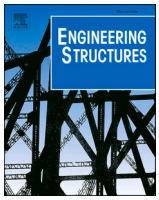
Engineering Structures 301 (2024) 117285

Contents lists available at ScienceDirect

Engineering Structures

journal homepage: www.elsevier.com/locate/engstruct



A deep learning-based method for structural modal analysis using computer vision

Yingkai Liu [a], Ran Cao [a,b], Shaopeng Xu [a], Lu Deng [a,b,*]

[a] College of Civil Engineering, Hunan University, Changsha, Hunan 410082, People's Republic of China
[b] Hunan Provincial Key Laboratory for Damage Diagnosis for Engineering Structures, Hunan University, Changsha, Hunan 410082, People's Republic of China

ARTICLE INFO

Keywords:
Computer vision
Modal parameter identification
CNN
LSTM

ABSTRACT

Structural modal analysis aims to determine a structure's natural frequency, damping ratio, and mode shape, helping with structural condition assessment and maintenance. In this study, a computer vision-based framework for the identification of structural modal parameters is developed, which consists of two main procedures: First, the one-dimensional (1D) vibration signals of edge pixels on the structure in the video are extracted via edge detection and optical flow theory. Second, a 1D convolutional neural network (CNN) coupled with long short-term memory (LSTM) is generated to extract structural modal parameters from the input 1D signal. The framework's performance has been validated through comparison with baseline values, which were obtained from contact sensors. Additionally, the model's robustness and extrapolability has been analyzed. The good performance of the computer vision-based approach confirms its potential for precise and dependable contact-free modal analysis.

1. Introduction

The modal characteristics of a structure, such as its natural frequency, mode shape, and damping ratio, are essential indicators of its response to external loads. Identifying these parameters accurately is crucial for damage detection and safety evaluation in the fields of civil, aerospace, and mechanical engineering. Since the 1960 s, studies on identifying modal parameters have extended beyond 60 years, and several algorithms have been created [1].

Traditionally, the modal parameters of a structure can be identified through the input excitation and output response. However, in practice, measuring the excitation signal is often unfeasible [2]. As a result, the output-only modal parameter identification algorithms have been proposed: the Ibrahim time domain method [3], the time series analysis method based on the ARMA model [4], the NExT technique that employs correlation functions [5], the ERA method that uses Hankel matrices and singular value decomposition [6], the EMD method based on localized features of the signals [7], and the BSS method that exploits the mutual independence of the signal sources [8]. Based on the above algorithms, a large number of researchers have investigated the modal parameter analysis of building structures [9–14]. For example, Siringoringo et al. [15] collected the ambient vibration response by installing 21 accelerometers on the Hakucho Suspension Bridge, and then used two output-only time-domain system identification methods (stochastic descent combined with the ITD method and NExT combined with ERA) to obtain the modal parameters of the bridge. It is accurate to state that the desired structural response can be captured using accelerometers. Nevertheless, the conventional contact sensors are often expensive to install and maintain [16].

Vision-based techniques offer non-contact advantages over conventional accelerometers, and many computer vision techniques have been developed for structural inspection and monitoring of civil infrastructure [17–19]. Based on computer vision, several algorithms can be used for displacement extraction of structures through template matching or target tracking [20–22]. Early applications of these algorithms focused on natural frequency estimation [23] and deformation measurements [24,25]. In recent years, the applications of these algorithms have been extended to modal analysis of structures. Schumacher and Shariati [26] employed targets in images as virtual visual sensors to conduct structural modal analysis. Yoon et al. [27] identified a laboratory-scale building using virtual sensors. Feng et al. [28] systematically identified laboratory structures by utilizing visual data obtained from unmanned aerial vehicles (UAVs) [29]. Hoskere et al. [30] recorded vibration videos of a structure with preset fiducials using a UAV camera

* Corresponding author at: College of Civil Engineering, Hunan University, Changsha, Hunan 410082, People's Republic of China.
E-mail address: denglu@hnu.edu.cn (L. Deng).

https://doi.org/10.1016/j.engstruct.2023.117285
Received 22 June 2023; Received in revised form 27 November 2023; Accepted 2 December 2023
Available online 19 December 2023

and then used NExT to identify the modal parameters of the structure. Although the studies were able to measure structural modes without contacting the object, they still required manually attaching markers to the surface of the structure. Furthermore, the studies primarily introduced vision-based procedures for acquiring signals of structural vibration, while the modal analysis process remains rooted in conventional algorithms.

The identification of structural modal parameters presents an inverse problem of structural dynamics that boils down to an optimization problem, while the deep neural network is a highly efficient optimization tool that outperforms traditional algorithms in terms of accuracy and cost-effectiveness [31]. Over the last few years, deep learning has become increasingly pivotal in the identification of structural modal parameters. Yue et al. [32] introduced a technique that integrates a priori knowledge and deep learning to precisely identify structural modes by analyzing transient vibration responses. Nevertheless, the accuracy of this approach heavily relies on the quality of the a priori information. Liu et al. [33] have developed an algorithm that employs a deep neural network to extract modal responses and vibration modes from raw response signals. However, traditional power spectral density methods still play a crucial role in the ultimate identification of modal parameters. In comparison, Kim et al. [34] introduced a faster R-CNN operational modal analysis algorithm that replaces manual determination in identifying natural frequencies from the spectrogram. Su et al. [35] employed an algorithm assisted with CNN for automated structural modal analysis. However, CNN was specifically used to substitute manual statistics in order to read natural frequencies and damping ratios from stability diagrams obtained by the SSI algorithm. Zhang et al. [36] suggested a 1D CNN detection algorithm to distinguish structural variations but it is unable to measure particular modes. In summary, these previous studies have all improved the accuracy of modal analysis through deep learning algorithms. However, these algorithms still require the use of traditional contact sensors to acquire the raw signals. Although contact sensors ensure reliability, they usually suffer from various problems such as limited range, complex procedures, high cost and low spatial resolution. Furthermore, within the aforementioned studies, deep learning serves primarily as a substitute for crucial steps of traditional algorithms and does not enable end-to-end structural modal analysis. Although accuracy is enhanced, such a deep-learning-based approach also results in a higher algorithmic complexity.

To decrease the complexity of the method, some scholars have started building deep neural networks by using raw signals and structural modal information directly [37,38]. Meanwhile, to overcome the disadvantages of contact sensors, the combination of deep neural networks with machine vision has been further explored [39,40]. In these studies, the neural networks were trained and tested using videos as opposed to 1D signals. For instance, Chao et al. [39] formulated a video recognition framework to ascertain faults in vibrating structures like bearings. The camera captured a video of the bearing vibration, and the frame-level features are separated using CNNs. These features were then characterized by LSTM units to identify the normal or faulty vibration frequency. Similar to Zhang et al. [36], their study classified structural modes as either normal or abnormal, without conducting regression analysis on the specific modal parameters. Yang et al. [41] introduced a method for capturing vibration videos of structures with a camera and determining natural frequencies from raw video using CNN-LSTM. Rather than traditional contact sensors, this study utilizes a video camera for non-contact modal analysis of a basic metal bar. However, to augment algorithmic autonomy and minimize human intervention, this approach utilizes raw vibration videos as training data. As the video is significantly larger in data size than that of the 1D signal obtained from the contact sensor, limitations of their method existed regarding the size of the structure being measured (less than 0.2 m) and the dimension of the video (video length less than 200 frames, size less than 64 $\times$64 $\times$3). Otherwise, as the dimensionality of the structure increases, the amount of data may become too large for standard computers to complete training and testing. Therefore, a balance between the autonomy of the algorithm and human involvement should be ensured in practical applications.

In this study, an algorithm for structural modal parameter identification based on computer vision and deep neural networks is proposed. In order to capture structural vibration videos at higher resolutions and longer video lengths, it is necessary to improve the training and detection efficiency and reduce network complexity. Therefore, rather than directly utilizing the recorded structural vibration videos as training and detection data, the raw videos were pre-processed to increase the density of valid information in the training and testing data. This preprocessing simply extracts the 1D vibration signals of the structural edge pixels from the video using edge detection and optical flow theory. As the training data consisted of 1D signals, a 1D CNN-LSTM with a more concise network structure has been implemented in the process of modal identification. This approach leads to significantly improved training and testing efficiency compared with the method that used video as direct training data.

The structure of this manuscript is explained as follows. Section 2 begins with a description of visual displacement extraction, where the established high-speed visual measurement system and Flownet2 are presented in detail. In Section 3, the proposed 1D-CNN-LSTM is introduced, followed by the dataset collection, model training, and coordinate recovery. In Section 4, parametric analyses are conducted to validate the accuracy, robustness, and extrapolability of the method. In Section 5, the final concluding remarks are presented.

2. Vision measurement of dynamic displacement

In the vision-based dynamic displacement measurement, the vision data contained in a sequence of image frames are acquired by a camera. Each image frame is a two-dimensional (2D) matrix of pixels with varying light intensity $I(\boldsymbol{x}, t)$ at different times t. This section outlines the approach to extract vibration responses of edge pixels from the visual data. As illustrated in Fig. 1, the corresponding flowchart outlines the following steps:

1. Acquire the Bayer8 images using the camera and convert them to RGB format.
2. Calculate the weighted average of the three-channel values to obtain the grayscale frames.
3. Smooth frames using a 3D Gaussian smoothing filter.
4. Extract edge pixels by using the Canny operator, and identify the full-field motion by using Flownet2.
5. Determine the motion of the edge pixels based on the extracted edge pixels and the full field motion.

2.1. High-speed vision-based measurement system

A high-speed vision measurement system (HVMS) has been developed, consisting of an industrial camera head, a zoom lens, and a computer workstation. The specific models of each component are listed in Table 1. As an image receiver, a high-performance charge-coupled system is installed into the camera head to capture up to 8-bit Bayer pictures (4096 $\times$ 3072 pixels) at a maximum frame rate of 409 fps (frames per second). Due to the requirement of real-time analyzing the video captured from the camera, a 10 GbE network internet card is needed to provide a steady stream of data from the camera to the computer.

2.2. Edge pixel motion

Each HVMS-obtained video frame comprises $M \times N$ pixels, with M pixels being oriented vertically and N pixels being oriented horizontally. After capturing the Bayer8 image frames, they are converted into RGB

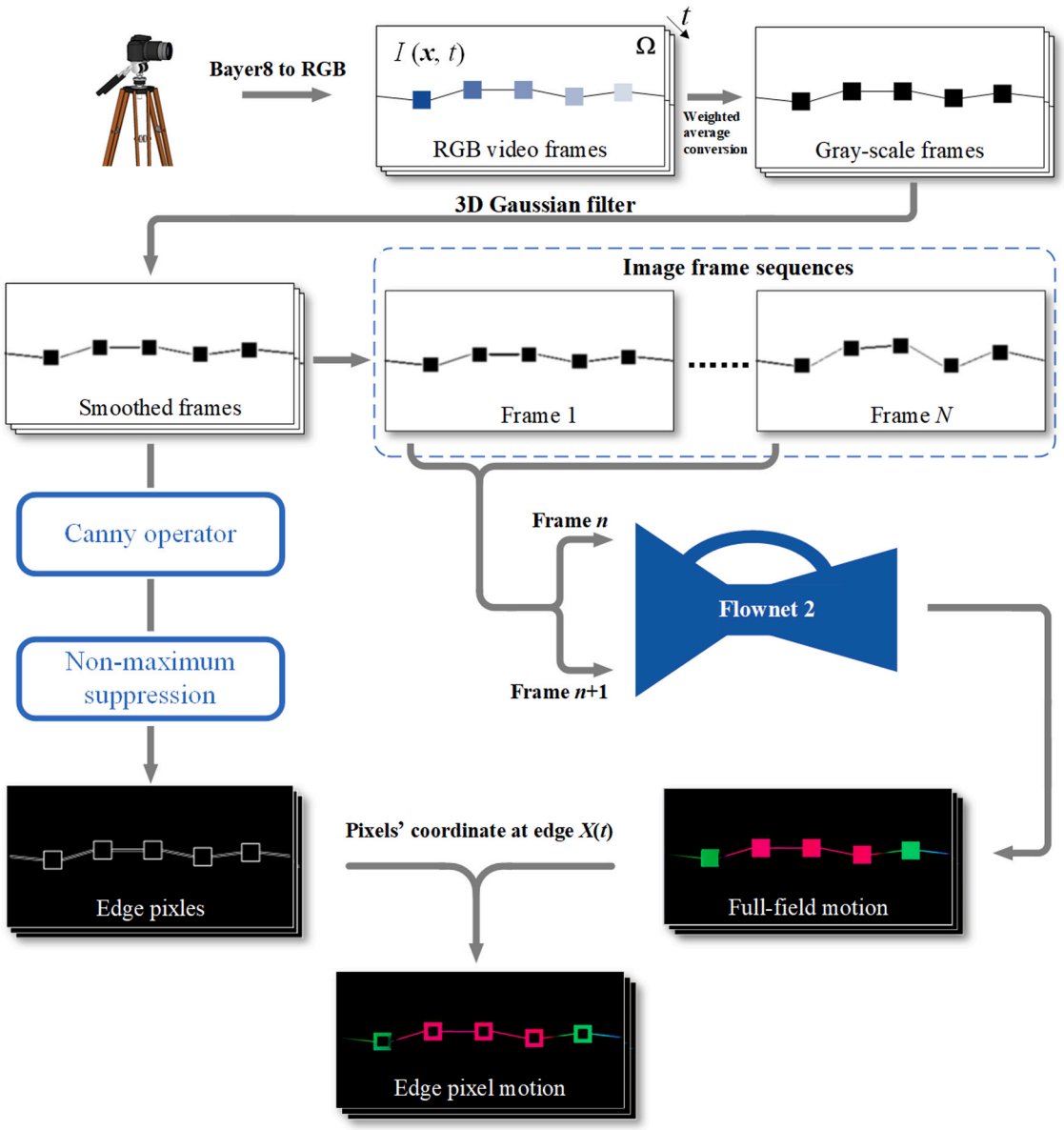


Fig. 1. Flowchart of the proposed vision-based dynamic displacement measurement system (illustrated in the figure with a five-DOF string-mass structure as an example).

format using differential conversion and then transformed into grayscale frames with a single-channel through weighted averaging, in the following manner:

$$\boldsymbol{Grey}(\boldsymbol{x},t) = [\boldsymbol{R}(\mathbf{x},t) \times 306 + \mathbf{G}(\mathbf{x},t) \times 601 + \mathbf{B}(\mathbf{x},t) \times 117] \gg 10 \tag{1}$$

The intensity of a pixel at a given time is denoted by $\boldsymbol{Grey}(\boldsymbol{x}, t)$. The pixel's location $\boldsymbol{x}$ is represented by $[x_1; x_2] \in \boldsymbol{\Omega}$, which falls within the entire image area denoted by symbol $\boldsymbol{\Omega}$. The RGB image's red, green, and blue channel values are respectively represented by $\boldsymbol{R}(\boldsymbol{x}, t)$, $\boldsymbol{G}(\boldsymbol{x}, t)$ and $\boldsymbol{B}(\boldsymbol{x}, t)$. The precision required is 10-bit, indicated by the notation "> >10".

