Measurement 218 (2023) 113158

Contents lists available at ScienceDirect

Measurement

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

Automated region-of-interest selection for computer-vision-based displacement estimation of civil structures

Jaemook Choi, Zhanxiong Ma, Kiyoung Kim, Hoon Sohn [*]

Department of Civil and Environmental Engineering, Korea Advanced Institute of Science and Technology, Daejeon, South Korea

ARTICLE INFO

Keywords:
Automated region of interest (ROI) selection
Structural displacement estimation
Computer vision

ABSTRACT

A recent trend in vision-based displacement measurement is to place a camera at the measurement point and capture the images of the surrounding areas. In this scheme, a proper region of interest (ROI) should be selected from the captured images. This paper proposes an automated ROI selection technique to improve displacement estimation accuracy. The image frames that capture larger movements of the surrounding areas were selected, and the features in the selected frames were grouped using clustering algorithms. The feature group with consistent movement and high density was finally selected as the optimum ROI. The proposed technique was validated through laboratory and field tests. A displacements estimation technique previously proposed by the authors were used to compared the optimum ROI and four intuitively selected ROIs. In all the tests, the displacement estimates from the optimum ROI showed a smaller RMSE (less than 2 mm) than those from other ROIs.

1. Introduction

The displacement of a civil structure plays a vital role in the monitoring and control of the structure, as it is directly related to the stiffness of the structure and intuitively represents the soundness of the structure. Several sensors have been introduced to measure or estimate the displacements of civil structures. Conventional sensors include linear variable differential transformers (LVDT) [1] and accelerometers [2,3]. However, installing an LVDT in the field is difficult because its ends should be connected at the measurement point and a stationary location. Acceleration measurements from accelerometers are converted to displacement through double integration, which causes serious errors in the low-frequency region.

In recent decades, several noncontact sensors, such as the real-time kinematic global navigation satellite system (RTK-GNSS) [4], laser Doppler vibrometer (LDV) [5] and radar systems [6,7], have been applied to structural displacement estimation. However, RTK-GNSS, the most common sensor for structural displacement monitoring, has a low sampling rate of up to 20 Hz and a limited accuracy of approximately 20 mm in the vertical direction. Although LDV and radar systems are capable of high-accuracy and high-sampling displacement measurements, these devices are expensive and require a stationary location for installation, making them impractical for civil infrastructure.

As an alternative to these traditional sensors, vision cameras have emerged as noncontact sensors for various applications [8–12], and recently combined with unmanned aerial vehicles (UAVs) to expand their usability [13]. When applied to structural displacement estimation, a vision camera is commonly installed at a stationary location and is aimed at artificial or natural targets on a target structure [14–16]. Various computer vision algorithms, such as template matching [17,18], optical flow [19,20] and feature matching algorithms [21] can be applied to vision images to extract the motion of targets. In this procedure, the field of view (FOV) of the vision camera is usually set to be sufficiently wide to cover artificial or natural targets, and the region of interest (ROI) is cropped from the vision images for displacement estimation.

However, this displacement estimation procedure has three significant limitations. First, the vision camera must be positioned at a stationary location in the proximity of a target structure, where the target structure has in-plane motion relative to the fixed vision camera. However, identifying such a location can be challenging when the structure is located offshore or in a metropolitan area. Second, the displacement in a physical length unit should be converted from the translation in pixel units by using a scale factor. The scale factor should be estimated manually by measuring the dimensions of the target [22] or the distance between the target and the vision camera [16] in physical

[*] Corresponding author.
E-mail address: hoonsohn@kaist.ac.kr (H. Sohn).

https://doi.org/10.1016/j.measurement.2023.113158
Received 2 March 2023; Received in revised form 10 May 2023; Accepted 31 May 2023
Available online 3 June 2023

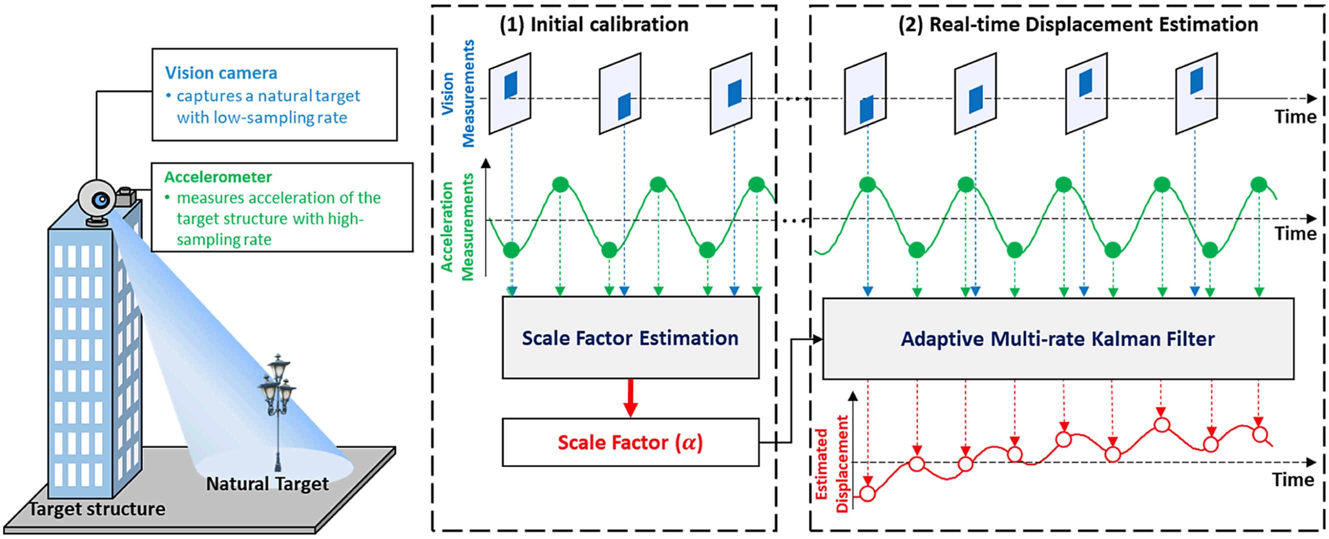
Vision camera
• captures a natural target with low-sampling rate
Accelerometer
• measures acceleration of the target structure with high-sampling rate
Natural Target
Target structure
(1) Initial calibration
(2) Real-time Displacement Estimation
Vision Measurements
Acceleration Measurements
Time
Time
Scale Factor Estimation
Scale Factor (α)
Adaptive Multi-rate Kalman Filter
Estimated Displacement
Time

Fig. 1. Existing displacement estimation technique fusing an asynchronous vision camera and accelerometer [24].

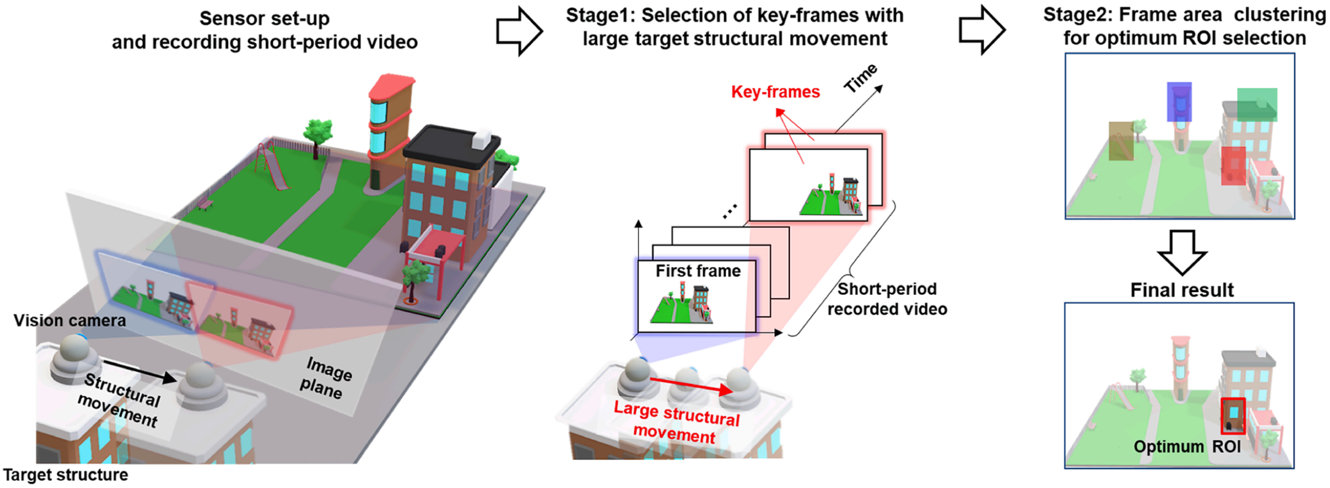
Sensor set-up and recording short-period video
Stage1: Selection of key-frames with large target structural movement
Stage2: Frame area clustering for optimum ROI selection
Vision camera
Structural movement
Image plane
Target structure
Key-frames
Time
First frame
Short-period recorded video
Large structural movement
Final result
Optimum ROI

Fig. 2. Overview of the automated ROI selection technique proposed in this study.

units in advance; however, this manual procedure is challenging in the field and can be a source of error for the estimated displacement. Finally, these vision-based techniques incur high computational costs, which hinder the real-time estimation of displacement with a high sampling rate.

Several studies have been conducted to resolve these shortcomings. The first shortcoming has been effectively addressed by installing a vision camera directly at the measurement point of the target structure [23]. In addition, the fusion of an accelerometer and a vision camera collocated on a target structure, as shown in Fig. 1 [24,25] can be a partial remedy for the second and third shortcomings in structural displacement estimation. This approach can enable autonomous scale factor estimation without any prior knowledge of the dimensions of a natural target or the target-to-camera distance and real-time high-sampling displacement estimation using asynchronous low-sampling vision images and high-sampling acceleration measurements based on an adaptive multi-rate Kalman filter. However, the ROI selection process was still manually performed by trial-and-error to cover one of the targets captured in the FOV of the camera. This ROI selection becomes cumbersome and time-consuming as the number of targets increases. In addition, it is difficult to identify the optimal position and size of an ROI because the ROI area is manually designated for each target. This imprecision in the manual ROI selection process can lead to inaccurate displacement estimation.

This paper proposes an automated ROI selection technique for structural displacement estimation using a vision camera. After a vision camera is mounted on the measurement point of a target structure, it starts recording nearby objects and/or structures and captures the vibration of the target structure through video images. The proposed technique selects several key-frames in which large movements of the target structure are captured. The key-frames and the first frame in the video are compared to construct several types of maps to assess the quality of the feature points detected in the key-frames. Two criteria, coefficient of variance (CV) of feature-based translations and feature density (FD), were adopted for assessing the feature point quality in the proposed technique, and the feature points were classified based on CV and FD by two unsupervised clustering algorithms. Finally, a cluster with a small CV and high FD was selected as the optimum ROI.

The main contributions of this study are as follows: (1) the proposed technique automates the optimum ROI selection process for computer-vision-based displacement estimation of civil structures, (2) it enables highly efficient optimum ROI selection by reducing computational cost through the automatic selection of key-frames with relatively large structural movements of a target structure, and (3) the optimum ROI can be selected much more precisely than existing manual processes by assessing feature points based on the CV of feature-based translation and

