Displacement measurement and nonlinear structural system identification: A vision-based approach with camera motion correction using planar structures

Jiao, Jian; Guo, Jia; Fujita, Kohei; Takewaki, Izuru


ISSUE DATE:
2021-08

URL:
http://hdl.handle.net/2433/269418

This is the peer reviewed version of the following article: [Jiao, J, Guo, J, Fujita, K, Takewaki, I. Displacement measurement and nonlinear structural system identification: A vision-based approach with camera motion correction using planar structures. Struct Control Health Monit. 2021; 28:e2761], which has been published in final form at https://doi.org/10.1002/stc.2761. This article may be used for non-commercial purposes in accordance with Wiley Terms and Conditions for Use of Self-Archived Versions. This article may not be enhanced, enriched or otherwise transformed into a derivative work, without express permission from Wiley or by statutory rights under applicable legislation. Copyright notices must not be removed, obscured or modified. The article must be linked to Wiley’s version of record on Wiley Online Library and any embedding, framing or otherwise making available the ...
Displacement measurement and nonlinear structural system identification: a vision-based approach with camera motion correction using planar structures

Jian Jiao\textsuperscript{1}, Jia Guo\textsuperscript{*,2}, Kohei Fujita\textsuperscript{1}, and Izuru Takewaki\textsuperscript{1}

\textsuperscript{1}Department of Architecture and Architectural Engineering, Kyoto University, Kyoto, Japan

\textsuperscript{2}International Research Institute of Disaster Science, Tohoku University, Sendai, Japan

Abstract

Knowledge of the camera motion is important in the applications of structural dynamic response measurement using a vision-based approach because in most of the field measurements, such motion can be non-trivial and the accuracy of dynamic response measurements is strongly affected by the camera motion. This paper presents a new framework for camera motion estimation and vision-based displacement measurement, which greatly lowers the barriers to the application of generally positioned cameras from strictly stationary cameras. The contributions in this paper are twofold. First, camera motion as well as reconstructed structural displacement are calculated based on reference planar structures visible in a given scene. In this case, a homography is more effective at describing view changes and planar geometric constraints can be incorporated early in the reconstruction process, thereby improving the quality and effectiveness of the estimates. The second contribution is that the homography of the planar structure is estimated by a newly proposed tracking algorithm that combines RANSAC algorithm and Efficient Second-order Minimization (ESM) technique, which refines the final estimates to sub-pixel accuracy and avoids tracking drift and non-smoothness effectively. Experimental results indicate that the quality of the camera motion estimation and displacement reconstruction can be significantly improved by the judgmatical use of the proposed algorithm for planar structure homography estimation. Furthermore, nonlinear structural system identification is carried out to additionally verify the proposed algorithm using unscented Kalman filter.

Keywords: vision-based measurement, camera motion estimation, planar structures, nonlinear system identification, image segmentation

1 Introduction

The field of vision-based optical techniques has attracted significant attention in the structural engineering community over the past decades, due to the prominent merits of this promising technique, such as remote contactless measurement, quick and automatic measurement, as well as real-time measurement and visualization \cite{1,2,3}. Camera motion and object motion are two main sources of dynamic information captured by a camera system in the vision-based techniques. The camera motion, comprising of pan, tilt, zoom and the combination of these basic components, is introduced by the annoying camera supported platform vibrations. Large camera motion results in strongly changing field of view between two consecutive images. The objective

\textsuperscript{*}Correspondence to: Jia Guo, Email: guo@irides.tohoku.ac.jp
motion, e.g. dynamic displacement responses of the structure, believes to be of importance to the evaluation of a mechanical system and thus is the key information to be identified from the images. Knowledge of the camera motion is crucial for the applications of structural dynamic response measurement, for the reason that in most of the field measurements, especially in a seismic-induced monitoring system, camera motion can be non-trivial and the accuracy of object motion estimation is strongly affected by the camera motion caused by the vibrating supported platform [4–6].

Much of the recent effort in vision-based measurement approach has been devoted to attempting to confront the challenges of camera motion correction and displacement reconstruction using un-stationary cameras. To overcome the slight camera position change due to wind, oscillations and the lack of stability of ground, Yoneyama et al. [7] applied a camera movement correction strategy using an equation of perspective transformation which describes the relationship between images before and after the camera movement. The effectiveness of this approach was validated by the deflection measurement of a wide-flange beam as well as the bridge deflection measurement. Cheng et al. [8] proposed a practical approach for high-rise building seismic behavior monitoring using surveillance camera with metric rectification technique during the Great East Japan Earthquake. Since it is just a preliminary study, the accuracy of structural seismic response was hardly fully assessed. Structural displacement measurement using un-stationary camera has also been applied with some success in Unmanned Aerial System (UAS). Reagan et al. [9] combined the use of the UAS and three-dimensional digital image correlation to perform full-field displacement measurements for long-term structural health monitoring of bridges and quantified the measuring capabilities with an extensive set of laboratory and bridge field tests. Yoon et al. [10, 11] and Hoskere et al. [12] presented a framework to achieve absolute displacement of a structure from a UAS equipped with consumer-grade camera. Displacement of a pinned-connected railroad truss bridge subjected to revenue-service traffic loading was shown to be recovered with reasonable accuracy by the proposed system. In the field of 3D reconstruction, Doerr et al. [13] developed a methodology for image-based tracking of seismically induced motions using multiple shaking high-speed cameras with a unique system to acquire, synchronize and time stamp image data at certain sustained rates.

