Smart Structures and Systems, *Vol. 27, No. 5 (2021) 803-818*
DOI: https://doi.org/10.12989/sss.2021.27.5.803

A novel DNN tracking algorithm for structural system identification

Sheng-Yun Peng [2,4a], Ling-Feng Yan [1b], Bin He [3c] and Ying Zhou[*1]

[1] *State Key Laboratory of Disaster Reduction in Civil Engineering, Tongji University, Shanghai 200092, China*
[2] *College of Civil Engineering, Tongji University, Shanghai 200092, China*
[3] *College of Electronic and Information Engineering, Tongji University, Shanghai 200092, China*
[4] *College of Computing, Georgia, Institute of Technology, Atlanta, GA 30332, USA*

(*Received July 11, 2020, Revised January 2, 2021, Accepted January 20, 2021*)

Abstract. In the field of structural health monitoring (SHM), cameras record videos and tracking methods can be applied to calculate the structural displacement. Commercial and unmanned aerial vehicle (UAV) cameras are promising non-contact sensors owning to their high availability and easy installation. However, effective tracking methods need to be developed. In this study, we firstly propose an end-to-end vision measuring framework with a novel deep neural network (DNN) tracker, named Siamese Single Decoder Network (SiamSDN). The system requires no target installation and uses cellphone cameras. For SiamSDN, the position and scale of bounding box are formulated through statistical parameter estimation. Unlike generative trackers, SiamSDN does not require manually extracted features or pre-defined motion areas. The tracking object is solely identified in the first frame. A shaking table test of a five-storey structure is carried out to demonstrate the efficiency. Besides, a UAV is used to simulate the field test. To minimize the error caused by the vibrations of UAV, digital video stabilization (DVS) is proposed to eliminate the drifts. Videos taken by both the commercial and UAV cameras are analyzed to calculate the displacements. Comparing our DNN tracker with feature point matching approach, SiamSDN improves the displacement measuring accuracy by 66.16% and 57.54%, respectively, and the frequency characteristics are obtained precisely.

Keywords: structural health monitoring; commercial camera; unmanned aerial vehicle; siamese network; frequency characteristics

1. Introduction

The structural health monitoring (SHM) of civil infrastructures mainly aims to monitor the structural condition, detect the structural abnormality, and evaluate the structural safety based on the long-term monitoring data from a variety of sensors installed on the structures (Ye *et al*. 2019). As a method of engineering structure evaluation, SHM has gradually become an important research area in civil engineering. In the SHM community, realizing the autonomous, accurate and robust data processing systems has been a great concern (Ye *et al*. 2019, Spencer *et al*. 2019).

Structural displacement responses to a variety of external loads are fundamental for structure status estimation and health monitoring. The displacement measurement systems can be divided into two types: contact and non-contact systems. The contact systems generally include high precision transducers, like LVDT and cable type. These systems require a stationary reference point to measure the relative displacement between the structure and the fixed point (Ribeiro *et al*. 2014). Meanwhile, it is difficult and almost impossible to install the measuring system on structures, which are super high-rises or are located over a watercourse (Feng *et al*. 2015).

The non-contact systems generally encompass laser, radar, GPS and camera systems. The measurement of the laser technique is very accurate, but the high intensity laser beam is detrimental to human health and the cost of the whole system is too high for regular SHM (Nassif *et al*. 2005). System that uses GPS technology can measure the displacement within 0.2 mm, but it requires a local station and the frequency is normally below 20 Hz (Jo *et al*. 2013).

Vision-based measuring methods have established a good compromise between the acquisition frame rate and the resolution (Mas *et al*. 2011). With the advent of artificial intelligence, more attention has been paid to explore the applications of deep learning-based image processing methods in the field of SHM (Spencer *et al*. 2019, Lee *et al*. 2017, Narazaki *et al*. 2019). This technique has the advantages of non-contact, non-destructive, long distance and immunity to electromagnetic interference (Ye *et al*. 2016). Liu *et al*. (2019b) proposed a machine learning method to measure the vibration frequency. The method was based on long short-term memory-recurrent neural networks (LSTM-RNN) and multi-target learning. A new model based on convolutional neural network (CNN) was established for SHM of tall buildings subject to wind loads (Oh *et al*. 2019). Zhang *et al*. (2019b) developed a simple one-dimensional CNN that detected tiny local structural

*Corresponding author, Professor,
E-mail: yingzhou@tongji.edu.cn
[a] M.S. Student, E-mail: shengyun@gatech.edu
[b] M.S. Student, E-mail: carlyan@tongji.edu.cn
[c] Professor, E-mail: hebin@tongji.edu.cn

Copyright © 2021 Techno-Press, Ltd.

ISSN: 1738-1584 (Print), 1738-1991 (Online)

stiffness and mass changes. The proposed framework was validated on T-shaped steel beam. Wahbeh *et al*. (2003) combined a highly accurate camera with a laser tracking reference to develop, calibrate, implement and evaluate the feasibility of obtaining the absolute displacement time history on a field test bridge. Busca *et al*. (2014) proposed and used two different types of cameras to monitor the responses of bridges as trains passed across them. Three different image processing techniques (pattern matching, edge detection, and digital image correlation) were applied to analyze the images. Results were compared to those obtained by a single-point laser interferometer. By comparing two types of non-contact measurement, Kohut *et al*. (2013) calculated displacement fields of structures with digital image correlation coefficients and measured the deflection of structures by using a radar interferometer. Dong *et al*. (2019) proposed a completely non-contact structural identification system, which targeted the identification of bridge unit influence line under operational traffic. (Lydon *et al*. 2019) provided a non-contact low-cost AI based solution for vehicle classification and associated bridge displacement using CNN methods. Jung *et al*. (2019) described three phases of a bridge inspection using UAVs. Also, three major challenges, which are related to a UAV's flight, image data acquisition, and damage identification, respectively, are identified and their possible solutions are discussed.

Hence, consumer-grade cameras can be utilized as non-contact sensors for civil structures inspections owning to their high availability, low cost and easy installation. Normally, abundant videos are taken by a single camera and object tracking methods are applied to figure out the structural displacement.

Template-matching techniques were adopted to measure the displacement time histories of structures in an indoor shaking table test and the dynamic response of built bridges in outdoor field tests (Fukuda *et al*. 2013). In the study of Feng and Feng (2016), the vision sensors were proposed based on two template matching schemes: the up-sampled cross correlation (UCC) and the orientation code matching (OCM). A new gradient-based computer vision technique edge enhanced matching (EEM), improved from OCM, was developed to measure displacements of low-contrast natural targets (Luo and Feng 2018). Two field tests were conducted to compare the tracking ability of EEM and OCM methods. An advanced video deflectometer using off-axis digital image correlation (DIC) was proposed for the measurement of vertical deflection of bridges. The inverse compositional Gauss-Newton algorithm was employed to achieve real-time displacement (Pan *et al*. 2016). Commercial cameras and optical flow technologies were used to calculate variation of optical intensity of an arbitrary selected region of interest (ROI) on the cable image sequence. The obtained optical flow vectors provided the direction of cable vibrations (Ji and Chang 2008). Yoon *et al*. (2016) used corner detection method suggested by Harris and Stephens (Harris and Stephens 1988) to extract the features within the ROI in the initial video frame. Identified corners in two images were matched with the help of the Kanade-Lucas-Tomasi Feature Tracker (Shi 1994) algorithm, and matching pairs were refined with the MLESAC algorithm. A simplified fast-Hessian detector and a pre-purification-based RANdom Sampling Consistency (RANSAC) were proposed to monitor bridge deflections. Compared with scale-invariant feature transform (SIFT), speeded-up robust features (SURF) and traditional template matching algorithms, the combined SURF-BRISK performed better, and the relative errors were within 10% (Yu and Zhang 2020).

In summary, the earliest stage of these tracking methods was built on template matching techniques, including DIC, edge detection, optical flow and OCM. Template matching methods could only detect motions moving in parallel, which were not suitable for changing light conditions, scale variances and fast motions. Then, feature point tracking methods like Harris corner detection, SIFT, SURF, and features from accelerated segment (FAST) were introduced in the SHM community. Feature tracking methods consisted of four steps: detection, description, matching, and purification. In the detection phase, the detector extracted all the potential points in a selected area. A selected region, which covers all the potential moving area, is defined before tracking. However, the displacements of structures vary under different external forces, which means the tracking object can randomly appear at any place in a video. Larger displacements require a larger pre-defined bounding box and more amount of time for feature point extraction and matching. Besides, these generative trackers construct templates based on intensive features inside the bounding box without considering the background information. A clear deficiency of using data exclusively from foreground is that comparatively simple model can be trained. The accuracy of the template matching techniques is largely dependent on the image quality, which is often difficult to guarantee in complex field environmental conditions such as illumination variation, partial target occlusion, fast motion, deformation and background clutter (Wu *et al*. 2015).

Nowadays deep learning has dominated various computer vision tasks including semantic segmentation (Liu *et al*. 2019a) and video object tracking (Zhang *et al*. 2019a, 2020, Peng *et al*. 2020). CNN automatically extracts features from raw input images rather than using human defined features. In this study, we firstly propose a vision structural system identification framework with a novel deep neural network (DNN) based object tracking algorithm, named Siamese Single Decoder Network (SiamSDN). Then a shaking table test of a five-storey structure is conducted to demonstrate the accuracy and efficiency of the approach. We also record the shaking table test with an unmanned aerial vehicle (UAV) camera to simulate the field test. To eliminate the small and random drifts, digital video stabilization (DVS) is proposed to obtain stable videos through motion compensation based on background stationary objects.