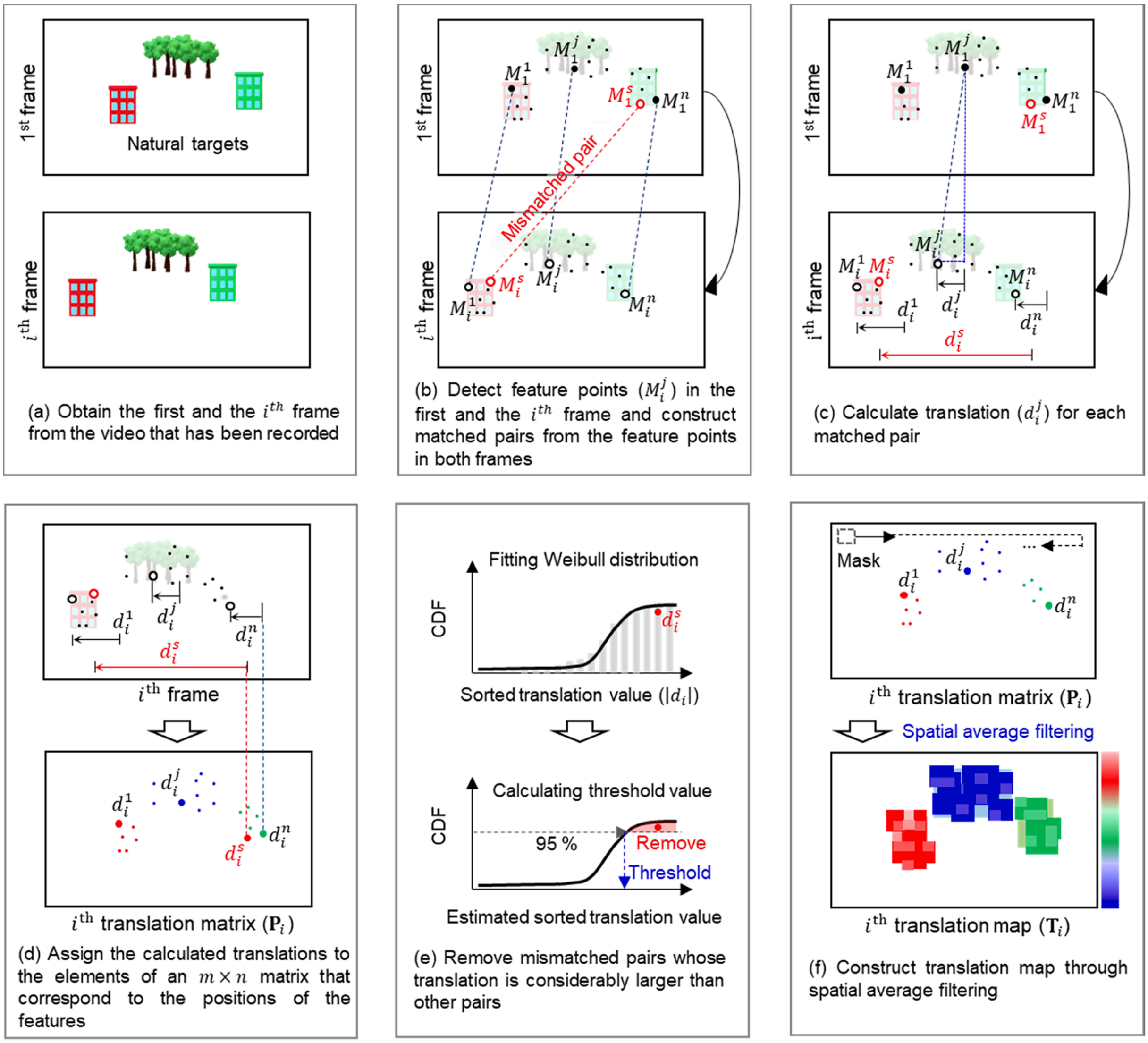
1st frame
Natural targets
ith frame
(a) Obtain the first and the ith frame from the video that has been recorded
M_1^1
M_1^j
M_1^s
M_1^n
Mismatched pair
M_i^1
M_i^s
M_i^j
M_i^n
(b) Detect feature points (M_i^j) in the first and the ith frame and construct matched pairs from the feature points in both frames
d_i^1
d_i^j
d_i^n
d_i^s
(c) Calculate translation (d_i^j) for each matched pair
ith frame
ith translation matrix ($\mathbf{P}_i$)
(d) Assign the calculated translations to the elements of an $m \times n$ matrix that correspond to the positions of the features
Fitting Weibull distribution
CDF
Sorted translation value ($|d_i|$)
Calculating threshold value
95 %
Remove
Threshold
Estimated sorted translation value
(e) Remove mismatched pairs whose translation is considerably larger than other pairs
Mask
ith translation matrix ($\mathbf{P}_i$)
Spatial average filtering
ith translation map ($\mathbf{T}_i$)
(f) Construct translation map through spatial average filtering

Fig. 3. Flowchart of translation map construction.

FD.

The remainder of this paper is organized as follows. The proposed ROI selection technique is described in Section 2. The performance of the proposed technique is validated using a laboratory test in Section 3, and a field test on a pedestrian bridge is presented in Section 4. The concluding remarks are presented in Section 5.

2. Development of the proposed automated ROI selection technique

In the configuration of structural displacement measurement using a vision camera mounted on a measurement point on a structure, the vision camera should keep track of stationary natural targets near the measurement point. As it is vital to select the optimum ROI that includes one of these natural targets, the proposed technique automates the ROI selection process by using a short-period video recorded after the installation of the vision camera. As shown in Fig. 2, the proposed technique consists of two stages: (1) the selection of key-frames with large target structural movement from the short-period video (i.e., 1 min) and (2) frame area clustering for optimum ROI selection. The working principles of Stages 1 and 2 will be explored in detail in Sections 2.1 and 2.2, respectively.

2.1. Stage 1: Selection of key-frames with large target structural movement

In Stage 1, key-frames, the frames in which the target structure has a relatively large movement in the recorded video, are selected. The selection of key-frames is an essential task because applying Stage 2 to all frames in the recorded video is time-consuming and requires a high computational cost. In addition, frames that contain relatively small structural displacements are of no use in Stage 2 because pixel discretization errors in these frames inhibit the accurate evaluation of feature-based translation properties. Stage 1 consists of three sub-steps: the construction of a translation map, active and inactive pixel counting, and the calculation of a motion index.

2.1.1. Translation map construction

In this sub-step, a translation map that illustrates the feature translations of a frame (i.e., the i^{th} frame) with respect to the first frame in the recorded video is constructed using feature matching and spatial average filtering.

First, the proposed technique detects feature points from the first and the i^{th} frame and matches the feature points in the two frames, as shown in Fig. 3(a) and (b), using speeded-up robust features (SURF) [26], a feature matching algorithm with a relatively lower computational cost

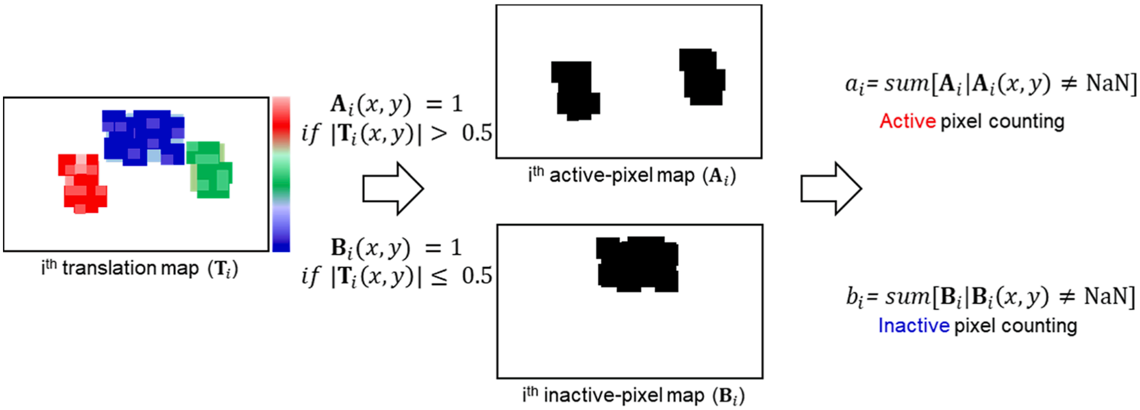
ith translation map ($\mathbf{T}_i$)
$\mathbf{A}_i(x,y) = 1$
$if\ |\mathbf{T}_i(x,y)| > 0.5$
$\mathbf{B}_i(x,y) = 1$
$if\ |\mathbf{T}_i(x,y)| \le 0.5$
ith active-pixel map ($\mathbf{A}_i$)
ith inactive-pixel map ($\mathbf{B}_i$)
$a_i = sum[\mathbf{A}_i|\mathbf{A}_i(x,y) \ne \mathrm{NaN}]$
Active pixel counting
$b_i = sum[\mathbf{B}_i|\mathbf{B}_i(x,y) \ne \mathrm{NaN}]$
Inactive pixel counting

Fig. 4. Active and inactive pixel counting.

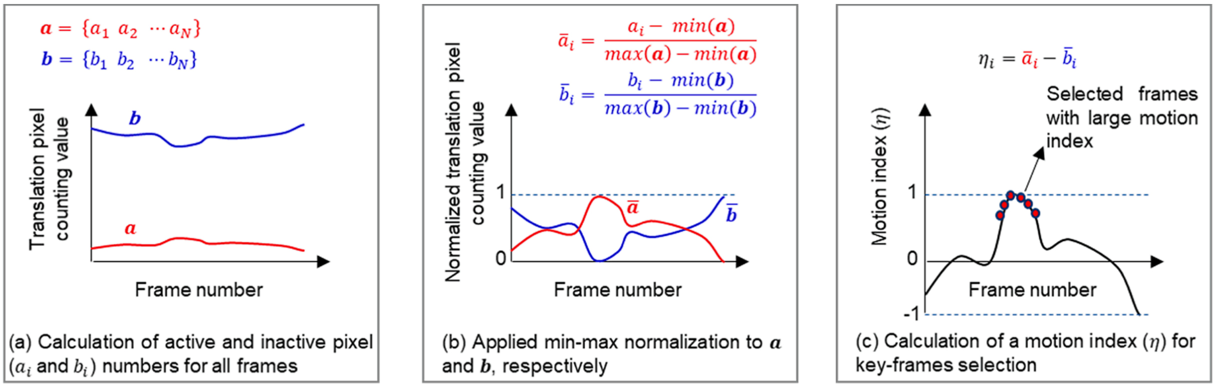
$\boldsymbol{a} = \{a_1\ a_2\ \cdots a_N\}$
$\boldsymbol{b} = \{b_1\ b_2\ \cdots b_N\}$
Translation pixel counting value
b
a
Frame number
(a) Calculation of active and inactive pixel (a_i and b_i) numbers for all frames
$\bar{a}_i = \frac{a_i - min(\boldsymbol{a})}{max(\boldsymbol{a}) - min(\boldsymbol{a})}$
$\bar{b}_i = \frac{b_i - min(\boldsymbol{b})}{max(\boldsymbol{b}) - min(\boldsymbol{b})}$
Normalized translation pixel counting value
1
0
$\bar{\boldsymbol{a}}$
$\bar{\boldsymbol{b}}$
Frame number
(b) Applied min-max normalization to $\boldsymbol{a}$ and $\boldsymbol{b}$, respectively
$\eta_i = \bar{a}_i - \bar{b}_i$
Selected frames with large motion index
Motion index (η)
1
0
-1
Frame number
(c) Calculation of a motion index (η) for key-frames selection

Fig. 5. Translation-based motion index calculation for key-frame selection.

and higher accuracy compared with other algorithms [27,28]. Let M_i^j denote the jth feature point in the ith frame, and let (M_1^j, M_i^j) denote the matched pair of the jth feature points in the first and ith frame. For each matched pair, the translation d_i^j is calculated as the distance between M_i^j and M_1^j in a pixel unit, as shown in Fig. 3(c). Assuming that the pixel dimensions of the frame $m \times n$, an $m \times n$ translation matrix $\mathbf{P}_i$ is constructed by assigning each d_i^j to the element corresponding to the position of M_1^j as shown in Fig. 3(d). Note that, when the position of an element in $\mathbf{P}_i$ does not correspond to any matched pair, its value is set as NaN.

Suppose that the translations of the feature points have a Weibull distribution. Some matched pairs, e.g., (M_1^s, M_i^s), with translations significantly larger than those of other matched pairs are regarded as mismatched pairs and removed, as shown in Fig. 3(e), to be ignored in the following steps. Note that mismatched pairs with small translations have a negligible impact on ROI selection after spatial average filtering and time averaging, which will be explored later in this section.

Subsequently, a spatial averaging filter with a moving mask is applied to $\mathbf{P}_i$ as shown in Fig. 3(f), and a translation map $\mathbf{T}_i$ can be constructed as follows:

$$\mathbf{T}_i\left(r_x, r_y\right) = E[\mathbf{P}_i(x,y)|(x,y) \in R^2, \mathbf{P}_i(x,y) \ne \mathrm{NaN} \tag{1}$$

where R^2 is the region of the spatial average filter mask, and is set to an odd positive integer. $\left(r_x, r_y\right)$ denotes the coordinates of the center of the mask, and $E[\bullet]$ denotes the expectation operator. While an element in $\mathbf{P}_i$ indicates the translation of a single feature point, the value of $\mathbf{T}_i$ becomes the average translation within the mask. Therefore, an element in $\mathbf{T}_i$ can be a number even though the corresponding element in $\mathbf{P}_i$ is NaN.

