021
Mstkwon (2008b) Sdof resonance vibration test. https://www.youtube.com/watch?v=LV_UuzEznHs
Nishi K, Matsuda Y (2017) Camera vibration measurement using blinking light-emitting diode array. Opt Express 25(2):1084–1105
Perry BJ, Guo Y (2021) A portable three-component displacement measurement technique using an unmanned aerial vehicle (uav) and computer vision: A proof of concept. Measurement 176:109222
Press WH, Teukolsky SA (1990) Savitzky-golay smoothing filters. Comput Phys 4(6):669–672
Rajaram S, Vanniamparambil P, Khan F et al (2017) Full-field deformation measurements during seismic loading of masonry buildings. Struct Control Health Monit 24(4):e1903
Ribeiro D, Santos R, Cabral R et al (2021) Non-contact structural displacement measurement using unmanned aerial vehicles and video-based systems. Mechanical Systems and Signal Processing 160:107869
Szeliski R (2010) Computer vision: algorithms and applications. Springer Science & Business Media, Berlin
White R, Alexander N, Macdonald J et al (2020) Characterisation of crowd lateral dynamic forcing from full-scale measurements on the clifton suspension bridge. Structures 24:415–425
Xiao P, Wu Z, Christenson R et al (2020) Development of video analytics with template matching methods for using camera as sensor and application to highway bridge structural health monitoring. J Civil Struct Health Monit 10(3):405–424
Yin Z, Wu C, Chen G (2014) Concrete crack detection through full-field displacement and curvature measurements by visual mark tracking: A proof-of-concept study. Struct Health Monit 13(2):205–218
Yoon H, Shin J, Spencer BF Jr (2018) Structural displacement measurement using an unmanned aerial system. Comput-Aided Civ Infrastruct Eng 33(3):183–192

Publisher's Note