Journal of Civil Structural Health Monitoring (2026) 16:24
https://doi.org/10.1007/s13349-025-01062-1

ORIGINAL PAPER

An automated vision-based approach for structural dynamic identification

Federico Ponsi[1] · Ghita Eslami Varzaneh[1] · Elisa Bassoli[1] · Loris Vincenzi[1]

Received: 9 October 2025 / Revised: 1 December 2025 / Accepted: 11 December 2025

Abstract

This paper presents an automated computer vision-based procedure for assessing structural displacements, aiming to perform modal identification and evaluate structural responses under dynamic loading. The proposed algorithm is designed to extract structural displacements from recorded videos by temporally tracking predefined points on the structure. The research addresses key methodological challenges in monitoring large structures, including detecting the relatively small displacements characteristic of structural vibrations under environmental conditions, correcting for typically unavoidable perspective distortions and camera movements, reconstructing the 3D displacements from single-camera videos and identifying global mode shapes. The effectiveness of the proposed procedure is first demonstrated through experiments on a laboratory-scale frame under controlled conditions. Its applicability to real-world scenarios and ability to identify global mode shapes are further evaluated using experimental data from a full-scale footbridge. In both cases, comparisons between vision-based monitoring results and reference measurements demonstrate the strong performance of the method. The proposed approach achieves an accuracy below one-tenth of a pixel, corresponding, in the investigated experimental case study, to a metric accuracy of approximately 0.1 mm at a 60 m camera-to-target distance and 0.03 mm at 25 m. This confirms the method's capability to provide reliable measurements for real-world structural monitoring applications.

Keywords Vision-based monitoring · Dynamic identification · Perspective correction · Camera shake effect · Laboratory experiments · Steel footbridge

1 Introduction

Vibration-based monitoring is essential across a wide range of structural applications, including system identification, model updating, health assessment, and damage detection (see, for instance, [1–5]). Traditionally, this process involves placing sensors, typically accelerometers, on the structure and connecting them via extensive wiring to a data acquisition system. This traditional approach, however, is costly, complex, and can compromise the structural operability. To overcome these challenges, the engineering community is increasingly turning to contact-less technologies (e.g. [6, 7]). These systems offer significant advantages in scenarios where minimal invasiveness is essential, such as preserving cultural heritage sites, when physical access is difficult, or when rapid structural assessment is required following extreme events.

Contact-less technologies for civil monitoring include Global Navigation Satellite Systems (GNSS) [8–10], satellite remote sensing [11, 12], terrestrial radar interferometry [13], and vision-based techniques [14, 15]. Among these, vision-based techniques stand out as the only type of remote sensing capable of reducing reliance on costly industrial products [16]. They have demonstrated significant potential even with consumer-grade devices like standard video

Ghita Eslami Varzaneh, Elisa Bassoli and Loris Vincenzi have contributed equally to this work.

✉ Elisa Bassoli
elisa.bassoli@unimore.it

Federico Ponsi
federico.ponsi@unimore.it

Ghita Eslami Varzaneh
ghita.eslami@unimore.it

Loris Vincenzi
loris.vincenzi@unimore.it

[1] Department of Engineering "Enzo Ferrari", University of Modena and Reggio Emilia, Via Vivarelli 10, 41125 Modena, Italy

Published online: 11 February 2026

cameras or smartphones [17–19]. This advancement is largely due to the development of low-cost technologies that offer high resolution and high frame rates, promising accurate monitoring of large structures in both static and dynamic contexts.

The core concept of vision-based monitoring is straightforward: it involves capturing a video of the structure and analyzing the frames, either in real-time or afterward, to extract motion data. This process relies on the assumption that, with a stationary camera, changes in the positions of key features between consecutive frames indicate the displacement of corresponding points on the structure over time. This generates displacement time histories, which can then be used to compute strains, velocities, and accelerations. Vision-based methods offer significant technical advantages, such as the direct measurement of displacements, which can be more readily related to structural damage. Additionally, a single camera sensor can provide distributed monitoring, allowing extraction of displacement data from multiple points on the structure within a single video recording. Furthermore, this approach results in significant cost savings and a substantially reduced setup effort. Given these advantages over traditional monitoring systems, vision-based techniques have been gaining increasing attention in civil engineering research. Recent applications are extensively reviewed in studies such as [16, 20–22], which include tests on bridges [23, 24] and footbridges [25, 26].

Besides the several advantages offered by vision-based monitoring, evaluating the accuracy in estimated structural displacements, and thus its application for modal and damage identification, remains a challenging aspect to assess. This is due to the fact that the achievable accuracy is influenced not only by the technical specifications of the camera but also by factors such as potential unwanted movements of its support, the camera-to-structure distance, perspective distortions, and uncertainties introduced by environmental conditions as well as by the adopted image processing procedure.

Regarding uncertainty from camera placement, cameras can be positioned flexibly depending on the application, but stability and sufficient distance from the structure are essential to prevent vibrations from affecting the camera. Any camera movement can be misinterpreted as structural movement, leading to inaccurate displacement assessments. Ensuring a stable camera base can be challenging, especially for monitoring small movements or short tests, where cost-effective, portable tripods are often preferred over rigid setups. In those cases, suitable procedures must be applied to filter out environmental vibrations and ensure accurate results. Camera-to-structure distance and angle should also be chosen to meet accuracy requirements. For large structures and infrastructure, multiple cameras can be employed, each focusing on a specific part of the structure, but this demands temporal and geometric alignment of results, making single-camera monitoring often more practical. Measurement points on the structure must be also selected in advance based on the required accuracy, and can be either artificial targets or distinct structural features such as corners, holes, or bolts [27]. Typically, artificial targets generally provide higher accuracy despite requiring access to the structure for installation.

As previously mentioned, environmental conditions are widely recognized as sources of errors and uncertainties. These include non-uniform air refraction caused by temperature differences between the camera and the monitored object, as well as variable weather and ambient light conditions. Literature on the assessment of environmental uncertainties in vision-based monitoring includes theoretical analyses and laboratory testing [21, 28]. However, the impact of these external factors on the accuracy of on-site tests remains not fully understood and is still under investigation. This gap is partly due to the limited number of full-scale outdoor tests, as vision-based techniques have only recently been adopted for monitoring large-scale civil structures. Recent advances in artificial intelligence (AI) offer a potential solution to some of these challenges. AI-aided vision-based monitoring, leveraging machine learning and deep learning techniques (see, e.g., [29, 30]), can improve robustness to noise, lighting variations, and occlusions, while enabling marker-less tracking. However, AI-methods also introduce challenges related to training data, transparency and computational costs.

In this context, the main objective of the paper is to propose a comprehensive procedure for assessing the dynamic behavior of full-scale structures using image processing techniques. Key steps emphasized in the paper include the use of effective image feature extraction methods that achieve sub-pixel accuracy, the elimination of camera movements to prevent errors or noise in the obtained displacements, and the development of an appropriate method to convert movements identified in the 2D image plane into 3D real-world displacements. This last step is particularly crucial when monitoring full-scale structures, where the involved dimensions require careful consideration of perspective effects, making the use of a simple scale factor, as suggested in other studies [31, 32], inadequate. Specifically, the Perspective-Three-Point (P3P) method [33] is adopted, enabling 2D-to-3D conversion from minimal input data using the law of cosines. Only the 2D projections of three points on the image and their mutual distances in the real world are required, allowing accurate estimation of the distance between the camera and the triplet of points, which serves to reconstruct their 3D motion. The proposed approach employs high-contrast targets (black-and-white

checkerboards), which provide a set of easily identifiable points (i.e., the square corners), enabling result averaging over multiple triplets and multiple corners, thereby enhancing accuracy. While 3D monitoring is typically performed with stereo vision setups (i.e., merging views from two cameras), e.g., [34, 35], the novelty of the proposed P3P-based method lies in achieving accurate 3D monitoring with a single camera.