2.1.2. Active and inactive pixel counting

In this sub-step, all the features in the translation map are divided into active and inactive pixels, and the numbers of active and inactive pixels are counted. Here, a pixel with a translation value ($\mathbf{T}_i(x,y)$) larger than the pixel discretization error (0.5 pixel unit) is defined as an active pixel; otherwise, it is defined as an inactive pixel. Fig. 4 shows the overall process of counting of the active and inactive pixels. First, a binary process is applied to $\mathbf{T}_i$ to construct the ith active pixel map ($\mathbf{A}_i$) and inactive pixel map ($\mathbf{B}_i$) using Equations (2) and (3):

$$\mathbf{A}_i(x,y) = \begin{cases} 1 & if|\mathbf{T}_i(x,y)|)0.5 \\ \mathrm{NaN} & otherwise \end{cases}, \tag{2}$$

$$\mathbf{B}_i(x,y) = \begin{cases} 1 & if|\mathbf{T}_i(x,y)| \le 0.5 \\ \mathrm{NaN} & otherwise \end{cases}. \tag{3}$$

When a large displacement of the target structure occurs in the ith frame, the features captured in the frame are also moved; however, the amount of movement of the features is not identical because the distance from the vision camera and the features are different. In this context, the elements in $\mathbf{A}_i$ whose value is 1 indicate that the corresponding natural targets are sufficiently close to the measurement point to observe an apparent translation. However, for natural targets that are far away from the target structure, 1 is assigned to the corresponding elements of $\mathbf{B}_i$. The number of active and inactive pixels of $\mathbf{A}_i$ and $\mathbf{B}_i$ are denoted as a_i and b_i, respectively, and calculated using the following equation:

$$a_i = \sum\{\mathbf{A}_i|\mathbf{A}_i(x,y) \ne \mathrm{NaN}\}, b_i = \sum\{\mathbf{B}_i|\mathbf{B}_i(x,y) \ne \mathrm{NaN}\}. \tag{4}$$

2.1.3. Motion index calculation

The previous two steps were repeated for all the frames in the recorded video, and then the active and inactive pixel numbers were

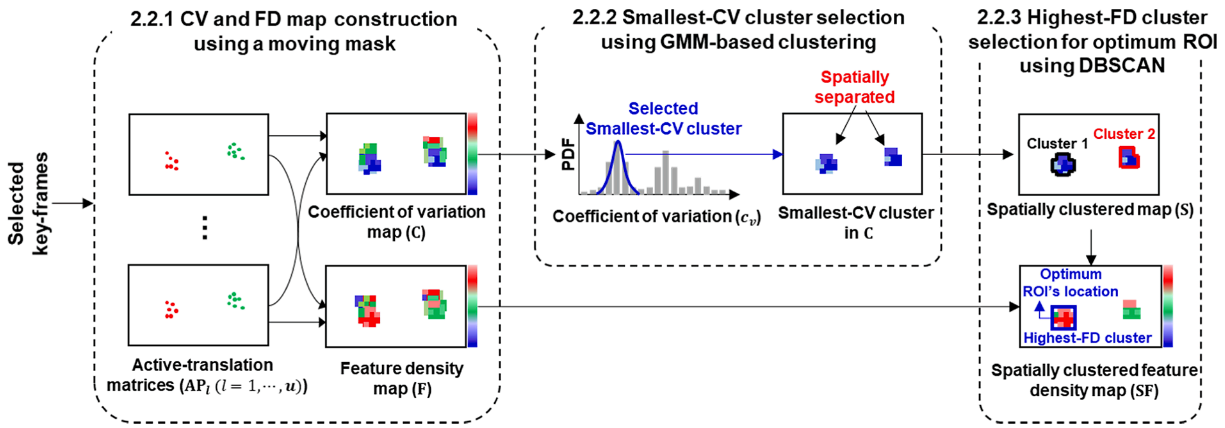
2.2.1 CV and FD map construction using a moving mask
2.2.2 Smallest-CV cluster selection using GMM-based clustering
2.2.3 Highest-FD cluster selection for optimum ROI using DBSCAN
Selected key-frames
Coefficient of variation map (C)
Active-translation matrices (AP$_l$ ($l = 1,\cdots,u$))
Feature density map (F)
PDF
Selected Smallest-CV cluster
Coefficient of variation (c_v)
Spatially separated
Smallest-CV cluster in C
Cluster 1
Cluster 2
Spatially clustered map (S)
Optimum ROI's location
Highest-FD cluster
Spatially clustered feature density map (SF)

Fig. 6. Overview of frame area clustering technique for the optimum ROI selection.

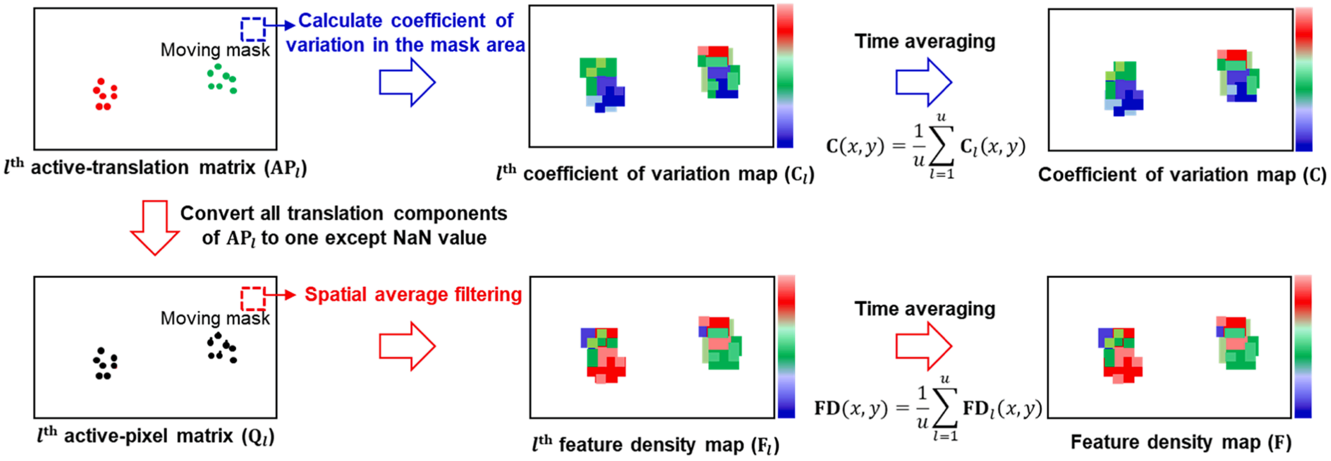
Moving mask
Calculate coefficient of variation in the mask area
l^{th} active-translation matrix (AP$_l$)
l^{th} coefficient of variation map (C$_l$)
Time averaging
$\mathrm{C}(x,y) = \frac{1}{u}\sum_{l=1}^{u} \mathrm{C}_l(x,y)$
Coefficient of variation map (C)
Convert all translation components of AP$_l$ to one except NaN value
Moving mask
Spatial average filtering
l^{th} active-pixel matrix (Q$_l$)
l^{th} feature density map (F$_l$)
Time averaging
$\mathrm{FD}(x,y) = \frac{1}{u}\sum_{l=1}^{u} \mathrm{FD}_l(x,y)$
Feature density map (F)

Fig. 7. Construction process of CV and FD maps.

counted for all the frames. In this sub-step, a motion index is calculated for each frame using its active and inactive pixel numbers, and several frames with large motion indices are selected. The motion index is a scalar value that indicates the amount of displacement that occurs in a specific frame.

Two vectors of active and inactive pixel numbers, denoted as **a** and **b**, respectively, are obtained, as shown in Fig. 5(a).

$$\mathbf{a} = \begin{pmatrix} a_1 & a_2 & \cdots & a_N \end{pmatrix}^T, \mathbf{b} = \begin{pmatrix} b_1 & b_2 & \cdots & b_N \end{pmatrix}^T, \tag{5}$$

where N is the number of frames in the recorded video. If several feature points are detected from natural targets farther from the target structure in practice, b_i is substantially greater than a_i in most frames. Therefore, a_i and b_i must be normalized to cope with the differences in active and inactive pixel numbers. Therefore, a min–max normalization in Equation (6) is applied to a_i and b_i to obtain their normalizations, denoted as $\overline{a}_i$ and $\overline{b}_i$, respectively (Fig. 5(b)):

$$\overline{a}_i = \frac{a_i - \min(\mathbf{a})}{\max(\mathbf{a}) - \min(\mathbf{a})}, \overline{b}_i = \frac{b_i - \min(\mathbf{b})}{\max(\mathbf{b}) - \min(\mathbf{b})}. \tag{6}$$

The motion index η_i is then defined as the difference between $\overline{a}_i$ and $\overline{b}_i$ as shown in Equation (7):

$$\eta_i = \overline{a}_i - \overline{b}_i. \tag{7}$$

As the amplitude of the target structure's movement increases, $\overline{a}_i$ becomes larger and $\overline{b}_i$ becomes smaller, since more pixels moves larger than the pixel discretization error in Equations (2) and (3). Therefore, η_i is proportional to the amplitude of the target structural movement and is in the range of [-1, 1]. $\eta_i = 1$ indicates that all the pixels in a frame are active pixels, indicating that a significant structural displacement occurs at the measurement point, whereas $\eta_i = -1$ indicates that all the pixels in a frame are inactive and the frame is not appropriate for ROI selection. Hence, several frames with the largest motion indices were selected as key-frames, as shown in Fig. 5(c).

2.2. *Stage 2: Frame area clustering for the optimum ROI selection*

Generally, a natural target area with high feature quality in the recorded frame is intuitively regarded as the ROI. In Stage 2, the optimum ROI is introduced based on the following criteria: (1) the translations of all feature points detected within the ROI are almost identical and (2) sufficient feature points are detected within the ROI.

The first criterion indicates that all the feature points in the optimum ROI need to have the similar distance to the vision camera. The criterion can be quantitatively assessed using the variation of the translation because different translation of two feature points indicates different distances from the vision camera. However, the variation of the translation can also be affected by the distance; it becomes smaller as the distance increases. The CV was used in this study to eliminate the dependency on distance. Here, the smaller the CV value, the more identical is the amplitude of translation. The ROIs chosen based on the first criterion were further examined using the second criterion, which sifts out the ROIs with more feature points. The second criterion is quantified using the FD, which denotes how many feature points are extracted in a unit area of an ROI. The higher the FD value, the more stable is the monitoring.

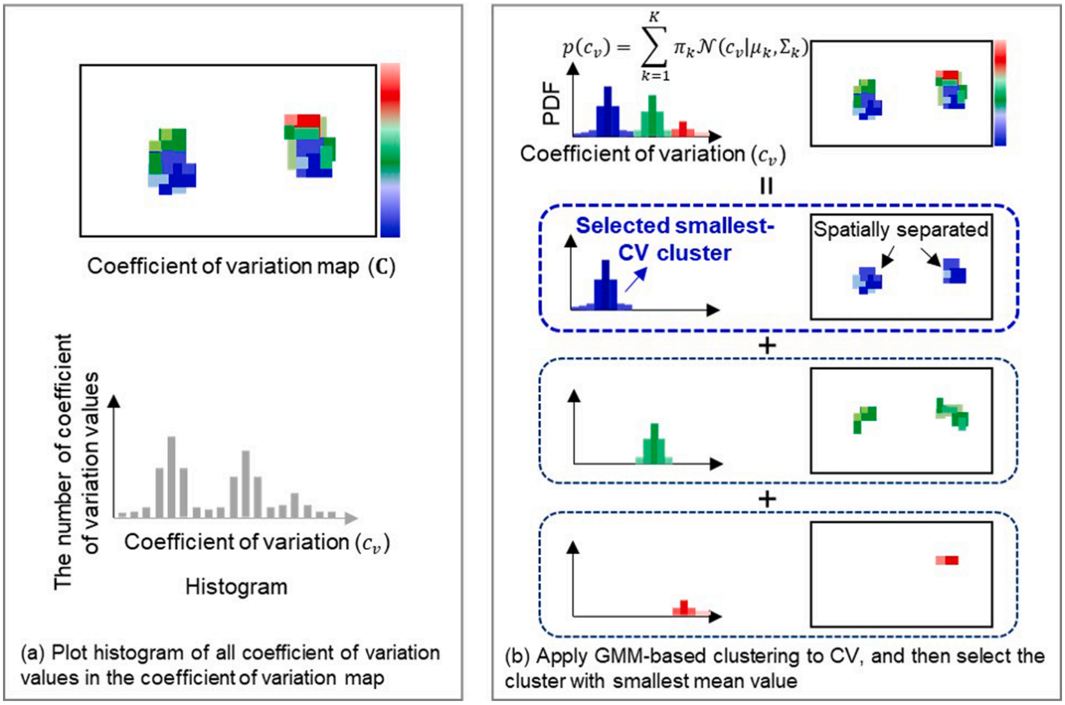
Coefficient of variation map (C)
The number of coefficient of variation values
Coefficient of variation (c_v)
Histogram
(a) Plot histogram of all coefficient of variation values in the coefficient of variation map
$p(c_v) = \sum_{k=1}^{K} \pi_k \mathcal{N}(c_v|\mu_k, \Sigma_k)$
PDF
Coefficient of variation (c_v)
Selected smallest-CV cluster
Spatially separated
+
+
(b) Apply GMM-based clustering to CV, and then select the cluster with smallest mean value

Fig. 8. Selection process of smallest-CV cluster.