In most of the previous works, to effectively reconstruct the motion of the object, two primary steps are involved: structure motion tracking within local image coordinate, and camera motion estimation which builds the correspondences between 2D images and the real 3D world. In the structure motion tracking step, natural features or artificial markers on the measurement point of the structure are normally tracked by template matching algorithms with subpixel accuracy. Camera motion estimation, on the other side, is always accomplished by tracking features rigid to the background of the image. The pose of the camera, however, is shown to be sensitive to feature detection and tracking, and the quality of its estimation might not have acceptable accuracy for images that do not contain enough/abundant sets of good features to track in the background [14]. Recent research has shown that using the planar structures visible in a given scene to compute the camera motion leads to an improvement of structure and motion reconstruction in terms of accuracy, stability and rate of convergence [15]. The reason is that prior geometric constraints introduced by the planar structures can be directly incorporated into the reconstruction process, thereby improving the quality of the estimates. As man-made environments usually contain various kinds of planar structures (e.g. building facades, floors), this approach has been rapidly expanded in recent years, especially in the field of robotics and Augmented Reality (AR) [16–18].

In this paper, we estimate the camera motion and reconstruct the structural displacement of a video sequence captured by the un-stationary camera by taking the planar structure based tracking framework of Mei et al. [19], and develop a novel homography estimation algorithm that combines RANSAC algorithm and Efficient Second-order Minimization (ESM) technique.
To validate the effectiveness and accuracy of the proposed framework, it is applied to a two-story base-isolated shear building with a shaking camera system. Our experimental results convincingly show that by using the above two strategies, tracking accuracy and robustness can be enhanced significantly.

The main contributions in this paper are twofold.

- The camera motion as well as the corresponding reconstructed structural displacement are directly calculated based on planar structure/template tracking. In this case, a homography is more effective at describing view changes than the associated epipolar geometry and planar geometric constraints are naturally included to improve the results with very little additional complexity. Besides, by tracking the visible planar templates in the scene, we take advantage of a rather simpler strategy for camera motion estimation (i.e. directly from a decomposition of the homography matrix) than that would be the case with a general structure-and-motion system.

- Homography estimation is conducted using the following phased approach: the initial guess of the homography is first obtained by RANSAC algorithm within each reference planar template. Next, the homography is being refined by the gradient-based ESM tracking algorithm. Since the ESM algorithm is based on minimization of the sum-of-squared-difference between two planar templates by only making use of the whole area’s intensity information, rather than the sparsely detected features, it achieves better accuracy. Although our proposed algorithm processes in an iterative minimization way, we show that the proposed algorithm has a higher convergence rate (like the newton method) and can be well adapted to real-time measurement applications.

The remainder of this paper is organized as follows. Section 2 establishes a framework for camera motion correction and vision-based measurement using homography-based tracking method. Section 3 then shows how homography can be accurately estimated with no loss of computational efficiency by our RANSAC-ESM combined method. Section 4 presents experimental results of a sliding structure to testify the proposed approach and nonlinear structural system identification is additionally carried out based on the Unscented Kalman filter (UKF) using the measured displacements obtained by vision system and the measured accelerations obtained by accelerometers, respectively. Final conclusions are drawn in Section 5.

2 Vision-based measurement system

In this section, we present the vision-based approach through which camera pose and 3D point position of the tracked object are recovered based on the reference planar structures in a video sequence. The basic idea is to recover the 3D world system by tracking the homography of each reference plane so that the relative motion between the world system and the object can
Figure 2 (a) Source image (b) Example of image segmentation

be calculated. Figure 1 illustrates the overall process of the proposed approach, and each step is described in detail in the following subsections.

2.1 Image segmentation

To reconstruct the 3D dynamic world system, the first step is to identify the region with planar structures on the full image. Homography estimation are then restricted to this region. This is always accomplished by the image segmentation technique. Image segmentation involves converting an image into a collection of regions of pixels by a labeled image, as shown in Figure 2. In this section, we present two segmentation methods to pave the way for planar structure identification.

- Normally used threshold segmentation
  Since different planes belonging to different objects always have sufficient contrast from each other in man-made environments (e.g. walls and floors), the simplest and most efficient threshold segmentation approach, which is to segment different regions using their pixel values, can be applied. Based on this approach, the pixel values below or above a threshold can be classified accordingly and the edges of different regions can be identified. Morphological filters are then adopted to remove noise from edges for each object. Complete procedure is described in detail in section 4.2.

- Automatic planar scene segmentation
  As planar structures are special geometries with multiple co-planar points, automatic planar segmentation can be applied directly. After generating multiple plane hypotheses using random sampling points, the RANSAC procedure [20] is introduced to give the homography that corresponds to the largest set of co-planar points and thus selects the most likely planes with respect to images automatically. Detailed discussion of these methods is beyond the scope of this paper. Readers are referred to literature [21].