The capability of the vision-based approach to identify global modal parameters of full-scale structures is also explored, aspect rarely addressed in the existing literature. Indeed, current vision-based methods have seldom captured complete bending mode shapes of real bridges [36] and research on torsion modes remains limited [37]. The potential and limitations of real-scale vision-based dynamic monitoring are evaluated through the case study of a steel footbridge in Modena (Italy). In this application, both bending and torsional modes are successfully identified and assessed against the results from a traditional monitoring system installed on the structure. Before addressing the full-scale application, the paper presents results from laboratory tests conducted to validate the proposed vision-based procedure. These experiments are carried out under controlled conditions on a laboratory-scale steel frame subjected to excitation by a shaking table. To assess the procedure accuracy, multiple experimental setups are designed, varying the camera-to-structure distance, as well as the incidence angle of the line of sight. The vision-based measurements are then compared to reference data obtained from displacement transducers installed for control purposes.

The paper is organized as follows. Section 2 presents the proposed vision-based monitoring approach. Section 3 describes the laboratory experiments, while Sect. 4 discusses the steel footbridge application. Finally, conclusions are drawn in Sect. 5.

2 Vision-based monitoring procedure

This section describes the vision-based procedure developed to quantify structural displacements from recorded videos through the temporal tracking of artificial checkerboard targets placed on the structure. The proposed approach can be divided into two main stages: the vision-based monitoring setup, which involves considerations prior to the monitoring campaign and is tailored to the specific structure, and video post-processing, which serves as the critical link between raw data (video frames) and the time series of structural displacements. The main steps of the proposed procedure are outlined in Sect. 2.1, while a more detailed description of each step can be found in Sects. 2.2 to 2.7.

2.1 Outline of the procedure

The vision-based monitoring procedure can be divided into the following steps:

- Step 1: Monitoring framework. This step involves the design of the experimental campaign based on the application scenario and the expected structural displacement. It includes selecting the camera based on its technical specifications and suitable lenses, and choosing the appropriate frame rate. In this paper, consumer-grade cameras are used to evaluate the effectiveness of low-cost monitoring systems for assessing the dynamic behavior of structures. Additionally, the camera position and measurement locations need to be pre-defined. In accordance with the procedure outlined in this paper, the placement of checkerboard artificial targets is required both on the monitored structure and at a distance from it, with the latter aimed at eliminating the effects of camera shake. The use of checkerboard targets enable the conversion of 2D image-plane displacements into 3D real-world displacements and improve the accuracy of structural displacement estimation by averaging the results from each square corner. Further details about this step are provided in Sect. 2.2.
- Step 2: Calibration of camera parameters. This process includes determining several factors, involving intrinsic parameters (focal length f, optical center O, lens distortion), extrinsic parameters (camera position and orientation relative to the scene being captured) and geometrical parameters (projection distortions introduced by the camera system). In the vision-based monitoring procedure presented in this paper, camera calibration is conducted using the approach proposed in [38], implemented in the MATLAB Computer Vision Toolbox, which is robust to varying imaging conditions and fully automatic. The camera parameters required for the presented approach are the lens distortion coefficients, the focal length f, and the pixel-to-mm transformation factor in the image plane, this latter related to the sensor dimension and the image resolution. For further information, please refer to Sect. 2.3.
- Step 3: Feature tracking. This step plays a critical role in estimating accurate displacements from videos in vision-based monitoring systems. Feature tracking is based on the identification of distinct points or patterns in the video frames, such as corners or edges, that can be followed over time. By consistently tracking these features between consecutive frames, the relative displacement (in pixels) of specific points on the structure can be determined. The accuracy of displacement estimates is highly reliant on the quality and reliability of feature

detection, extraction, and tracking. Even minor errors in those tasks can result in substantial inaccuracies in displacement estimation, particularly when dealing with small displacements. Moreover, feature tracking methods must be able to account for variations in lighting, perspective, and other environmental factors. Several feature tracking methods are proposed in literature, as reported in the review studies presented in [39, 40]. In this paper, the Kanade-Lucas-Tomasi algorithm [41, 42] is used to track the movement of the checkerboard corners over time, which are initially identified using the Harris-Stephens algorithm [43]. Additional details are provided in Sect. 2.4.

- Step 4: Coordinate transformation from the image plane to the real world. The square corner locations obtained from feature tracking are referenced to the 2D image plane and measured in pixel units, whereas real-world coordinates represent physical locations in three-dimensional space. In this step, corner locations are converted into metric units within a 3D reference system and adjusted for perspective distortion and camera orientation effects. The structural displacement is then evaluated over time by tracking changes in the real-world coordinates calculated at each time step. The procedure proposed in this article involves the use of the Perspective-Three-Point (P3P) method [33] to identify both the relative movements between the camera and the target placed on the structure, and the relative displacements between the camera and fixed targets positioned far from the monitored structure, so as to estimate the absolute structural displacements. The P3P method is chosen because it allows obtaining 3D-to-2D point correspondences analytically, without the need for numerical optimization or minimization procedures that could introduce errors and noise into the resulting world coordinates. The reader is referred to Sect. 2.5 for further details.
- Step 5: Calculation of displacements relative to the structure reference system. To fully understand the structural behavior, it is essential to reference the structural displacements to a coordinate system aligned with the main directions of the structure, rather than the camera perspective. To do this, the structure reference system can be defined with two axes aligned to the checkerboard directions, closely matching the directions of the structural movement to be measured, and the third axis perpendicular to the target plane. A matrix transformation is then applied to rotate and translate the camera reference system into the structure reference frame, ensuring proper alignment of the measured displacements with the structural coordinate system. This process is detailed in Sect. 2.6.
- Step 6: Filtering camera shake effects. This step is essential for estimating absolute structural displacements. In the outlined procedure, the filtering relies on the presence of ground-based targets, which are assumed to be stationary. Any displacements detected for these targets are attributed solely to unintended camera movements, not actual structural motion. By subtracting these undesired movements, the procedure ensures that only the absolute displacements of the structure are captured, free from camera movements (see Sect. 2.7).

2.2 Monitoring framework

A vision-based monitoring system typically consists of one or more video cameras equipped with zoom lenses, tracking either artificial targets or intrinsic details of the structure, such as corners, holes, or bolts. In relation to the procedure outlined in this paper, the structure is monitored using a single camera, and the points where measurements are acquired are marked on the structure with checkerboard targets. Several key factors must be considered in the design of the monitoring campaign to ensure accurate and reliable measurements. These factors include image resolution, camera-to structure-distance, zoom factor, target size and frame acquisition rate. While the latter relates to the range of natural frequencies that may be identified, the combination of the other factors determines the potential accuracy that can be achieved.

Measuring small displacements requires a high pixel density in the target being tracked, which is achieved by focusing the camera view on the measured position. However, widening the field of view allows for monitoring more points on the structure and analyzing its overall behavior, albeit at the cost of reducing the number of pixels defining each target. Therefore, the intrinsic parameters of the camera (such as resolution and focal length range), along with the camera-to-structure distance, zoom factor, and target size, must be carefully chosen to achieve the optimal balance between these conflicting factors, based on the expected structural displacement. Drawing on the experience acquired by the authors in the applications presented in this paper, it is recommended that each square of the checkerboard target should measure at least a few pixels (around 3 or 4) per side in the acquired frames to ensure accurate displacement measurements. Furthermore, the laboratory experiments presented in Sect. 3 demonstrates that the noise in the signal (calculated from data collected without input forces) has a standard deviation of approximately 0.008 pixel. Therefore, it is recommended that the expected displacement be at least 0.08 pixel to ensure it is clearly detectable above the measurement noise.

To minimize unwanted camera shaking, the camera should be placed in a stable, sheltered location away from the vibrating structure, and a remote controller should be used to start and stop video recordings to prevent vibrations from manual intervention. Despite these measures, unintended camera movements may still occur due to wind or other uncontrollable external factors. This makes filtering out camera shake effects a key task for reliably estimating structural displacements.

2.3 Camera calibration

Camera calibration is performed to determine the camera parameters, which are essential for accurate 3D measurements. Among these parameters, the focal length has a strong impact on the metric accuracy. To assess its uncertainty, the calibration procedure can be repeated several times using different frames extracted from the same calibration video to determine the focal length. The deviation of the average estimated focal length from its extreme values can then be quantified as an uncertainty level. Based on experience from the authors, this results in an uncertainty of approximately 1.5% under constant lighting conditions and a standard focusing quality. Poor lighting or inadequate focus may increase the uncertainty associated with the focal length estimation.