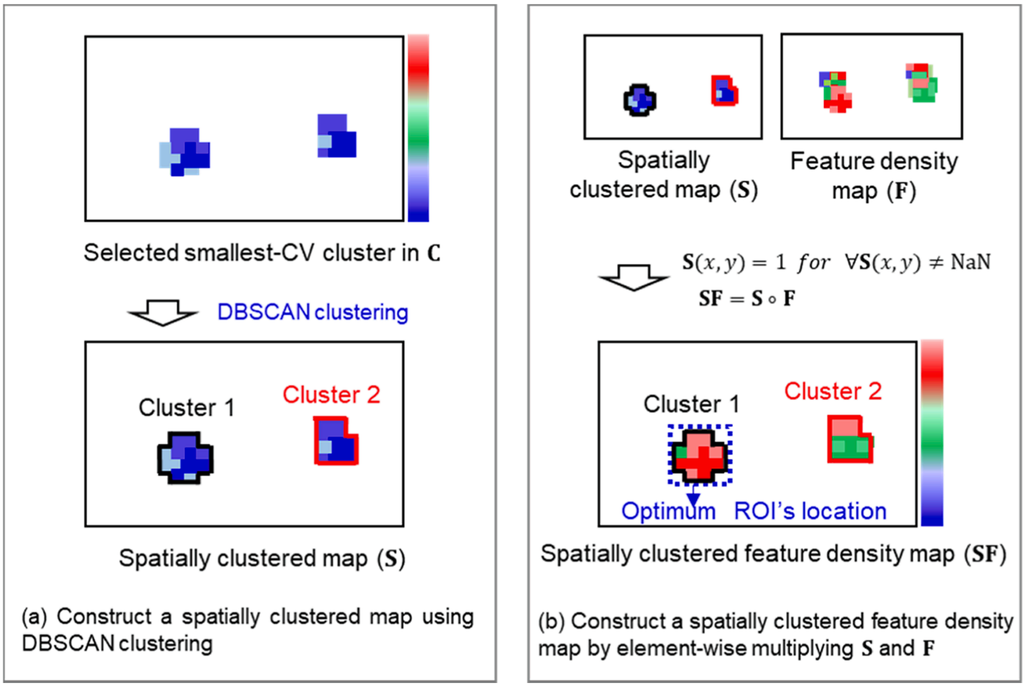
Selected smallest-CV cluster in C
DBSCAN clustering
Cluster 1
Cluster 2
Spatially clustered map (S)
(a) Construct a spatially clustered map using DBSCAN clustering
Spatially clustered map (S)
Feature density map (F)
$\mathbf{S}(x, y) = 1$ for $\forall \mathbf{S}(x, y) \neq \mathrm{NaN}$
$\mathbf{SF} = \mathbf{S} \circ \mathbf{F}$
Cluster 1
Cluster 2
Optimum ROI's location
Spatially clustered feature density map (SF)
(b) Construct a spatially clustered feature density map by element-wise multiplying S and F

Fig. 9. Construction of spatially clustered FD map.

In this section, the procedures for selecting the optimum ROI with a small CV and high FD values are explored in detail based on unsupervised clustering techniques, as shown in Fig. 6. Let u frames be selected as key-frames after Stage 1, active-translation matrices (AP) are generated by setting all the elements of the key-frames to NaN, except for active pixel elements. After the CV and FD maps were constructed from these active-translation matrices (Section 2.2.1), the part of the CV map with the smallest CV value was determined using Gaussian mixture model (GMM)-based clustering (Section 2.2.2). When the cluster with the smallest CV is separated into different pieces, it is further clustered by the density-based spatial clustering of applications with noise (DBSCAN). Finally, a spatially clustered FD map was constructed using the spatially clustered map and FD map, and the cluster with the highest FD was selected as the optimum ROI (Section 2.2.3).

2.2.1. CV and FD map construction using a moving mask

In Section 2.2.1, CV and FD maps were constructed by applying a moving mask to the CV and FD values, which are important indicators of the quality of an ROI (Fig. 7). First, translation matrices ($\mathbf{P}_l(l = 1, \cdots u)$) are constructed for u selected key-frames as shown in Fig. 2(d), and

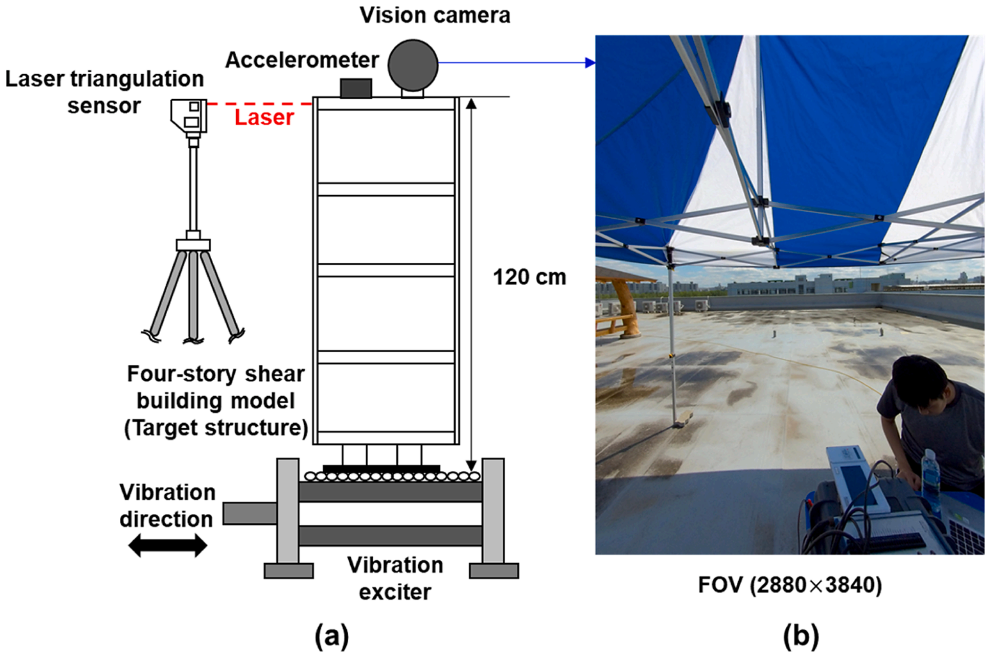

Fig. 10. Overview of the lab-scale test: (a) experimental setup of a four-story shear building model and (b) view from the vision camera.

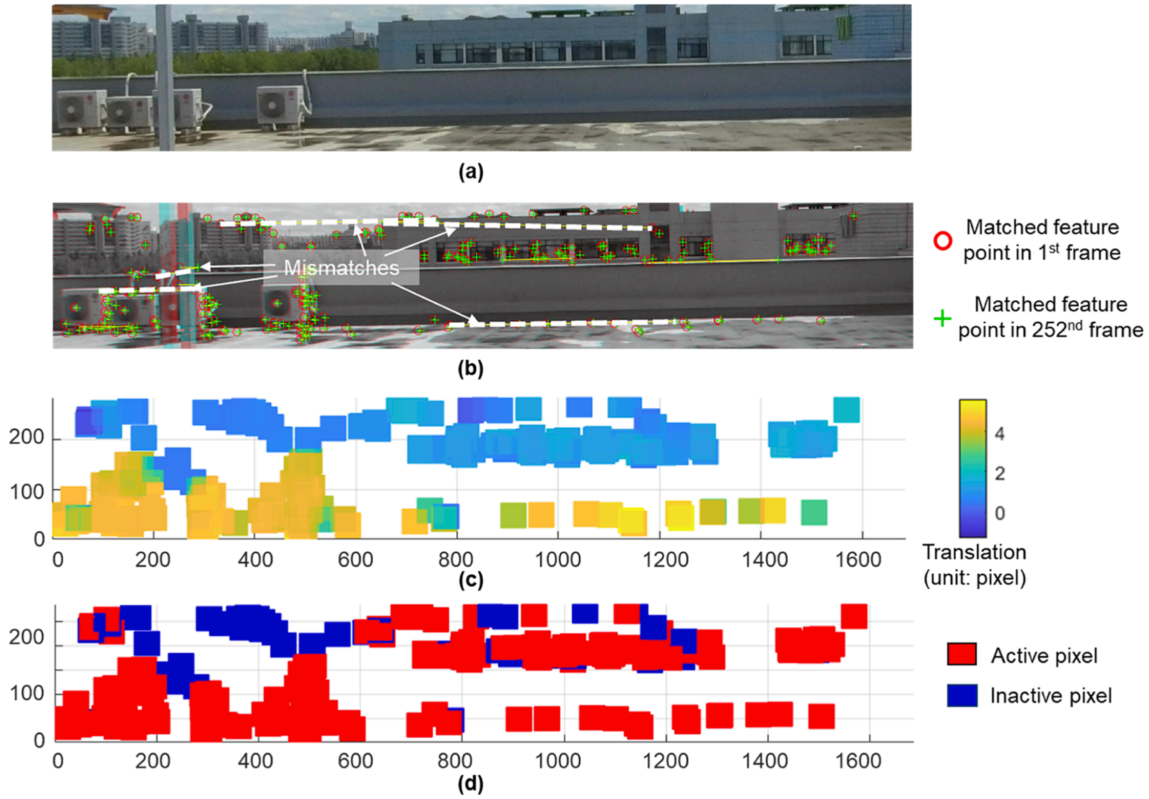

Fig. 11. Constructed translation, active-pixel, and inactive-pixel maps using the 1st and 252nd frames: (a) cropped FOV, (b) overlapped feature matching image between the 1st and 252nd frames, (c) 252nd translation map ($\mathbf{T}_{252}$), and (d) overlapped map between the 252nd active-pixel and inactive-pixel maps ($\mathbf{A}_{252}$ and $\mathbf{B}_{252}$).

active-translation matrices ($\mathbf{AP}_l(l = 1, \cdots u)$) are created by leaving only the active pixels. For each $\mathbf{AP}_l$, a CV map denoted as $\mathbf{C}_l$ was constructed using Equation (8):

$$\mathbf{C}_l\left(r_x, r_y\right) = \frac{\sigma\left[\mathbf{AP}_l(x,y)\middle|(x,y)\in R^2, \mathbf{AP}_l(x,y)\neq \mathrm{NaN}\right]}{E\left[\mathbf{AP}_l(x,y)\middle|(x,y)\in R^2, \mathbf{AP}_l(x,y)\neq \mathrm{NaN}\right]} \tag{8}$$

where $\sigma[\bullet]$ denotes the standard deviation. Using CV values as a criterion for a good ROI reduces the influence of the target-to-camera distance because $\sigma[\mathbf{AP}_l(x,y)]$ and $E[\mathbf{AP}_l(x,y)]$ are reduced by the same proportion as the distance increases. Then, an averaged CV map, denoted as $\mathbf{C}$, was obtained by time-averaging all the CV maps using Equation (9):

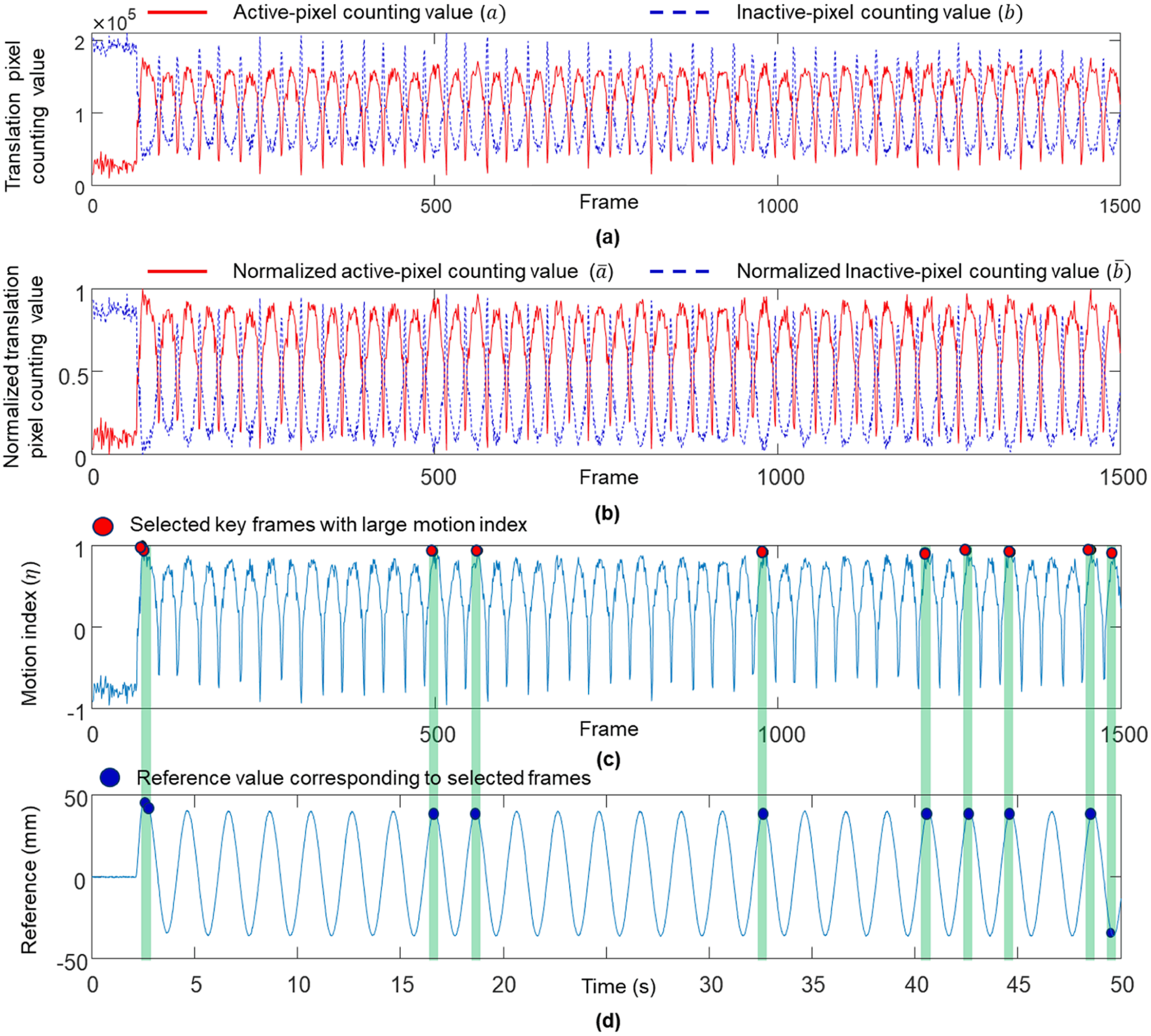

Fig. 12. Motion index calculation results: (a) active- and inactive-pixel numbers, (b) normalized active- and inactive-pixel numbers, (c) calculated motion index, and (d) reference displacement.