The regions of the image where the intensity changes sharply in a specific direction, known as edge pixels, are identifiable via the derivative peak of the image's intensity field $\boldsymbol{Grey}(\boldsymbol{x}, t)$. However, because of the noise present in the raw video, it is imperative to initially convolve $\boldsymbol{Grey}(\boldsymbol{x}, t)$ with a Gaussian filter in time and space. Thus, the noise can be reduced prior to motion detection of the edge pixels [42]:

$$\mathbf{S}(\mathbf{x},t) = G(\mathbf{x},t) \otimes \mathbf{Grey}(\mathbf{x},t) \tag{2}$$

where $\boldsymbol{S}(\boldsymbol{x},t)$ denotes the smoothed image intensity field and $G(\boldsymbol{x},t)$ represents a 3D Gaussian kernel [43]:

$$G(\mathbf{x},t) = u\left(\frac{\sigma}{u}\right)^{3/2} e^{-\sigma\left(x^2+y^2+u^2t^2\right)} \tag{3}$$

where u is the scale velocity and σ is the standard deviation.

The image's high-frequency noise can be removed adeptly with Gaussian filtering. Following that, the gradient vector ΔG of the 3D Gaussian function $G(\boldsymbol{x},t)$ on $\boldsymbol{S}(\boldsymbol{x}, t)$ can be determined by taking the finite difference of first-order partial derivatives:

$$\Delta G = \begin{bmatrix} \frac{\partial G(\mathbf{x},t)}{\partial x} \\ \frac{\partial G(\mathbf{x},t)}{\partial y} \end{bmatrix} \tag{4}$$

Due to the separability of the Gaussian filter, ΔG can be decomposed into row and column filters, respectively:

Table 1
Technical specifications of the proposed HVMS.

Component	Model	Technical Specifications
Video camera		Maximum resolution: 4096 × 3072 Maximum frame rate: 409 fps Chroma: Bayer8 Pixel size: 3.2 μm× 3.2 μm Lens mount: C-mount
	MindVision/MV-XG1205GC/M	Sensitivity: 4050 mV 1/30 s
Optical lens		Focal length: 8 to 90 mm Aperture: F1.6~F22 Mount: C-mount
	MindVision/MV-LD-8-5 M-C	
Computer workstation		Intel(R) Core (TM) i9-11900 H CPU @ 4.9 GHz NVIDIA GeForce RTX 3050 Ti 4 GB GDDR6
	Lenovo/ThinkStation P350	
Network Card		Type of network: 10 Gigabit Ethernet Interface speed: 10 G Frame cache: 4 G byte
	MindVision/MV-CPE40	Driver support: Windows 7/ Windows 10/Ubuntu

$$\frac{\partial G}{\partial x} = -2\sigma u\left(\frac{\sigma}{u}\right)^{3/2} e^{-\sigma\left(x^2+y^2+u^2t^2\right)} = n_1(x)n_2(x) \tag{5}$$

$$\frac{\partial G}{\partial y} = -2\sigma u\left(\frac{\sigma}{u}\right)^{3/2} e^{-\sigma\left(x^2+y^2+u^2t^2\right)} = n_1(y)n_2(y) \tag{6}$$

Then convolve Eqs. (5) and (6) with $\boldsymbol{S}(\boldsymbol{x}, t)$, respectively:

$$H_x = \frac{\partial G}{\partial x} \otimes \mathbf{S}(x, y, t) \tag{7}$$

$$H_y = \frac{\partial G}{\partial y} \otimes \mathbf{S}(x, y, t) \tag{8}$$

Consequently, the gradient values $A(i, j)$ and the orientation angle $\partial(i, j)$ of pixel (i, j) on image $\boldsymbol{\Omega}$ can be determined as:

$$\begin{cases} A(i,j) = \sqrt{H_x^2(i,j) + H_y^2(i,j)} \\ \partial(i,j) = \arctan\left|\dfrac{H_y(i,j)}{H_x}\right| \end{cases} \tag{9}$$

Each element in matrix $A(i, j)$ corresponds to the gradient value of a specific pixel in the image. However, the value alone is insufficient to ascertain whether the pixel qualifies as an edge pixel. Consequently, to remove non-edge pixels reliably and preserve the candidate ones, a non-maximum suppression along the gradient direction is required. Specifically, the example shown in Fig. 2 depicts a 3 × 3 pixel window. The red line in the figure indicates the gradient direction $\partial(i, j)$ of the central pixel $O(i, j)$. To determine if point O is an edge pixel, the first step is to compare the gradient values of point O and its eight neighboring pixels. If the gradient value of point O is the largest, it can be concluded that the local maximum of the region is located on the red line. In addition to point O, the gradient values of intersection points C_1 and C_2 may also be local maxima. Next, the gradient value of point O is compared with the gradient values of points C_1 and C_2. If point O's gradient value is greater than either of these two intersection points, it becomes a candidate edge pixel. Conversely, if point O's gradient value is lower than either of these two intersection points, it is excluded from the edge pixel.

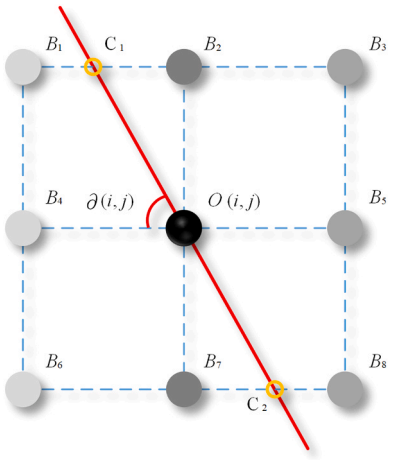


Fig. 2. A 3 × 3 pixels window with O as the central pixel, B_1-B_8 as the neighbor pixels, and the red line as the gradient direction.

After non-maximum suppression, there are still spurious edge pixels among the candidate pixels, which can be removed using high and low thresholds: 1. If the gradient value of a pixel surpasses the high threshold, it is considered as an edge pixel; 2. If the gradient value of a pixel is below the low threshold, it is excluded; 3. If the gradient value falls between the two thresholds, the pixel is retained only when it is adjacent to another pixel with a value above the high threshold. The ratio of high-to-low threshold varies between 2:1 and 3:1, depending on the specific Canny operator used [44]. Following the aforementioned edge detection procedures, edge pixels $\boldsymbol{x}^{\text{edge}}$ within the region $\boldsymbol{\Omega}$ can be chosen as:

$$\mathbf{x}^{\text{edge}} \in \mathbb{R}^{\text{edge}}, \quad \forall t \tag{10}$$

To obtain the full-field motion of all image frames, a pre-trained end-to-end deep learning network, known as Flownet2 [45], is employed. As depicted in Fig. 3, the input of the model are two successive grayscale image frames $\boldsymbol{S}(\boldsymbol{x}, t)$ and $\boldsymbol{S}(\boldsymbol{x}, t + \partial t)$, and the output is the displacement field $D(\boldsymbol{x}, t)$ of all pixels $\boldsymbol{x}$ in the image at time t after a time interval of $\partial t = 1/f$ (where f is the camera frame rate). Consequently, the motion data $\boldsymbol{d}(\boldsymbol{x}^{\text{edge}}, t)$ constrained to the edge pixels are collected as:

$$\mathbf{d}\left(\mathbf{x}^{\text{edge}}, t\right) = \left\{D(\mathbf{x}, t), x^{\text{edge}} \in \boldsymbol{\Omega}\right\} \in \mathbb{R}^{\text{edge}}, \forall t \tag{11}$$

The implementation and pre-trained models of Flownet are detailed in the research of Ilg et al. [46].

3. Vision modal analysis by 1D-CNN-LSTM

In this section, a 1D-CNN-LSTM architecture is constructed to identify the modal parameters of the structure from the vibration signals of the collected edge pixels, as shown in Fig. 4. CNN has the ability to extract features from signals, while LSTM can fully explore the spatio-temporal dependencies from sequential data. In the following subsections, the network architecture, dataset collection, and coordinate recovery will be presented in detail.

3.1. 1D-CNN

In the 1D-CNN module, it is apparent that every feature value in the maximum pooling layer is the highest value amongst the set of neurons detected in the previous convolutional layer, as demonstrated in Fig. 5. The fundamental composition of the internal layer is shared within all CNN models and includes the input layer, the convolutional layer, the maximum pooling layer, and the flattened layer.

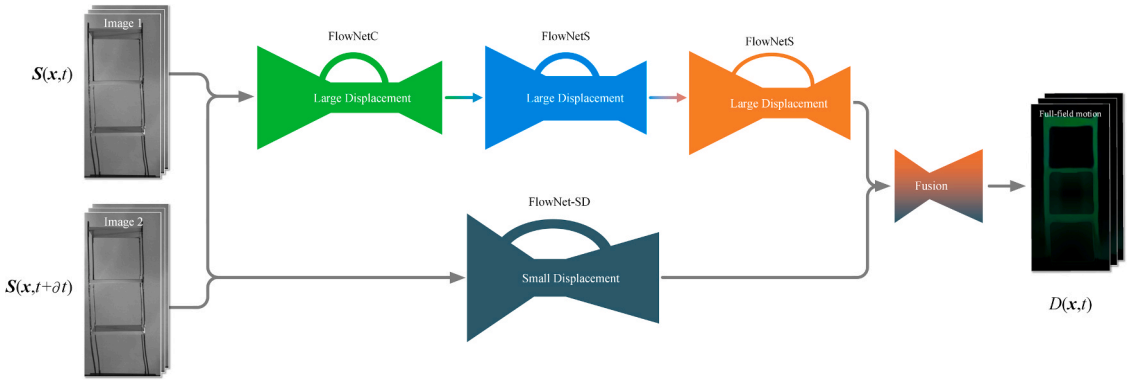


Fig. 3. Schematic view of complete Flownet2 architecture [45].

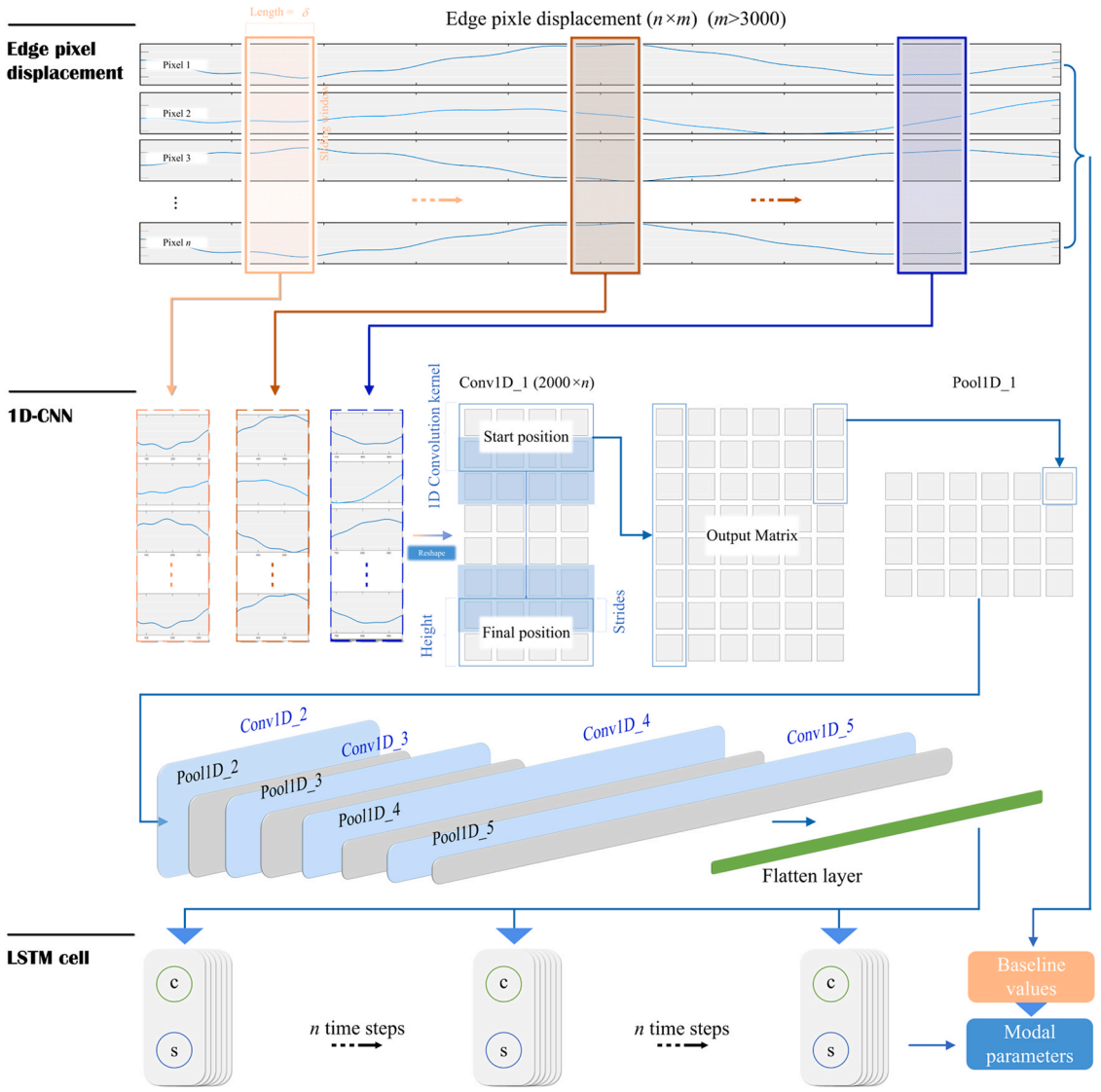


Fig. 4. The 1D-CNN-LSTM architecture.