2.4 Feature tracking

Feature tracking involves monitoring the movement of specific features of the structure over time, which are identified as key visual patterns or attributes through feature detection and extraction.

Various methods in the literature address the identification and comparison of distinct, repeatable, and stable features across different views or frames [16]. The procedure proposed in this paper relies on the Kanade-Lucas-Tomasi (KLT) algorithm, a well-established technique for optical flow and feature tracking, particularly in visual tracking applications [41, 42]. Initially, the KLT algorithm identifies reliable feature points in the first frame of the video using a feature point matching method. Then, it tracks the movement of these points frame by frame through optical flow estimation.

Feature point matching is based on sparse feature points, also referred to as key-points, which represent the tracked objects. A key-point is a small region (e.g., 5×5 pixels) characterized by unique and invariant features, such as distinctive structural (or architectural) elements, or artificial targets. This approach involves two components: a feature detector and a feature descriptor. The feature detector identifies a sub-region of the image selected by the algorithm, while the feature descriptor is a matrix or vector that encodes the characteristics of that sub-region, based on the shape and appearance of the surrounding area of the key-point.

The most effective features are those with strong local intensity gradients, as these tend to remain stable across frames. The Harris-Stephens algorithm [43], in particular, is designed for corner and edge detection and has shown superior performance in the applications discussed in this paper. For this reason, it is employed in the proposed procedure to detect the corners of checkerboard targets. The approach is robust to scale, rotation, and illumination changes, enabling accurate sub-pixel displacement estimation and making it suitable for tasks such as object detection, motion tracking, and 3D reconstruction.

The results presented in the following sections are obtained adopting the KLT tracker and the Harris-Stephens corner detection algorithm, both implemented in the MATLAB Computer Vision Toolbox. The camera parameters required in this step are the lens distortion parameters.

To increase the speed of automated analysis, the operational area within video frames is restricted by defining Regions Of Interest (ROI) around the checkerboard targets, selected from the initial frames. The defined ROI are handled as a matrix of pixels, where each pixel is defined by its 2-D coordinates (expressed in pixels and related to the upper-left vertex of the frame) and by a unique RGB intensity level.

The output of this step consists of a time series of displacement for each square corner of every checkerboard target. The displacement values $u(t)$ and $v(t)$, expressed in pixels, are referenced to the image plane π and describe the relative motion between the camera and the square corners.

The adopted approach allows estimating sub-pixel displacements. In particular, the laboratory experiment presented in Sect. 3 demonstrates that the noise in the signal, calculated from data collected in the absence of input forces, exhibits a standard deviation of approximately 0.008 pixel, emphasizing the high measurement accuracy.

2.5 Coordinate transformation from the image plane to the real world

Displacement time series evaluated as shown in Sect. 2.4 are transformed from the image plane (π) to the real world (W) by solving the so-called Perspective-n-Point (PnP) problem for the feature coordinates calculated at each frame.

Solving the PnP problem means to determine the 3D position and orientation of a camera based on a set of 2D image points and their corresponding 3D world coordinates. This fundamental problem was first explored in photogrammetry and later adopted in computer vision. In the context of this research, however, the PnP problem is used to estimate the

position and orientation of the checkerboard targets, rather than those of the camera, as also performed in [44].

In the early 1980 s, Fischler and Bolles [45] proposed the PnP procedure to estimate the position and orientation of a calibrated camera using known 3D-to-2D point correspondences between a 3D model and its projections in the image. Later, several numeric algorithms were proposed to efficiently solve the PnP problem, including Direct Linear Transform (DLT), EPnP (Efficient PnP), and RPnP (Robust PnP) [46]. Each method to solve the PnP problem efficiently leverages the law of cosines to bridge the gap between two-dimensional image data and three-dimensional spatial information, enabling accurate calculation of object positions in real space.

The P3P (Perspective-Three-Point) method is a specific case of the PnP problem, where only three 3D-to-2D point correspondences are used to calculate the camera pose, or, as in this application, the position and orientation of the artificial targets. The solution of the P3P used here has its origins in the studies of Grunert, dating back to 1841, and for nearly two centuries it has remained relevant in various researches and applications [33].

The P3P method enables the calculation of the distance between each square corner and the camera optical center O. The inputs required for this task are the focal length f, the corner coordinates in the image plane π, the mutual distance between corners in the real world and the pixel-to-mm conversion factor. This latter is related to the sensor dimension (in mm) and the image resolution (in pixel) and allows transforming the focal length and the distances between corners in the image plane from pixels to millimeters. The mutual distance between corners in the real world is known from the checkerboard geometry while the focal length is obtained through the camera calibration process. Finally, the corner coordinates in the image plane π are the result of the feature tracking process described in section 2.4 (specifically $u(t)$ and $v(t)$ at the given time instant/video frame).

The intuitive procedure is based on applying the law of cosines, which is essential for calculating distances and angles within both the image plane and the real-world coordinate system. The P3P-based proposed framework is mathematically detailed below, geometrically illustrated in Fig. 1 and graphically summarized in the flowchart of Fig. 2.

With reference to Fig. 1, let O be the optical center, f the focal length, and P_1, P_2 and P_3 three checkerboard corners, whose coordinates in the image plane π at the given time instant (evaluated from Sect. 2.4) are:

$$P_{1,\pi} = (u_1, v_1), P_{2,\pi} = (u_2, v_2), P_{3,\pi} = (u_3, v_3) \tag{1}$$

while the corresponding -unknown- coordinates in real world read:

$$P_{1,W} = (x'_1, y'_1, z'_1),\ P_{2,W} = (x'_2, y'_2, z'_2),\ P_{3,W} = (x'_3, y'_3, z'_3) \tag{2}$$

First, the coordinates in the π image plane, u_i and v_i, are converted from pixels to millimeters using the appropriate transformation factor, leading to $\hat{u}_i$ and $\hat{v}_i$. Next, the mutual distances between the corners are calculated from their 2D coordinates in the image plane. Given the focal length f, the 2D checkerboard corner coordinates and the mutual distances in the image plane, the distances along the lines of