2.2 Camera motion estimation

In the following expressions, we define one 3D coordinate system and two 2D coordinate systems as shown in Figure 3: a 3D dynamic world coordinate system, where points are denoted by homogeneous coordinates with upper case letters $X = (X, Y, Z, 1)^T$; two 2D image plane coordinate systems, representing camera at different time (e.g. first $I_0$ for $t = 0$ and current $I_k$ for $t = t_k$), where points are denoted with lower case letters $x = (x, y, 1)^T$. In the dynamic world system, to obtain the camera motion as well as the corrected 3D object motion of point $\tilde{X}$, the rigid reference plane $\pi_i$ ($i = 1, \cdots, N$) is introduced, which is first identified by image segmentation strategy according to section 2.1 and then tracked throughout the video sequences.
Figure 3 Overview of the dynamic coordinate system

Under the pinhole camera model, if the world coordinate system is aligned with the first image plane system, the point \( X^i = (X^i, Y^i, Z^i, 1)^T \) on the \( i \)th reference plane is mapped to \( x^i_0 \) on \( I_0 \) and \( x^i_k \) on \( I_k \) by the following relations

\[
x^i_0 = P_0 X^i_0 = K (I_0) X^i_0 \quad x^i_k = P_k X^i_k = K [R \ t] X^i_k
\]  

(1)

where \( X^i_0 = X^i_k = X^i \), since \( X^i \) lies on the reference plane rigid to the world system. \( K \) is the camera intrinsic matrix, and \([R \ t]\) are the camera extrinsic parameters. The parameters \( R \) and \( t \) relate the current camera orientation and position with respect to the first one. \( P = K [R \ t] \) is the \( 3 \times 4 \) camera projection matrix. According to Eq. (1), it is possible to obtain the fundamental relation that links the projection of \( X^i \) on both image planes:

\[
x^i_k = KR K^{-1} x^i_0 + (Z^i)^{-1} K t
\]  

(2)

Supposing that the point \( X_i \) is on the plane \( \pi_i \), we have \( n^T_i (X^i, Y^i, Z^i)^T = n^T_i Z^i K^{-1} x^i_0 = -d_i \), where \( n_i \) is the unit normal and \( d_i \) is the distance from the origin to the plane. The reference plane can thus be represented as \( \pi_i = [n^T_i, d_i]^T \) in the world coordinate system and \( Z^i \) is given by

\[
(Z^i)^{-1} = -n^T_i K^{-1} x^i_0 / d_i
\]  

(3)

By plugging Eq. (3) into Eq. (2), the homography \( H^i_k \) induced by plane \( \pi_i \) is achieved by [22]

\[
x^i_k = H^i_k x^i_0
\]  

(4)

where

\[
H^i_k = K (R - t n^T_i / d_i) K^{-1}
\]  

(5)

If the image points are normalized with respect to camera internal matrix \( K \), the homography becomes

\[
H^i_k = R - t n^T_i / d_i
\]  

(6)

We present the algorithm that accurately computes the homography between two image planes in the next section. Having \( H^i_k \) of the form \( H^i_k = R - t n^T_i / d_i \), it can be decomposed into parameters \( \{R, t, n^T_i / d_i\} \). There are in general four solutions to this decomposition and at most two are physically valid by imposing the positive depth constraint. The ambiguity between the two remaining solutions can be solved by using several frames [23]. For the general case with multiple reference planes, the motion of the camera \( R \) and \( t \) can be either obtained by averaging over the individual motions from each plane homography [24] or be from the decomposition of a united homography with the constraint of the same camera motion [19].
2.3 3D point reconstruction

Once the camera motion $[R \ t]$ is determined, the camera projection matrix $P$ can be estimated by $P = K[R \ t]$ and the 3D coordinate of the tracking point $X$ from the object is computed using a stereo algorithm as described below.

Supposing $P_k$ and $P'_k$ are projection matrices of two synchronous cameras at time $t_k$, the 3D coordinate of the tracked point $X_k = (X, Y, Z)^T$ and its projected image correspondences $\tilde{x}_k = (\tilde{x}, \tilde{y})^T$ and $X'_k = (X'_1, X'_2, X'_3, X'_4)^T$ in Euclidean space have the following relation

$$
\begin{bmatrix}
(p_{31} \tilde{x} - p_{11}) & (p_{32} \tilde{x} - p_{12}) & (p_{33} \tilde{x} - p_{13}) \\
(p_{31} \tilde{y} - p_{21}) & (p_{32} \tilde{y} - p_{22}) & (p_{33} \tilde{y} - p_{23}) \\
(p_{31} \tilde{x}' - p'_{11}) & (p_{32} \tilde{x}' - p'_{12}) & (p_{33} \tilde{x}' - p'_{13}) \\
(p_{31} \tilde{y}' - p'_{21}) & (p_{32} \tilde{y}' - p'_{22}) & (p_{33} \tilde{y}' - p'_{23})
\end{bmatrix}
\begin{bmatrix}
X \\
Y \\
Z
\end{bmatrix} =
\begin{bmatrix}
p_{14} - p_{34} \tilde{x} \\
p_{14} - p_{34} \tilde{y} \\
p'_{14} - p'_{34} \tilde{x}' \\
p'_{14} - p'_{34} \tilde{y}'
\end{bmatrix}
$$

(7)

where

$$
P_k = \begin{bmatrix}
p_{11} & p_{12} & p_{13} & p_{14} \\
p_{21} & p_{22} & p_{23} & p_{24} \\
p_{31} & p_{32} & p_{33} & p_{34}
\end{bmatrix}, \quad P'_k = \begin{bmatrix}
p'_{11} & p'_{12} & p'_{13} & p'_{14} \\
p'_{21} & p'_{22} & p'_{23} & p'_{24} \\
p'_{31} & p'_{32} & p'_{33} & p'_{34}
\end{bmatrix}.
$$

(8)

In this way, the 3D position of the tracked point in the dynamic word coordinate system is reconstructed and displacement of the object at time $t_k$ with respect to $t_0$ has the form of $d_k = X_k - X_0$.