2. System model overview

In this section, we provide an overview of the proposed

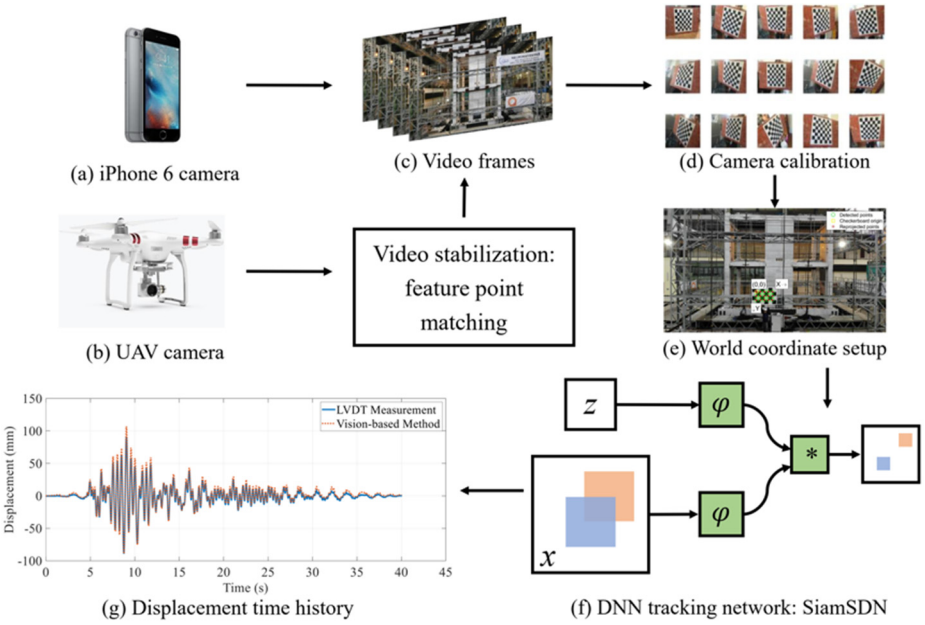

Fig. 1 The flowchart of vision-based displacement measurement system. Original videos are taken by (a) commercial cameras and (b) UAVs. UAV videos are stabilized to eliminate the vibrations of UAV itself. Then, (c) video frames need to rectify lens distortions through (d) camera calibration. Camera parameters and (e) world coordinate systems are also established. Raw video images are sent to our firstly proposed (f) DNN tracking network: SiamSDN. Finally, we transform the (g) displacement time histories from pixel level to metric system

vision-based displacement measurement system. The method mainly consists of four parts: (1) original video taken by commercial cellphone cameras and preprocessed video taken by the UAV; (2) camera system setups including calibrating cameras and building reference coordinates; (3) DNN based video object tracking algorithm and (4) displacement measurements and system identifications. The technical flowchart is shown in Fig. 1.

Original videos are taken by consumer grade cameras: iPhone 6 rear camera and Da Jiang Innovations (DJI) UAV. A UAV has small degrees of vibrations even if it is stationary at one point during its flight. The small vibrations lead to huge error while calculating the motions of structures. Therefore, the UAV videos need to be stabilized to eliminate the small and random drifts. Then, videos are divided into subsequent frames for future processing. Since regular commercial cameras use wide angle lens, the captured images encompass distortions. It is crucial to calibrate the lens distortion and rectify the video frames. The intrinsic and extrinsic parameters are calculated, and the world coordinate system (WCS) and local coordinate system (LCS) are settled down to transform the pixel level into metric system. The preprocessed video frames are the input of the end-to-end DNN object tracking pipeline, named SiamSDN. The tracking object is identified solely in the first frame. Given a sequence of input images,

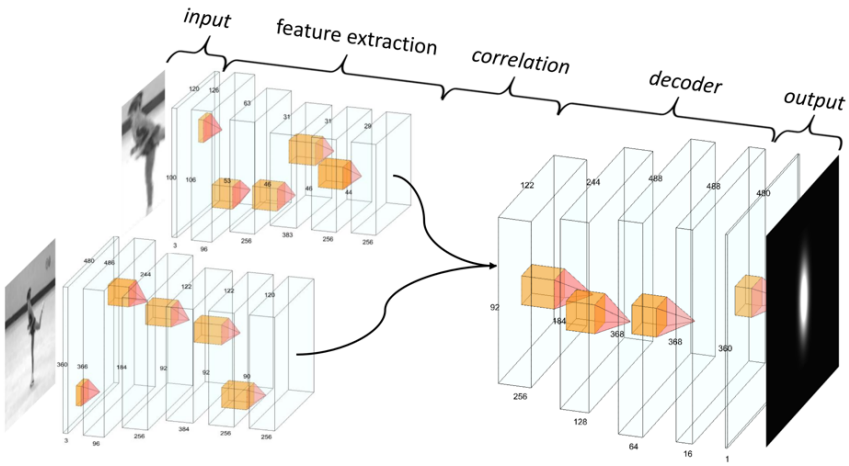

Fig. 2 Main framework of SiamSDN. From left to right: input image pair, feature extraction subnetwork, correlation operator, decoder and output score map. Full scale exemplar and instance raw images are fed into the template and search branches respectively. Single decoder is used for restoring both position and scale information

SiamSDN utilizes a class specific detector to accurately predict the motion state (location, size or orientation) of the object in each frame. Finally, the pixel displacement time histories are obtained and transformed into the metric displacements via the calculated camera parameters.

3. Vision-based displacement measurement

3.1 SiamSDN framework

In this section, we provide details of the proposed SiamSDN framework. As shown in Fig. 2, the proposed network includes five typical parts: input image pairs, feature extraction subnetwork, matching function, decoder and output score map. The matching function is also known as the correlation operator. Conventional discriminative trackers are based on correlation filter, which trains a regressor by exploiting the properties of circular correlation and performing operations in the Fourier domain. It can do online tracking and update the weights of filters at the same time efficiently (Bertinetto *et al*. 2016). SiamSDN combines the advantage of CNN and correlation filter. It uses deep features to improve the accuracy.

3.1.1 Input image pairs

Since the network is an end-to-end design, raw images are fed into the two input branches. Meanwhile, the fully-convolutional framework accepts arbitrary size of instance and exemplar pairs as illustrated in Fig. 3. Traditional Siamese based trackers fill the missing portions with the mean RGB value when a sub-window extends beyond the extent of the image. In this study, to ensure that all the image pairs are reshaped into a fixed size, the resizing of image is not included. Instead, the original image without padding is treated as instance. Then, exemplar images are up-sampled to the maximum size in a batch, which means the exemplar size is randomly altered according to each batch.

3.1.2 Feature extraction subnetwork

The feature extraction subnetwork is fully-convolutional. Similar to VGG, we only use an architecture with very small (3× 3) convolution filters. The network can achieve same receptive field with low burden of parameters.

Two branches compose the subnetwork. The template branch receives the exemplar patch (denoted as z). The search branch receives the full-scale instance patch (denoted as x). The two feature extraction branches share the same parameters. Thus, the same types of features can be compared in the following network. Let L_t represents the extraction operator $L_t x[u] = x[u - t]$, single padding introduced into the network with stride k can be defined as

$$h\,(L_{kt}x) = L_t h(x) + b \tag{1}$$

3.1.3 Correlation operator

Correlation operator is a batch processing function, which compares the Euclidean distance or similarity metric between $\varphi(z)$ and $\varphi(x)$. Here $\varphi(z)$ and $\varphi(x)$ denote

Fig. 3 Training pairs extracted from the same video: the first row and the second row display the exemplar and the instance, respectively. The third row is the corresponding label

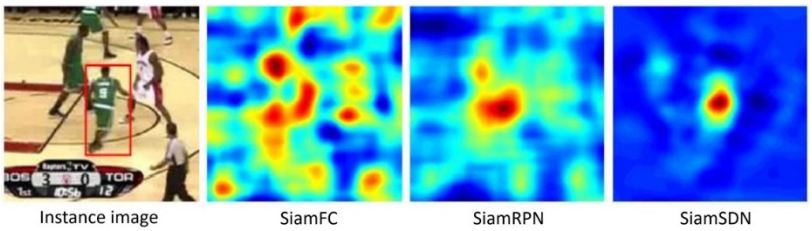

Fig. 4 Visualization of output heat map of an instance image. The results of SiamSDN are more accurate due to the single decoder layer

the outputs of template and search branches. Combining deep features in a higher dimension is equivalent to dense sampling around the bounding box and evaluating similarity after each feature extraction. However, the former method is more efficient due to the smaller scale of high dimension feature. For convenience, let $u(\varphi(z), \varphi(x))$ denotes the output of correlation function. The output is a multi-layer two-dimension raw map.

Various deep features are combined by the subsequent decoder network. To obtain a single channel response map for target localization, SiamFC (Bertinetto *et al*. 2016) utilizes a cross-correlation layer to obtain a single channel response map for target localization. In SiamRPN (Li *et al*. 2018), cross-correlation is extended to embed much higher level information such as anchors. Simply adding up these features will mix features from different aspects. The objects in the same category have high response on same channels, while responses of the rest channels are suppressed. Each channel represents the semantic information in accordance with the class. Thus, element-wise correlation is proposed for future decoder analysis.