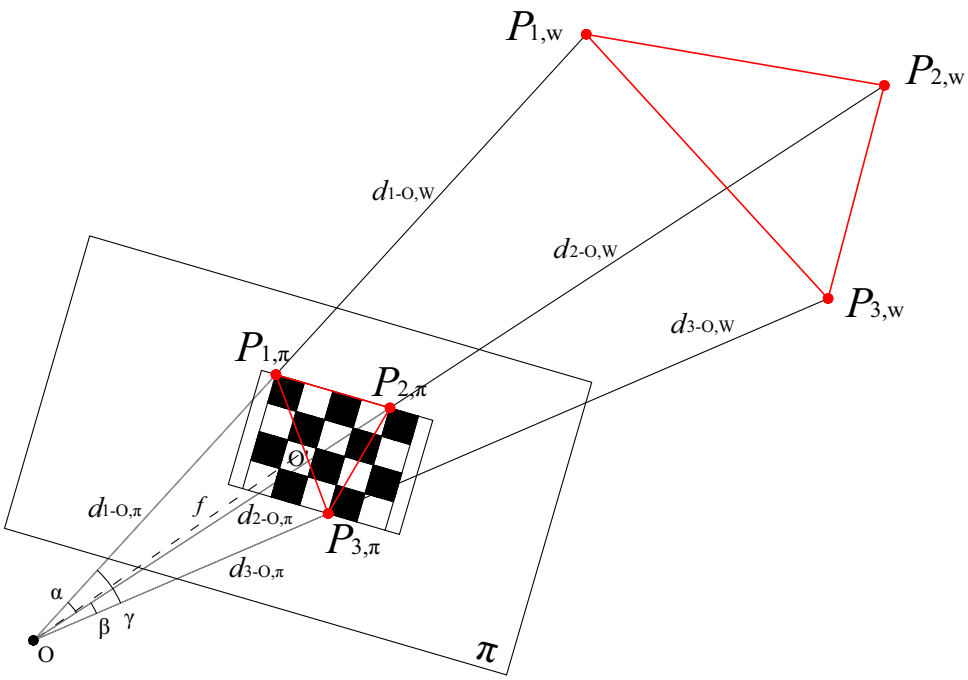

Fig. 1 Geometric interpretation of the P3P method applied to target positioning

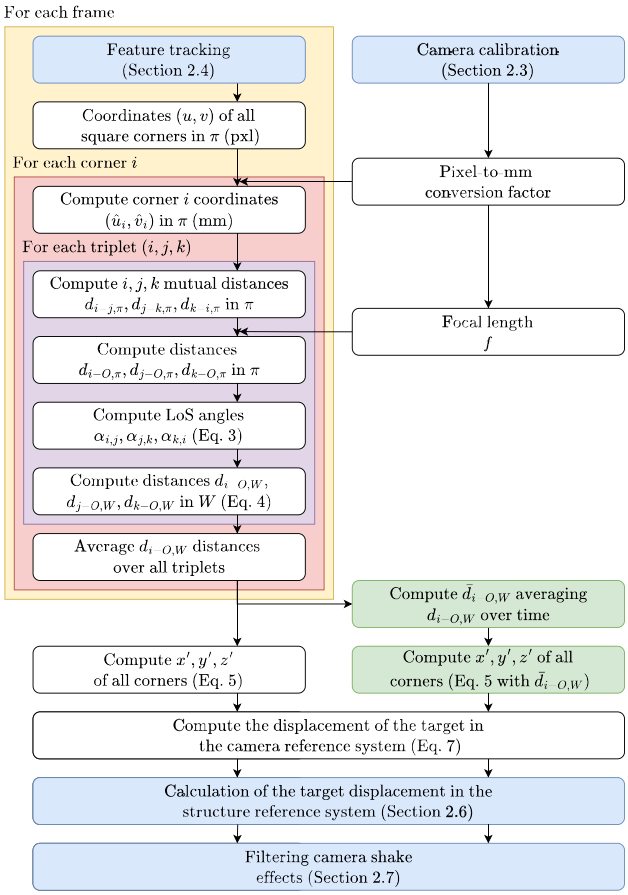

Fig. 2 Workflow of the overall vision-based procedure, with focus on the conversion from image to camera-centered (real-world) reference systems (Sect. 2.5). Green steps are optional

sight between the camera optical center O and the square corners in the image plane can be easily computed. Such distances are indicated as $d_{i-O,\pi}$ with $i = 1, 2, 3$, while the mutual distance between corners in the image plane are referred to as $d_{i-j,\pi}$ with $i, j = 1, 2, 3$ and $i \neq j$. Then, the law of cosines is used to calculate the cosines of the angles ($\alpha_{i,j}$, with $i, j = 1, 2, 3$ and $i \neq j$) between the lines of sight ($d_{i-O,\pi}$ and $d_{j-O,\pi}$):

$$cos\alpha_{i,j} = \frac{d_{i-O,\pi}^2 + d_{j-O,\pi}^2 - d_{i-j,\pi}^2}{2 d_{i-O,\pi} d_{j-O,\pi}} \tag{3}$$

Once the angles between the lines of sight are determined, the law of cosines is applied in the real world reference system to compute the actual distances $d_{i-O,W}$ between the i-th corner and the optical center O, based on their known mutual distances $d_{i-j,W}$, where $i = 1, 2, 3$ and $i \neq j$. This results in a nonlinear system of three equations with three unknowns, which can be written as follows:

$$\begin{cases} d_{1-O,W}^2 + d_{2-O,W}^2 - 2\, d_{1-O,W}\, d_{2-O,W}\, cos\alpha_{1,2} - d_{1-2,W}^2 = 0 \\ d_{2-O,W}^2 + d_{3-O,W}^2 - 2\, d_{2-O,W}\, d_{3-O,W}\, cos\alpha_{2,3} - d_{2-3,W}^2 = 0 \\ d_{1-O,W}^2 + d_{3-O,W}^2 - 2\, d_{1-O,W}\, d_{3-O,W}\, cos\alpha_{1,3} - d_{1-3,W}^2 = 0 \end{cases} \tag{4}$$

The P3P problem can yield multiple solutions due to its non-linear nature. The geometric interpretation of this is made clear through [47]. Here, the set of non-linear equations is solved implementing Finsterwalder's solution, as reported in [48]. Finsterwalder proposed to introduce a new variable to transform a fourth-degree non-linear equation into directly solvable second-degree equations, leading to greater accuracy in the results since no numerical approximations are involved. The choice of this algorithm comes from a comparison of methods for solving the P3P problem, where it proved to have higher accuracy. This is especially important for structural monitoring purposes as small angles are generally involved when the targets are far from the camera.

For each target, the P3P method is applied to every combination of triplets of square corners, enabling the calculation of the distance between the optical center and each corner in the real world multiple times. Finally, the results related to the same corner are averaged to enhance the accuracy of the distance estimate. The procedure is repeated for each video frame (i.e., for each time instant), resulting in the displacements $x'(t)$, $y'(t)$ and $z'(t)$ of each square corner i relative to the camera coordinate system at each time step:

$$\begin{aligned} x_i' &= \frac{d_{i-O,W}}{\sqrt{f^2 + \hat{u}_i^2 + \hat{v}_i^2}}\, \hat{u}_i \\ y_i' &= \frac{d_{i-O,W}}{\sqrt{f^2 + \hat{u}_i^2 + \hat{v}_i^2}}\, \hat{v}_i \\ z_i' &= \frac{d_{i-O,W}}{\sqrt{f^2 + \hat{u}_i^2 + \hat{v}_i^2}}\, f \end{aligned} \tag{5}$$

In particular, z' represents the axis along the camera line of sight, while x' and y' are the horizontal and vertical axes of the camera. These displacement values are then used to compute the target overall displacement and rotation as follows. To identify the target plane, the equation of a plane is calibrated to best fit the coordinates of the key points in the 3D space of the target. The equation of the fitting plane is:

$$ax' + by' + cz' = 1 \tag{6}$$

Given a set of n square corners for a checkerboard target, and considering the measurement and positioning errors in real condition, the coefficient a, b, and c can be obtained by means of the least-square method. For this purpose, the matrix $\mathbf{P_c}$ containing the corner coordinates is first constructed:

$$\mathbf{P_c} = \begin{bmatrix} x'_1 & x'_2 & \dots & x'_n \\ y'_1 & y'_2 & \dots & y'_n \\ z'_1 & z'_2 & \dots & z'_n \end{bmatrix}^T \tag{7}$$

and coefficients are obtained as:

$$\begin{bmatrix} a \\ b \\ c \end{bmatrix} = \left(\mathbf{P_c}^T \mathbf{P_c}\right)^{-1} \mathbf{P_c} \mathrm{U} \tag{8}$$

where U is unit vector of dimension $n \times 1$. After identifying the target plane through the coefficients $[a \quad b \quad c]$, the target displacement over time is determined by the movement of its centroid. Meanwhile, the target rotations appear as variation in the plane normal and/or the checkerboard directions.

Note that this approach offers much more flexibility than the simple scale factor, which only provides accurate results when the camera-to-target line of sight is perpendicular to the target. In contrast, the proposed method can adjust for perspective distortions caused by varying angles, which are common when monitoring large structures and/or when there are unavoidable restrictions on camera positioning.

This paper presents two possible strategies for evaluating the target displacement. The first involves accounting for the variation over time of the distance between the camera and each square corner, namely $d_{i-O,W}(t)$, allowing for the evaluation of the target displacements in all three directions, as well as the three rotations. The second averages the distances between each square corner and the camera across the entire time period, resulting in average distances over the full acquisition window $\bar{d}_{i-O,W}$. While the latter enhances the accuracy of the structural displacement calculation by reducing noise through averaging, some components of the target rotation may be lost. In particular, if the camera-to-square corner distance is averaged over time, only displacements and rotations within the target plane can be identified. Otherwise, if this distance varies over time, also out-of-plane displacements and rotations can be computed.