2.4 A degenerate 2D case

In this section, we illustrate a degenerate case in which the object only has 2D in-plane motion within the rigid reference plane, as shown in Figure 4. The solving process of this case is consistent with the reference [7]. In such situation, the computation of the reconstructed object motion can be much simplified using only one camera as follows.

If the homography $H_k$ between the first frame $I_0$ and the current frame $I_k$ is obtained from the rigid reference region in the plane, the camera motion is still able to be acquired according to Eq. (6). The coordinate of $\tilde{x}_k$ aligned to $I_0$ is directly given by

$$
\tilde{x}_{0k} = H_k^{-1} \tilde{x}_k
$$

(9)

Thus, the corrected displacement of the point becomes $d = \tilde{x}_{0k} - \tilde{x}_0$, with respect to the first image plane system. If the coordinate of first image plane system is not parallel with the reference plane, perspective correction is necessary in the initialization process by applying the image registration method. Note also that there is still an absolute scale ambiguity that cannot be recovered for all the above coordinate systems without additional metric scene measurements. Conventionally, the scale factor is introduced to solve this problem. It remains fixed over the video sequence and can be approximated by a known physical dimension on the object surface and its corresponding image dimension in pixels.
3 Homography estimation

To obtain the homography matrix $H$, each image projected point correspondence $(x_0, x_k)$ generates two equations, then $n \geq 3$ points generate $2n$ linear equations which are sufficient to solve for $H$. To make this estimation process less sensitive to the mismatches of the point correspondences, the usual approach is to introduce the robust RANSAC algorithm. The estimation process is accomplished with the following steps: (a) randomly selecting a subset of the point correspondences, (b) fitting the camera motion model to the selected subset, (c) determining the number of outliers, and (d) repeating steps (a)-(c) for a prescribed number of iterations.

Normally, there are two ways to estimate the homography matrix $H_k$ between the reference planar regions from $I_k$ and $I_0$ using RANSAC algorithm. If $H_k^{k-1}$ is the homography between the regions from two consecutive frames $I_{k-1}$ and $I_k$, $H_k$ can be obtained by using the listed two algorithms in Table 1. In general, $H_k$ obtained from either Algorithm 1 or Algorithm 2 should be the exact value for the transformation between the $k$th frame and the first frame. However, due to the existence of noise, Algorithm 1 always results in undesired non-smoothing jitter while Algorithm 2 encounters the drift problems due to the accumulated error. We demonstrate these two phenomena in detail in the following laboratory test.

In this section, we propose an alternative homography estimation algorithm (Algorithm 3) that combines RANSAC algorithm and Efficient Second-order Minimization (ESM) technique. Unlike the RANSAC algorithm where features such as points and edges are being used, ESM technique is commonly a template-based approach which only makes use of image intensity information of selected templates. ESM technique is one kind of the extensions of the Lucas-Kanade algorithms, which formulates the homography searching problem into a nonlinear optimization problem. It minimizes a dissimilarity measure called the sum of squared differences (SSD) between the reference template and the current template. Gradient descent is the most popular approach to solve this kind of nonlinear optimization problem due to its speed and simplicity. ESM technique belongs to the gradient-based method but it has a high convergence rate like the Newton method. It tries to update each iteration by using the mean of the initial and current Jacobians ($J_{mom}$) from the templates in $I_0$ and $I_k$. A simple process of the ESM technique is shown in Table 2 (step 5). The reader is referred to [19, 25] for the comprehensive study about this approach.

As is well know, however, nonlinear optimization method has the disadvantage of potentially becoming trapped in local minima, which makes the appropriate initial guess for the searching process a crucial issue. In our algorithm, we use the homography matrix obtained by the RANSAC algorithm as the initial guess and refine the homography using the gradient-based ESM algorithm. This procedure is described in detail in Table 2. Since the initial guess $H_k$ obtained by RANSAC algorithm is a much closer estimation to the true value, far fewer iterations are required for convergence and the accumulated error which causes the final drift
Table 2 Proposed approach for homography estimation

<table>
<thead>
<tr>
<th>Algorithm 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Initial algorithm setting:</td>
</tr>
<tr>
<td>1. set ESM tracking templates $T_0$ in $\pi_0$ of $I_0$</td>
</tr>
<tr>
<td>2. define the convergence tolerance TOL or the maximum number of iterations $n_{\text{max}}$ for the ESM algorithm</td>
</tr>
</tbody>
</table>

for each frame $k$:
1. Reference region $\pi_{k-1}, \pi_k \leftarrow$ Image segmentation on $I_{k-1}$ and $I_k$
2. $(x_{k-1}, x_k) \leftarrow$ Feature detection in $\pi_{k-1}$ and $\pi_k$
3. $\hat{H}^{k-1}_k \leftarrow$ RANSAC$(x_{k-1}, x_k)$
4. Initial guess $\hat{H}_k \leftarrow \hat{H}^{k-1}_k H_{k-1}$
5. Final estimation $H_k \leftarrow$ ESM($\hat{H}_k, T_0, \text{TOL}, n_{\text{max}}$) \[19\]

while ($\| t_0 \| < \text{TOL}$) and $i < n_{\text{max}}$ do
- Calculate $J_{\text{esm}}$ in $T_0$
- $t_0 = -(J_{\text{esm}}^{T} J_{\text{esm}})^{-1} J_{\text{esm}}^{T} f(0)$
  - $f(0)$ is a column vector containing the template intensity differences
- $H_k \leftarrow \hat{H}^{k-1}_k \exp(A(t_0))$, $H_k$ is parameterized by the Lie algebra of $\text{SE}(3)$ with $A$ as generators of the algebra and being independent of parameters $t_0$

for the tracking can be properly avoided due to the ESM-based refinement process.