The algorithm for a 1D convolutional layer performs two operations on the input array, as illustrated in Fig. 6. The input sub-array is multiplied by the kernel element-wise. The resulting products and bias values are then entered into the activation function, producing the output values. 1D Convolutional kernels share the same width as the input array. To clarify the convolutional layer's presentation, integer values are displayed in Fig. 6. However, the parameters in actual CNN model convolutional layers consist of real values.

The 1D max pooling layer uses a 1D pooling operation to extract the highest output from a specified neighborhood along the time series direction. This neighborhood is determined by a sliding window that moves along the time dimension. An example of 1D max pooling is

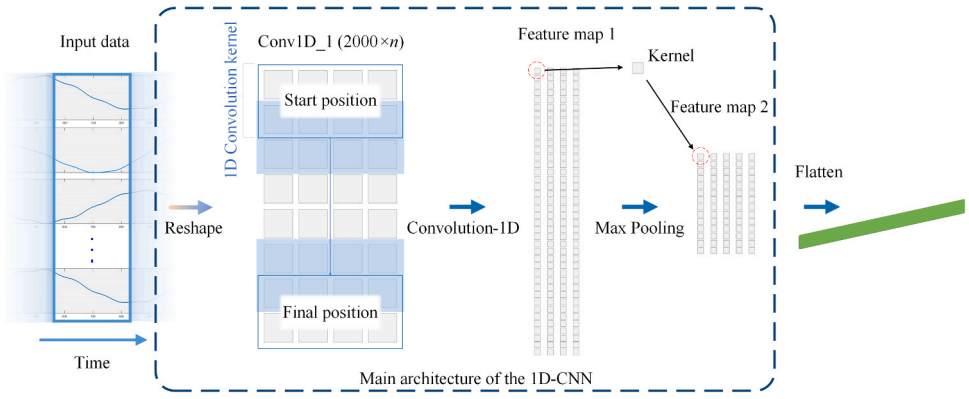


Fig. 5. Architecture of the 1D-CNN.

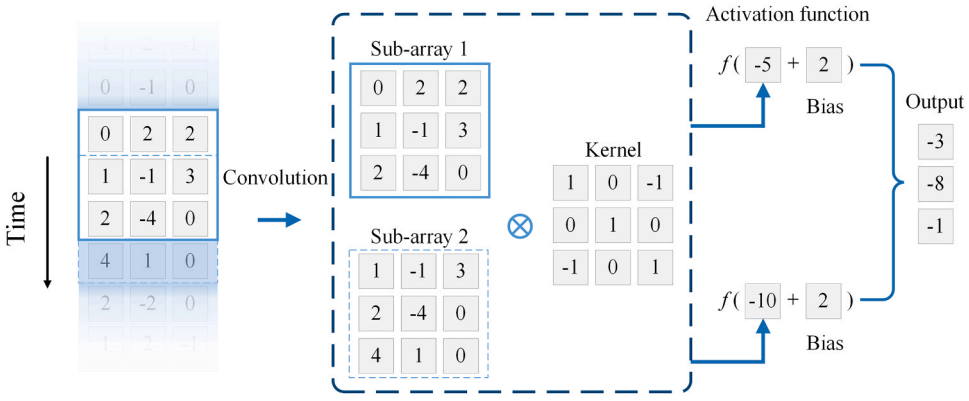


Fig. 6. Demonstration of a 1D convolutional layer.

depicted in Fig. 7. Following pooling, the data representation is essentially unaltered by small changes in input data, indicating that while data size is reduced, the data features remain consistent. 1D max-pooling is a valuable down-sampling technique to enhance the statistical efficiency and computational speed of a neural network. Subsequently, a flattened layer follows the last max pooling layer, the output of which can be used as the input to the next LSTM cell.

3.2. LSTM

The architecture of the LSTM cell is illustrated in Fig. 8. Based on the original short-term memory s_t, a memory c_t indicating the current state of the cell is adopted for the storage of long-term memory. It is shown that in each time step, the LSTM cell receives three inputs: the current moment's input x_t, the state c_{t-1}, and the last moment's output s_{t-1}, and is updated as follows:

$$f_t = \sigma\left(W_{xf}x_t + W_{sf}s_{t-1} + b_f\right) \tag{12}$$

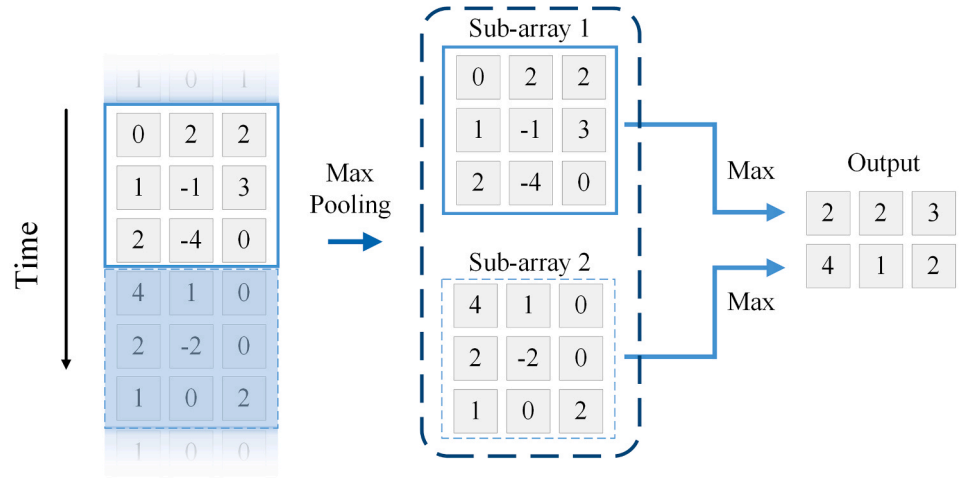


Fig. 7. Demonstration of a 1D max pooling layer.

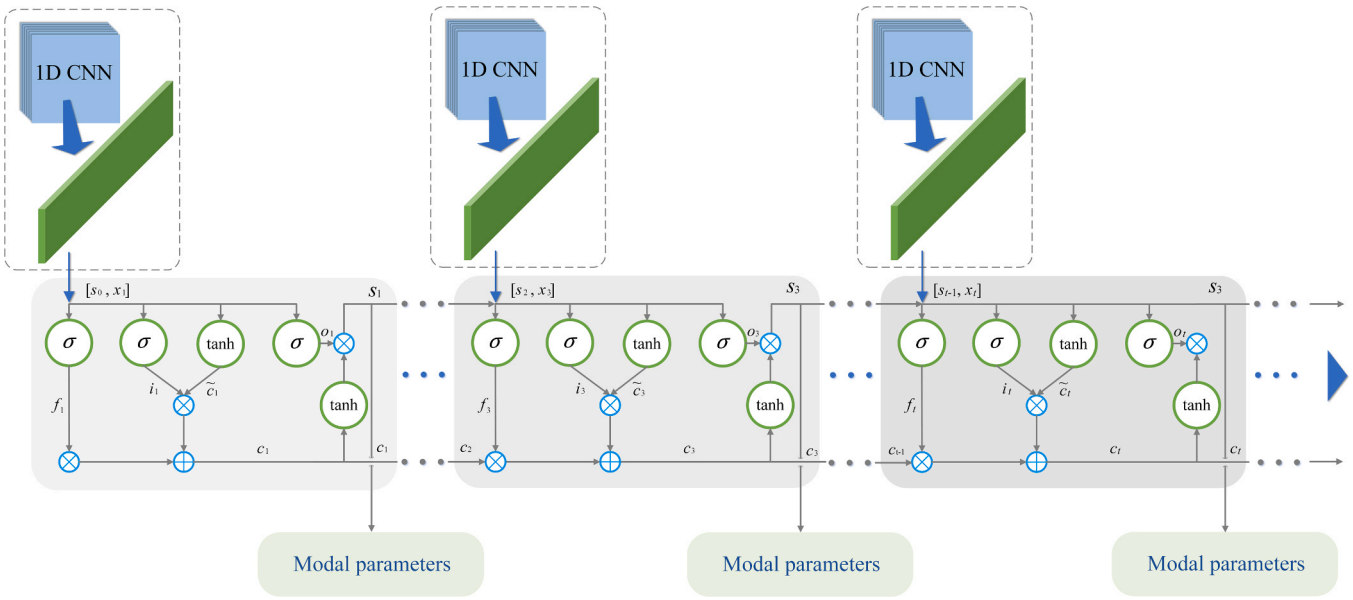


Fig. 8. The LSTM cell's architecture.

$$i_t = \sigma(W_{xi}x_t + W_{si}s_{t-1} + b_i) \quad (13)$$

$$c_t^* = \tanh(W_{xc}x_t + W_{sc}s_{t-1} + b_c) \quad (14)$$

$$o_t = \sigma(W_{xo}x_t + W_{so}s_{t-1} + b_o) \quad (15)$$

$$c_t = f_t \otimes c_{t-1} + i_t \otimes c_t^* \quad (16)$$

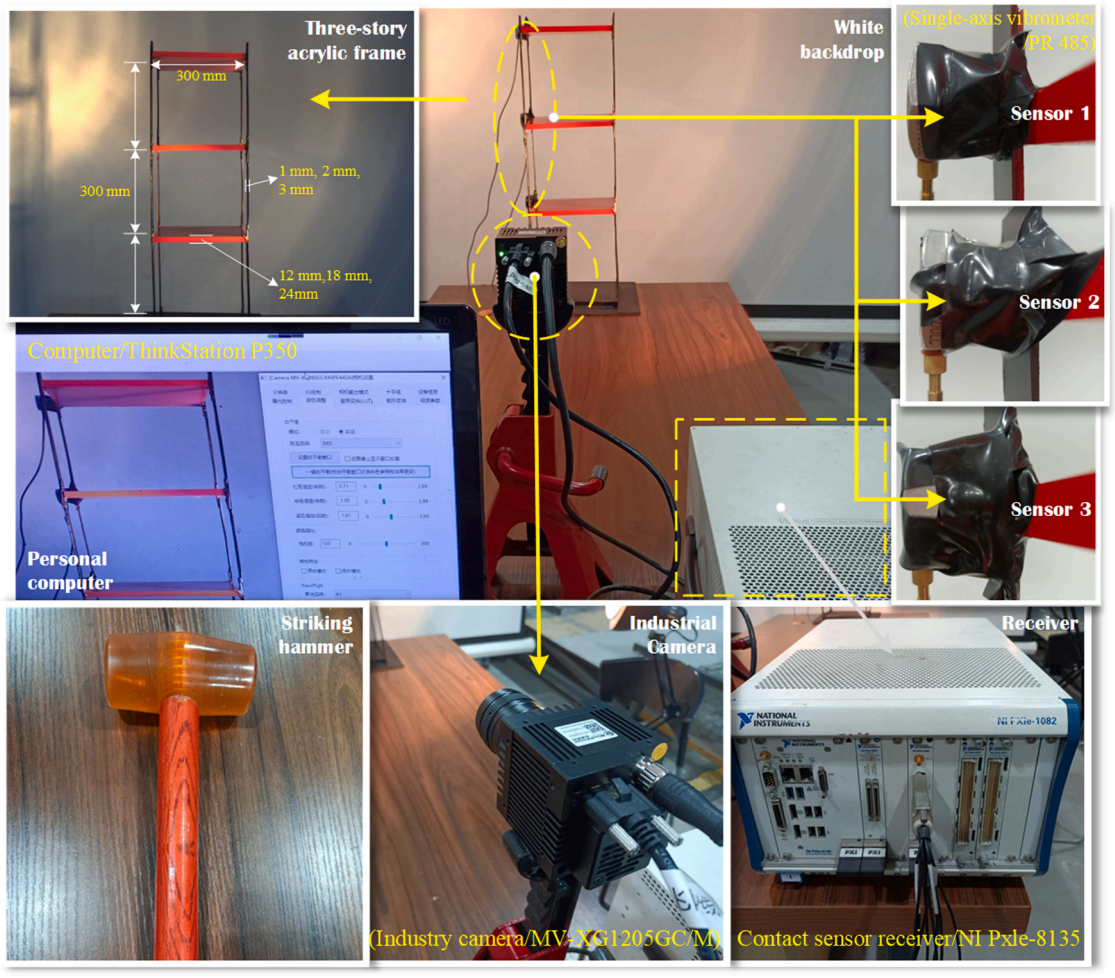


Fig. 9. Vibration data collection setup: a three-story acrylic frame; white backdrop; contact sensors; hammer; industrial camera; and receiver.

$$s_t = o_t \otimes \tanh(c_t) \tag{17}$$

where the subscript t is the timestep; W_{xf}, W_{sf}, W_{xi}, W_{si}, W_{xc}, and W_{sc} indicates the weighting factor that is used to enhance the potential cell status; b_f, b_i, b_c, b_o represent the bias matrix; $\boldsymbol{\sigma}$ represents a sigmoid function; $\otimes$ denotes the convolution operator; i_t, f_t and o_t are input gate, forget gate and output gate, respectively. At any timestep t, the forget gate controls the extent to which memories are forgotten, while the input gate controls how much new information is written into the long-term memory.