$$\mathbf{C}(x,y) = \frac{1}{u}\sum_{l=1}^{u}\mathbf{C}_l(x,y). \tag{9}$$

The next step is to construct an averaged FD map denoted as **F**. First, active-pixel matrices ($\mathbf{Q}_l(l = 1,\cdots u)$) are constructed for u selected key-frames by converting all non-NaN entries of active-translation matrices ($\mathbf{AP}_l(l = 1,\cdots u)$) to 1 as follows:

$$\mathbf{Q}_l(x,y) = \begin{cases} 1 if \mathbf{AP}_l(x,y) \neq NaN \\ NaN otherwise \end{cases}, \; l = 1,\cdots u. \tag{10}$$

$\mathbf{Q}_l$ indicates the position of active pixels in a key-frame and is used to calculate the matched FD because the translation amplitude information is replaced by 1. FD maps ($\mathbf{F}_l(l = 1,\cdots u)$) are constructed by spatial average filtering $\mathbf{Q}_l(l = 1,\cdots u)$:

$$\mathbf{F}_l(r_x,r_y) = E\left[\mathbf{Q}_l(x,y)\middle|(x,y)\in R^2, \mathbf{Q}_l(x,y)\neq \mathrm{NaN}\right], \; l = 1,\cdots u. \tag{11}$$

Finally, **F** is obtained by time averaging all the FD maps:

$$\mathbf{F}(x,y) = \frac{1}{u}\sum_{l=1}^{u}\mathbf{F}_l(x,y). \tag{12}$$

2.2.2. Smallest-CV cluster selection using GMM-based clustering

In Section 2.2.2, the GMM technique is applied to create CV clusters in **C** and to select the smallest-CV cluster among them. Fig. 8 illustrates the construction process of the GMM-based clustered map. The first step is to generate a histogram of all the components in **C** as shown in Fig. 8 (a). Assuming that the CV values of natural targets follow a Gaussian distribution, the GMM is applied to fit the histogram data to cluster CV values using Equation (13) (Fig. 8(b)):

$$p(c_v) = \sum_{k=1}^{K}\pi_k\mathcal{N}\left(c_v|\mu_k,\sum_k\right), \tag{13}$$

where K is the number of Gaussian components determined using the Bayesian information criterion [29]. μ_k and $\sum_k$ denote the mean and covariance matrices of the k^{th} Gaussian component, respectively. $\pi_k(k = 1,\cdots,K)$ are the coefficients of the GMM, which satisfy $0 \leq \pi_k \leq 1$ and $\sum_{k=1}^{K}\pi_k = 1$. The values of μ_k, $\sum_k$, and π_k can be automatically determined using the expectation maximization (EM) algorithm [30]. The GMM clustering divides various CV values into k clusters, and the cluster with the smallest mean value is selected (Fig. 8(b)). Note that every part of a natural target may have different CV values and that different targets may have similar CV values. This indicates that the selected cluster may correspond to different parts of different targets and may be spatially spread out within **C**.

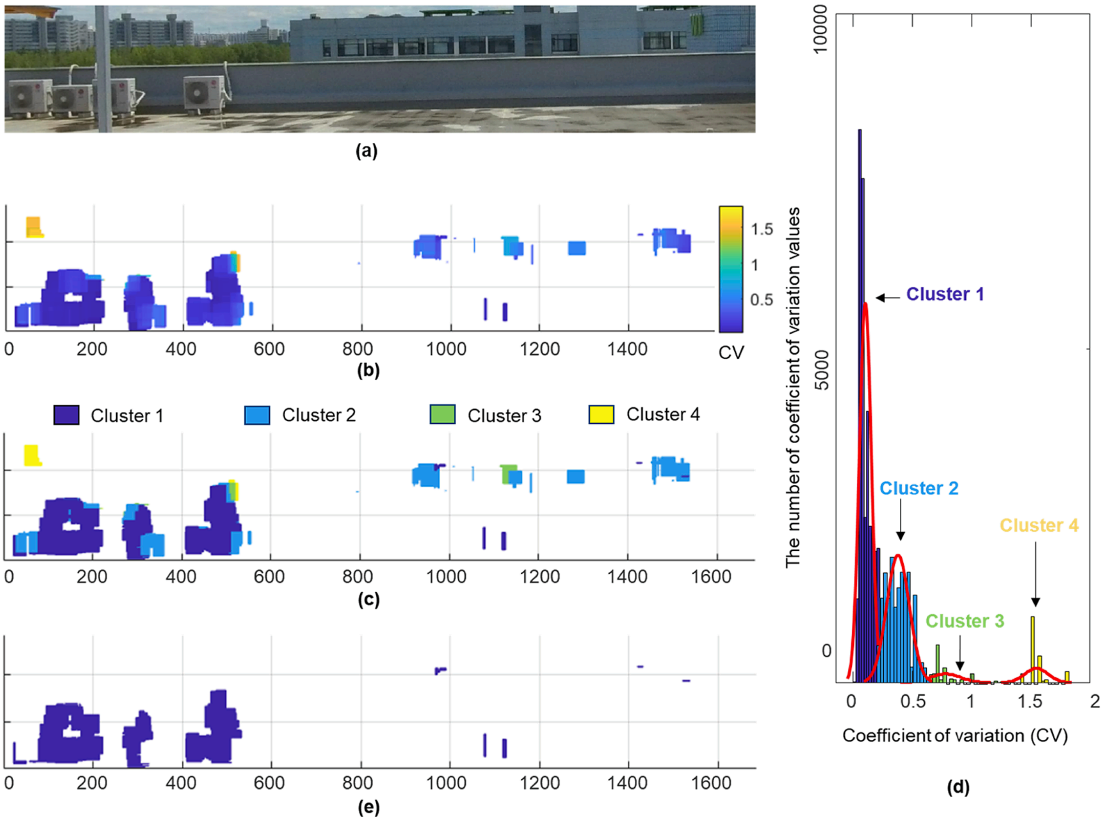

Fig. 13. GMM-based CV clustering under 0.5 Hz sinusoidal signal excitation: (a) cropped FOV, (b) CV map (**C**), (c) GMM-based clustered CV map, (d) histogram of clustered CV map, and (e) selected smallest-CV cluster.

2.2.3. Highest-FD cluster selection for optimum ROI using DBSCAN-based clustering

When the selected smallest-CV cluster area of **C** is separated into pieces, the cluster was further clustered using its FD, and the cluster with the highest FD was selected as the optimum ROI. The selected cluster with the smallest mean CV value may include spatially different FOV areas, as shown in Fig. 8(b). Therefore, DBSCAN [31] was applied to the selected smallest-CV cluster of **C** to construct a spatially clustered map (**S**) as shown in Fig. 9(a). All the clusters of **S** have a similar CV, and the FDs of **S** are then compared for the selection of the optimum ROI. First, all the entries of **S** are converted to 1, except for the NaN values, using Equation (14):

$$\mathbf{S}(x,y) = 1 for \forall \mathbf{S}(x,y) \neq \mathrm{NaN} \tag{14}$$

Then, a spatially clustered FD map (**SF**) is constructed by multiplying each entity of **S** and the corresponding entity of **F** using element-wise multiplication called the Hadamard product [32] as shown in Fig. 9(b):

$$\mathbf{SF} = \mathbf{S}^{\circ}\mathbf{F} \tag{15}$$

where $^{\circ}$ is the Hadamard product operator. To select the optimum ROI, the mean values of all the clusters in **SF** were compared, and the cluster with the highest mean value of FD was selected as the optimum ROI, as shown in Fig. 9(b). Note that, as an ROI should have a rectangular shape, the minimum-bounding rectangle of the selected cluster is set as the boundary of the optimum ROI.

3. Lab-scale test on a four-story shear building model

Fig. 10 shows the overall configuration of the lab-scale test conducted to examine the performance of the proposed ROI selection technique. A four-story shear building model was used as the test structure and placed on an ELECTRO-SEIS APS 400 vibration exciter. An Insta360 Pro 2 vision camera and EpiSensor ES-U2 uniaxial force balance accelerometer were mounted on top of the building model, as shown in Fig. 10(a). Note that the acceleration data from the accelerometer were not used in the proposed ROI selection technique but were used for displacement estimation using the existing technique [24]. The reference displacement was measured using an Optex CD5-W500 laser triangulation sensor with a resolution of 10 μm. The acceleration and reference displacement were digitized and recorded at a sampling rate of 100 Hz using a National Instrument USB-6366 data acquisition system. The vision camera recorded a series of images at a sampling rate of 29.97 Hz with an image size of 2880 × 3840 pixels. To simulate seismic events, three different signals were applied to the vibration exciter to shake the building model in the horizontal direction: (1) a 0.5 Hz sinusoidal signal, (2) a multitone signal having a frequency bandwidth of 0.5–2.5 Hz with a frequency step of 0.5 Hz, and (3) a pseudo-static signal.

3.1. Stage 1: Selection of key-frames with large target structural movement

Vision measurements under 0.5 Hz sinusoidal signal excitation were used for the step-by-step verification of the proposed technique. Fig. 12 shows the constructed translation, active-pixel, and inactive-pixel maps using the 1st and 252nd frames. The FOV was cropped with a size of 1687 × 285 pixels so that undesired features, such as testing equipment and tents in Fig. 10(b), do not interfere with the accuracy (Fig. 11(a)). Fig. 11(b) shows an overlapped feature-matching image between the 1st and 252nd. Here, red circles and green crosses represent matched feature points from the 1st and 252nd frames, respectively. The yellow lines represent properly matched points, and the white dotted lines correspond to extremely mismatched features that induce significantly large translations as outliers. These outliers were removed via the

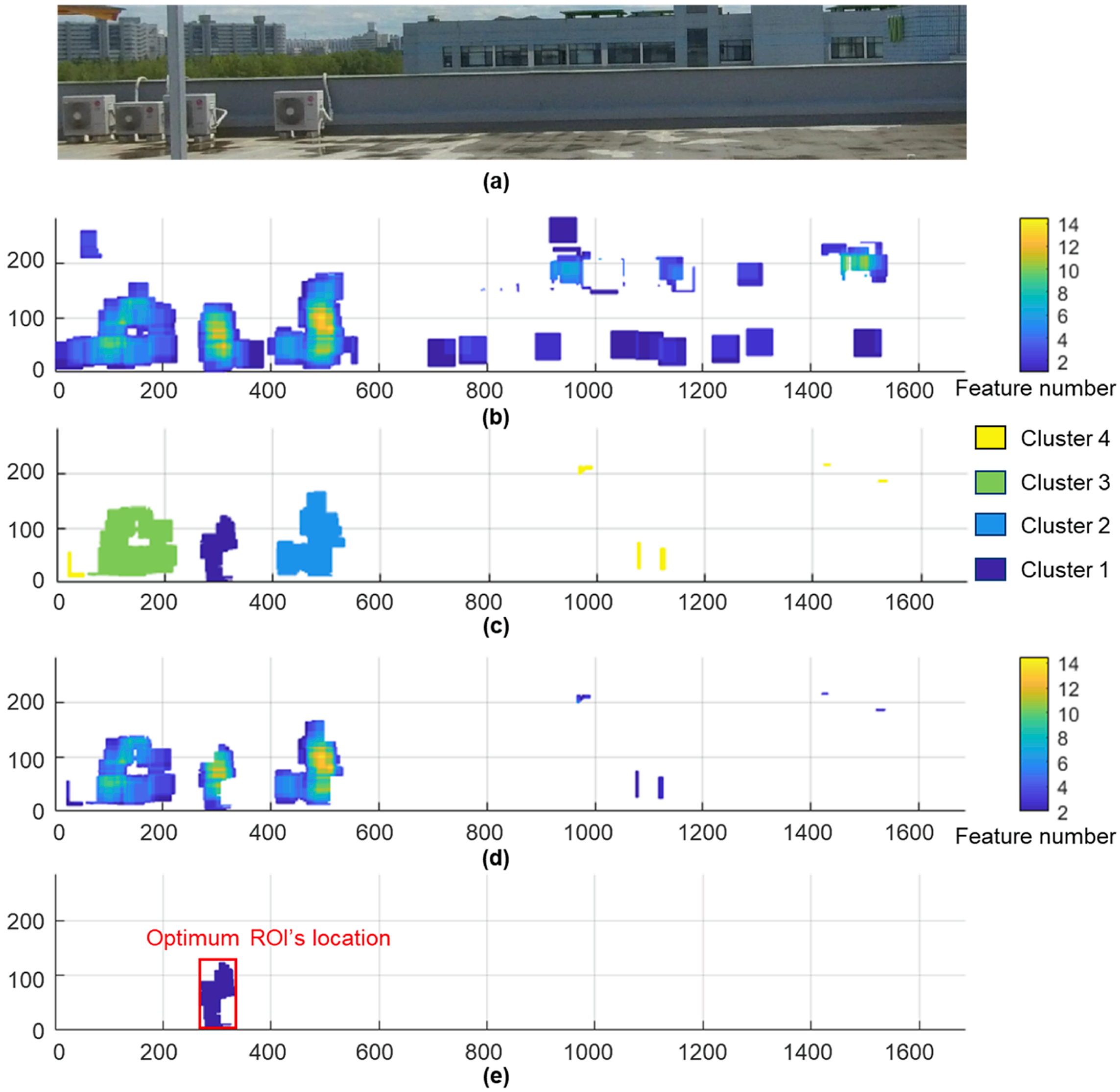

Fig. 14. DBSCAN-based FD clustering under 0.5 Hz sinusoidal signal excitation: (a) cropped FOV, (b) FD map (**F**), (c) spatially clustered map (**S**), (d) spatially clustered FD map (**SF**), and (e) selected highest-FD cluster with the location of the optimum ROI.