4 Laboratory test of two-story base-isolated building

4.1 Test setup

To validate the accuracy and effectiveness of the proposed vision-based framework, dynamic tests of a two-story base-isolated building were conducted in the structure laboratory in Kyoto University. The test environment (see Figure 5(a)) consisted of one camera system and two shaking tables. Technical specifications of the camera system are shown in Table 3. As shown in Figure 5(b), the base-isolated building was bolted on Shaking table 1, where an accurate scaled ground acceleration recorded from the 2016 Kumamoto earthquake (K-NET http://www.kyoshin.bosai.go.jp) was selected as the input ground motion along the X direction.

![Figure 5](image-url)
### Table 3 Technical specifications of the camera system used in the experiment

<table>
<thead>
<tr>
<th>Component</th>
<th>Model</th>
<th>Technical Specifications</th>
</tr>
</thead>
<tbody>
<tr>
<td>High-speed camera</td>
<td>FASTCAM SA6 75K-M3IT</td>
<td>Maximum resolution: 1920×1440</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Frame rate: 250 FPS</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Chroma: Mono</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Pixel size: 20×20 μm</td>
</tr>
<tr>
<td>Optical lens</td>
<td>AI AF Zoom-Nikkor</td>
<td>Aperture: f/8</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Focal length: 24-85 mm</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Field angle: 33.4°</td>
</tr>
<tr>
<td>LED light panel</td>
<td>VLP-10000X L26995</td>
<td>Lumen: 4680 Lux/m</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Power: 54W</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Size: 260×188×40 mm</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Color Temperature: 5500K</td>
</tr>
</tbody>
</table>

(see the 3D world coordinate system in Figure 5(b)). Structure motion of the base-isolated building was tracked by eight artificial circular markers with high contrast closed boundaries. The center of gravity of each marker was then calculated to represent its position within the image. The 3D world system was built on Shaking table 1, that is to say, all the displacements measured by the vision-based system were with respect to the Shaking table 1. Two reference planes were rigidly attached to Shaking table 1 and served as ESM templates to be tracked using the algorithm proposed in Section 3. The high-speed camera was placed on the manually excited Shaking table 2. The motion of the camera and Shaking table 2 was mainly horizontal. Full motion of the camera was recovered in Section 4.2. The initial camera axis was perpendicular to the reference planes (along Z-axis), and thus the Shaking table 1’s direction was aligned with the first image plane.

The absolute horizontal displacements and accelerations of the base-isolated building as well as the Shaking table 1 in the reference plane were additionally measured by four laser sensors and four accelerometers, respectively. Besides, to validate the camera motion estimating results, the absolute horizontal displacement of the camera was also measured by a laser sensor, as shown in Figure 5. Final laser-measured data being used in the subsequent sections to compare with the results of the vision-based system was obtained by subtracting Shaking table 1’s absolute displacement from the building’s and camera’s absolute displacements.

Figure 6(a) shows the overview of the two-story base-isolated shear building. The mass of the isolation story (base), first story and second story are 2.100 kg, 3.247 kg and 1.531 kg, respectively. The stiffnesses of the first story and second story are about $2.3 \times 10^4$ N/m and $7.5 \times 10^3$ N/m. A BSG-H10 slider (Figure 6(c)) with frictional interfaces served as the seismic isolator of the structure. The maximum sliding displacement of this isolator is 11 mm. Details of the slider can be referred to [26]. Tension springs in Figure 6(d) were inserted into the sliding to provide linear stiffness ($K_b = 2.3 \times 10^3$ N/m) for the isolator. More dynamic characteristics of the two-story base-isolated shear building can be found in [27]. Supposing the friction force within the interface follow the Coulomb friction law, the analytical model and equations of motion for a typical base-isolated building structure are described in detail in Appendix A.

#### 4.2 Camera motion estimation

In this experiment, threshold segmentation was first applied to separate the reference plane region from the object region in the full image. Several steps are included in the segmentation process:

- change the original image (Figure 7(a)) into a binary image using the Otsu’s method with
For Peer Review

Figure 6 (a) Overview of the two-story base-isolated shear building, (b) story block, (c) BSG-H10 slider, (d) details of the isolation story

Figure 7 Process of image segmentation: (a) original image, (b) binary image with noise, (c) binary image without noise, (d) labeled image

- a threshold value $T_h = 0.28$, as shown in Figure 7(b).
- simplify the image with morphological filter using square linear structuring elements, as shown in Figure 7(c).
- return the corresponding label matrix for the image, as shown in Figure 7(d).

Based on the image segmentation results, corner features belonging to the reference plane regions were detected using the Harris-Stephens algorithm. Features from the current image and the previous image were then matched by computing the pairwise distance between feature vectors. After determining the initial value for the homography by using the matched features between two frames, the ESM algorithm was then called to refine and update the homography matrix.