2.6 Calculation of displacements relative to the structure reference system

To ensure a clear interpretation of the structural behavior, the displacements calculated in Sect. 2.5 are subsequently converted from the camera-centered reference system $(x'\, y'\, z')$ to the structure reference system $(x\, y\, z)$. By positioning the target so that the directions of the checkerboard align as closely as possible with the directions of the structural movement to be measured, the structure reference system can be defined with two axes aligned to the checkerboard directions, and the third axis perpendicular to the target plane.

The three orthonormal vectors defining the reference system are extracted by means of a Principal Component Analysis (PCA), that enables the computation of a 3×3 change-of-basis matrix from the camera reference system to that of the structure, thereby ensuring the identification of the three principal directions as an orthonormal basis. Finally, the change-of-basis matrix is applied to roto-translate the displacement from the camera reference frame to the structure reference frame.

2.7 Filtering camera shake effects

The procedure described so far allows evaluating the relative movements between the camera and the monitored positions. To assess absolute movements, however, the effects of camera movement caused by wind, environmental noise, possible transmission of structural vibrations, or other factors need to be filtered out. For this purpose, a set of artificial targets should be placed in a stable position (e.g., on the ground, far from the monitored structure). Assuming these ground-based targets remain stationary over time, any movement of the camera will be identified as apparent movement of the ground-based targets. Once the displacement and rotation over time of the ground-based targets are identified in the same way as for the target on the structure, their averaged movements can be subtracted from the relative movements of the targets on the structure, resulting in the absolute structural movements.

3 Laboratory experiments

Laboratory tests have been performed to validate the proposed vision-based procedure. The experiments were carried out under controlled conditions, using a steel frame subjected to excitation from a shaking table (see Fig. 3). High-contrast checkerboard targets were fixed at the top and base of the frame to capture the corresponding displacement time histories. Moreover, a third checkerboard placed on the laboratory floor is used to detect any potential camera movements. Videos were captured using a Panasonic Lumix GH6 camera, recording in 4K at 50 Frames Per Second (FPS). For comparison and validation purposes, Linear Variable Displacement Transducers (LVDTs) were installed on the steel frame. The adopted LVDTs are able to measure displacements in the range [0, 100 mm] with a sensitivity equal to 80 mV/V and an excitation voltage of 10 V.

Several tests were performed, varying the input excitation, the camera-to-frame distance and the angle of incidence β, where β represents the angle between the camera

Fig. 3 Laboratory experiment framework: **a** steel frame, **b** equipped steel frame

Table 1 Laboratory experiment results: measured camera-to-target distance d_{meas} and tilt angle β; standard deviation of the vision-based signal in unforced conditions in pixels ($\sigma^{pxl}_{(0-5)\text{s}}$) and millimeters ($\sigma^{d}_{(0-5)\text{s}}$ and $\sigma^{\overline{d}}_{(0-5)\text{s}}$), based on instantaneous (d) or time-averaged ($\overline{d}$) camera-to-corner distance; RMSE between reference and vision-based displacements during forced conditions, $\text{RMSE}^{d}_{(5=25)\text{s}}$ and $\text{RMSE}^{\overline{d}}_{(5-25)\text{s}}$, with superscripts indicating the type of camera-to-corner distance; averaged post-processed camera-to-target distance $\overline{\overline{d}}$, and its deviation ϵ_d from the measured distance d_{meas}

Test	d_{meas} (m)	β (°C)	$\sigma^{\text{px}}_{(0-5)\text{s}}$ (pixel)	$\sigma^{d}_{(0-5)\text{s}}$ (mm)	$\sigma^{\overline{d}}_{(0-5)\text{s}}$ (mm)	$\text{RMSE}^{d}{}_{5-25\text{s}}$ (mm)	$\text{RMSE}^{\overline{d}}{}_{5-25\text{s}}$ (mm)	$\overline{\overline{d}}$ (m)	ϵ_d (%)
1	1.84	0	0.0082	0.0072	0.0066	0.2220	0.2216	1.85	0.54
2	10.90	0	0.0077	0.1033	0.0145	0.1382	0.0857	10.94	0.37
3	25.21	0	0.0073	0.1904	0.0313	0.3807	0.1986	25.57	1.43
4	1.85	45	0.0094	0.0549	0.0087	0.3423	0.2058	1.87	1.08

Line of Sight (LoS) and the normal to the target plane. Four test configurations were considered, referred to as Test 1 to 4. In the first three tests, the camera LoS was perpendicular to the target plane ($\beta = 0°$) and the camera-to-frame distance was increased from approximately 2 m to 25 m. In Test 4, the camera was placed about 2 m far from the frame, with the LoS tilted at a 45 °C angle relative to the normal of the target plane. The four test configurations are listed in Table 1. The excitation applied to the frame base included both sinusoidal base displacements and simulated earthquake ground motions. Among them, the results presented in the following refer to the Irpinia ground motion [49].

As an example, results presented below show the displacements obtained from the target positioned at the base of the shaking table, the same target against which the structure reference system has been defined (with its x and y axes shown in Fig. 3(b)). In relation to the adopted reference system, the actuator applied displacements in the x direction. Results presented hereinafter have been corrected for potential camera vibrations using the procedure outlined in this paper, although this correction has minimal impact in a laboratory setting.

First, Fig. 4a and b display the displacements $u(t)$ and $v(t)$ in pixels, referred to the image plane, obtained from Test 3 and Test 4, respectively. It can be observed that in Test 3, where the camera LoS was approximately perpendicular to the target plane, only the horizontal displacement component ($u(t)$) was identified. On the contrary, when the camera LoS was tilted relative to the normal of the target plane, as in Test 4, both displacement components are present due to perspective effects. Moreover, the difference in the displacement amplitude is also related to the camera-to-frame distance, which was approximately 25 m in Test 3 and 2.5 m in Test 4. Displacements converted from the image plane to the structure-referenced real world result in the displacements shown in Fig. 4c and d for Test 3 and 4, respectively, both in the x direction. These figures also compare the estimated displacements with the reference