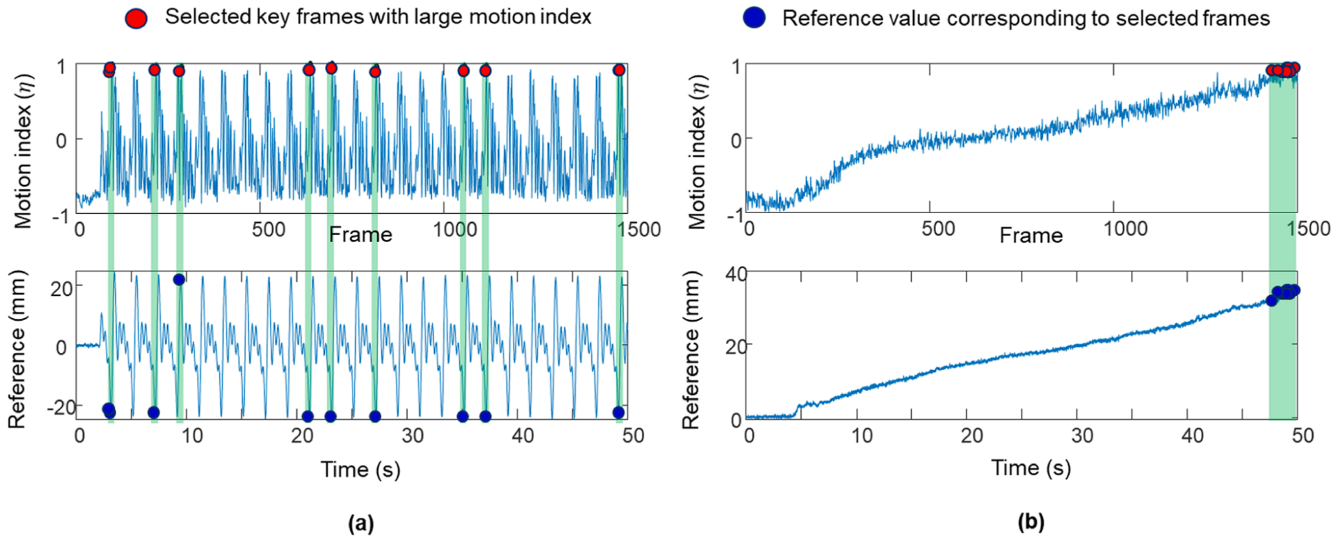

Fig. 15. Monotonic relationship between motion index and reference under: (a) multi-tone and (b) pseudo-static base excitations.

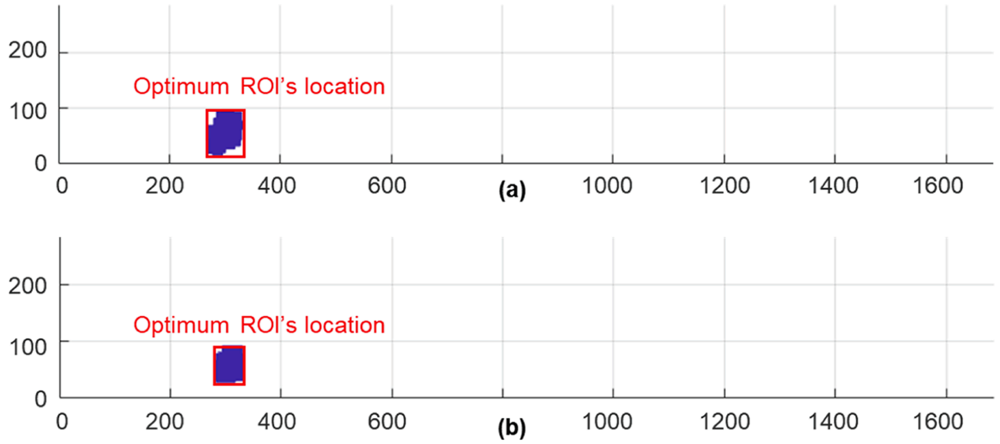

Fig. 16. Selected highest-FD cluster with the location results of the optimum ROI under: (a) multi-tone and (b) pseudo-static base excitations.

Weibull distribution, as they exhibited different colors in the translation map and adversely affected further processing. Fig. 11(c) shows the 252nd translation map ($\mathbf{T}_{252}$), and natural targets have different colors depending on their distances, without the color representing an extremely large translation. Note that the construction of the translation map required a moving mask. If the mask size is too small, the translation map becomes sparse making later clustering difficulty. On the other hand, if the mask size is too large, different natural targets tend to be merged, so its size was manually set to 51 × 51. Fig. 11(d) shows the overlapped map between the active pixel map ($\mathbf{A}_{252}$) and inactive pixel map ($\mathbf{B}_{252}$). The valid elements (not NaN) of $\mathbf{A}_{252}$ correspond to the natural targets close to the building model, whereas the valid elements of $\mathbf{B}_{252}$ correspond to the natural targets at a relatively large distance from the target structure.

By repeating the active-pixel and inactive-pixel counting process described in Section 2.1.2, the numbers of active and inactive pixels were counted for all the frames, as shown in Fig. 12(a). Fig. 12(b) shows the normalized active and inactive pixel numbers ($\overline{a}$ and $\overline{b}$, respectively). Then, the motion index (η) was calculated, as shown in Fig. 12(c), which was compared with the reference displacement (Fig. 12(d)). Here, the green shade area shows that the selected keyframes correspond to large values of reference displacement. A large η corresponds to large structural movement, indicating that η can be used to select key frames captured with large structural movements. In this study, 10 frames corresponding to the largest 10 motion indices were selected, as shown in Fig. 12(c).

3.2. Stage 2: Frame area clustering for the optimum ROI selection

3.2.1. Smallest-CV cluster selection using GMM-based clustering

Fig. 13 shows the GMM-based frame area clustering under 0.5 Hz sinusoidal signal excitation. The natural targets had various CV values in the CV map ($\mathbf{C}$), as shown in Fig. 13(b). The CV values of the outdoor units of air conditioners in Fig. 13(a) are smaller than those of the other areas, indicating that air conditioners have better feature quality than the others. Various CV values in $\mathbf{C}$ were clustered using the GMM with the EM algorithm [30] as shown in Fig. 13(c). Four clusters were obtained, and the mean CV values were 0.1, 0.36, 0.65, and 1.55 for Clusters 1–4, respectively. The histogram of clustered $\mathbf{C}$ is shown in Fig. 13(d), where Cluster 1 has the smallest CV mean value. Fig. 13(e) shows the selected smallest-CV cluster (i.e., Cluster 1). Note that the cluster was spatially separated into several different mapped areas in the cropped FOV.

3.2.2. Highest-FD cluster selection for optimum ROI using DBSCAN-based clustering

Fig. 14 shows the DBSCAN-based feature clustering under 0.5 Hz sinusoidal signal excitation. The FD map ($\mathbf{F}$) based on the key-frames selected in Section 3.2 is shown in Fig. 14(b). Denser feature points were detected from the outdoor units of air conditioners. In the next step, the DBSCAN algorithm was applied to the smallest-CV cluster selected in Section 3.2.1 to construct a spatially clustered map ($\mathbf{S}$), as shown in Fig. 14(c). Clusters 1, 2, and 3 were outdoor units with different air conditioners. The other areas were clustered into Cluster 4. A spatially clustered FD map ($\mathbf{SF}$) is constructed, as shown in Fig. 14(d). The average FDs of Clusters 1 and 2 were higher than those of the other two clusters. The mean FD values were calculated as 7.09, 6.85, 5.29, and 2.52 for Clusters 1–4, respectively. Considering that Cluster 1 had the highest mean FD value, the minimum-bounding rectangle of Cluster 1 was selected as the optimum ROI, as shown in Fig. 14(e).

The proposed technique selects only a few key-frames (Stage 1) from the recorded video and then applies the clustering algorithms (Stage 2) to these selected key-frames instead of all the frames in the recorded video. A total of 1500 frames were obtained during 50-second video recording of the experiment with a frame per second (FPS) of 30. The proposed technique used only 10 selected key-frames, and the optimum ROI selection took about 5 min using a personal computer equipped with Intel i7-6700 3.4 GHz CPU and 8 GB RAM. If all 1500 frames were used, the optimum ROI selection could have taken over 750 min.

3.3. Results of stages 1 and 2 under multi-tone and pseudo-static base excitations

Fig. 15 shows the calculated η and reference displacement under multi-tone and pseudo-static excitations. The figure shows that η is proportional to the amplitude of movement of the target structure not only in multi-tone excitation, where various frequencies are mixed, but also in pseudo-static excitation with extremely low-frequency displacement only. Consequently, the proposed technique effectively chooses appropriate key-frames even though a target structure has low-frequency vibration only, which is one of the most common types of dynamic behavior of typical civil structures [33]. Fig. 16 shows the optimum ROIs selected under multi-tone and pseudo-static signal excitations. The proposed technique obtained consistent results under different structural motions, and the selected optimum ROIs were similar to that under 0.5 Hz sinusoidal signal excitation.

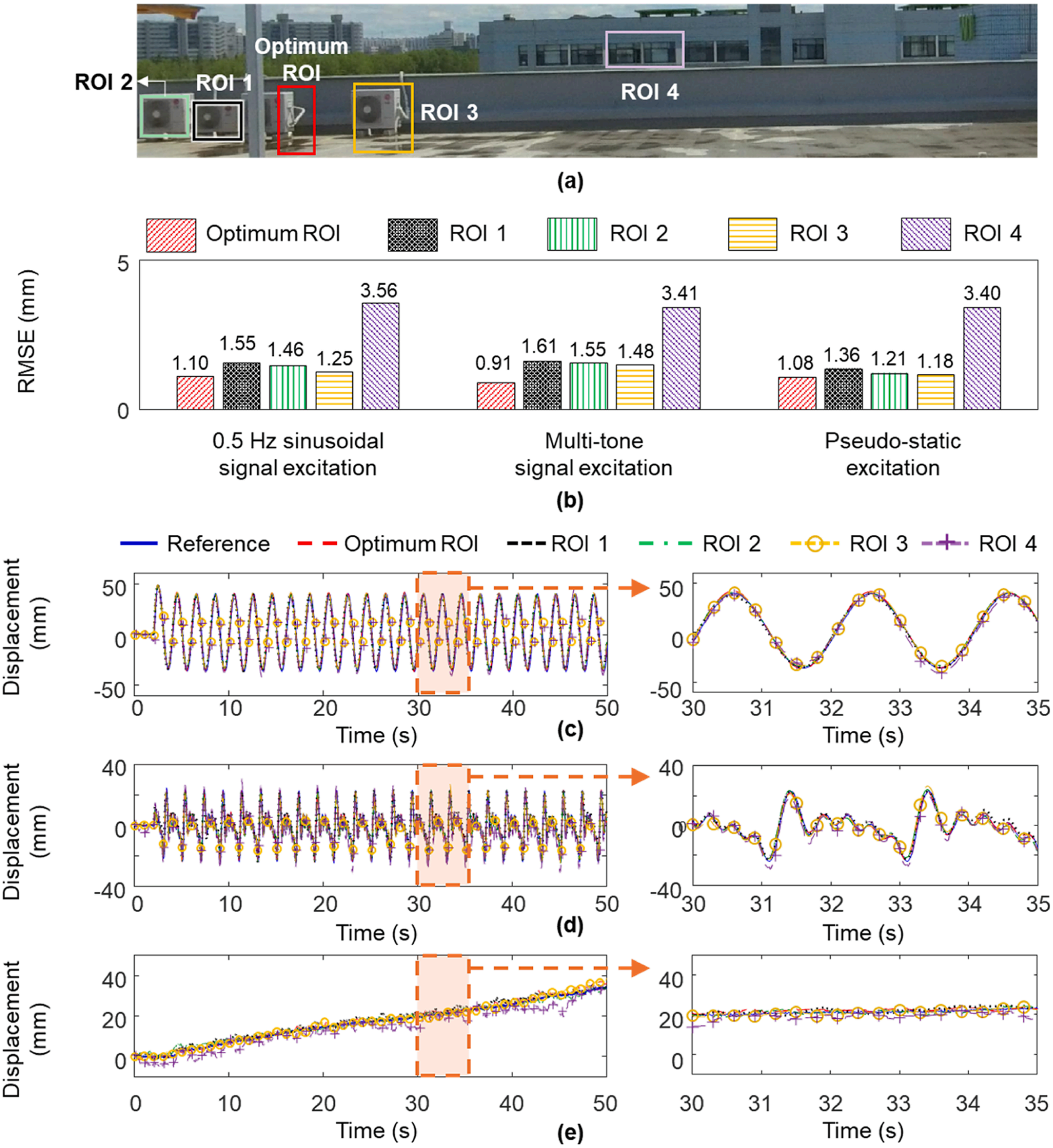

Fig. 17. Displacement of a four-story shear building model estimated using the optimum ROI and other four ROIs: (a) Cropped FOV and selected multiple ROIs, (b) RMSEs of the displacements and translations estimated using different ROIs, and (c)–(e) displacements estimated using different ROIs under 0.5 Hz sinusoidal, multi-tone, and pseudo-static excitations.