Full camera motion during the experiment was recovered from the homography of the planar structure according to Eq. (6) in Section 2.2. The final estimating results were smoothed by the moving average window and represented in Figure 8. Rotations of the camera were expressed using Euler angles. Figure 8 shows that although the motion of the camera was mainly horizontal (X direction), translations in other directions and rotations also happened. Validation of the above camera motion estimation process was performed by comparing the horizontal moving displacement of the camera measured by a laser displacement sensor with that obtained from the homography decomposition. Since the laser pointing direction was along the X-axis of the first image plane system, the above mentioned two displacements were directly comparable with a scale factor. Clearly, the recovered camera motion fits reasonably well with the displacement measured by the laser sensor, as shown in Figure 8(a).
4.3 Displacement reconstruction

The displacements of the two-story base-isolated shear building relative to Shaking table 1 were then reconstructed according to Section 2. Figure 9(a)-(c) compare the building’s horizontal displacements in X direction of the first story $u_1$, second story $u_2$ and base $u_b$ measured by the camera system (before and after camera motion correction) with the laser sensors, while Figure 9(d)-(f) represent the vertical displacements in Y direction of the first story $v_1$, second story $v_2$ and base $v_b$. Note that the ground truth value of $v$ is equal to zero. As is seen, in contrast to the displacements abstracted from the original video sequence without any camera motion correction, the reconstructed displacements agree well with the displacements measured by the laser sensor and the ground truth value, which shows the success of the reconstruction process using the proposed algorithm.

To assess the accuracy of the newly proposed algorithm 3, displacement results by using Algorithm 1 and Algorithm 2 in Section 3 were also calculated, as shown in Figure 10. An obvious tracking drift can be found in Algorithm 2, which made this algorithm impossible to be used in this case. This drift was attributed to the continuously increasing accumulated error when estimating $H_k$ using $H_{k-1}$. This is the reason why the refinement step with ESM tracking should be included in Algorithm 3. The displacement results obtained by Algorithm 1 seemed to behave as well as the results obtained by Algorithm 3, except for a relative larger error in the displacements $u_2$ and $v_2$. Larger error in Algorithm 1 was because of the lack of motion smoothing between consecutive frames. Due to the limitation of the experimental facilities, our camera went through a slight vibration during the experiment. It is predictable that this error would become more serious for the camera with large motion, which might result in tracking failure eventually.

Computing the root mean square error (RSME) between the reference template and the current template in each iteration step also allowed us to quantitatively assess the refining contribution of the ESM tracker and its computational efforts in Algorithm 3. As shown in
Figure 9 Comparison of each story’s displacement: (a) $u_b$, (b) $u_1$, (c) $u_2$, (d) $v_b$, (e) $v_1$, (f) $v_2$

Figure 10 Comparison of each story’s displacement with different algorithms: (a) $u_b$, (b) $u_1$, (c) $u_2$, (d) $v_b$, (e) $v_1$, (f) $v_2$
Figure 11. Example convergence for the ESM tracker.

The EMS algorithm converged at a very high speed: only two iterations were required in most of the frames. This proved the second-order convergence property for the ESM tracker and its little computational requirement. It took about $5 \sim 7$ ms for each iteration by a computer with an Intel i7-8700 processor with 16 GB of RAM. The above efficiency practically guarantees the real-time application of the proposed algorithm.

4.4 Nonlinear structural system identification

In this section, the nonlinear structural system identification was further achieved by a stochastic filtering technique, referred to as the unscented Kalman filter (UKF). A brief review of the UKF can be found in Appendix B. The readers are also referred to [28–31] for comprehensive study about this technique. Since UKF works in a two-step recursive process, i.e., predicting the current state variables according to the discrete-time state equation, and updating the state variables by the use of the observation equation, the state equation and the observation equation for the base-isolated building are defined as follows.

According to Appendix A, the equation of motion can also be expressed in the state space as follows:

$$\dot{X} = \begin{pmatrix} \dot{u} \\ \ddot{u} \\ \dot{\theta} \end{pmatrix} = \begin{pmatrix} M^{-1} [-M g \ddot{u}_g - F_{nl}(u, \dot{u}, \theta) - C \dot{u} - K u] \\ 0 \end{pmatrix} = f(X, \ddot{u}_g)$$  \quad (10)

where $f(\cdot)$ denotes a vector nonlinear function of the augmented state vector $X$ and $X$ is defined as

$$X = [x_1, x_2, x_3, x_4, x_5, x_6, x_7, x_8]^T = [u_1, u_2, u_b, \ddot{u}_1, \ddot{u}_2, \ddot{u}_b, \nu N_0, \epsilon]^T$$  \quad (11)

where $[\nu N_0, \epsilon]$ are the unknown structural parameters that parameterize the nonlinear regularized Coulomb friction force $f_{nl}$, which has been described in detail in Appendix A. The initial values for the state variables were $X^0 = [0, 0, 0, 0, 0, 2.1, 0.01]^T$ in this example. $\epsilon^0$ was estimated according to reference [32]. The state equation in Eq. (10) can be formulated in a discrete form as

$$X_{k+1} = F(X_k, \ddot{u}_g,k) + w_k$$  \quad (12)

where $w$ is the process noise assumed to follow a Gaussian distribution with zero mean and a covariance matrix $Q$. The function $F$ can be obtained with the following integration.

$$F(X_k, \ddot{u}_g,k) = X_k + \int_{k \Delta t}^{(k+1) \Delta t} f(X, \ddot{u}_g) \, dt$$  \quad (13)
As can be seen, when using the state space representation, the original 2nd order ordinary differential equation (ODE) can be converted into an equivalent 1st order ODE system and any 1st order ODE integration method such as fourth-order Runge-Kutta method [33] may be used to obtain an approximate solution.