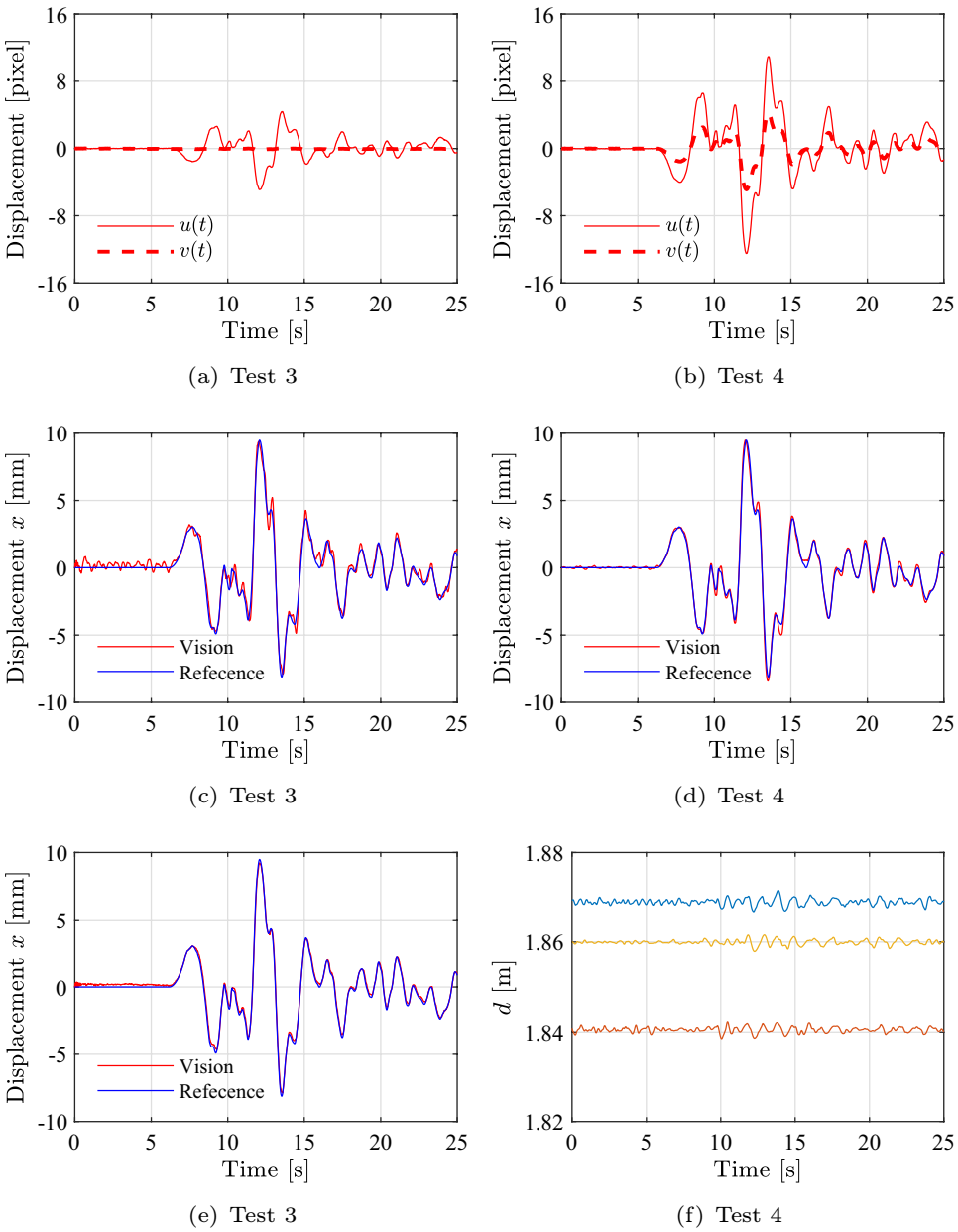

Fig. 4 Laboratory experiment results: (**a**,**b**) image-plane displacement in pixel units of Test 3 and Test 4; (**c**, **d**) frame system referenced displacements in millimeters with varying camera-to-corner distance d of Test 3 and Test 4; (**e**) frame system referenced displacements in millimeters with averaged camera-to-square corner distance $\overline{d}$ of Test 3; (**f**) time variation of the camera-to-corner distances d

displacements, demonstrating excellent agreement. Figure 4dalso highlights the capability of the proposed procedure to correct the effects resulting from the inclination of the camera LoS in relation to the monitored object. Such a correction would not have been possible using a simple scaling factor.

Results of Fig. 4c and d are obtained under the assumption that the distance from each square corner to the camera changes over time. As an example, the time history of this distance for three square corners during Test 4 is shown in Figure 4f, where it can be observed that its variation is affected by both the actual movement of the target and by identification errors, which introduce noise into the distance estimate. Indeed, during periods of greater target displacement, particularly between 10 and 15 s, the distance exhibits increased variability. On the contrary, if the distances between each corner and the camera are averaged over time, the accuracy in the displacement estimation is increased thanks to the noise reduction at the expense of the possibility of estimating rotations out of the target plane. Figure 4e displays the displacement for Test 3 obtained assuming time-averaged camera-to-corner distances, which is significantly less noisy than the one in Figure 4c.

The main findings of the conducted tests are summarized in Table 1. This table shows the standard deviation of the signal recorded during the initial 5 s of the test, a period without any external excitation. Therefore, the data from this interval only reflect the signal noise. The standard

deviation σ is computed for the signal in both pixel and metric units. In the metric case, calculations are performed using the time-averaged camera-to-corners distance ($\overline{d}$) as well as the time-dependent distance (d). The standard deviation in pixel units is denoted as σ^{px}, while the values in metric units as $\sigma^{\overline{d}}$ for the averaged distance and σ^{d} for the time-varying distance.

The average signal noise estimated for the four tests is about 0.008 pixels, ($\sigma^{\mathrm{px}}_{(0-5)\mathrm{s}}$ in Table 1), indicating that displacements with amplitudes of at least 0.08 pixels can be reliably detected. Regarding displacement in metric units, the signal noise increases with both the camera-to-frame distance as well as the tilt angle. This increase is less pronounced when the time-averaged camera-to-corner distance is used. Specifically, when the time-dependent distance is considered, the signal noise $\sigma^{d}_{(0-5)\mathrm{s}}$ rises from 0.0072 mm to 0.1904 mm as the distance increases from approximately 2 to 25 meters. These values decrease significantly to a range of $0.0066-0.0313$ mm when the time-averaged distance is used ($\sigma^{\overline{d}}_{(0-5)\mathrm{s}}$). At a fixed camera-to-frame distance of about 2 m, the noise standard deviation increases from 0.0072 mm to 0.0549 mm as the tilt angle increases from 0 °C to 45 °C. For the case of a 45 °C tilt angle, the standard deviation is substantially reduced to 0.0087 mm when the time-averaged distance is applied.

Table 1 also presents a comparison between the displacements estimated using the proposed vision-based method and the reference measurements during the forced vibration phase. The comparison is based on the Root Mean Square Error (RMSE), defined as the standard deviation of the differences between the two time series over the interval from 5 to 25 s. Analogous to the evaluation of signal noise, this is assessed under two conditions: when the distance from the camera to each square corner is averaged over time ($\mathrm{RMSE}^{\overline{d}}_{(5-25)\mathrm{s}}$), and when it varies with time ($\mathrm{RMSE}^{d}_{(5-25)\mathrm{s}}$). In the first case, with the exception of Test 2, the RMSE generally increases with both the camera-to-frame distance and the tilt angle, ranging from about 0.22 mm to 0.38 mm. When the distance is averaged over time, all of these values are reduced to about 0.20 mm. Notably, Test 2 yields even lower discrepancies between the estimated and reference displacements. Given that the maximum displacement is approximately 10 mm, a RMSE amounting to about 2% of the signal amplitude demonstrates excellent result accuracy. Finally, the table also reports the camera-to-target distance measured with a consumer-grade laser distance meter (d_{meas}) and the one estimated by averaging the mean distance of each square corner of the target ($\overline{d}$). The comparison shows very good agreement, with a maximum difference ϵ_d of 1.43%.

4 Steel footbridge experiments

The effectiveness and applicability of the proposed procedure for dynamic identification purposes are evaluated through a real case study involving a 3D truss-girder steel footbridge spanning the Panaro River in Modena, Italy (see Fig. 5a, b). This structure is characterized by a slender and lightweight design, which makes it highly deformable and sensitive to dynamic vibrations induced by pedestrians, cyclists, wind, and other environmental factors. With a total length of 160 meters divided into three spans (45 m, 70 m, and 45 m), the footbridge features a box cross-section of 3.00×3.20 meters, constructed from truss girders made of hollow tubular steel elements.

To assess the dynamic behavior of the footbridge, seven cross sections within the central span are instrumented for monitoring. These include the midspan and six additional sections, symmetrically distributed at 7.5-meter intervals from the center, as illustrated in Fig. 5c. At each of these locations, checkerboard targets labeled with identification numbers (IDs) 1 to 10 in Fig. 5b and c are installed to enable displacement tracking. To compensate for possible camera movement caused by environmental factors, ten ground-fixed targets are also placed in the riverbed. These serve as stable points for filtering camera shake effects and are identified by IDs 11 to 20 in Fig. 5b. Note that in this case several ground-based targets are used for redundancy; however, only two or three are typically enough to compensate for camera shake.

The vision-based monitoring system employs a Nikon D7500 camera positioned on the dry riverbed to capture a transverse view of the footbridge, as depicted in Fig. 5b. Video recordings are made in 4K resolution at 30 FPS. To eliminate disturbances associated with manual camera operation, the initial and final seconds of each video, corresponding to the start-up and shutdown phases, are excluded from the analysis.

An accelerometer-based monitoring system has been installed on the footbridge for comparison purposes. Specifically, 10 biaxial accelerometers are positioned at the same locations as the checkerboard targets to measure vertical and transverse horizontal (lateral) vibrations. The system, developed by Teleco SpA, consists of a control unit and Micro-Electro-Mechanical Systems (MEMS) accelerometers, as described in [50]. Data acquisition is carried out at a sampling frequency of 80 Hz. A laser scanner survey was also performed to obtain accurate measurements of all relevant geometries.