3.1.4 Decoder

Since correlation operator is adopted into the network, decoder is needed to interpret the comparison results, namely raw maps, from higher into lower dimension. Conventional Siamese trackers simply interpolate the raw map into a larger size. This operation cannot fully decode the positions and details of correlation operator. A decoder with three interpolating layers and three deconvolution layers is adopted, as illustrated in Fig. 2. Correlation operator offers the maximum response position and to what extent these two features are similar to each other. Simple interpolating can only provide an area in the neighbourhood around the object center. Moreover, a small scale (17×17) score map cannot properly reserve the scale information, and the similarity between the exemplar and the instance is diluted. As shown in Fig. 4, the heat map of simple up-sampling of the correlation operator output is less accurate both in position and scale than the decoder subnetwork. Thus, a deconvolution layer is needed after interpolating. A single decoder branch is adopted taking these into consideration. Other decoders like regional proposal, multi-scale test and online fine-tuning can be discarded.

3.1.5 Output score map

The output score map shares the same size as instance image. In order to combine the classification and bounding box regression tasks of tracking together, the label is designed to obey two-dimension normal distribution as illustrated in Fig. 3. The mean value is the centre of bounding box. According to the three-sigma rule (Pukelsheim 1994), the probability for X falling away from its mean by more than 3 standard deviation is at most 5% if X obeys the normal distribution. Thus, the standard deviation in our label is one sixth of the width and height. The area outside of the bounding box is set to zero. The response value intensifies with the increase of overlapping area between the exemplar and the instance. Hence, the score around the edge of bounding box should be lower than the center part.

The tracking problem can be reinterpreted from a binary classification into a simple regression issue. The centre of the tracking object is obtained by mean values of the score map and the scale can be obtained by standard deviation. With single branch decoder, we integrate the foreground and background classification and bounding box regression into one regression issue.

Then, the loss function for each pair computes the distance between the label and the score map

$$l(y, y_{pred}) = (y - y_{pred})^2 \tag{2}$$

where y_{pred} is the response value in the score map. To exploit the fully-convolutional nature of our model, the decoder will produce a map of scores D. Define the loss of a score map to be the sum of individual losses

$$L(y, y_{pred}) = \frac{1}{|D|}\sum_{i \in D} l(y[i], y_{pred}[i]) \tag{3}$$

Notice that the positive area is far more less than the negative area. To balance the unevenness, the loss function can be redefined as

$$\begin{aligned} L(y, y_{pred}) = &\frac{\lambda_1}{|D_{pos}|}\sum_{i \in D_{pos}} l(y[i], y_{pred}[i]) \\ &+ \frac{\lambda_2}{|D_{neg}|}\sum_{i \in D_{neg}} l(y[i], y_{pred}[i]) \end{aligned} \tag{4}$$

where λ_1 and λ_2 are proportions of negative and positive areas over the entire score map. The loss of positive and negative area carries the same weight. Finally, the parameters in the network are obtained by applying Stochastic Gradient Descent (SGD)

$$argmin\theta\ L(y, y_{pred}) \tag{5}$$

3.1.6 Implementation details

The parameters of feature extraction subnetwork and decoder are found by minimizing Eq. (5) with SGD. The initial values of the parameters follow a Gaussian distribution, and are scaled according to the Kaiming method (He *et al*. 2015). 50 epochs are performed and the learning rate is decreased geometrically at 30 epoch from 10^{-3} to 10^{-5}. The gradients for each iteration are calculated using mini batches of size 128. We use GOT-10k (Huang *et al*. 2019) for training and OTB (Wu *et al*. 2015) and VOT dataset for testing in order to verify the feasibility and efficiency of our model. During tracking phase, the template branch is not updated online. For evaluation metrics, precision plot and success plot are two metrics to evaluate the tracker. The precision rate calculates the percentage of frames whose centre distances between the target and the ground-truth are within the given threshold. Normally, the threshold is 20 pixels. The success rate computes the quotient of the intersection and union areas. The area under curve of success plot is used to rank all the trackers.

3.2 Camera calibration

This study uses consumer-grade cameras (iPhone6 rear camera) to capture images. Since this kind of camera uses wide-angle lens, the captured images involve some distortions. Thus, the parameters of the camera and lens have to be calibrated first. It is crucial to obtain the optical parameters to calculate the structural displacement.

Under ideal conditions, imaging of objects in cameras can be described as using the "pinhole imaging" model, which is mathematically expressed as (Forsyth and Ponce 2002)

$$s\begin{bmatrix}u\\v\\1\end{bmatrix} = K_{3\times3}[R_{3\times3} \quad t_{3\times1}]\begin{bmatrix}X\\Y\\Z\\1\end{bmatrix} \tag{6}$$

Eq. (6) describes the relationship between a point $p\ (X,Y,Z,1)^T$ using the homogeneous coordinates in three-dimensional space and the corresponding point $p\ (u,v,1)^T$ on the image plane. In particular, K is called the intrinsic matrix, which is only related to the camera; R and t are extrinsic parameter-matrices; s is the scaling factor. The intrinsic and extrinsic matrix are calculated to transform the pixel level displacements into metric level displacements. Calibrations are conducted in Figs. 5 and 6 to fit the five independent parameters (α,β,c,u_0,v_0) in the intrinsic matrix K. The lens distortion coefficients have been acquired. $k_i\ (i=1,2,3)$, and $p_j\ (j=1,2)$ denotes the distortion coefficients. The results of parameter calibration are as shown in Tables 1 and 2.

3.3 Digital video stabilization

As UAV vibrates while hovering over a stationary point, this section introduces the method to eliminate displacements caused by the UAV. Although the tripod head can substantially reduce the jitter of cameras, small motions or rotations of camera will lead to huge error when measuring the structural displacements. Hence, the raw videos taken by the UAV still have to undergo digital video stabilization (DVS). Yoon *et al*. (2018) calculated the non-stationary motion of the UAV camera by using background information and integrated the relative displacement with the camera motion to acquire the absolute displacement. DVS aims to obtain stable videos through vibration compensation based on image processing algorithms.

Motions between two video frames can be divided into global motions and local motions. Global motions happen in most of the video frames, while the latter one only involves a small portion of the video. Global motions generally represent motions of the camera itself, whereas local motions mainly represent motions of foreground objects, as illustrated in Fig. 7. In the research scene of this study, global motions are vibrations of the UAV camera, whereas local motions are the vibrations of the structural model being shot. Since DVS aims to eliminate the jitter of the camera, its main task includes global motion estimation, motion compensation, and image generation.

DVS includes 2D, 3D, and 2.5D algorithms according to the differences between motion models. The 2D algorithm assumes that all points in an image only show planar motion and optical flow is determined by estimating the translation, scaling, and rotation between adjacent frame images. The model of the 2D algorithm is relatively simple. Stable output results can be obtained successfully in relatively

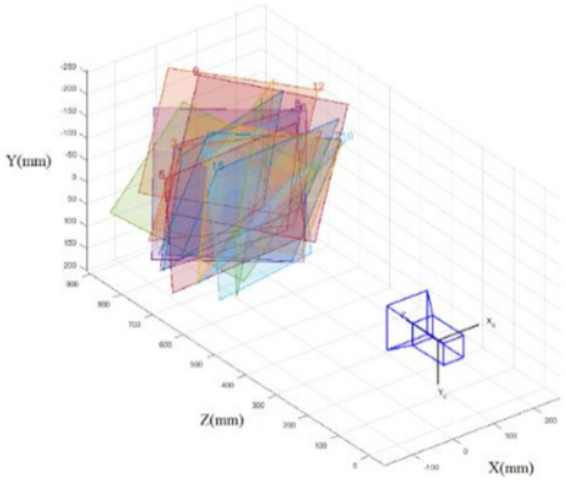

Fig. 5 Calibration process: camera-centered

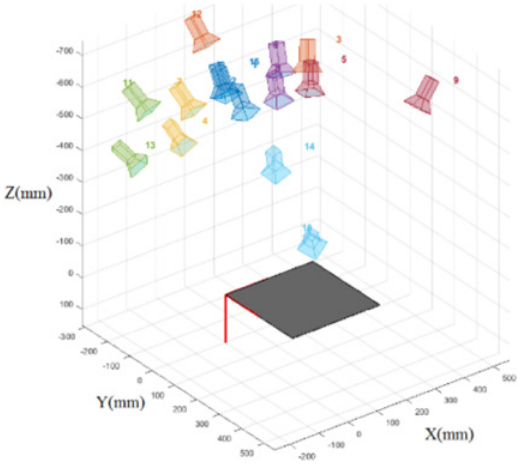

Fig. 6 Calibration process: chessboard-centered

Table 1 Internal parameters of iPhone 6 rear camera

Parameter	α	β	c	u_0	v_0
Value	2821.7	2815.78	-0.3299	1627.8	1201.1

Table 2 Distortion coefficient of iPhone 6 rear camera

Parameter	Parameter	k_1	k_2	k_3	p_1
Value	0.0617	0.125	-0.605	-9.738×10^{-4}	2.508×10^{-4}

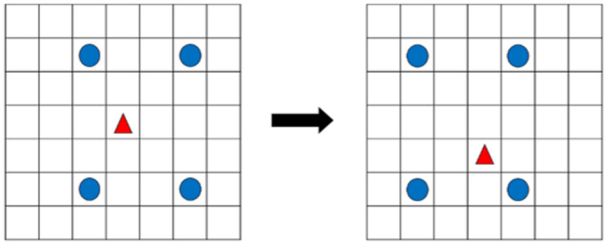

Fig. 7 Global motions and local motions

Fig. 8 SIFT point detection

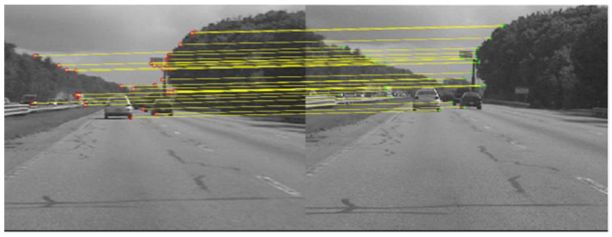

Fig. 9 SIFT point matching

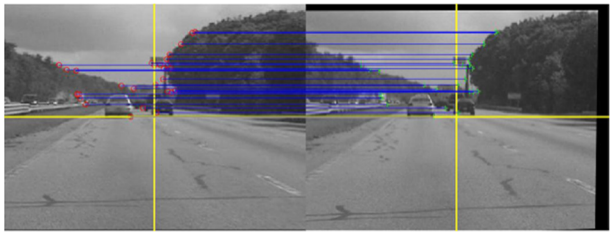

Fig. 10 Motion compensation

simple motion cases. For the application studied in this paper, when the UAV records videos of a structural model, the jitter of the camera is relatively small. In addition, tracking object scales are relatively small in comparison to the imaging field of the camera. Therefore, the 2D algorithm is employed to perform video stabilisation on videos recorded.