Throughout the process, various features can be extracted from the input data depending on different model parameters. To calculate the natural frequencies f_i, damping ratios ξ_i, and mode shapes $\boldsymbol{\phi}_i$ of the structure, three 1D-CNN-LSTM architectures are trained separately. The inputs are the displacements of each edge pixel determined in Section 2, while the outputs are the modal parameters. For engineering structure, the vast majority of dynamic characteristics are often contained in its first three orders of modes. Therefore, in this study, the proposed method is illustrated with a 3-DOFs structure, i.e., only the first three orders of natural frequencies, damping ratios, and mode shapes are considered.

3.3. Modal Parameter dataset collection

Laboratory experiments are conducted on a three-story acrylic frame to collect the ModalParameter dataset, which consists of the pixel displacements of selected edge pixels as input and corresponding modal parameters as output. Fig. 9 depicts the whole experimental setup, which includes a three-story acrylic frame, a personal computer, a white backdrop, an industrial camera, vibrometers, a receiver, and a hammer. To enrich the dataset, the three-story acrylic frame is assembled from components of different sizes, as shown in Fig. 10. Depending on the column size and the floor mass, there are a total of 81 combinations, whose specific values are listed in Table 2. The frame is anchored to the foundation and excited at the third story with a hammer. For the robustness of the model, it is not necessary to keep the magnitude of the strike force constant, but it is necessary to ensure that the amplitude is detectable. At each story of the frame, a vibrometer is attached to capture the horizontal displacement for comparison with the HVMS measurements. Note that during vibration data acquisition, the camera should simultaneously capture vibration signals as the contact vibrometer is installed. If the camera is used to capture vibration signals as a training and test data set when the vibrometer is not installed, subsequent signals collected after the installation of the vibrometer will be unable to function as a benchmark due to deviations in the vibration characteristics before and after vibrometer installation.

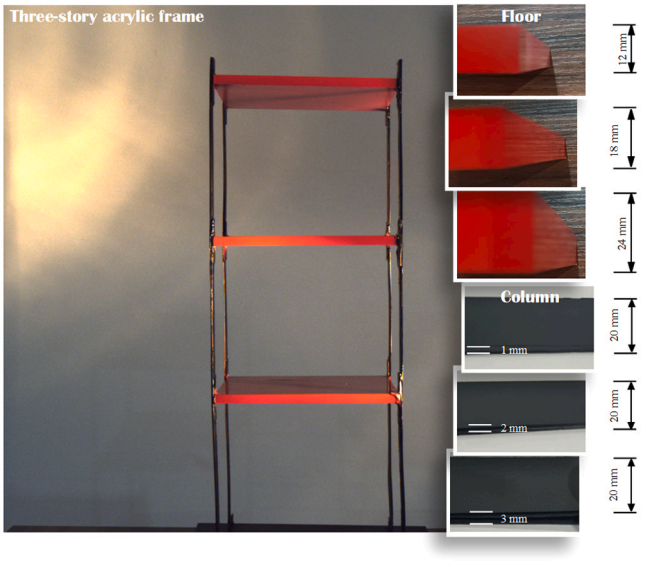


Fig. 10. Different component sizes used for data collection.

The HVMS is employed to capture edge pixel displacements of the vibrating structure at 300 fps. For each of the 81 specimen structures, 5 different videos and the corresponding natural frequencies, damping ratios, and mode shapes are recorded. Inevitably, the video background and illumination conditions produce different noises in the five separate videos of the same structure. Each video lasts for 1 min, and only the portion that contains perceivable vibrations longer than 10 s is of interest. For each structure, 4 videos are the training dataset and the remaining one is the validation dataset. A total of 405 videos, each lasting for 10 s, are gathered, where 324 videos are for training and 81 videos are for hyperparameter tuning. It should be noted that the network input is a pre-extracted 1D vibrational signal of the edge pixels, with a size of 10×300, rather than the $10 \times 1920 \times 1080$ pixel raw images. The test set is collected from an unseen new structure, details of which are given in Section 6.5. The number of train, validation, and test dataset samples is further augmented by using a sliding window of length δ to move over the raw signal, thus obtaining adequate data for better model training. The effect of the signal length δ on the model accuracy is analyzed in Section 3.

For the convenience of using video directly as the input source for Flownet, a module named Flownet2_easy_to_handle is provided and located together with the database in [47].

3.4. Network training

1D-CNN-LSTM learns to recognize structural modalities through two alternating steps: (1) training and (2) validation. During training, the network weight assignments are updated batch by batch by evaluating and reducing the deviations between the predicted (output) and actual (labeled) values of local structural modal parameters. The consistency and deviation between the labels and the network outputs are called accuracy and loss, respectively. During validation, for a more efficient and scientific training of the model and a better use of the dataset, the current training status should be evaluated by observing the loss and accuracy of the training process and promptly adjusting the hyperparameters. This study considers a range of hyperparameters including signal length for training, deep learning model architecture (such as number of layers, layer size, and filter size), learning rate, batch size, and overfitting reduction methods. The length of the training signal has the greatest impact, as outlined in Section 4.2, in obtaining optimal values. This section addresses the tuning of the ideal 1D-CNN-LSTM architecture, as well as the learning rate and batch size.

3.4.1. Determination of the optimal 1D-CNN-LSTM architecture

As shown in Table A1 in the Appendix, three different architectures of 1D-CNN-LSTM are designed to process the collected 1D edge pixel displacement signals to determine the optimal architecture. During the training process, the same hyperparameters (loss function: MSE, optimizer: adaptive moment estimation (Adam), batch size $= 8$, epoch $= 1800$) are used for the three designed networks and their mean relative error (MRE) and training times are compared as shown in Table 3. It is shown that the medium 1D-CNN-LSTM can better balance the identification accuracy and computational efficiency, and thus this network is employed in the following analysis.

In terms of training time, in the study by Yang et al. [48], it took 3.5 h for the training of 100×100 pixels RGB images and 8.5 h for 128×128 pixels RGB images (with 306 training samples in the training dataset). The images collected in the present study are 1920×1080 pixels (324 training videos in the training dataset), and the 2D neural network simply cannot sustain the training. Therefore, in this study, a 1D-CNN-LSTM network is developed by training with 1D vibrational signals extracted from the videos. It can be seen from Table 3 that the

Table 2
Parameters and dimensions of 81 specimen structures.

Column		1 mm × 20 mm			2 mm × 20 mm			3 mm × 20 mm		
Story mass		1st story/kg			1st story/kg			1st story/kg		
2nd story/kg	3rd story/kg	0.5	0.75	1.0	0.5	0.75	1.0	0.5	0.75	1.0
0.5	0.5	#1	#2	#3	#4	#5	#6	#7	#8	#9
	0.75	#10	#11	#12	#13	#14	#15	#16	#17	#18
	1.0	#19	#20	#21	#22	#23	#24	#25	#26	#27
0.75	0.5	#28	#29	#30	#31	#32	#33	#34	#35	#36
	0.75	#37	#38	#39	#40	#41	#42	#43	#44	#45
	1.0	#46	#47	#48	#49	#50	#51	#52	#53	#54
1.0	0.5	#55	#56	#57	#58	#59	#60	#61	#62	#63
	0.75	#64	#65	#66	#67	#68	#69	#70	#71	#72
	1.0	#73	#74	#75	#76	#77	#78	#79	#80	#81

Table 3
The MRE and training time of the three designed 1D-CNN-LSTMs for modal parameters identification.

Designed model	MRE (%)	Training time (s/epoch)
Shadow	2.23	1.23
Medium	0.73	1.77
Deep	1.32	2.56

training speed of a medium 1D-CNN-LSTM can reach 1.77 *s*/epoch. The training loss is stable around 800 epochs, thus the training time is about 0.4 h. The reason for the significantly higher training efficiency is that the convolutional kernel is 1D, and the size of the 1D signal is obviously smaller than that of the 2D signal. Consequently, not only the training time is significantly reduced, but also the speed of testing new signals is faster compared to the 2D networks.

3.4.2. Determination of the optimal hyperparameters

The loss function, optimizer, and batch size of the network can also have a significant impact on the performance of 1D-CNN-LSTM. In this study, 24 combinations of the above three hyperparameters are considered, as shown in Table A2 in Appendix. In particular, the variation of training loss on epoch for the case with loss function of MSE, optimizer of Adam, and batch size of 8 is presented in Fig. 11. From the figure, it can be seen that the training loss decreases dramatically at the beginning and gradually stabilizes when epoch reaches about 800. Therefore, the epoch is considered to be 800 in the present study. In this section, the mean relative errors (MRE) of the structural natural frequencies (24 different combinations in Table A2) identified by 1D-CNN-LSTM are also provided, as shown in Fig. 12. It can be seen that the minimum MRE for the first order is 1.54% (case 4), the second order is 2.75% (case 4), and the third order is 3.62% (case 18). Therefore, case 4 is chosen as the optimal model and its loss function, optimizer, and batch size of MSE, Adam, and 64, respectively.

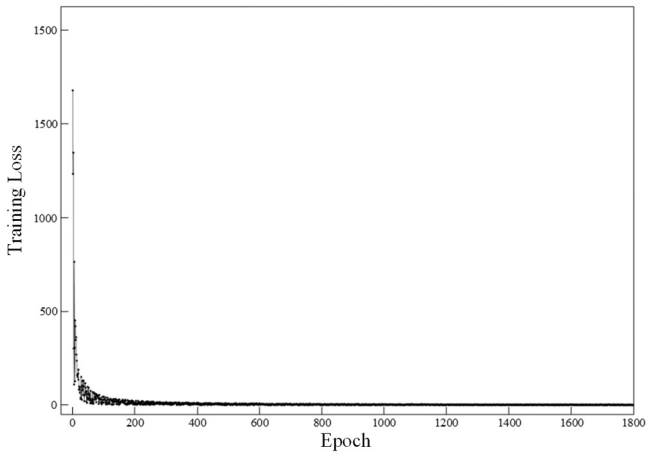


Fig. 11. The variation of the training loss on the epochs with loss function of MSE, optimizer of Adam, and batch size of 8.

Furthermore, the length of the training signal, the Gaussian smoothing parameter, and the structural amplitude are additional hyperparameters that can have an effect on the accuracy and automation of the extracted 1D signals. Thus, these hyperparameters will be further analyzed in the subsequent section. Following the determination and appropriate adjustment of these hyperparameters, the proposed framework will be able to operate automatically.

3.5. Structural mode shape recovery from pixel coordinate

In structural damage detection [49] and health monitoring [50], only the natural frequencies and damping ratios of the structure are usually required. Therefore, the image processing procedure of recovering the true displacement from the pixel coordinate is no longer needed. Unless the real mode shapes are required, the proposed 1D-CNN-LSTM framework-based visual modal analysis would be rather simple.

The real mode shapes $\boldsymbol{\phi}_i(\boldsymbol{X})$ have the following functional relationship with the visual mode shapes φ_i in pixel coordinate [51]:

$$\nabla \mathbf{Grey}_0(\mathbf{x})\mathbf{R}\boldsymbol{\phi}_i(\mathbf{X}) = -\varphi_i(\mathbf{x}) \qquad \forall i \in [1,2,\ldots,n] \tag{18}$$

where $\boldsymbol{Grey}_0(\boldsymbol{x})$ is the image intensity of the initial structure when the displacement value is 0, $\nabla\, \boldsymbol{Grey}_0(\boldsymbol{x}) = [\partial \boldsymbol{Grey}_0(\boldsymbol{x})/\partial x_1, \partial \boldsymbol{Grey}_0(\boldsymbol{x})/\partial x_2]$, and $\mathbf{R}$ is the transform matrix. For a given position x_0, $\varphi_i(x_0)$ is a scalar, $\mathbf{R}\boldsymbol{\phi}_i(\boldsymbol{X})$ is a vector, and several non-unique solutions can be derived from Eq. (18). Therefore, it is assumed that the transformed image mode shapes $\mathbf{R}\boldsymbol{\phi}_i(\boldsymbol{X})$ remain consistent near x_0, denoted by $V(x_0)$, which can then be recovered at x_0 using the least squares method:

$$\begin{aligned}&\mathbf{R}\boldsymbol{\phi}_i(\mathbf{X})\\ &= \arg\min_{\mathbf{z}\in R} \sum_{\mathbf{x}\in V(x_0)\subset\boldsymbol{\Omega}} \left(\varphi_i(\mathbf{x}) + \nabla \mathbf{Grey}_0(\mathbf{x})\mathbf{z}\right)^2\\ &= -\left[\sum_{\mathbf{x}\in V(x_0)} \left(\nabla \mathbf{Grey}_0(\mathbf{x})\right)^{\mathrm{T}} \nabla \mathbf{Grey}_0(\mathbf{x})\right]^{-1} \left[\sum_{\mathbf{x}\in V(x_0)} \left(\nabla \mathbf{Grey}_0(\mathbf{x})\right)^{\mathrm{T}} \varphi_i(\mathbf{x})\right]\end{aligned} \tag{19}$$

where, for in-plane motion, matrix $\mathbf{R}$ is invertible and the structural mode shape $\boldsymbol{\phi}_i(\boldsymbol{X})$ can be inversely derived from Eq. (19). That is, when the optical axis of the camera is perpendicular to the structural motion plane, $\mathbf{R}\boldsymbol{\phi}_i(\boldsymbol{X})$ is exactly the real mode shape of the structure. Otherwise, the transformation matrix $\mathbf{R}$ is irreversible and visual measurements from different perspectives are required to jointly recover $\boldsymbol{\phi}_i(\boldsymbol{X})$ [52].