Experimental tests were performed under both forced vibrations and ambient conditions. Preliminary ambient vibration tests based on acceleration responses identified the fundamental vertical and torsional modes of the footbridge,

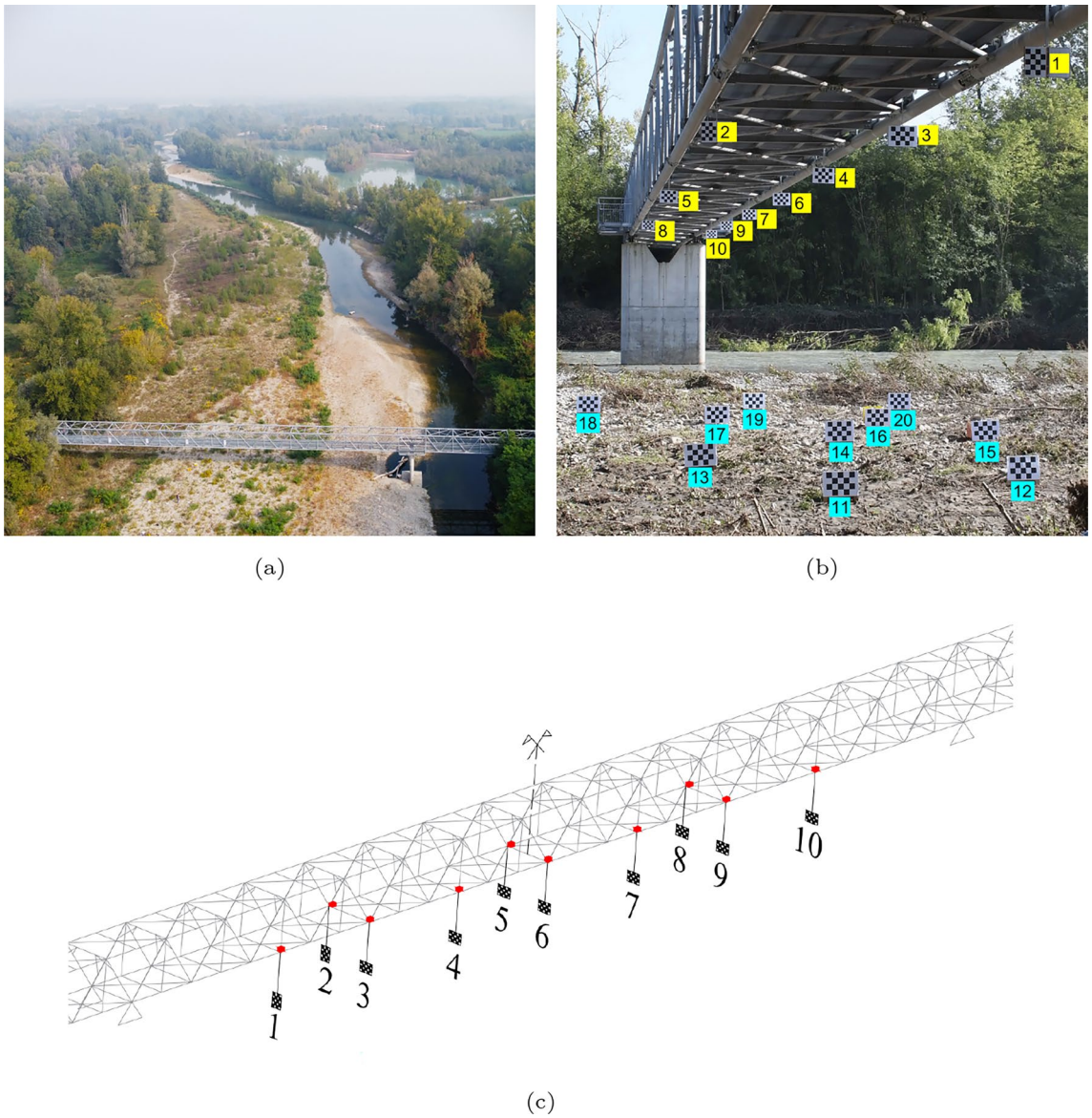

Fig. 5 Footbridge experiment framework: **a** wide view of the footbridge; **b** checkerboard targets on the footbridge (yellow) and the ground (cyan); **c** schematic representation of the measuring positions

with natural frequencies of approximately 2.32 Hz and 4.40 Hz. Based on these findings, experiments under forced conditions were performed by jumping at 138 and 130 Beats Per Minute (BPM), regulated by a metronome, to target the primary vertical modes. The recorded video includes a few seconds before the excitation, a 10-second jumping phase, and the subsequent two minutes to observe the decay of structural motion after the excitation.

The results presented below refer to vibrations induced by jumping at 138 BPM (Test A) as well as to vibrations recorded under ambient conditions (Test B). The outcomes of Test A are dealt with first. Figure 6 presents the midspan (ID6) vertical and lateral displacements estimated using the proposed vision-based approach in comparison with those derived from measured accelerations, as well as their corresponding frequency content. In the vision-based results shown in Fig. 6, displacements are estimated under the assumption of constant distances for the square corners, and they are corrected by subtracting the average apparent displacement measured from the ground targets. Structural displacements are also derived from measured accelerations using an enhanced double integration procedure. This method, proposed in [51], minimizes the squared error between the measured accelerations and the second-order time derivative of the reconstructed displacements. The algorithm was chosen for its accuracy when the overall temporal history of observed displacement has zero mean.

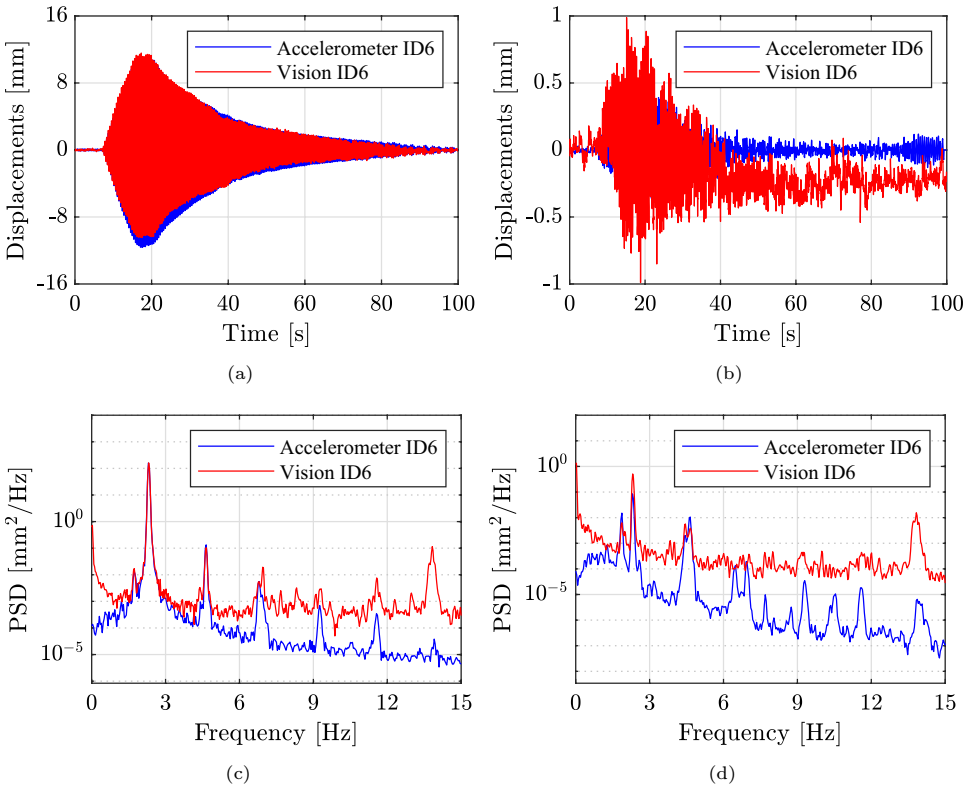

Fig. 6 Footbridge experiments: results of Test A. (**a**–**c**) Vertical and (**b**–**d**) lateral displacement and corresponding PSD at midspan (ID6), as estimated from the vision-based approach (red) and the accelerometer-based system (blue)

The comparison reveals strong agreement between the displacements estimated using the vision-based approach and those obtained from accelerometer data, particularly for vertical vibrations, which exhibit maximum amplitude of about 12 mm. As expected, the lateral vibrations induced by jumping are significantly smaller, with maximum amplitude lower that 0.5 mm, as evaluated from the post-processing of measured accelerations. The lower signal-to-noise ratio associated with lateral vibrations together with possible unwanted movements of the target result in significantly noisier vision-based displacements. Nevertheless, the frequency content of the two displacement estimates is still comparable, although the peaks are more clearly identifiable in the acceleration-based data, especially for lateral vibrations.