Table 1
RMSE reduction achieved by the optimum ROI compared with the other four ROIs in the four-story shear building model test.

Excitations	RMSE reduction (%) compared with different ROIs			
	ROI 1	ROI 2	ROI 3	ROI 4
0.5 Hz sine	29.03	24.66	12	69.1
Multi-tone	43.48	41.29	38.51	73.31
Pseudo-static	20.59	10.74	8.47	68.24
Average	31.03	25.56	19.66	70.22

3.4. Displacement estimation results

Fig. 17 compares the displacement estimated by one of the displacement estimation techniques [24] using the optimum ROI automatically selected by the proposed technique and four other ROIs. To confirm the results of the proposed technique step-by-step in the lab-scale test, the video images captured by the vision camera were cropped, and the cropped FOV with a size of 1687 × 285 pixels is shown in Fig. 17(a). Although ROI 1 and 2 were intuitively selected by the human eye, ROI 3 had the second highest FD in the spatially clustered FD map (**SF**), and ROI 4 contained the cluster with the second smallest mean CV in the CV map (**C**). Fig. 17 (b) shows that the displacements estimated using the optimum ROI had the smallest root mean square errors (RMSEs) compared with those estimated using other ROIs. All ROIs, except ROI 4, were included in the smallest-CV cluster. Considering that CV is defined as the ratio of standard deviation of translation (σ) to the mean of translation (μ) (Equation (8)), the clusters within smaller CVs tend to be close to the camera than these with larger CVs. Since ROI 4 is farther away (approximately 65 m) from the camera than that (approximately 15 m) of the other ROIs, the largest displacement estimation error occurred when the ROI 4 was used. The other four ROIs showed similar results due to their similar distances to the camera. However, the best displacement estimation performance was obtained using the optimum ROI, because the number of the detected feature points within the optimum ROI was largest as shown in Fig. 14(d). Using the optimum ROI, the proposed technique achieved translation errors in the range of [0.11 pixels, 0.15 pixels] and displacement errors in the range of [0.91 mm, 1.10 mm]. Note that here displacement errors were relative to the reference displacements measured using a laser triangulation sensor (LTS), but the translation errors were relative to the translations calculated by dividing the reference displacements by the scale factor. The estimated displacement results are compared in Fig. 17 (c), (d), and (e) to demonstrate the superiority of the displacement estimated using the optimum ROI. The RMSE reductions achieved by the optimum ROI are compared with those of the other four ROIs in Table 1, and up to 70% of the errors were reduced.

4. Field test

Fig. 18 shows an overview of the field test performed on a pedestrian bridge. The bridge shown in Fig. 18(a) is located in Daejeon, Korea, and has a length of 45 m and a width of 8 m. A vision camera and uniaxial force-balance accelerometer identical to those used in the lab-scale test were installed at approximately 1/4 of the span length of the bridge, as shown in Fig. 18(b). A Polytec RSV-150 laser Doppler vibrometer was installed at a stationary location under the bridge to measure the reference displacement. Fig. 18(c) shows the first frame of vision measurement. The pedestrian bridge was excited by (1) four people jumping near the measurement point and (2) 16 people walking slowly across the bridge.

To validate the superiority of the optimum ROI selected using the proposed technique further, vision-based displacements were estimated using the optimum ROI and three other ROIs, and the displacement estimation results were compared with the reference displacement, as shown in Fig. 19 and Table 2. The locations of the four ROIs are shown in Fig. 19(a). Similar to the lab-scale test, ROI 1 was intuitively selected by the naked eye, and ROI 2 corresponded to the cluster with the second-

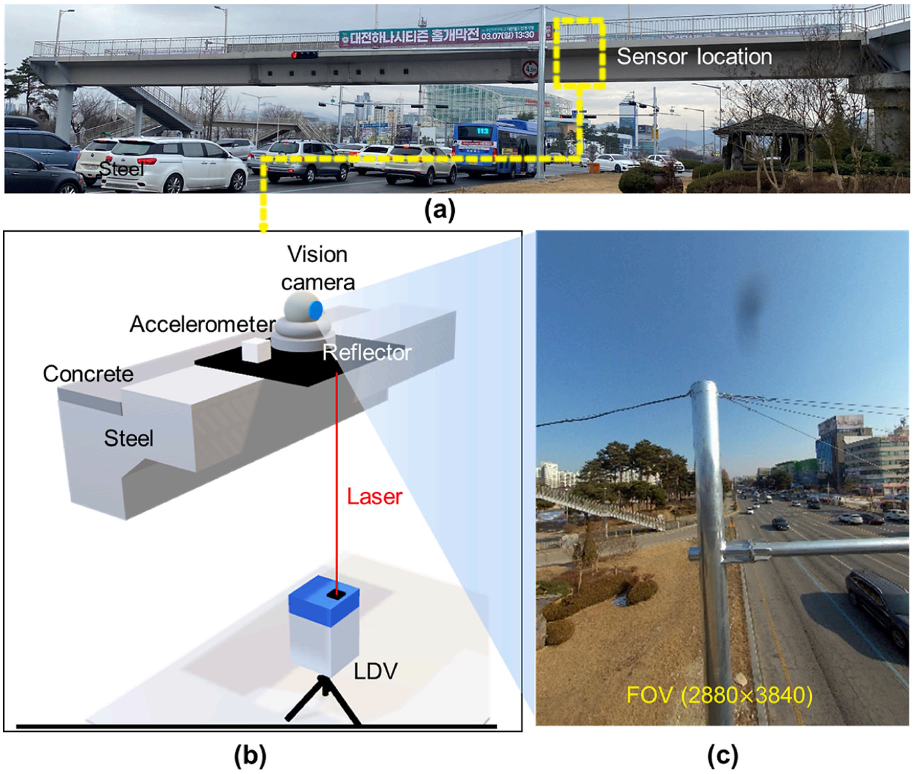

Fig. 18. Overview of the field test: (a) pedestrian steel box girder bridge, (b) sensor setup on the bridge, and (c) view from the vision camera.

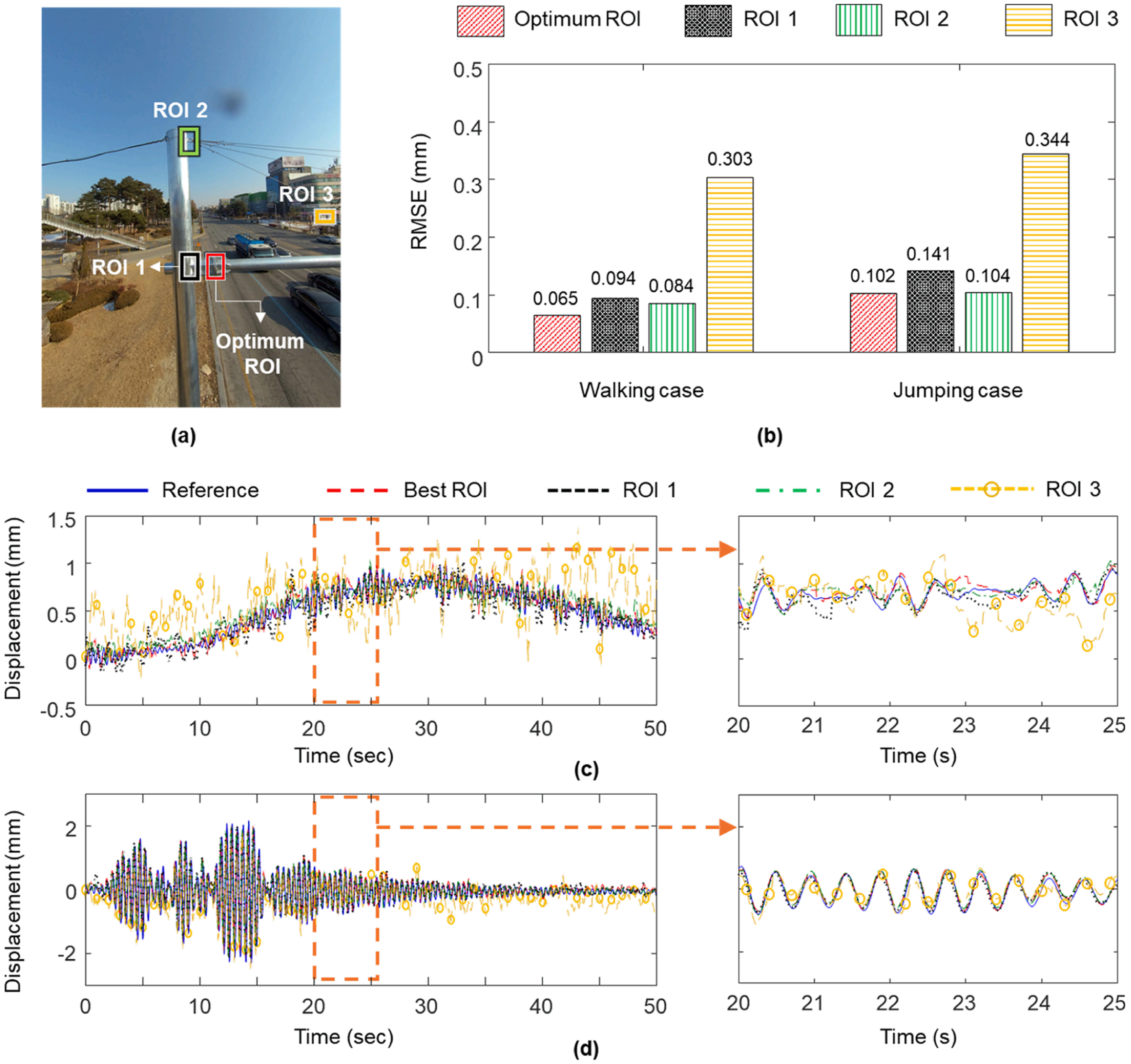

Fig. 19. Estimation results of pedestrian bridge displacement: (a) FOV and selected multiple ROIs, (b) RMSEs of the displacements estimated using different ROIs, (c) and (d) displacements estimated using different ROIs in walking and jumping cases.

Table 2
RMSE reduction achieved by using the optimum ROI compared with other ROIs in the pedestrian bridge field test.

Excitations	RMSE reduction (%) compared with different ROIs		
	ROI 1	ROI 2	ROI 3
Walking	30.85	22.62	78.55
Jumping	27.66	1.92	70.35
Average	29.26	12.27	74.45

highest FD in $\mathbf{SF}_{\text{jumping}}$. As the other clusters (i.e., Cluster 2 in $\mathbf{SF}_{\text{walking}}$ and Cluster 3 in $\mathbf{SF}_{\text{jumping}}$) were too small to act as effective ROIs, a distant target from a nearby building was selected as ROI 3. In both cases, the optimum ROI reduced the RMSEs of the estimated displacements by more than 70% compared with ROI 3. Compared with ROI 1 and 2, the optimum ROI shows 29.26% and 12.27% reductions in RMSE, respectively, although the three ROIs have similar CV values.

Fig. 20(a) and (b) depict the numbers of active and inactive pixels before and after min–max normalization, respectively. As several feature points are detected as distant objects from the bridge, inactive pixels are significantly more abundant than active pixels, as shown in Fig. 20(a). This issue was addressed as shown in Fig. 20(b) after the min–max normalization process for calculating the motion index (η). The calculated η was then compared with the reference displacement (Fig. 20(c)). The key-frame selection results in the figure demonstrate that frames with a large movement of the target structure were successfully chosen by the proposed technique.

Fig. 21(b) and (c) show the spatially clustered FD map results for the walking and jumping cases, respectively. In both cases, the consistency of the proposed technique was confirmed by overlapping the selected optimum ROIs, as shown in Fig. 21(a). In Fig. 21(c), the second-highest cluster indicates the position of ROI 2 in Fig. 19(a), and it shows the second-best displacement estimation result, as shown in Fig. 19(b).