4. Method evaluation

To evaluate and compare the performance of the proposed method, the effects of Gaussian smoothing parameters, displacement amplitude,

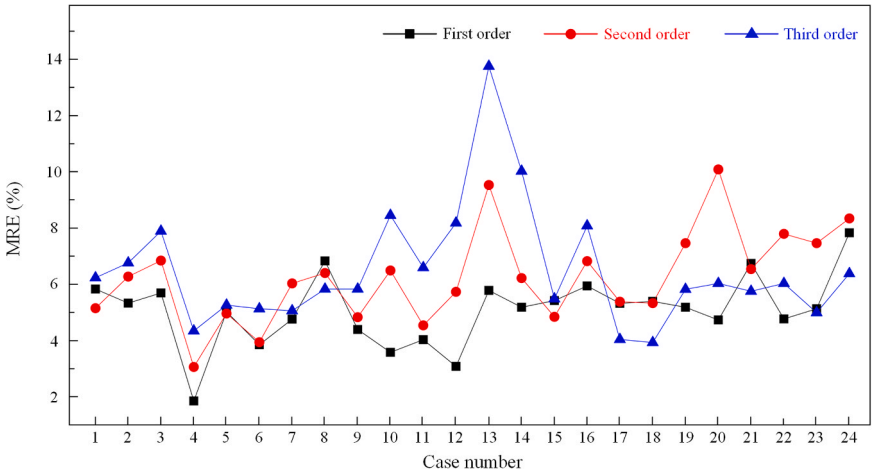


Fig. 12. MREs obtained with 24 different hyperparameters.

training signal length, robustness, and extrapolability are considered. In the present section, these parameters are presented separately, and the calculation results and analysis will be given in detail in Section 6.

In this study, the accuracy of the identified natural frequencies and damping ratios are evaluated by mean squared error (MSE), and the consistency between the identified structural mode shapes and the baseline is quantified by the modal assurance criteria (MAC).

$$\mathrm{MSE}\left(f^{\mathrm{id}},f\right)=\frac{\sum_{i-1}^{n}\left(f_i^{\mathrm{id}}-f_i\right)^2}{n} \tag{20}$$

$$\mathrm{MAC}\left(\boldsymbol{\phi}_i^{\mathrm{id}}(\mathbf{X}),\boldsymbol{\phi}_i(\mathbf{X})\right)=\frac{\left|\boldsymbol{\phi}_i^{\mathrm{T}}(\mathbf{X})\cdot\boldsymbol{\phi}_i^{\mathrm{id}}(\mathbf{X})\right|}{\|\boldsymbol{\phi}_i(\mathbf{X})\|\cdot\left\|\boldsymbol{\phi}_i^{\mathrm{id}}(\mathbf{X})\right\|}\quad(i=1,2,\ldots,n) \tag{21}$$

where the baselines are determined from the vibration signals collected by the contact vibrometer using the state variable method (SVM) [53].

SVM is an algorithm that can accurately identify the modal parameters of a structure by solving a dynamic inverse problem, therefore in this study, the results of SVM are used as a baseline for comparison with the results of the proposed method. Taking structure #81 as an instance, during the construction of a Hankel matrix in SVM, the matrix is comprised of 200 rows and 1000 columns, with a specified order of 18 (the structure has only 3 possible orders, and the remaining orders can provide outlets for noise modes). Modal assurance criteria have been established at a 1% error rate for natural frequency, a 5% error rate for damping ratio, and a mode shape assurance criterion of 98% to create a stability diagram, as exhibited in Fig. 13. The figure presents a superimposed power spectral density curve of the vibration response signals collected by the three contact sensors. "○" symbolizes the stable point, while "× " represents the unstable point. The stable points' concentration increases with mode order, which aligns with the power spectral density curve's peak position. The natural frequencies, damping ratios, and mode shapes, estimated by the SVM method, are presented in Table 4, serving as a baseline for subsequent work.

To verify the precision of the proposed method, it is necessary to compare the results obtained with the benchmark. The method comprises of two parts, namely, edge pixel displacement extraction and

Table 4
The natural frequencies, damping ratios, and mode shapes of structure #81 estimated by SVM.

Mode	Modal parameters		
	Nature frequencies (Hz)	Damping ratio	Mode shapes
1	1.136	0.459%	(−0.037, −0.064, −0.083)
2	2.992	0.106%	(0.081, 0.032, −0.079)
3	4.312	0.150%	(0.061, −0.095, 0.019)

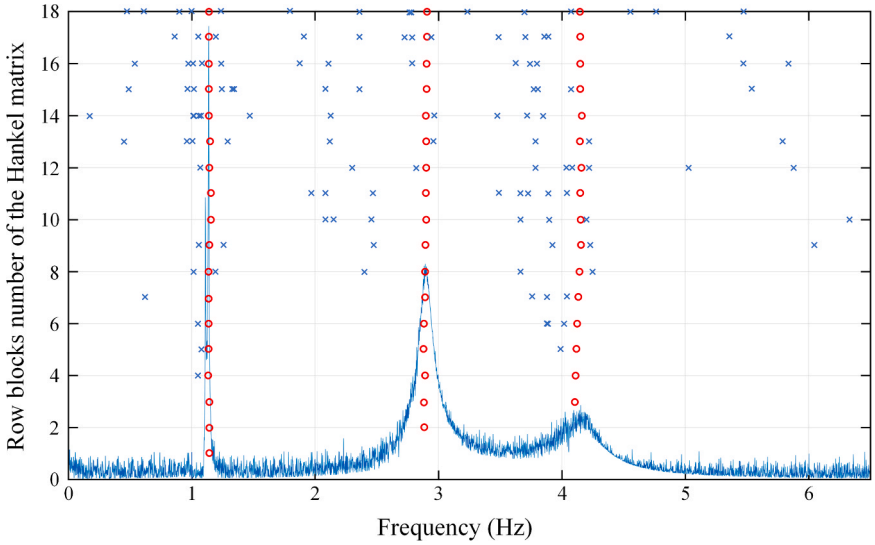


Fig. 13. Stability diagram obtained by SVM method.

modal identification, and both of these are validated in the current study. The edge pixel displacements can be validated using the benchmark, which is the horizontal vibration signals recorded by the contact sensors. Owing to the scanty amount of contact sensors and their limited spatial resolution, a comparison is made by selecting the edge pixels at their respective locations, illustrated in Fig. 14.

Fig. 14 (2) shows the time history of a part of the edge pixels (e.g., the edges circled by the red dashed lines in Fig. 14 (1)) obtained according to the proposed algorithm. As can be seen from the figure, the magnitude of the displacement of each pixel points decreases with time. The bottom-up pixel displacement along the three-story frame has a more pronounced peak at the three mass aggregation points, which coincides with the fixed position of the sensors. A comparison of the displacement time-history curves recorded by the contact sensor with the computer vision technique is shown in Fig. 14(2). The agreement between the curves indicates the accuracy of the proposed edge pixel displacement signal extraction method. The accuracy of the modal parameter identification will be verified in the subsequent sections and extrapolability analyses, and the modal parameters obtained from the SVM, as shown in Table 4, can be used as the true values for the comparison.

4.1. Effect of Gaussian smoothing parameter and displacement amplitude

Prior to implementing the proposed method for conducting structural modal analysis, it is imperative to capture photographs of the structure. Due to the variation in camera resolution and structure amplitude under different working conditions, the effects of the Gaussian smoothing parameter σ and the displacement amplitude D on the performance of the HVMS are first investigated. For illustration, the effects of several displacement amplitudes (D = 1.95, 4.0, 9.5, 11.5 pixels) and the Gaussian smoothing parameter (σ = 0.1, 1, 3, and 10) on the edge-pixel displacements are investigated using structure #81 in Table 2 as an example. It is worth noting that this process requires the use of human experience and judgement. However, the preprocessing process solely influences extraction quality of 1D signals and does not impede the subsequent modal recognition process. Hence, the feature generation procedure and hyperparameters constructed manually in this step will not impact the modal identification network that has already been well trained.

In practical experiments, it is necessary to smooth the raw video, and proper working of Flownet2 requires an appropriate displacement

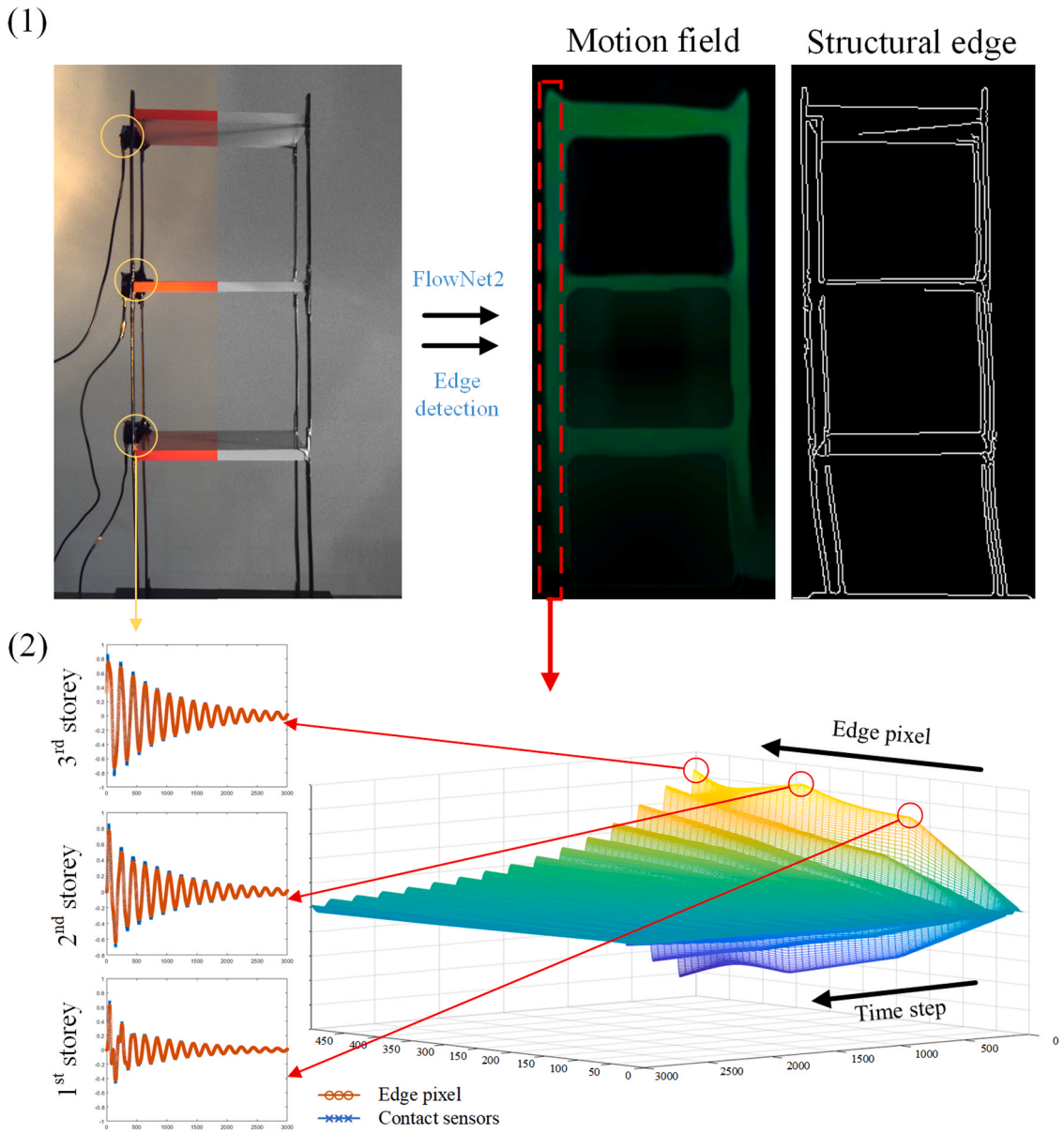


Fig. 14. Edge pixel displacement extraction results and comparison.

amplitude D. Therefore, the effects of the Gaussian smoothing parameter σ and the displacement amplitude D on the performance of the HVMS are first investigated. For illustration, the effects of several displacement amplitudes ($D = 1.95$, 4.0, 9.5, 11.5 pixels) and the Gaussian smoothing parameter ($\sigma = 0.1$, 1, 3, and 10) on the edge-pixel displacements are investigated using structure #81 in Table 2 as an example.

For comparison purposes, an additional vibrometer is installed on each floor to record the horizontal displacement time histories. Fig. 15 shows the FFT spectrum of the third-floor vibrations identified by the HVMS and the vibrometer for different Gaussian smoothing parameters σ and displacement amplitudes D. For $D = 1.95$, 4.0 pixels, the peaks are less pronounced. Thus, it is reasonably deduced that the first three-order modal parameters of the frame cannot be correctly identified when $D < 5$ pixels. Fig. 15 shows that a larger σ should be chosen as D increases. This is because the Gaussian smoothing filter acts as a low-pass filter with an effective spatial frequency ω that is inversely proportional to σ. The amplitude increases while the peaks in the spectrogram decrease when σ is higher. Therefore, as the σ increases, the displacement amplitude also increases. The image depicting $D = 11.5$ pixels illustrates that the edge pixels become more accurate but discontinuous as Gaussian smoothing increases. Nevertheless, this discontinuity does not considerably interfere with frequency domain feature extraction. Each edge pixel equates to a sensor with a much higher spatial sensing resolution than conventional contact sensors, despite the sparse distribution.

For a quantitative analysis of the values of D/pixel and σ, Table 5 compares the MAC of the initial mode shapes identified with different D

Table 5
The MAC values of the first mode shapes identified with different D and σ.

D/pixel	MAC			
	$\sigma = 0.1$	$\sigma = 1$	$\sigma = 3$	$\sigma = 10$
1.95	-	-	-	-
4.0	-	0.5014	-	-
9.5	0.7542	0.8257	0.5841	-
11.5	0.9909	0.9997	0.9912	0.8631

Note: '-' indicates non-recognition, and the underlined '_' indicates the best recognition at the specified displacement amplitude D.