In the 2D video stabilisation algorithm, estimation of camera global motion parameters mainly includes gray projection (Lu and Du 2011), block matching (Bierling 1988) and feature point matching (Rosten and Drummond 2005). The feature point matching method obtains the optical flow of global motions in videos through extracting and matching of feature points. During video stabilization, the first frame is used as a reference. SIFT Point detection is carried out (Fig. 8) for two consecutive image frames. The vehicles in the video are foreground tracking objects. Since motion compensations are based on fixed points in two neighbouring frames, more feature points are detected in the background than foreground. Subsequently, feature descriptions and feature point matching are conducted (Fig. 9). Consistent refinement is employed to estimate motion models. During motion estimation, it is assumed that the images satisfy the refined model, the matrix of which can be expressed as

$$H_{affine} = \begin{bmatrix} a_1 & a_2 & 0 \\ a_3 & a_4 & 0 \\ t_x & t_y & 1 \end{bmatrix} \tag{7}$$

where a_i $(i = 1, 2, 3, 4)$ is related to scaling, rotation, and cropping in image transformation, and t_x and t_y are translation components. Motion compensation for each image is achieved through calculated motion models (Fig. 10).

4. Shaking table test and dynamic performance evaluation

This section provides the detail design of the shaking table test of a five-storey structure and the corresponding dynamic performance evaluation. The proposed vision framework is applied to the measurement process to verify its accuracy and robustness.

4.1 Test design

A five-storey structure is designed and deployed in the test. The dimensions of the model and the arrangement of transducers are shown in Fig. 11. Two types of video recording equipment are used in the test. One of them is the iPhone 6 rear camera, with a frame rate of 30 fps and a resolution of 1920×1080 pixels. The other one is a DJI UAV with a frame rate of 30 fps and a resolution of 3968×2976 pixels. The test conditions are listed in Table 3. The Linear Variable Differential Transformer (LVDT) displacement transducers used for reference have a sampling rate of 256 Hz, and are placed on each floor and on the base of the model. The experimental setup is shown in Figs. 13 and 14.

4.2 Commercial camera test results

Videos recorded by iPhone 6 are separated into sequences of video frames. Then, the tracking object of each floor is selected in a bounding box in the first frame, as shown in Fig. 12. The bounding box position and scale are detected automatically in the following frames. After the end-to-end SiamSDN pipeline, we obtain pixel level displacement time histories. By measuring pixel distances of objects in the calibrated images, scale factors between actual physical coordinates and pixel coordinates (pixel resolution) are obtained. Scale factors of the commercial cameras are listed in Table 4. The pixel distance and scale factor are multiplied to give the final displacement time

Table 3 Test conditions

No.	Condition	Amplitude (mm)	Frequency (Hz)	Duration (s)	Measuring tool
1	Sine	5	8	60	iPhone
2	Sine	5	11	60	iPhone
3	Sine	5	8	60	UAV
4	Sine	5	11	60	UAV
5	Linear sweep	10	1~50	180	iPhone

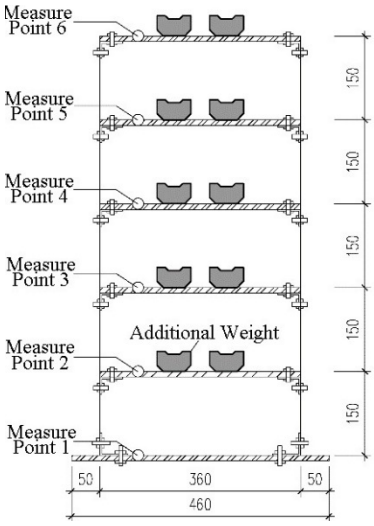

Fig. 11 Diagram of model design and transducer arrangement

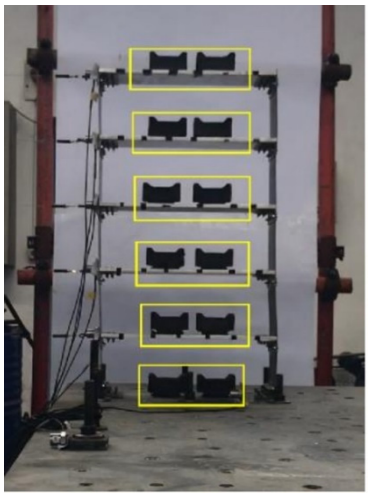

Fig. 12 Tracking setup: bounding boxes defined in the first frame

history curves.

Figs. 15 and 16 show the displacement time histories of testing scenarios 1 and 2. For a better comparison, one segment of the time histories is enlarged. The measurement results based on the proposed vision method agree well with those obtained by the LVDT displacement transducers. To quantify the precision of the vision-based system, error analysis is conducted using the root mean square error (RMSE) in Eq. (8) and the normalized root mean square error (NRMSE) in Eq. (9) (Feng *et al*. 2015)

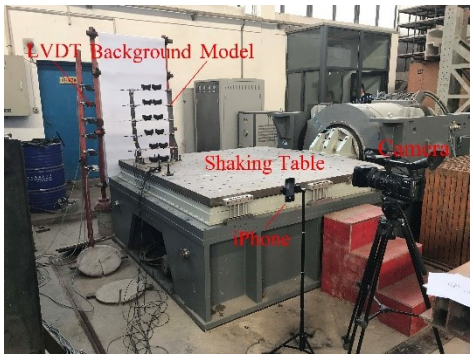

Fig. 13 Setup of the commercial camera test

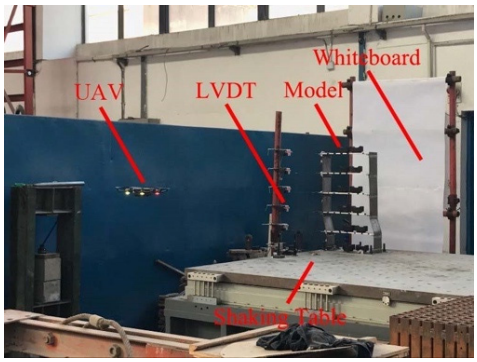

Fig. 14 Setup of the UAV test

Table 4 Scale factors of each floor from camera calibrations

No.	Ground	1F	2F	3F	4F	5F
1	1.96	1.81	1.81	1.81	1.79	1.82
2	1.25	1.48	2.32	2.41	1.39	1.88
3	1.33	1.30	1.30	1.22	1.22	1.23
4	1.62	1.52	1.52	1.52	1.45	1.45
5	1.26	1.17	1.17	1.20	1.19	1.19

Table 5 Measurement errors of each floor in the shaking table test with commercial cameras

Condition	Algorithm	Error type	1F	2F	3F	4F	5F
Scenario 1	SiamSDN	RMSE (mm)	0.11	0.14	0.07	0.25	0.18
		NRMSE (%)	1.06	1.56	3.23	4.50	2.01
	SIFT	RMSE (mm)	0.73	1.85	0.08	0.68	0.40
		NRMSE (%)	6.87	20.59	4.00	12.07	4.37
		Improvement (%)	84.57	92.42	19.25	62.72	54.00
Scenario 2	SiamSDN	RMSE (mm)	0.62	0.27	0.61	0.09	0.30
		NRMSE (%)	3.90	1.24	4.42	1.97	2.03
	SIFT	RMSE (mm)	1.77	3.31	0.94	0.62	0.99
		NRMSE (%)	11.11	15.00	6.89	14.29	6.76
		Improvement (%)	64.90	91.73	35.85	86.21	69.97