5. Conclusion

This paper proposed an automated ROI selection technique for computer-vision-based structural displacement estimation. The proposed technique first selected several key-frames, and CV and FD maps were constructed using the selected key-frames. These two maps were clustered using unsupervised GMM and DBSCAN clustering algorithms, and the cluster with the smallest CV value and the highest FD value was selected as the best ROI. The performance of the proposed technique was

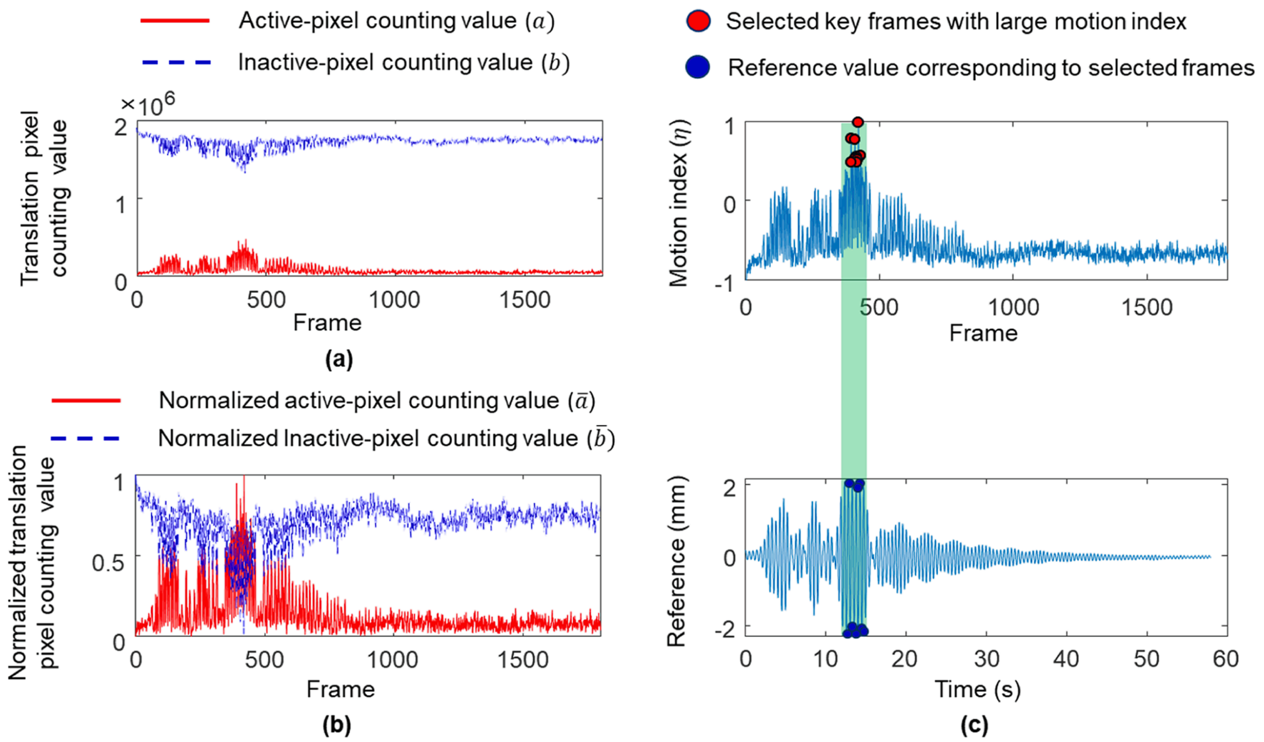

Fig. 20. Motion index calculation results in the jumping case: (a) active- and inactive-pixel numbers, (b) normalized active- and inactive-pixel numbers, and (c) calculated motion index and reference displacement.

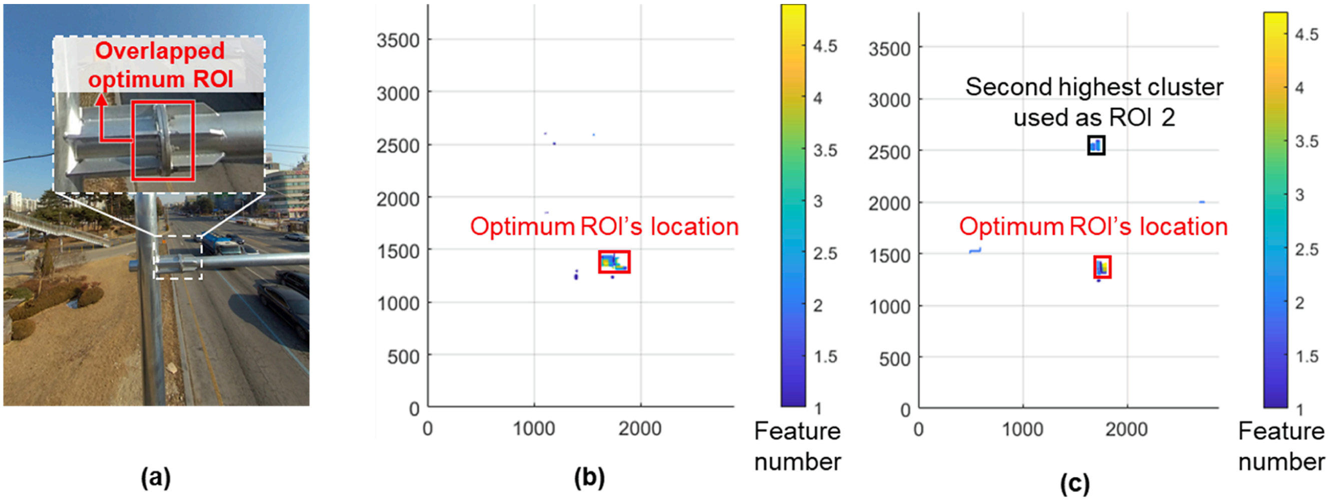

Fig. 21. Spatially clustered FD maps of the field test: (a) view from the vision camera with overlapped optimum ROI of both cases, (b) walking case ($\mathbf{SF}_{\text{walking}}$), and (c) jumping case ($\mathbf{SF}_{\text{jumping}}$).

validated through a laboratory test on a four-story shear building and a field test on a pedestrian bridge. The selected optimum ROIs enabled the best displacement estimation performance compared with other intuitively selected ROIs, and the overall RMSEs of displacements estimated using the optimum ROIs were less than 2 mm in both tests, indicating that the proposed technique successfully selected a high-reliability ROI. However, the proposed technique requires manual determination of the mask size and assumes that all natural targets within the FOV are stationary. Future works are warranted to optimize the mask size and consider non-stationary target with the FOV. We also plan to improve the applicability and efficiency of the proposed technique further in the future. As the optimum ROI cannot be updated in this study after a short-time vision measurement is completed, the optimum ROI updating technique is being further studied for time-variant ROI detection. In addition, a new formulation of the vision camera position is explored to incorporate the rotation of the camera in the proposed technique.

CRediT authorship contribution statement

Jaemook Choi: Conceptualization, Methodology, Software, Validation, Writing – original draft. **Zhanxiong Ma:** Validation, Writing – review & editing. **Kiyoung Kim:** Writing – review & editing. **Hoon Sohn:** Supervision, Writing – review & editing, Funding acquisition.

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.

Acknowledgement

This work was supported by National Research Foundation of Korea (NRF) grants funded by the Korean government (MSIT) [grant numbers 2017R1A5A1014883 and 2022R1C1C2008186].

References

[1] K. Santhosh, B.K. Roy, Online implementation of an adaptive calibration technique for displacement measurement using LVDT, Appl. Soft Comput. 53 (2017) 19–26.

[2] M. Gindy, R. Vaccaro, H. Nassif, J. Velde, A state-space approach for deriving bridge displacement from acceleration, Comput. Aided Civ. Inf. Eng. 23 (2008) 281–290.

[3] D. Hester, J. Brownjohn, M. Bocian, Y. Xu, Low cost bridge load test: Calculating bridge displacement from acceleration for load assessment calculations, Eng. Struct. 143 (2017) 358–374.

[4] J. Yu, X. Meng, X. Shao, B. Yan, L. Yang, Identification of dynamic displacements and modal frequencies of a medium-span suspension bridge using multimode GNSS processing, Eng. Struct. 81 (2014) 432–443.

[5] H.H. Nassif, M. Gindy, J. Davis, Comparison of laser Doppler vibrometer with contact sensors for monitoring bridge deflection and vibration, NDT and E Int. 38 (2005) 213–218.

[6] C. Gentile, G. Bernardini, Output-only modal identification of a reinforced concrete bridge from radar-based measurements, NDT and E Int. 41 (2008) 544–553.

[7] C. Gentile, G. Bernardini, An interferometric radar for non-contact measurement of deflections on civil engineering structures: laboratory and full-scale tests, Struct. Infrastruct. Eng. 6 (2010) 521–534.

[8] D.T. Bartilson, K.T. Wieghaus, S. Hurlebaus, Target-less computer vision for traffic signal structure vibration studies, Mech. Syst. Sig. Process. 60 (2015) 571–582.

[9] Y.-J. Cha, J.G. Chen, O. Büyüköztürk, Output-only computer vision based damage detection using phase-based optical flow and unscented Kalman filters, Eng. Struct. 132 (2017) 300–313.

[10] D. Feng, M.Q. Feng, Computer Vision for Structural Dynamics and Health Monitoring, John Wiley & Sons, 2021.

[11] D. Jana, S. Nagarajaiah, Computer vision-based real-time cable tension estimation in Dubrovnik cable-stayed bridge using moving handheld video camera, Struct. Control Health Monit. 28 (2021) e2713.

[12] Z. Ma, J. Choi, H. Sohn, Noncontact cable tension force estimation using an integrated vision and inertial measurement system, Measurement 199 (2022), 111532.

[13] Y. Han, G. Wu, D. Feng, Vision-based displacement measurement using an unmanned aerial vehicle, Struct. Control Health Monit. 29 (2022) e3025.

[14] Y. Han, G. Wu, D. Feng, Structural modal identification using a portable laser-and-camera measurement system, Measurement 214 (2023), 112768.

[15] D.V. Jáuregui, K.R. White, C.B. Woodward, K.R. Leitch, Noncontact photogrammetric measurement of vertical bridge deflection, J. Bridg. Eng. 8 (2003) 212–222.

[16] T. Khuc, F.N. Catbas, Computer vision-based displacement and vibration monitoring without using physical target on structures, Struct. Infrastruct. Eng. 13 (2017) 505–516.

[17] L. Luo, M.Q. Feng, Edge-enhanced matching for gradient-based computer vision displacement measurement, Comput. Aided Civ. Inf. Eng. 33 (2018) 1019–1040.

[18] P.L. Reu, D.P. Rohe, L.D. Jacobs, Comparison of DIC and LDV for practical vibration and modal measurements, Mech. Syst. Sig. Process. 86 (2017) 2–16.

[19] D. Diamond, P. Heyns, A. Oberholster, Accuracy evaluation of sub-pixel structural vibration measurements through optical flow analysis of a video sequence, Measurement 95 (2017) 166–172.

[20] S. Bhowmick, S. Nagarajaiah, Spatiotemporal compressive sensing of full-field Lagrangian continuous displacement response from optical flow of edge: Identification of full-field dynamic modes, Mech. Syst. Sig. Process. 164 (2022), 108232.

[21] C.-Z. Dong, O. Celik, F.N. Catbas, Marker-free monitoring of the grandstand structures and modal identification using computer vision methods, Struct. Health Monit. 18 (2019) 1491–1509.

[22] J.J. Lee, M. Shinozuka, A vision-based system for remote sensing of bridge displacement, NDT and E Int. 39 (2006) 425–431.

[23] S. Yu, J. Zhang, Fast bridge deflection monitoring through an improved feature tracing algorithm, Comput. Aided Civ. Inf. Eng. 35 (2020) 292–302.

[24] Z. Ma, J. Choi, H. Sohn, Real-time structural displacement estimation by fusing asynchronous acceleration and computer vision measurements, Comput. Aided Civ. Inf. Eng. 37 (2022) 688–703.

[25] Z. Ma, J. Choi, P. Liu, H. Sohn, Structural displacement estimation by fusing vision camera and accelerometer using hybrid computer vision algorithm and adaptive multi-rate Kalman filter, Autom. Constr. 140 (2022), 104338.

[26] H. Bay, T. Tuytelaars, L.V. Gool, Surf: Speeded up robust features, in: European Conference on Computer Vision, Springer, 2006, pp. 404–417.

[27] D.G. Lowe, Distinctive image features from scale-invariant keypoints, Int. J. Comput. Vis. 60 (2004) 91–110.

[28] E. Rublee, V. Rabaud, K. Konolige, G. Bradski, ORB: an efficient alternative to SIFT or SURF, in: 2011 International Conference on Computer Vision, IEEE, 2011, pp. 2564–2571.

[29] G. Schwarz, Estimating the dimension of a model, Ann. Stat. (1978) 461–464.

[30] C.M. Bishop, N.M. Nasrabadi, Pattern Recognition and Machine Learning, Springer, 2006.

[31] D. Birant, A. Kut, ST-DBSCAN: an algorithm for clustering spatial–temporal data, Data Knowl. Eng. 60 (2007) 208–221.

[32] R.A. Horn, The hadamard product, Proc. Symp. Appl. Math. (1990) 87–169.

[33] J. Lovse, W. Teskey, G. Lachapelle, M. Cannon, Dynamic deformation monitoring of tall structure using GPS technology, J. Surv. Eng. 121 (1995) 35–40.