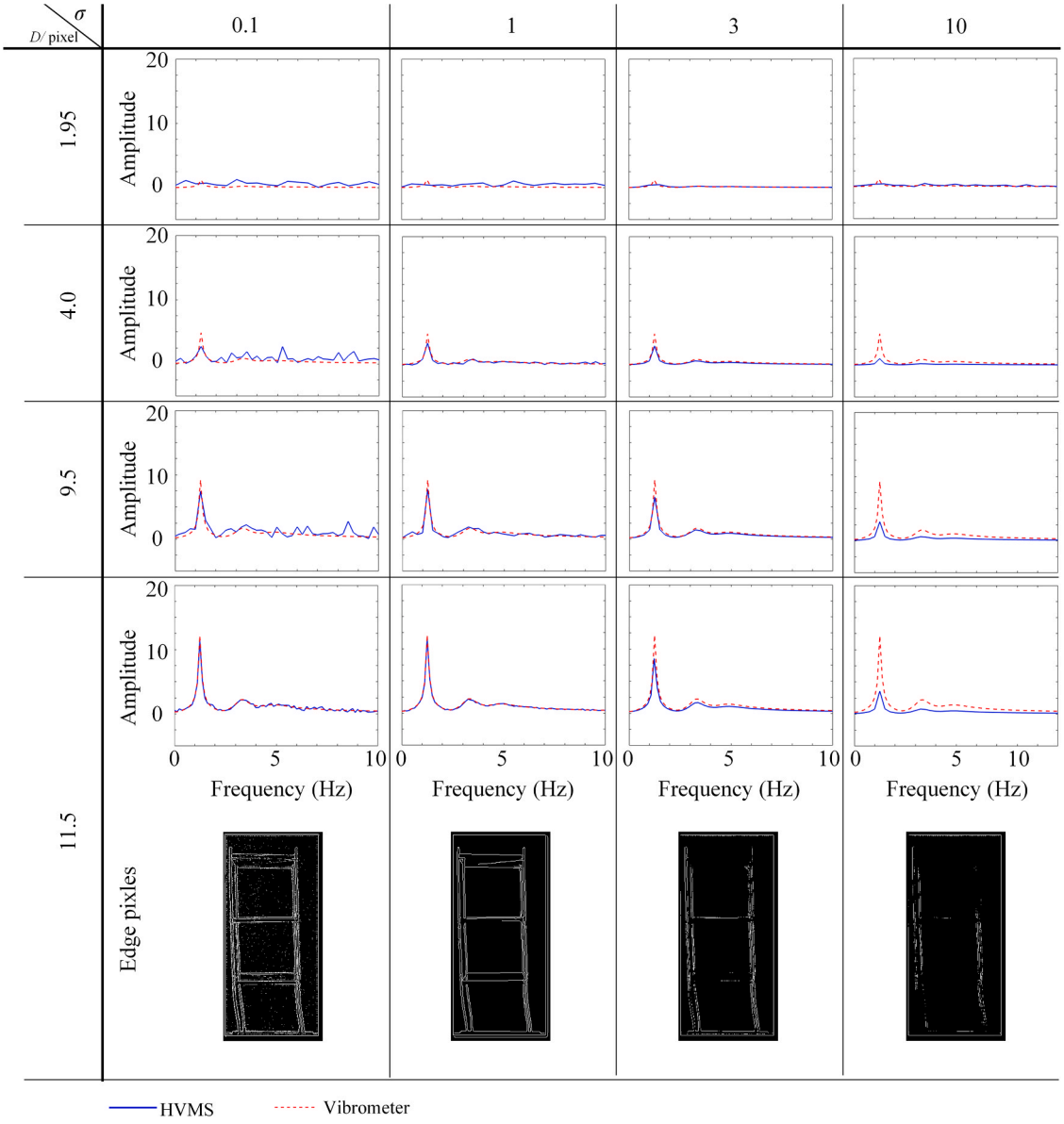


Fig. 15. The FFT spectrum of the 3rd-story vibration identified by HVMS and vibrometers under different Gaussian smoothing parameters σ and displacement amplitudes D.

and σ. The results show that the first mode shape of the structure is almost indistinguishable when D/pixel = 1.95. The best identification with the highest MAC values is achieved when D/pixel is greater than or equal to 4, and $\sigma = 1$. Admittedly, there might be more suitable σ values in the range of 0.1 to 3 since the values of D/pixel and σ are both discontinuous. However, it is readily discernible from Table 5 that with D/pixel = 11.5, σ values of 0.1, 1, and 3 all allow for successful identification of the structure's mode shape (MAC > 0.99). Within the present study, a displacement amplitude D equal to or greater than 11.5 pixels and a smoothing parameter of $\sigma = 1$ were chosen. In actual measurements, variations in shooting distance, camera resolution, and structure scale have an inevitable impact on the correlation between pixel size and structure amplitude in the captured image. Therefore, the appropriate smoothing parameter σ must be set on a case-by-case basis.

4.2. Effect of training signal length

Structured videos are divided into short segments to enrich the dataset. However, reducing the signal length may result in the loss or reduction of the initial features [49]. Consequently, analyzing the impact of various signal lengths on model performance is crucial. Using structure #81 as an example, the training signal length δ increases from 800 to 3000. The prediction frequency of the training model with different δ is shown by the blue line in Fig. 16. As δ increases, the FFT spectrum progressively aligns with the actual curve. At $\delta = 2000$, the first three orders of frequency almost completely match the true value, with an MSE < 0.1. Thus, it is imperative to pre-select a sliding window of the appropriate length to intercept the original signal with the correct frequency domain characteristics. The present study opts for a sliding window of $\delta = 2000$ to augment the training dataset and guarantee recognition accuracy. Using structure #81 as an illustration, Table 6 displays the MSE and MAC values of the predicted outcomes for $\delta = 2000$. The table reveals that the natural frequencies, damping ratios, and mode shapes can be identified almost entirely while the relative error remains below 1%, the absolute error below 0.1%, and the MAC value above 0.995.

4.3. Robustness

In the vibration video, the three-story acrylic frame was subjected to different noises (e.g., lighting conditions and shooting distance). To assess the robustness of the model, videos of each specimen were taken from a random location between 2 m and 3 m directly in front of the structure as shown in Fig. 17 (1). Note that the data here is collected at different distances by field acquisition, rather than using the random zoom-in/zoom-out data enhancement functionality. The primary

Table 6
Predicted modal parameters by the proposed method with comparison to the baseline from the contact vibrometers (structure #81 when $\delta = 2000$).

Mode	Baseline		Predicted modal parameters by the proposed method				
	f (Hz)	ξ	f^{id} (Hz)	ξ^{id}	RE (f^{id})	AE (ξ^{id})	MAC
1	1.136	0.459%	1.139	0.458%	0.26%	0.001%	0.9998
2	2.992	0.106%	3.015	0.117%	0.77%	0.009%	0.9981
3	4.312	0.150%	4.277	0.138%	0.81%	0.012%	0.9974

Note: RE is the relative error and AE is the absolute error.

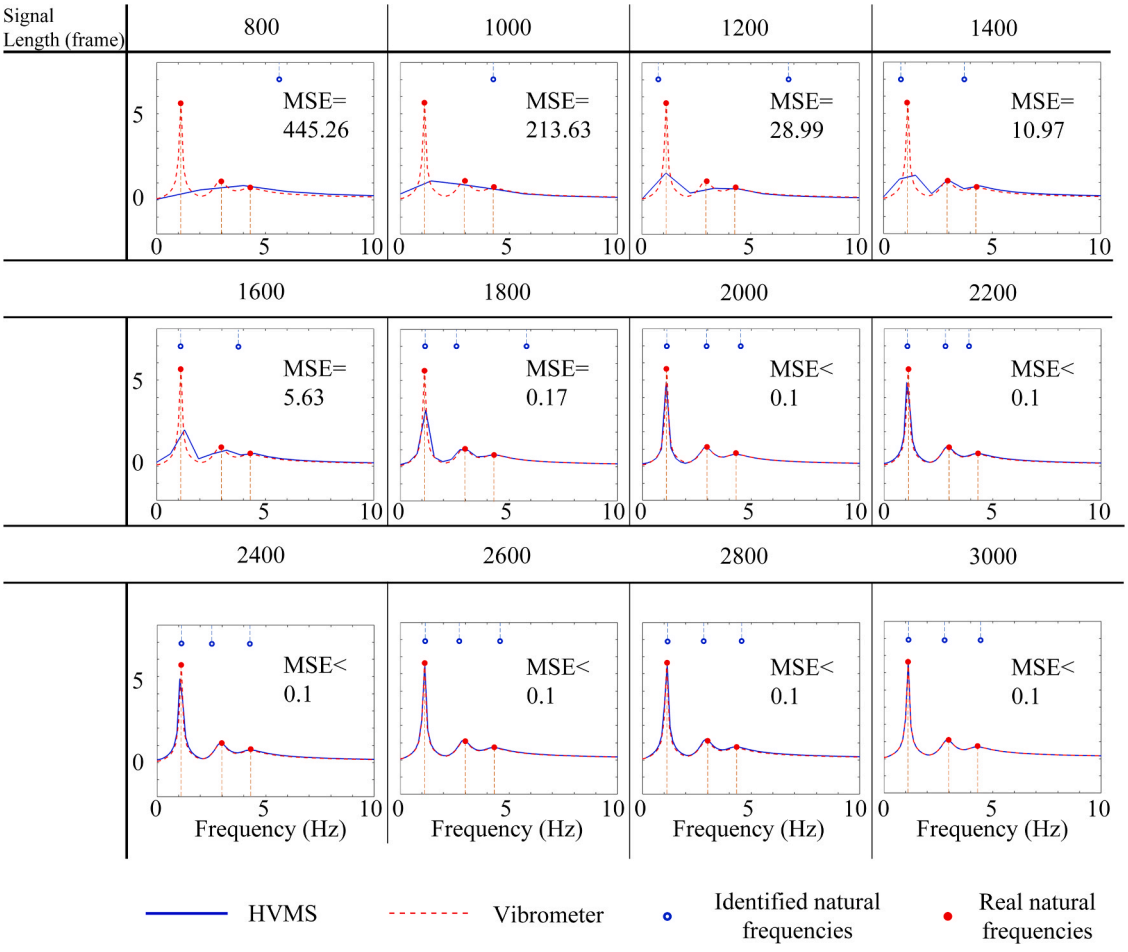


Fig. 16. Illustrated with #81 structure, the predicted natural frequencies, and the corresponding MSE values of the models trained with different signal lengths δ.

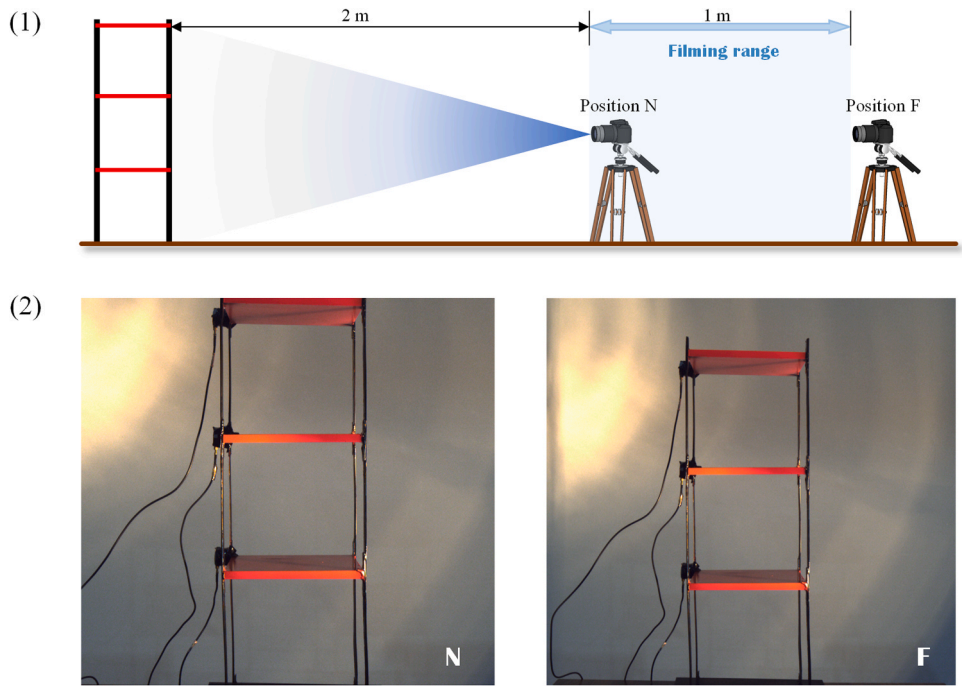


Fig. 17. (1) Take videos from a random position between 2 m and 3 m directly in front of the acrylic frame; (2) Front views of the acrylic frame taken from positions **N** and **F**, respectively.

consideration is that the background of the structure will alter as per the shooting distances, owing to perspective imaging. Such actual changes cannot be completely replaced by data enhancement. In order to evaluate the performance of the model, the modal parameters of the same structure were determined and compared from the videos taken at the furthest and nearest points within the shooting range (referred to as **F** and **N** positions, respectively), as depicted in Fig. 17 (2). The outcomes of the modal parameter identification for videos captured at the farthest point **F** (3 m) and the nearest point **N** (2 m) (known as video *F* and video *N*, correspondingly) are examined via one-sided violin plots shown in Fig. 18. It is evident that video *N* (right half) yields better identification results. Furthermore, Fig. 19 displays the results of comparing the modal parameter identification using high-resolution video *F* and low-resolution video *N*, where the resolution was artificially adjusted while keeping the same *D*/pixel ratio relationship. As depicted in the figure, both videos yield similar accuracies. This is consistent with the finding in Section 4.1 that the precision of extracting edge pixel displacements reduces as the *D*/pixel ratio decreases.