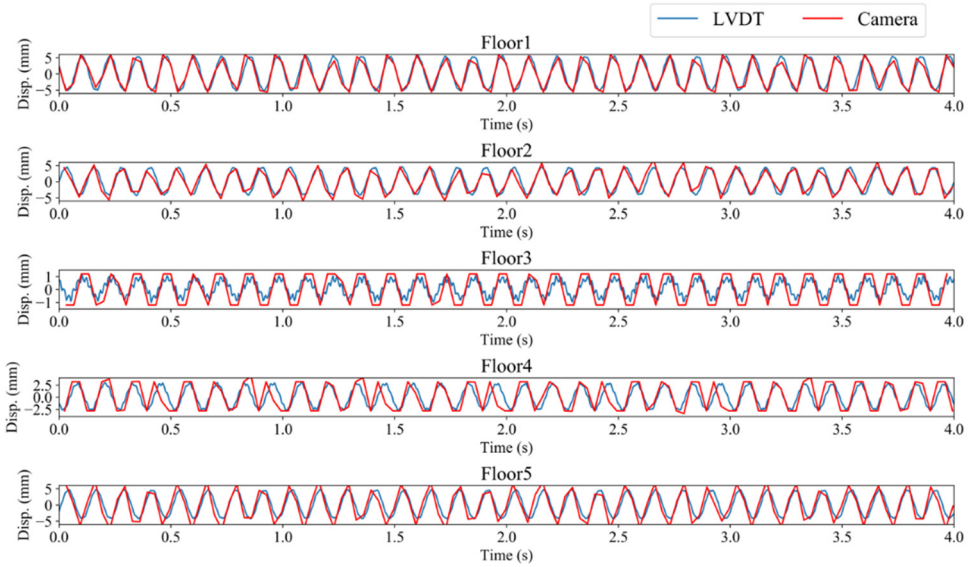

Fig. 15 Testing scenario 1: comparison of displacements by the proposed SiamSDN system and LVDT transducers

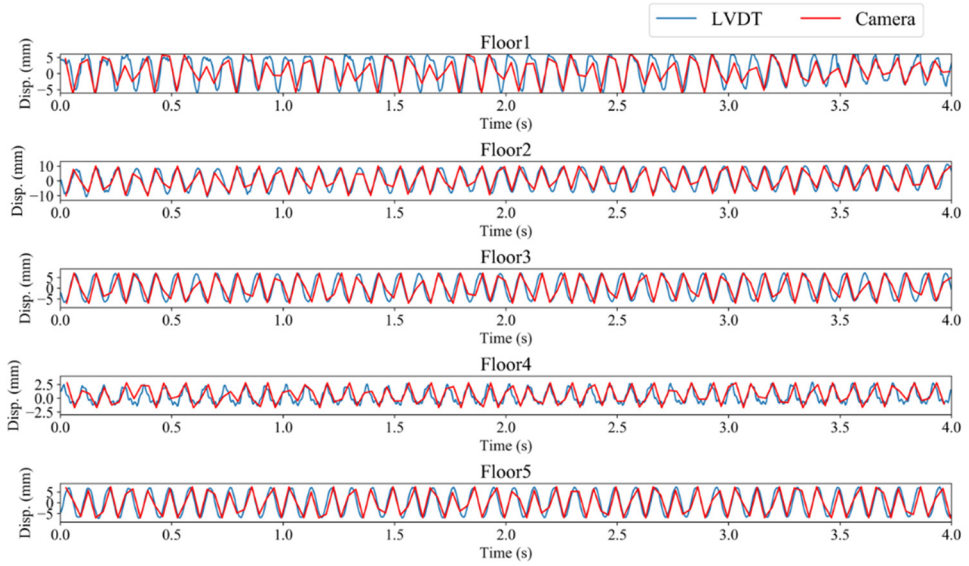

Fig. 16 Testing scenario 2: comparison of displacements by the proposed SiamSDN system and LVDT transducers

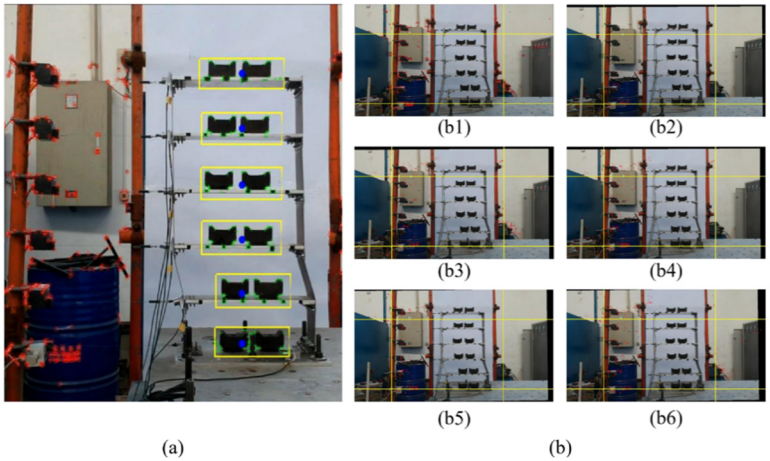

Fig. 17 DVS: (a) Feature points extracted from the background in the first frame; (b) Adjusted images after motion compensation: b1, b2, b3, b4, b5 and b6 are frame 1, 200, 500, 800, 1000 and 1400 from the original video

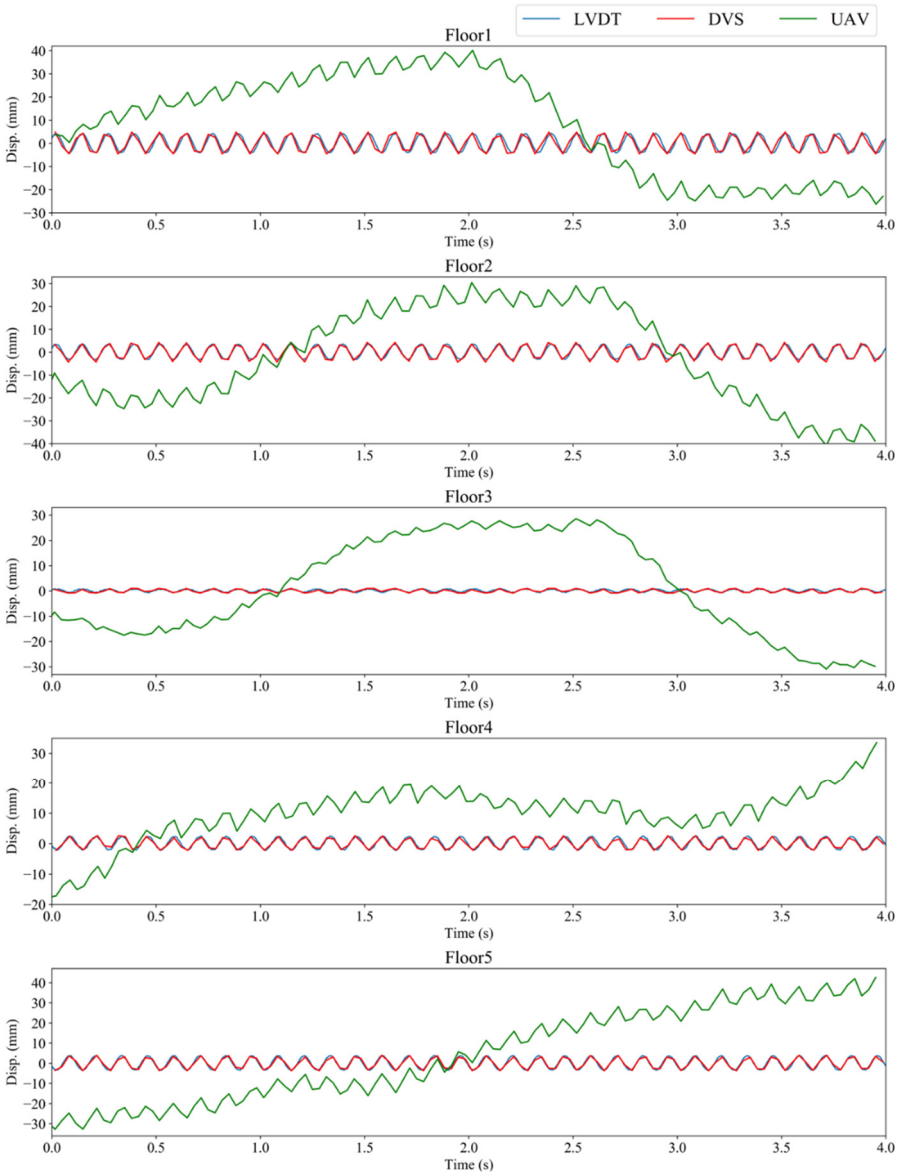

Fig. 18 Testing scenario 3: comparison of displacements by the proposed SiamSDN system and LVDT transducers. The green curve is calculated from the original UAV video, the red curve is obtained from the video after DVS processing, and the blue one is the LVDT displacement

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - y_i)^2} \tag{8}$$

$$NRMSE = \frac{RMSE}{y_{max} - y_{min}} = \frac{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - y_i)^2}}{y_{max} - y_{min}} \times 100\% \tag{9}$$

where n is the number of data points, x_i and y_i are the metric displacement measured by the vision and LVDT systems at time t_i, repectively.

Errors between the vision method and LVDT sensors are listed in Table 5. For the vision method, both SIFT point matching and the proposed SiamSDN tracking network are considered for comparison. As shown in Table 5, SiamSDN provides a more accurate tracking result than feature point matching (SIFT). The NRMSE of SiamSDN has improved from 19.25% (3F) to 92.42% (2F) in testing scenario 1, and also increased from 35.85% (3F) to 91.73% (2F) in testing scenario 2. The average improvement is 66.16%. The precision drops when the testing frequency increases. This is mainly because the frequency of test condition 2 is 11 Hz, which reaches the theoretical sampling limits of iPhone 6.