The observation equation has the form

$$Y_k = h(X_k, \bar{u}_{g,k}) + v_k$$

where $Y_k$ is the measurement vector at time $t_k$, $h$ is the nonlinear function for the observation equation and $v$ is the measurement noise pertaining to the Gaussian distribution with zero mean and a covariance matrix $E[v_k v_k^T] = R$ with $E[\cdot]$ denoting the expectation operator. Since the displacements $u_1$, $u_2$ and $u_b$ were measured by the camera system, the observation equation is expressed as $Y = [u_1 + u_b, u_2 + u_b, u_b]^T + v$. In this research, we assumed that in the measured data the errors in displacement responses at different time points and different positions of the structure were independent and followed a zero mean white Gaussian noise. The output measurement noise level was estimated as 20% RMS noise to signal ratio.

Figure 12 shows the comparison of the initial and estimated displacement and acceleration results using the UKF method. As is seen, by using the displacement results obtained from the vision-based system as the measured data, the UKF estimated accelerations corresponded with the accelerometer-measured results very well. This again indicated the effectiveness of vision-based system and the propose algorithm.

Unknown parameters $\theta = [\nu N_0, \epsilon]^T$ for the tanh-regularized Coulomb friction was also determined using the UKF. To validate the efficiency of the results, $\theta$ was additionally estimated.
in a different way, that is, by using the acceleration obtained from the accelerometer as measured data in the UKF. Figure 13 compares the final estimated results using different measured data. Overall, the parameters \( \nu N_0 \) and \( \epsilon \) for the tanh-regularized Coulomb friction based on different measured data converged to the values close to each other very well, which proved that not only the dynamic responses but also the physical parameters of the nonlinear structural system could be appropriately updated. A small deviation was observed in the regularization parameter \( \epsilon \) since this parameter was highly sensitive to the measurement noise. As a final note, it is worth mentioning that this information is very important to accurately predict the remaining useful life of the structure as well as the reliability and risk of operation.

5 Conclusions and future work

This paper proposes a novel homography estimation algorithm for camera motion estimation and displacement reconstruction when using un-stationary cameras in the vision-based system. Whereas most algorithms operate on the premise that the image point correspondences are provided with good accuracy, our method directly calculates the camera motion and structural displacement using the estimated homography of reference planes in images. To overcome the possible tracking drift and non-smoothness problems, our homography estimation algorithm combines RANSAC algorithm and ESM technique, and achieves the reconstruction results accurately and efficiently. Experiments demonstrated the effectiveness of the proposed algorithm and nonlinear structural system identification was also conducted successfully using the measured results.

Possibilities for future work include the comparison of the proposed algorithm to conventionally feature tracking algorithms both in theory and through experiments in a more realistic 3D object tracking situation. Another possible avenue to evaluate the performance of the proposed algorithm is to apply this method in structural health monitoring system for real-time wind/seismic-induced deformation measurement with strong camera vibration of a diverse set of structures such as high-rise buildings and long-spanned bridges.

Appendix A Analytical model and equations of motion

Let’s assume the superstructure of a base-isolated building to remain elastic. The equation of motion, governing a typical superstructure building with total number of \( n \) floors, is written as

\[
M_s \ddot{u}_s + C_s \dot{u}_s + K_s u_s = -M_s r ( \ddot{u}_b + \ddot{u}_y )
\] (15)
where $M_k$, $C_k$ and $K_k$ are the mass, damping and stiffness matrices of the typical shear building with dimensions $n \times n$ and $r$ represents the location matrix (matrix dimension $n \times 1$) for base acceleration $\ddot{u}_b$ and ground motion acceleration $\dddot{g}$. $\ddot{u}_b$, $\dddot{u}_s$ and $u_s$ are, respectively, the vectors of acceleration, velocity and displacement of the superstructure relative to the base. The equation of dynamic equilibrium of the base is:

$$r^T M_s [\dddot{u}_s + r(\dddot{u}_b + \dddot{g})] + M_b (\dddot{u}_b + \dddot{g}) + C_b \ddot{u}_b + K_b u_b + f_{nl}(u_b, \dot{u}_b, \theta) = 0$$  

(16)

where $\dddot{u}_b$ and $u_b$ represent the base velocity and displacement with respect to the ground and $M_b$, $C_b$ and $K_b$ are the mass, damping and linear stiffness of the base, respectively. $\theta$ is the unknown structural parametric vector that parameterizes the nonlinear restoring force $f_{nl}$ of the isolation system. If $f_{nl}$ follows the tanh-regularized Coulomb dry friction law

$$f_{nl}(u_b, \dot{u}_b, \nu N_0, \epsilon) = \nu N_0 \tanh\frac{\dot{u}_b}{\epsilon}$$  

(17)

where $N_0$ is the normal load, $\nu$ is the dynamic friction coefficient and $\epsilon$ is the regularization parameter [32], then the unknown structural parametric vector can be set as $\theta = [\nu N_0, \epsilon]^T$.

Combining Eq. (15) and Eq. (16), the following system of equations is derived

$$M \ddot{u} + C \dot{u} + K u + F_{nl}(u, \dot{u}, \theta) = -M_g \dddot{g}$$  

(18)

where

$$M = \begin{pmatrix} M_s & M_r \\ r^T M_s & r^T M_s r + M_b \end{pmatrix}, \quad C = \begin{pmatrix} C_s & 0 \\ 0 & C_b \end{pmatrix}, \quad K = \begin{pmatrix} K_s & 0 \\ 0 & K_b \end{pmatrix}$$

$$u = \begin{pmatrix} u_s \\ u_b \end{pmatrix}, \quad F_{nl} = \begin{pmatrix} 0 \\ f_{nl} \end{pmatrix}, \quad M_g = \begin{pmatrix} 0 \\ M_r \end{pmatrix}$$

(19)