4.4. Extrapolability

To assess the model's performance on data beyond its training range (extrapolability), this study cites a video of the new test structure [51], illustrated in Fig. 20. Moreover, two comparable deep neural networks,

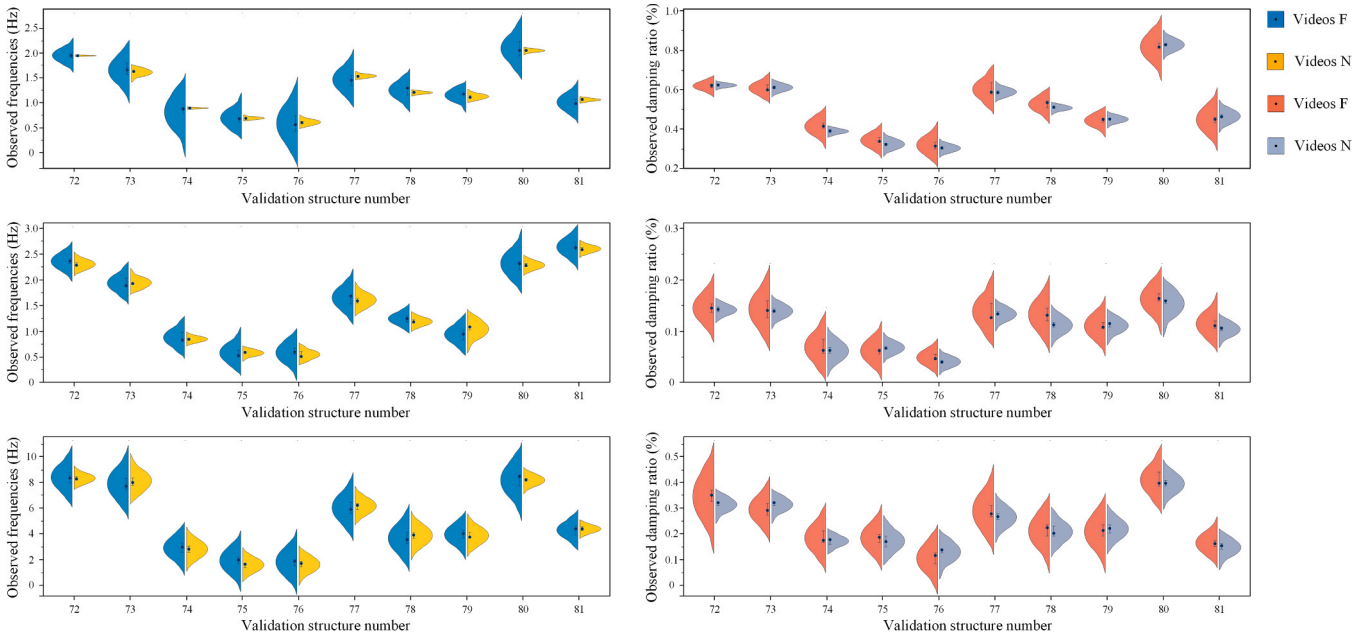


Fig. 18. One-sided violin plots of the modal parameter identification results for #72-#81 structures from videos *F* and *N*.

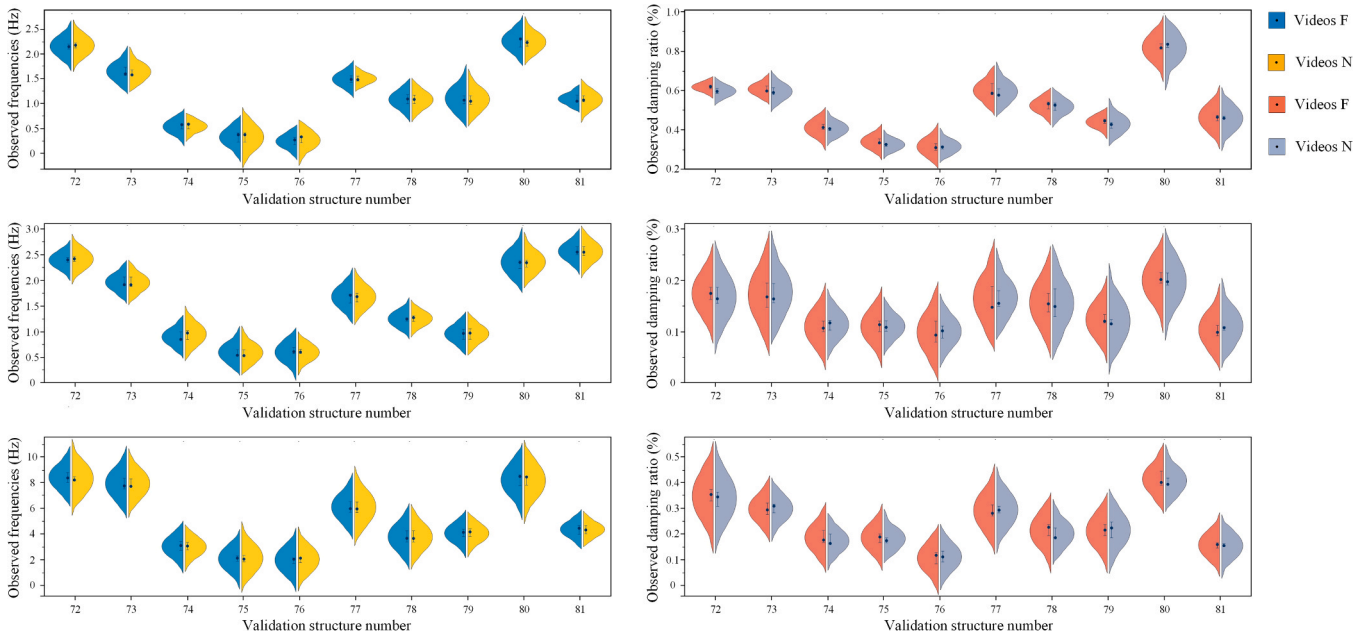


Fig. 19. Video *N* reduces its resolution to maintain the *D*/pixel ratio of Video *F*, and the modal parameter identification results are subsequently compared using one-sided violin plots.

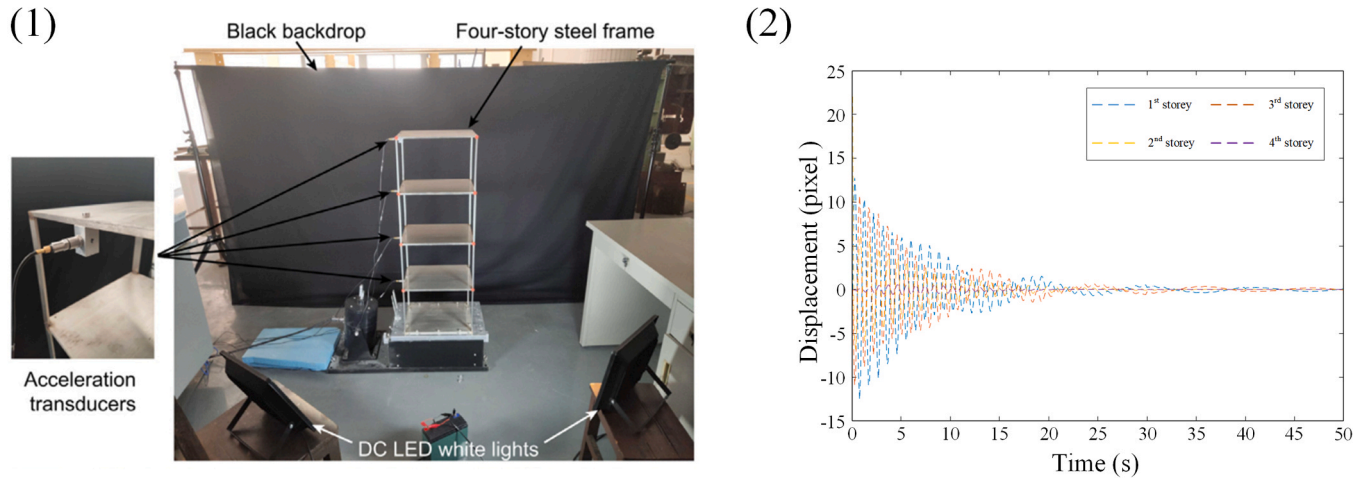


Fig. 20. (1) The 4-story steel frame in the study of Lu et al. [51]; (2) Measured pixel displacement with the proposed method.

1D-CNN-Recurrent Neural Network (RNN) [54] and 1D-CNN-Gated Recurrent Unit (GRU) [55], were also trained using the same dataset and individually compared to the developed 1D-CNN-LSTM. Firstly, the edge pixel displacements were obtained using the proposed method, with corresponding pixel displacements for four floors illustrated in Fig. 20(2). Subsequently, the trained 1D-CNN-LSTM, 1D-CNN-RNN and 1D-CNN-GRU were applied to identify the modal parameters of the test structure and compare the outcomes. Natural frequency and damping ratio for the first three orders can be found in Table 7, while the first three orders of modal shapes are illustrated in Fig. 21. The figure displays the baseline value outlined in literature [51] as a red line, while the first, second and third order modal shapes' determined values are depicted as green, yellow and blue dashed lines, respectively. The 1D-CNN-LSTM model accurately tracks the baseline, validating the model's extrapolability.

Table 7
For the test structure, the modal parameter results of 1D-CNN-LSTM are compared with those of the conventional methods.

Parameter	Mode	Baseline	1D-CNN-LSTM	1D-CNN-RNN	1D-CNN-GRU
Natural frequency (Hz)	First	7.83	7.64	7.41	7.19
	Second	24.57	23.86	23.23	23.53
	Third	39.97	38.21	37.55	36.35
Damping ratio	First	0.23%	0.28%	0.38%	0.14%
	Second	0.11%	0.27%	0.39%	0.30%
	Third	0.25%	0.53%	0.78%	0.64%

4.5. Limitations and future discussion

After conducting the above parametric analysis, the accuracy, robustness, and extrapolability of the proposed method are verified. Nevertheless, the current study has certain limitations. Firstly, the order of structural modal identification is restricted. Secondly, the sensor connecting the wire may act as a moving object in the background and thus affect the outcomes. Thirdly, Lack of validation on real buildings. To address the previously mentioned limitations, future research

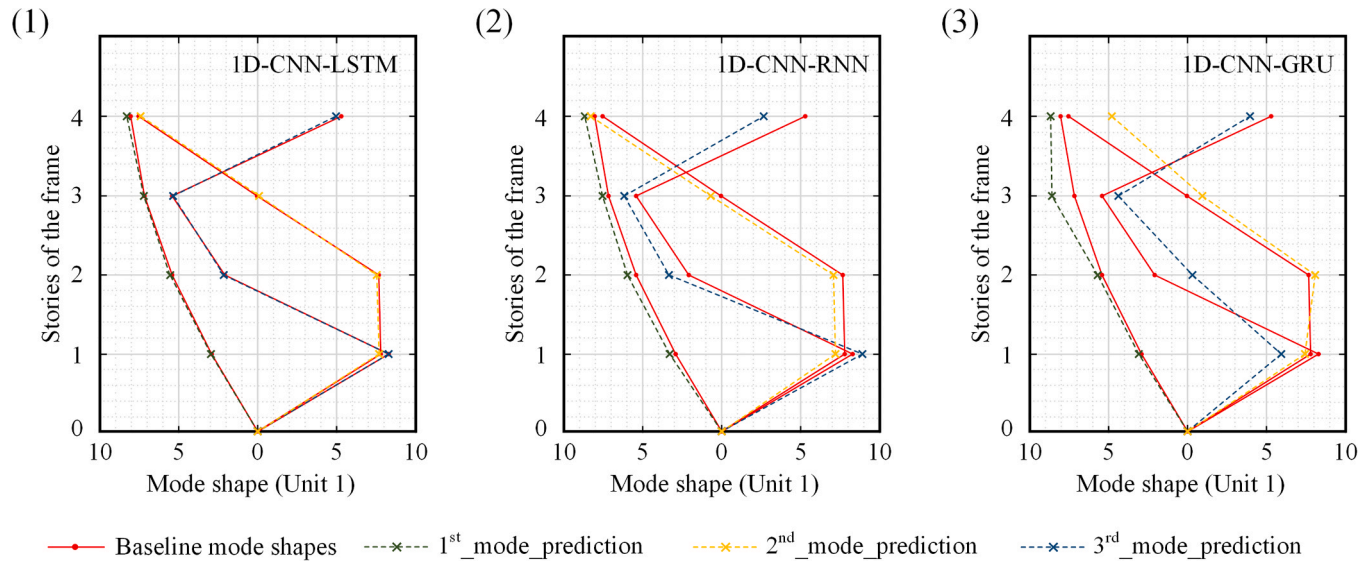


Fig. 21. Baselines and identification values for the first four orders of modal shapes.

solutions include (1) The addition of adaptive mechanisms to the network. (2) Background effects were eliminated in this study by selecting ROIs. For further improvement, additional laser vibrometer equipment is needed to replace the contact vibrometer. (3) Addition of excitation devices for real buildings and improved algorithms to determine the modal parameters of structures under environmental excitation.

5. Concluding remarks

In this study, a framework based on computer vision was developed for the automatic identification of structural modal parameters. The framework consisted of two main components: 1. The extraction of edge pixel vibration signals from videos using edge detection and Flownet2; 2. The identification of structural modal parameters from 1D signals utilizing the developed 1D-CNN-LSTM deep learning model. To establish the efficiency and precision of the suggested model, an open-source video was utilized to conduct comparative testing with two other models. Moreover, the study analyzed the impacts of the Gaussian coefficient, structural amplitude, noise, and training signal length separately. The findings suggest that.