4.3 UAV camera test results

The vibrations of the UAV camera induce global motions. Therefore, DVS has to be conducted before tracking the motions of the target. First, the background region is selected in the first frame for key points detection (Fig. 17). Then, the feature points in the chosen background area are detected in subsequent frames. By matching feature

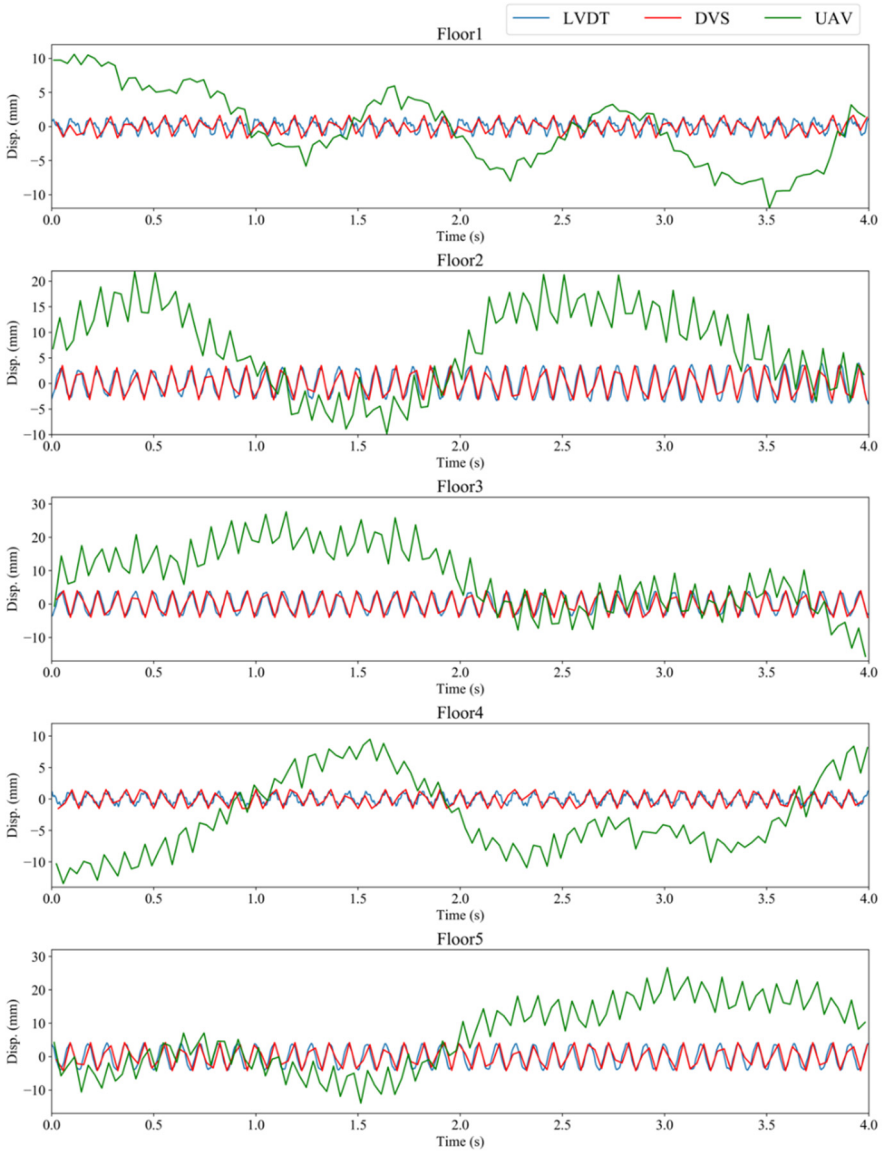

Fig. 19 Testing scenario 4: comparison of displacements by the proposed SiamSDN system and LVDT transducers. The green curve is calculated from the original UAV video, the red curve is obtained from the video after DVS processing, and the blue one is the LVDT displacement

points of two adjacent frames, the projection transformation model between the two images is calculated. The projection transformation matrix of the i th frame is obtained by successive multiplication of the projection transformation matrices of the previous $i-1$ frames

$$H_{cmu,i} = \prod_{j=0}^{i-1} H_j \tag{10}$$

Finally, using the first frame as the stabilisation reference, a stabilized image sequence is obtained.

Figs. 18 and 19 are the final tracking result of the testing scenario 3 and 4. The original UAV video tracking results are also displayed. As shown in the figures, the UAV camera vibrates as it is hovering over a fixed point, which largely affects the displacement analysis adversely. And the proposed DVS method has eliminated global motions caused by the camera. The motions of UAV camera in scenarios 3 and 4 are shown in Figs. 20 and 21. The largest displacements of the UAV are 180 mm and 60 mm in two scenarios. Frequency domain analyses reveal that the vibrations of the UAV cause energy concentration of measurement results in the low-frequency range below 1.5 Hz. Errors between the proposed method and LVDT sensors are listed in Table 6. Again, for the vision method, both SIFT point matching and the proposed SiamSDN tracking network are considered for comparison.

As shown in Table 6, all the tracking results of SiamSDN are better than the SIFT method. The NRMSE of SiamSDN has improved from 22.70% (3F) to 91.38% (2F) in testing scenario 3, and also increased from 1.01% (3F) to 85.36% (2F) in testing scenario 4. The average improvement is 57.54%. Comparing scenario 1 and 3, 2 and 4, the RMSE of the UAV is smaller than the commercial cameras under the same shaking table input. This is mainly because the UAV camera has a higher resolution than iPhone 6, which provides a better tracking accuracy. Through DVS and the novel DNN object tracking algorithm, the UAV measuring system can effectively and accurately restore the actual displacement of the target structure.

4.4 Analysis of frequency characteristics

Figs. 22, 23, 24 and 25 manifest the power spectrum density (PSD) computed from the displacements in testing conditions 1, 2, 3 and 4 respectively. Comparing the results of the vision and LVDT methods, we can observe that the frequency component obtained from both displacements agree well with each other. In testing scenario 5, we analyze the dominant frequencies and other high-order frequencies from the vision measurement system (Fig. 26). The higher-order frequencies are calculated from the EMD technique (Huang *et al*. 1998), and the results are listed in Table 7. Comparisons of the test results reveal that the non-contact vision measurement method is as accurate as the traditional LVDT methods.

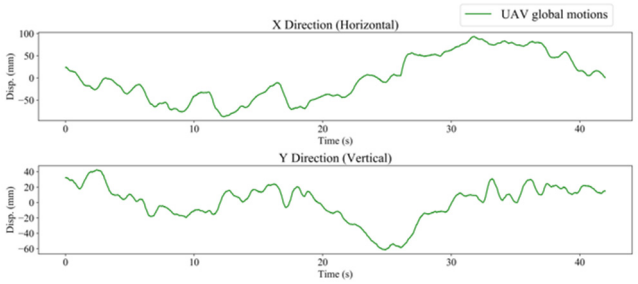

Fig. 20 Testing scenario 3: global motions of the UAV camera in both x and y directions

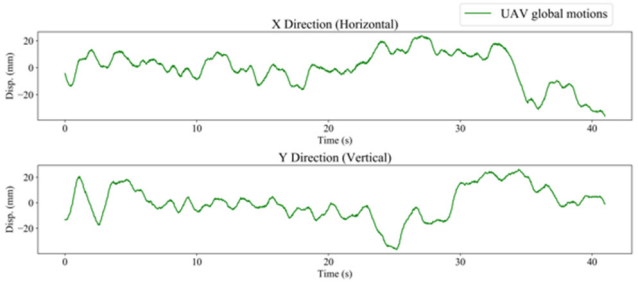

Fig. 21 Testing scenario 4: global motions of the UAV camera in both x and y directions

Table 6 Measurement errors of each floor in the shaking table test with UAV cameras

Condition	Algorithm	Error type	1F	2F	3F	4F	5F
Scenario 3	SiamSDN	RMSE (mm)	0.38	0.09	0.05	0.03	0.16
		NRMSE (%)	4.52	1.32	4.12	0.70	2.22
	SIFT	RMSE (mm)	1.18	1.05	0.06	0.31	0.29
		NRMSE (%)	13.91	15.32	5.33	6.65	4.01
		Improvement (%)	67.51	91.38	22.70	89.47	44.64
Scenario 4	SiamSDN	RMSE (mm)	0.20	0.19	0.44	0.04	0.27
		NRMSE (%)	6.30	2.44	5.91	1.69	3.33
	SIFT	RMSE (mm)	0.25	1.32	0.45	0.21	0.82
		NRMSE (%)	8.07	16.67	5.97	8.42	9.94
		Improvement (%)	21.93	85.36	1.01	79.93	66.50

Table 7 Natural frequencies identification from testing scenario 5

Measurement	1st	2nd	3rd
LDVT (Hz)	2.000	3.379	10.948
SiamSDN (Hz)	1.997	3.354	11.052