### Appendix B Parameter estimation using UKF algorithm

From a Bayesian perspective, the recursive nonlinear filtering problem of determining the estimated of $X_k$ based on the sequence of all available measurements up to time $k$, $Y_{1:k} := \{Y_1, \cdots, Y_k\}$, is to compute the posterior probability density function (PDF) $p(X_k|Y_{1:k})$. Assuming that the posterior PDF at time $k-1$, $p(X_{k-1}|Y_{1:k-1})$, is known, the posterior PDF at time $k$, $p(X_k|Y_{1:k})$, can be obtained through a Bayesian prediction-correction scheme. Since the analytic solution of the above Bayesian scheme is intractable, the UKF is presented as follows to resort to approximations of the original problem. The UKF approximates the posterior PDF $p(X_k|Y_{1:k})$ by a Gaussian distribution and the mean and covariance of the Gaussian distribution is captured by a set of deterministically sigma points. On considering this, in the UKF algorithm, the first step is to generate a set of $2N + 1$ sigma points using the unscented transform (UT) as [34]

$$\hat{X}_k = \left[ \hat{X}_k, \chi_k \right] + \left[ \begin{pmatrix} \sqrt{(N+\lambda)} P_k \end{pmatrix}_i \right]_i, \quad i = 1, \cdots, N$$

(20)

where $N$ is the number of the factors in the augmented state vector and $\lambda$ is a gain parameter with a definition related to the dimension and the distribution of the state vector. $\hat{X}_k = \mathbb{E}[X_k]$ and $P_k = \mathbb{E}[(X_k - \hat{X}_k)(X_k - \hat{X}_k)^T]$. $\left[ \begin{pmatrix} \sqrt{(N+\lambda)} P_k \end{pmatrix}_i \right]_i$ is the $i$th row of the matrix square root. Robust calculation for the square root of the covariance matrix $P_k$ can be obtained by singular value decomposition (SVD). The predicted sigma points of the $(k+1)$th step can be calculated as:

$$\chi_{k+1|i} = F(\chi_k, u_k, w_k)$$

(21)
where \( \mathbf{u} \) denotes all the inputs of the nonlinear structural system. Then the predicted mean \( \hat{\mathbf{x}}_{m+1} \) and the predicted covariance \( \hat{\mathbf{P}}_{X,m+1} \) are approximated using a weighted sample mean and covariance of the posterior sigma points

\[
\hat{\mathbf{x}}_{m+1} = \sum_{i=0}^{2N} c_i \hat{\mathbf{X}}_{i,m+1|k}
\]

\[
\hat{\mathbf{P}}_{X,m+1} = \sum_{i=0}^{2N} c_i \left[ \hat{\mathbf{X}}_{i,m+1|k} - \hat{\mathbf{x}}_{m+1} \right] \left[ \hat{\mathbf{X}}_{i,m+1|k} - \hat{\mathbf{x}}_{m+1} \right]^T + Q_{m+1}
\]

where \( c_i \) are the weights for the predicted mean and covariance, respectively \([35]\).

The predicted measurement vector \( \mathbf{y}_{m+1} \) and its predicted covariance matrix \( \hat{\mathbf{P}}_{Y,m+1} \) are given by

\[
\hat{\mathbf{y}}_{m+1|k} = \mathbf{H}(\hat{\mathbf{x}}_{m+1|k}, \mathbf{u}_k, \mathbf{v}_k)
\]

\[
\hat{\mathbf{y}}_{m+1} = \sum_{i=0}^{2N} c_i \hat{\mathbf{y}}_{i,m+1|k}
\]

\[
\hat{\mathbf{P}}_{Y,m+1} = \sum_{i=0}^{2N} c_i \left[ \hat{\mathbf{y}}_{i,m+1|k} - \hat{\mathbf{y}}_{m+1} \right] \left[ \hat{\mathbf{y}}_{i,m+1|k} - \hat{\mathbf{y}}_{m+1} \right]^T + R_{m+1}
\]

\[
\hat{\mathbf{P}}_{XY,m+1} = \sum_{i=0}^{2N} c_i \left[ \hat{\mathbf{x}}_{i,m+1|k} - \hat{\mathbf{x}}_{m+1} \right] \left[ \hat{\mathbf{y}}_{i,m+1|k} - \hat{\mathbf{y}}_{m+1} \right]^T
\]

Then the measurement update using the Kalman filtering is as follows:

\[
\mathbf{K}_{m+1} = \hat{\mathbf{P}}_{XY,m+1} \hat{\mathbf{P}}_{Y,m+1}^{-1}
\]

\[
\hat{\mathbf{x}}_{m+1} = \hat{\mathbf{x}}_{m+1} + \mathbf{K}_{m+1} (\mathbf{y}_{m+1} - \hat{\mathbf{y}}_{m+1})
\]

\[
\hat{\mathbf{P}}_{m+1} = \hat{\mathbf{P}}_{m+1} - \mathbf{K}_{m+1} \hat{\mathbf{P}}_{Y,m+1} \mathbf{K}_{m+1}^T
\]

Note that throughout the process of the UKF algorithm, the derivation of Jacobians is not required to linearize the nonlinear system. The insight behind the UKF algorithm is that it is easier to approximate a probability distribution than it is to approximate an arbitrary nonlinear function. The approximation of the statistics of the random variable which undergoes a nonlinear transformation is calculated by the UT. Approximations taken with the UT can be accurate to the third order for Gaussian inputs and at least the second-order for non-Gaussian input for all nonlinearities.

References