1. When applying a displacement amplitude of $D \geq 11.5$ pixels and using a Gaussian smoothing parameter of $\sigma = 1$ during preprocessing, an accurate vibration signal can be extracted with a MAC value greater than 0.99 for the first-order mode shape.
2. Additionally, when the video is recorded at 300fps and the training signal's step size is set to $d = 2000$, the well-trained model predicts the natural frequency with an RE of less than 1%, the damping ratio with an AE of less than 0.1%, and the mode shape with a MAC value greater than 0.995.
3. The proposed method accurately identifies the modal parameters of an unseen structure and outperforms both 1D-CNN-RNN and 1D-CNN-GRU architectures.

The outlined study presents two advantages over existing modal parameter identification methods. Firstly, in terms of "signal acquisition", a computing balance is achieved between the convenience of machine vision and the usage of high-capacity video for modal analysis. This is achieved by employing preprocessing to increase the effective information density. Secondly, regarding "modal analysis", the novel contribution to the automated approach is the reduction of method complexity via deep neural networks, without adding further clustering centers or thresholds to conventional algorithms. The proposed framework's superiorities support adopting a non-contact computer vision-based approach to real-time structural modal analysis.

CRediT authorship contribution statement

Deng Lu: Supervision, Resources, Project administration. **Liu Yingkai:** Writing – original draft, Software, Methodology, Investigation, Conceptualization. **Cao Ran:** Writing – review & editing, Data curation. **Xu Shaopeng:** Visualization, Validation, Investigation.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data Availability

Data will be made available on request.

Acknowledgements

The present investigation was performed under the support of Hunan Provincial Science and Technology Innovation Leader Project (Grant No. 2021RC4025) and National Natural Science Foundation of China (Grant No. 52278177).

Appendix

Table A1
Three designed architectures of 1D-CNN-LSTMs for modal parameters identification.

Configuration	Kernel information	Shadow	Medium	Deep
Conv1D_1	Kernel number	16	16	16
	Kernel size	64	64	64
Pooling1D_1	Kernel size	2	2	2
Conv1D_2	Kernel number	32	32	32
	Kernel size	3	3	3
Pooling1D_2	Kernel size	2	2	2
Conv1D_3	Kernel number	64	64	64
	Kernel size	3	3	3
Pooling1D_3	Kernel size	2	2	2
Conv1D_4	Kernel number	128	128	128
	Kernel size	3	3	3
Pooling1D_4	Kernel size		2	2
Conv1D_5	Kernel number		256	256
	Kernel size		3	3
Pooling1D_5	Kernel size		2	2
Conv1D_6	Kernel number			256
	Kernel size			3
Pooling1D_6	Kernel size			2
Conv1D_7	Kernel number			256
	Kernel size			3
Pooling1D_7	Kernel size			2
LSTM Layer	Number of nodes	20	20	20
Dense Layer	Number of nodes	9	9	9

Table A2
Parameters considered and corresponding MREs for the predicted natural frequencies of the first three orders.

Cases	Loss function	Optimizer	Batch size	MREs (%) of natural frequencies at different orders.			MRE (%)
				1	2	3	
1	MSE	Adam	8	5.52	4.84	5.92	5.42
2			16	5.02	5.96	6.45	5.81
3			32	5.38	6.53	7.58	6.49
4			64	**1.54**	**2.75**	4.03	**2.77**
5			128	4.71	4.65	4.95	4.77
6			256	3.54	3.63	4.82	3.99
7		RMSProp	8	4.45	5.72	4.75	4.97
8			16	6.52	6.09	5.52	6.04
9			32	4.08	4.52	5.52	4.70
10			64	3.27	6.18	8.14	5.86
11			128	3.72	4.23	6.28	4.74
12			256	2.77	5.42	7.87	5.35
13	MAE	Adam	8	5.47	9.22	3.44	9.37
14			16	4.87	5.91	9.71	6.83
15			32	5.11	4.53	5.17	4.93
16			64	5.63	6.51	7.77	6.63
17			128	5.01	5.07	3.73	4.60
18			256	5.08	5.02	**3.62**	4.57
19		RMSProp	8	4.87	7.15	5.51	5.84
20			16	4.42	9.77	5.72	6.63
21			32	6.43	6.23	5.44	6.03
22			64	4.46	7.48	5.72	5.88
23			128	4.82	7.15	4.68	5.55
24			256	7.52	8.03	6.07	7.20

References

[1] Reynders E. System identification methods for (operational) modal analysis: review and comparison. Arch Comput Method E 2012;19:51–124.
[2] Yang Y, Nagarajaia HS. Blind modal identification of output-only structures in time-domain based on complexity pursuit. Earthq Eng Struct D 2013;42(13):1885–905.
[3] Ibrahim SR, Mikulcik EC. A time domain modal vibration test technique. Shock Vib 1973;43(4):21–37.
[4] Erdem E, Shi J. ARMA based approaches for forecasting the tuple of wind speed and direction. Appl Energ 2011;88(4):1405–14.
[5] G.H. James, T.G. Carne, J.P. Lauffer, The Natural Excitation Technique for Modal Parameters Extraction from Operating Wind Turbines. Report No. SAND92–1666, UC 261 (1993).
[6] Juang JN, Pappa RS. An eigensystem realization algorithm for modal parameter identification and model reduction. J Guid Control Dynam 1985;8(5):620–7.
[7] Huang NE, Shen Z, Long SR, et al. The empirical mode decomposition and the hilbert spectrum for nonlinear and non-stationary time series analysis. Proc R Soc Lond Ser A Math Phys Eng Sci 1998;454(1971):903–95.
[8] Antoni J. Blind separation of vibration components: principles and demonstrations. Mech Syst Signal Pr 2005;19(6):1166–80.
[9] Farrar CR, James GH. System identification from ambient vibration measurements on a bridge. J Sound Vib 1997;205(1):1–18.
[10] J.P. Lynch, A. Sundararajan, K.H. Law, et al., Field validation of a wireless structural health monitoring system on the Alamosa Canyon Bridge. Proceedings of Smart Structures and Materials 2003: Smart Systems and Nondestructive Evaluation for Civil Infrastructures. SPIE 5057 (2003) 267–278.
[11] Jang S, Jo H, Cho S, et al. Structural health monitoring of a cable-stayed bridge using smart sensor technology: deployment and evaluation. Smart Struct Syst 6 (5-6) 2010:439–59.

[12] Luo K, Kong X, Wang X, et al. Cable vibration measurement based on broad-band phase-based motion magnification and line tracking algorithm. Mech Syst Signal Pr 2023;200:110575.
[13] Luo K, Kong X, Zhang J, et al. Computer vision-based bridge inspection and monitoring: A review. Sensors 2023;23(18):7863.
[14] Cho S, Giles RK, Spencer BF. System identification of a historic swing truss bridge using a wireless sensor network employing orientation correction. Struct Control Health Monit 2015;22(2):255–72.
[15] Siringoringo DM, Fujino Y. System identification of suspension bridge from ambient vibration response. Eng Struct 2008;30(2):462–77.
[16] Spencer BF, Hoskere V, Narazaki Y. Advances in computer vision-based civil infrastructure inspection and monitoring. Engineering 2019;5(2):199–222.
[17] V. Hoskere, Y. Narazaki, T.A. Hoang, et al., Vision-based structural inspection using multiscale deep convolutional neural networks. arXiv preprint arXiv: 1805.01055, 2018.
[18] Spencer BF, Hoskere V, Narazaki Y. Advances in computer vision–based civil infrastructure inspection and monitoring. Engineering 2019;5(2):199–222.
[19] Y. Narazaki, V. Hoskere, T.A. Hoang, et al., Automated bridge component recognition using video data. arXiv preprint arXiv:1806.06820, 2018.
[20] Tomasi C, Kanade T. Detection and tracking of point. Int J Comput Vis 1991;9: 137–54. 3.
[21] Yoo JC, Han TH. Fast normalized cross-correlation. Circ Syst Signal Pract 2009;28: 819–43.
[22] Wadhwa N, Rubinstein M, Durand F, et al. Phase-based video motion processing. ACM T Graph 2013;32(4):1.
[23] Nogueira FMA, Barbosa FS, Barra LPS. Evaluation of structural natural frequencies using image processing. Proc EVACES 2005.
[24] Chang CC, Xiao XH. An integrated visual-inertial technique for structural displacement and velocity measurement. Smart Struct Syst 2010;6(9):1025–39.
[25] Fukuda Y, Feng MQ, Narita Y, et al. Vision-based displacement sensor for monitoring dynamic response using robust object search algorithm. IEEE Sens J 2013;13(12):4725–32.
[26] Schumacher T, Shariati A. Monitoring of structures and mechanical systems using virtual visual sensors for video analysis: Fundamental concept and proof of feasibility. Sensors 2013;13(12):16551–64.
[27] Yoon H, Elanwar H, Choi H, et al. Target-free approach for vision-based structural system identification using consumer-grade cameras. Struct Control Health Monit 2016;23(12):1405–16.
[28] Feng D, Feng MQ. Vision-based multipoint displacement measurement for structural health monitoring. Struct Control Health Monit 2016;23(5):876–90.
[29] Yoon H, Hoskere V, Park JW, et al. Cross-correlation-based structural system identification using unmanned aerial vehicles. Sensors 2017;17(9):2075.
[30] Hoskere V, Park JW, Yoon H, et al. Vision-based modal survey of civil infrastructure using unmanned aerial vehicles. J Struct Eng 2019;145(7): 04019062.
[31] Han J, Jentzen A, Weinan E. Solving high-dimensional partial differential equations using deep learning. Proc Natl Acad Sci 2018;115(34):8505–10.
[32] Yue Z, Liu L. A single-mode recursive validation method for modal identification of linear time-varying structures based on prior knowledge. Struct Control Health Monit 2021;28(12):e2845.
[33] Liu D, Tang Z, Bao Y, et al. Machine-learning-based methods for output-only structural modal identification. Struct Control Health Monit 2021;28(12):e2843.
[34] Kim H, Sim SH. Automated peak picking using region-based convolutional neural network for operational modal analysis. Struct Control Health Monit 2019;26(11): e2436.
[35] Su L, Zhang JQ, Huang X, et al. Automatic operational modal analysis of structures based on image recognition of stabilization diagrams with uncertainty quantification. Multidim Syst Sign Pract 2021;32:335–57.
[36] Zhang Y, Miyamori Y, Mikami S, et al. Vibration-based structural state identification by a 1-dimensional convolutional neural network. Comput-Aided Civ Inf 2019;34(9):822–39.
[37] Yun DY, Shim HB, Park H. SSSI-LSTM network for adaptive operational modal analysis of building structures. Mech Syst Signal Pr 2023;195:110306.
[38] Shu JP, Zhang CG, Gao YF, et al. A multi-task learning-based automatic blind identification procedure for operational modal analysis. Mech Syst Signal Pr 2023; 187:109959.
[39] Zhang C, Wang WZ, Zhang C, et al. Extraction of local and global features by a convolutional neural network–long short-term memory network for diagnosing bearing faults. Proc Inst Mech Eng, Part C: J Mech Eng Sci 2022;236(3):1877–87.
[40] Dizaji MS, Mao Z, Haile MA. Hybrid-attention-ConvLSTM-based deep learning architecture to extract modal frequencies from limited data using transfer learning. Mech Syst Signal Pr 2023;187:109949.
[41] Yang R, Singh SK, Tavakkoli M, et al. CNN-LSTM deep learning architecture for computer vision-based modal frequency detection. Mech Syst Signal Pr 2020;144: 106885.
[42] Bruhn A, Weickert J, Schnorr C. Lucas/Kanade meets Horn/Schunck: combining local and global optical flow methods. Int J Comput Vis 2005;61(3):211–31.
[43] Buxton B, Buxton H. Monocular depth perception from optical flow by space time signal processing. Proc R Soc Lond B 1983;218(1210):27–47.
[44] Raman SV, Sharkar S, Boyer KL. Tissue boundary refinement in magnetic resonance images using contour-based scale space matching. IEEE T Med Imaging 1991;10(2):109–21.
[45] E. Ilg, N. Mayer, T. Saikia, M. Keuper, A. Dosovitskiy, T. Brox, FlowNet 2.0: Evolution of optical flow estimation with deep networks, Proc. 2017 IEEE Conf. Computer Vision and Pattern Recognition (2017) 1647–1655.
[46] https://github.com/NVIDIA/flownet2-pytorch.
[47] www.kaggle.com/liuyingkai/video-2-flow-and-dataset.
[48] Mottershead JE, Link M, Friswell MI. The sensitivity method in finite element model updating: a tutorial. Mech Syst Signal Pr 2011;25(7):2275–96.
[49] Wen P, Khan I, He J, Chen QF. Application of improved combined deterministic-stochastic subspace algorithm in bridge modal parameter identification. Shock Vib 2021;2021:8855162.
[50] He XH, Hua XG, Chen ZQ, Huang FL. EMD-based random decrement technique for modal parameter identification of an existing railway bridge. Eng Struct 2011;33 (4):1348–56.
[51] Lu ZR, Lin GF, Wang L. Output-only modal parameter identification of structures by vision modal analysis. J Sound Vib 2021;497:15949.
[52] Liang CJ, Deng HX, Zhang J, Yu LD. Vibration studies of simply supported beam based on binocular stereo vision. Proc SPIE 2015;9446:94463M.
[53] Zheng MY, Peng P, Zhang BJ, Zhang N, Wang LF, Chen YC. A new physical parameter identification method for two-axis on-road vehicles: simulation and experiment. Shock Vib 2015:191050.
[54] Liang Y, Wu D, Liu G, Li Y, Gao C, Ma ZJ, et al. Big data-enabled multiscale serviceability analysis for aging bridges. Digit Commun Netw 2016;2(3):97–107.
[55] Z. Zhang, D. Robinson, J. Tepper, Detecting hate speech on Twitter using a convolution-GRU based deep neural network, European semantic web conference. Cham: Springer (2018), 745–760.