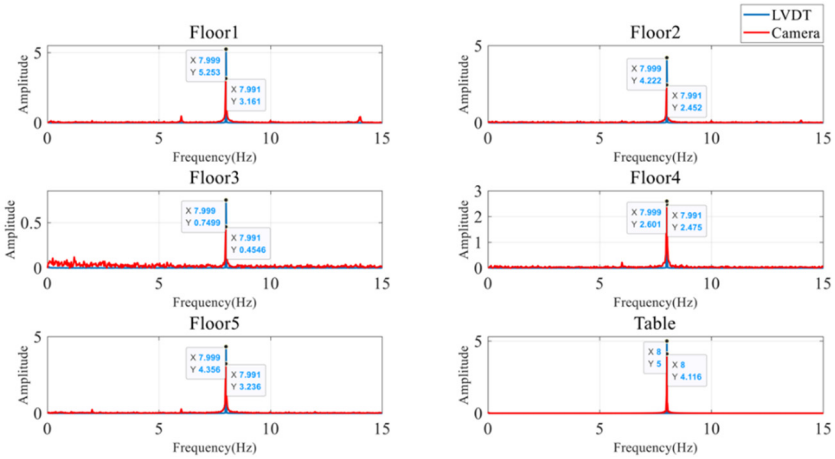

Fig. 22 Testing scenario 1: comparisons of the power spectrum density

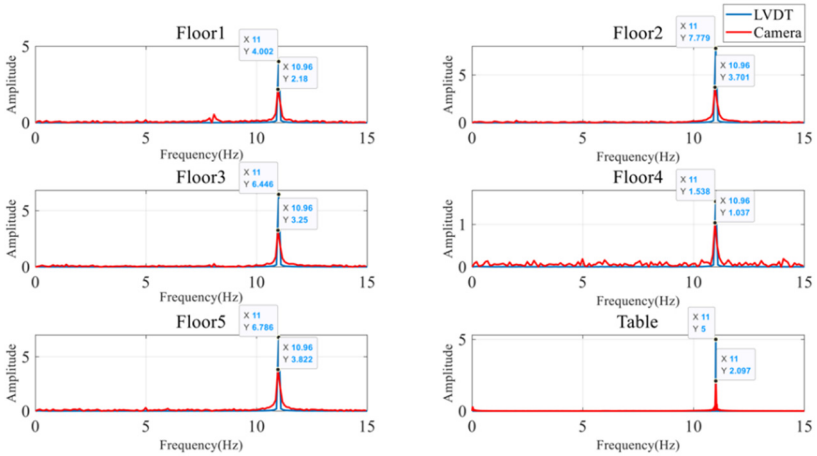

Fig. 23 Testing scenario 2: comparisons of the power spectrum density

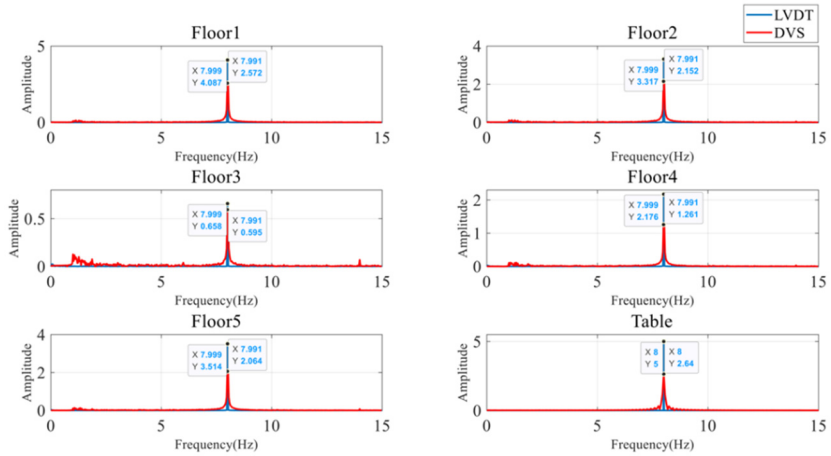

Fig. 24 Testing scenario 3: comparisons of the power spectrum density

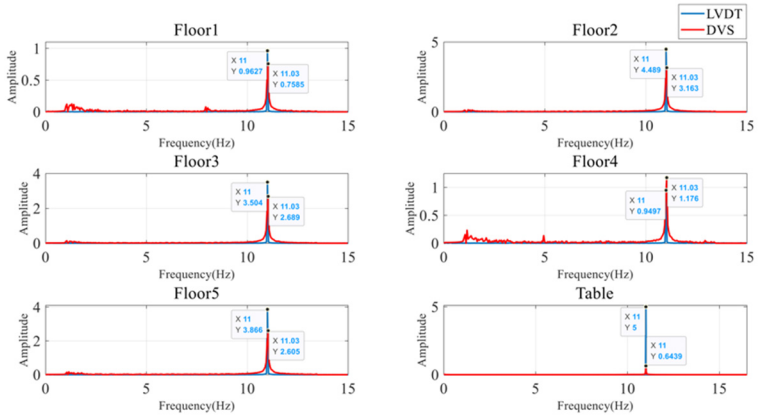

Fig. 25 Testing scenario 4: comparisons of the power spectrum density

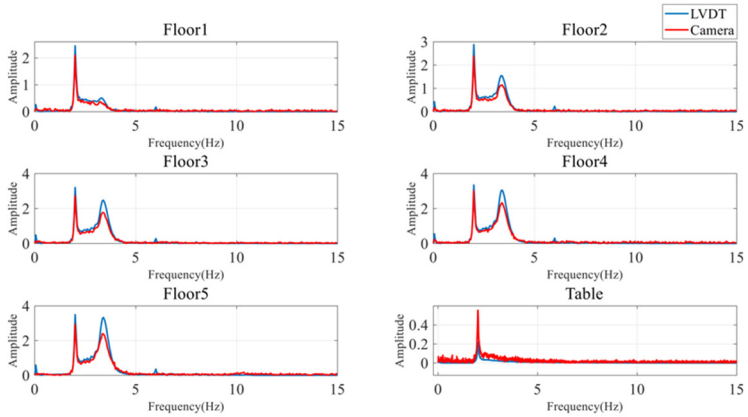

Fig. 26 Testing scenario 5: comparisons of the power spectrum density

5. Conclusions

This study proposes a vision structural system identification framework with a novel DNN tracker SiamSDN. The proposed system requires no target installation and utilizes consumer-grade cellphone cameras. This study firstly explores the application of an end-to-end DNN tracking pipeline, which is designed based on the Siamese network, in the field of structural system identification. A shaking table test of a 5-storey structure is carried out to demonstrate the efficiency. Besides, a UAV camera is used to simulate the field test. To minimize the vibrations of UAV, DVS is proposed to eliminate the drifts through motion compensation. Videos taken by both the commercial and UAV cameras are analyzed to calculate the displacement time histories, and the frequency characteristics are obtained from the displacements. The following conclusions can be made after analyzing the test results.

- The accuracy and robustness of the proposed vision measuring framework have been proved. The shaking table test has simulated the field test of measuring the displacements of tall buildings via UAV camera with no manually installed target. SiamSDN does not require manually extracted features or pre-defined motion areas. The tracking object is solely identified in the first frame. Given a sequence of input images, SiamSDN uses a class-specific detector to predict the states of an object in each frame.
- The tracking precision of DNN tracker is more accurate than conventional generative trackers. CNN automatically extracts and selects the features from the tracking object, which is more preferable for tracking the small displacements of the structure. SiamSDN has improved the displacement measuring accuracy by 66.16% through analyzing results of the commercial camera. Besides, SiamSDN operates at 67 frames per second (fps), while SIFT operates at 0.8 fps.
- Original videos recorded by the UAV contain the structural displacement and the global motion of UAV itself. Frequency domain analysis reveals that vibrations of the UAV are mainly below 1.5 Hz. DVS has successfully eliminated drifts induced by the UAV vibration, and restore the absolute displacement of the structure. It is crucial when using UAV as a camera carrier for SHM in the field test.
- After DVS processing, the displacement time histories of the structure can be obtained. SiamSDN has improved the displacement measuring accuracy by 57.54% through analyzing results of the UAV

camera. Dominant and higher-order frequencies are calculated from the displacements.

The high availability and low cost of this framework largely facilitate the building monitoring. It has good potential for the applications in SHM. In future work, stereo-vision can be adopted to further improve the accuracy of measurement.

Acknowledgments

This research work is supported by the National Nature Science Foundation of China (Grant No. 52025083, 51878449) and the Fundamental Research Funds for the Central Universities.

References

Bertinetto, L., Valmadre, J., Henriques, J.F., Vedaldi, A. and Torr, P.H.S. (2016), "Fully-convolutional siamese networks for object tracking", *Lecture Notes in Computer Science*: *Computer Vision - ECCV 2016 Workshops*, Vol. 9914, pp. 850-865.

Bierling, M. (1988), "Displacement estimation by hierarchical blockmatching", *Proceedings of Visual Communications and Image Processing '88*: *Third in a Series*, November, Cambridge, MA, USA, Vol. 1001, pp. 942-953. https://doi.org/10.1117/12.969046

Busca, G., Cigada, A., Mazzoleni, P. and Zappa, E. (2014), "Vibration monitoring of multiple bridge points by means of a unique vision-based measuring system", *Experim. Mech.*, **54**(2), 255-271. https://doi.org/10.1007/s11340-013-9784-8

Dong, C.Z., Bas, S. and Catbas, F.N. (2019), "A completely non-contact recognition system for bridge influence line using portable cameras and computer vision", *Smart Struct. Syst.*, *Int. J.*, **24**(5), 617-630. https://doi.org/10.12989/sss.2019.24.5.617

Feng, D.M. and Feng, M.Q. (2016), "Vision-based multipoint displacement measurement for structural health monitoring", *Struct. Control Health Monitor.*, **23**(5), 876-890. https://doi.org/10.1002/stc.1819

Feng, M.Q., Fukuda, Y., Feng, D.M. and Mizuta, M. (2015), "Nontarget vision sensor for remote measurement of bridge dynamic response", *J. Bridge Eng.*, **20**(12), 04015023. https://doi.org/10.1061/(ASCE)BE.1943-5592.0000747

Forsyth, D.A. and Ponce, J. (2002), *Computer Vision*: *A Modern Approach*, Prentice Hall Professional Technical Reference.

Fukuda, Y., Feng, M.Q., Narita, Y., Kaneko, S. and Tanaka, T. (2013), "Vision-based displacement sensor for monitoring dynamic response using robust object search algorithm", *IEEE Sensors J.*, **13**(12), 4725-4732. https://doi.org/10.1109/JSEN.2013.2273309

Harris, C. and Stephens, M. (1988), "A combined corner and edge detector", *Proceedings of 4th Alvey Vision Conference*.

He, K., Zhang, X., Ren, S. and Sun, J. (2015), "Delving deep into rectifiers: surpassing human-level performance on imagenet classification", *Proceedings of the IEEE International Conference on Computer Vision* (*ICCV*), December, pp. 1026-1034.

Huang, N.E., Shen, Z., Long, S.R. and Wu, M.C. (1998), "The empirical mode decomposition and the Hilbert spectrum for nonlinear and non-stationary time series analysis", *Proceedings of the Royal Society of London. Series A*: *Mathematical, Physical and Engineering Science*, **454**(1971), 903-995. https://doi.org/10.1098/rspa.1998.0193

Huang, L., Zhao, X. and Huang, K. (2019), "Got-10k: A large high-diversity benchmark for generic object tracking in the wild", *IEEE Transact. Pattern Anal. Mach. Intell.*, **43**(5), 1562-1577. https://doi.org/10.1109/TPAMI.2019.2957464

Ji, Y.F. and Chang, C.C. (2008), "Nontarget image-based technique for small cable vibration measurement", *J. Bridge Eng.*, **13**(1), 34-42. https://doi.org/10.1061/(ASCE)1084-0702(2008)13:1(34)

Jo, H., Sim, S.H., Tatkowski, A., Spencer, B.F. and Nelson, M.E. (2013), "Feasibility of displacement monitoring using low-cost GPS receivers", *Struct. Control Health Monitor.*, **20**(9), 1240-1254. https://doi.org/10.1002/stc.1532

Jung, H.J., Lee, J.H., Yoon, S. and Kim, I.H. (2019), "Bridge Inspection and condition assessment using Unmanned Aerial Vehicles (UAVs): Major challenges and solutions from a practical perspective", *Smart Struct. Syst., Int. J.*, **24**(5), 669-681. https://doi.org/10.12989/sss.2019.24.5.669

Kohut, P., Holak, K., Uhl, T., Ortyl, L., Owerko, T., Kuras, P. and Kocierz, R. (2013), "Monitoring of a civil structure's state based on noncontact measurements", *Struct. Health Monitor.*, **12**(5-6), 411-429. https://doi.org/10.1177/1475921713487397

Lee, J.H., Jung, C.Y., Choi, E. and Cheung, J.H. (2017), "Vision-based multipoint measurement systems for structural in-plane and out-of-plane movements including twisting rotation", *Smart Struct. Syst., Int. J.*, **20**(5), 563-572. https://doi.org/10.12989/sss.2017.20.5.563

Li, B., Yan, J.J., Wu, W., Zhu, Z. and Hu, X.L. (2018), "High performance visual tracking with siamese region proposal network", *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR*), June, pp. 8971-8980.

Liu, C.X., Chen, L.C., Schroff, F., Adam, H., Hua, W., Yuille, A.L. and Li, F.F. (2019a), "Auto-deeplab: Hierarchical neural architecture search for semantic image segmentation", *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR*), June, pp. 82-92.

Liu, J.T., Yang, X.X. and Li, L. (2019b), "VibroNet: Recurrent neural networks with multi-target learning for image-based vibration frequency measurement", *J. Sound Vib.*, **457**, 51-66. https://doi.org/10.1016/j.jsv.2019.05.027

Lu, L. and Du, W.T. (2011), "The vehicle-borne electronic image stabilization system based on Gray Projection Algorithm", *Proceedings of International Conference on Electric Information and Control Engineering*, April, Wuhan, China, pp. 4687-4690.

Luo, L.X. and Feng, M.Q. (2018), "Edge-enhanced matching for gradient-based computer vision displacement measurement", *Comput.-Aided Civil Infrastruct. Eng.*, **33**(12), 1019-1040. https://doi.org/10.1111/mice.12415

Lydon, D., Taylor, S.E., Lydon, M., Rincon, J.M. and Hester, D. (2019), "Development and testing of a composite system for bridge health monitoring utilizing computer vision and deep learning", *Smart Struct. Syst., Int. J.*, **24**(6), 723-732. https://doi.org/10.12989/sss.2019.24.6.723

Mas, D., Ferrer, B., Espinosa, J., Perez Rodriguez, J., Roig Hernandez, A.B. and Illueca Contri, C. (2011), "High speed imaging and algorithms for non invasive vibrations measurement", *EVACES 2011 – Proceedings of the 4th International Conference on Experimental Vibration Analysis for Civil Engineering Structures*, Varenna, Italy, October.

Narazaki, Y., Hoskere, V., Eick, B.A., Smith, M.D. and Spencer, B.F. (2019), "Vision-based dense displacement and strain estimation of miter gates with the performance evaluation using physics-based graphiscs models", *Smart Struct. Syst., Int. J.*, **24**(6), 709-721. https://doi.org/10.12989/sss.2019.24.6.709

Nassif, H.H., Gindy M. and Davis, J. (2005), "Comparison of laser Doppler vibrometer with contact sensors for monitoring bridge deflection and vibration", *NDT & E International*, **38**(3), 213-

218. https://doi.org/10.1016/j.ndteint.2004.06.012
Oh, B.K., Glisic, B., Kim, Y. and Park, H.S. (2019), "Convolutional neural network-based wind-induced response estimation model for tall buildings", *Comput.-Aided Civil Infrastruct. Eng.*, **34**(10), 843-858. https://doi.org/10.1111/mice.12476
Pan, B., Tian, L. and Song, X.L. (2016), "Real-time, non-contact and targetless measurement of vertical de- flection of bridges using off-axis digital image correlation", *NDT & E Int.*, **79**, 73-80. https://doi.org/10.1016/j.ndteint.2015.12.006
Peng, S.Y., Yu, Y.X., Wang, K. and He, L. (2020), "Accurate Anchor Free Tracking", *arXiv*, 2006.07560.
Pukelsheim, F. (1994), "The three sigma rule", *The American Statistician*, **48**(2), 88-91.
Ribeiro, D., Calçada, R., Ferreira, J. and Martins, T. (2014), "Non-contact measurement of the dynamic dis- placement of railway bridges using an advanced video-based system", *Eng. Struct.*, **75**, 164-180. https://doi.org/10.1016/j.engstruct.2014.04.051
Rosten, E. and Drummond, T. (2005), "Fusing points and lines for high performance tracking", *Proceedings of the 10th IEEE International Conference on Computer Vision*, *ICCV 2005*, pp. 1508-1515.
Shi, J.B. (1994), "Good features to track", *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, Seattle, WA, USA, June, pp. 593-600. https://doi.org/10.1109/CVPR.1994.323794
Spencer, B.F., Hoskere, V. and Narazaki, Y. (2019), "Advances in computer vision-based civil infrastructure inspection and monitoring", *Engineering*, **5**(2), 199-222. https://doi.org/10.1016/j.eng.2018.11.030
Wahbeh, A.M., Caffrey, J.P. and Masri, S.F. (2003), "A vision-based approach for the direct measurement of displacements in vibrating systems", *Smart Mater. Struct.*, **12**(5), 785-794. https://doi.org/10.1088/0964-1726/12/5/016
Wu, Y., Lim, J. and Yang, M. (2015), "Object tracking benchmark", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **37**(9), 1834-1848. https://doi.org/10.1109/TPAMI.2014.2388226
Ye, X.W., Dong, C.Z. and Liu, T. (2016), "A review of machine vision-based structural health monitoring: methodologies and applications", *J. Sensors*, 7103039. https://doi.org/10.1155/2016/7103039
Ye, X.W., Jin, T. and Yun, C.B. (2019), "A review on deep learning-based structural health monitoring of civil infrastructures", *Smart Struct. Syst., Int. J.*, **24**(5), 567-585. https://doi.org/10.12989/sss.2019.24.5.567
Yoon, H., Elanwar, H., Choi, H., Golparvar-Fard, M. and Spencer, B.F. (2016), "Target-free approach for vision-based structural system identification using consumer-grade cameras", *Struct. Control Health Monitor.*, **23**(12), 1405-1416. https://doi.org/10.1002/stc.1850
Yoon, H., Shin, J. and Spencer, B.F. (2018), "Structural displacement measurement using an unmanned aerial system", *Comput.-Aided Civil Infrastruct. Eng.*, **33**(3), 183-192. https://doi.org/10.1111/mice.12338
Yu, S.S. and Zhang, J. (2020), "Fast bridge deflection monitoring through an improved feature tracing algorithm", *Comput.-Aided Civil Infrastruct. Eng.*, **35**(3), 292-302. https://doi.org/10.1111/mice.12499
Zhang, X.C., Ye, P., Peng, S.Y., Liu, J., Gong, K. and Xiao, G. (2019a), "SiamFT: An RGB-infrared fusion tracking method via fully convolutional siamese networks", *IEEE Access*, **7**, 122122-122133. https://doi.org/10.1109/ACCESS.2019.2936914
Zhang, Y.Q., Miyamori, Y., Mikami, S. and Saito, T. (2019b), "Vibration-based structural state identification by a 1-dimensional convolutional neural network", *Comput.-Aided Civil Infrastruct. Eng.*, **34**(9), 822-839. https://doi.org/10.1111/mice.12447
Zhang, X.C., Ye, P., Peng, S.Y., Liu, J. and Xiao, G. (2020), "DSiamMFT: An RGB-T fusion tracking method via dynamic Siamese networks using multi-layer feature fusion", *Signal Process.: Image Commun.*, **84**, 115756. https://doi.org/10.1016/j.image.2019.115756

BS