Main findings from Test A are summarized in Table 2. For each target location, where IDs 1 to 10 correspond to targets on the footbridge and the remaining ones are on the ground, the camera-to-target distance measured via the laser scanner survey (d_{meas}) is reported alongside the distance estimated using the vision-based procedure ($\overline{d}$). A strong agreement between these two values is observed, with a maximum difference (ϵ_d) of 1.3%. Moreover, the side lengths of the squares composing each checkerboard target are reported in both millimeters and pixels. All targets from IDs 3 to 20 share the same square side length in millimeters, whereas the first two targets are smaller. In contrast, the dimensions in pixels vary depending on both the actual size of the squares and the camera-to-target distance. As expected, the square side length in pixels decreases with increasing camera-to-target distance.

$\sigma^{d}_{(0-5)s}$ and $\sigma^{\overline{d}}_{(0-5)s}$ denote the standard deviation of the vertical displacement measured during the first 5 s of the test under unforced conditions in metric units, considering the time-varying camera-to-square corner distance d and the average distance $\overline{d}$, respectively. The displacement standard deviation for the targets on the structure (IDs 1 to 10) is computed from the signal corrected for camera movement, which was quantified from the ground targets. Overall, the signal standard deviation of the structure-mounted targets (IDs 1 to 10) is higher than that of the ground-based targets (IDs 11 to 20), since the former are influenced by environmental vibrations, while the latter primarily reflect measurement and procedural noise, in addition to any residual camera movements. For the same reason, the signal standard deviation for the targets on the structure is higher close to the midspan, where structural vibrations are greatest. As for the ground-based targets, which mainly reflect noise in the signal, it tends to increase with the camera-to-target distance, while it typically decreases when the average camera-to-corner distance is considered.

Finally, $\sigma^{\text{px}}_{u(0-5)s}$ and $\sigma^{\text{px}}_{v(0-5)s}$ represent the standard deviation of the signal along the directions u and v of the camera reference frame during the first 5 s of the test, expressed in pixel units. In this case, the u and v directions approximately correspond to the horizontal and vertical directions,

Table 2 Footbridge experiments: results of Test A

ID	d_{meas} (m)	$\bar{\bar{d}}$ (m)	ϵ_d (%)	Square side length (mm)	Square side length (pixel)	$\sigma^d_{(0-5)s}$ (mm)	$\sigma^{\bar{d}}_{(0-5)s}$ (mm)	$\sigma^{\text{px}}_{v(0-5)s}$ (pixel)	$\sigma^{\text{px}}_{u(0-5)s}$ (pixel)
1	14.93	14.99	+0.40	50	20.6	0.0372	0.0376	0.0148	0.0066
2	21.59	21.73	+0.65	50	14.0	0.0327	0.0325	0.0113	0.0085
3	21.92	21.92	+0.00	66	18.4	0.0522	0.0519	0.0141	0.0057
4	29.58	29.43	−0.51	66	13.4	0.0682	0.0684	0.0176	0.0086
5	36.48	36.30	−0.49	66	11.2	0.0702	0.0707	0.0138	0.0099
6	36.69	36.67	−0.05	66	10.5	0.0660	0.0658	0.0129	0.0116
7	44.45	44.25	−0.45	66	9.5	0.0652	0.0644	0.0120	0.0155
8	52.24	51.56	−1.30	66	7.7	0.0507	0.0519	0.0106	0.0072
9	52.37	51.79	−1.11	66	8.0	0.0816	0.0847	0.0134	0.0158
10	59.36	58.85	−0.86	66	6.7	0.1032	0.0977	0.0201	0.0145
11	17.99	17.99	+0.00	66	23.2	0.0219	0.0203	0.0075	0.0068
12	19.37	19.56	+0.98	66	20.9	0.0256	0.0185	0.0073	0.0094
13	20.08	19.98	−0.50	66	20.9	0.0245	0.0231	0.0085	0.0077
14	21.84	21.85	+0.05	66	18.7	0.0260	0.0206	0.0069	0.0061
15	22.83	22.84	+0.04	66	17.7	0.0291	0.0256	0.0109	0.0100
16	24.95	25.07	+0.48	66	16.4	0.0358	0.0334	0.0089	0.0054
17	26.22	26.10	−0.46	66	14.9	0.0289	0.0279	0.0076	0.0074
18	27.86	27.55	−1.11	66	14.9	0.0232	0.0233	0.0088	0.0103
19	29.17	29.27	+0.34	66	13.9	0.0308	0.0314	0.0093	0.0081
20	28.51	28.68	+0.60	66	13.9	0.0375	0.0336	0.0087	0.0106

Target identification, measured camera-to-structure distance d_{meas}, averaged post-processed camera-to-structure distance $\bar{\bar{d}}$, percentage difference ϵ_d between them, square side length (pixels and mm), standard deviation of the signal in unforced conditions considering the camera-to-square corner distance varying over time ($\sigma^d_{(0-5)s}$) or averaged ($\sigma^{\bar{d}}_{(0-5)s}$) in mm, standard deviation in pixel units along the *v* and *u* directions of the camera reference frame, denoted as $\sigma^{\text{px}}_{v(0-5)s}$ and $\sigma^{\text{px}}_{u(0-5)s}$, respectively

respectively. The results from the structure-mounted targets show a higher standard deviation in the *v* direction, indicating that environmental vibrations are more pronounced vertically. On the contrary, the ground-based targets exhibit comparable standard deviations in both directions, which is attributed to measurement and procedural noise. In line with the laboratory results, the standard deviation associated with measurement and procedural noise, derived from the ground-based target signal, averages about 0.008 pixel. Consequently, a displacement must reach about 0.08 pixel to be reliably detected. This explains why the lateral displacement of the midspan in Fig. 6bcannot be clearly identified. Based on a rough estimate of the ratio between the square side length in millimeters and pixels for ID6 (66 mm / 10.5 pixels = 6.3 mm/pixel), the minimum detectable displacement of 0.08 pixel corresponds to approximately 0.51 mm. Therefore, the lateral displacement of the midspan remains below the detection threshold.

Regarding the agreement between vision- and acceleration-based measurements during the forced condition, the vertical displacements of the example target ID6 during Test A (please see Fig. 6a) result in a vision-to-reference RMSE of 1.26 mm. However, it should be noted that the displacements derived by accelerations do not capture static displacements associated with jumping, since double integration of acceleration data filters out low-frequency or static components. In contrast, the vision-based displacement estimates are unaffected by such conditions and remain unbiased in all scenarios. Consequently, the RMSE value is reported only qualitatively, to broadly confirm the consistency of the results obtained with the two measurement technologies.

As far as Test B is considered, the vibrations measured in ambient conditions are analyzed to identify the footbridge natural modes. Results of Test B are shown in Figs. 7 and 8 as well as in Table 2. First of all, Fig. 7a and b show the effect of camera movement. While vertical camera motion proved negligible, it noticeably affected the lateral vibration measurements. Specifically, Figure 7acompares the midspan lateral displacement estimated through the vision-based approach with the average movement of the ground-based targets, which represents the camera motion. The midspan lateral displacement corrected for camera motion is shown in Fig. 7b, along with the corresponding displacement derived from the measured acceleration. Although the effect of camera shake has been successfully removed, the lateral displacement estimated via the vision-based approach can not be accurately determined due to its very small amplitude (approximately 0.15 mm, as estimated from the measured acceleration).

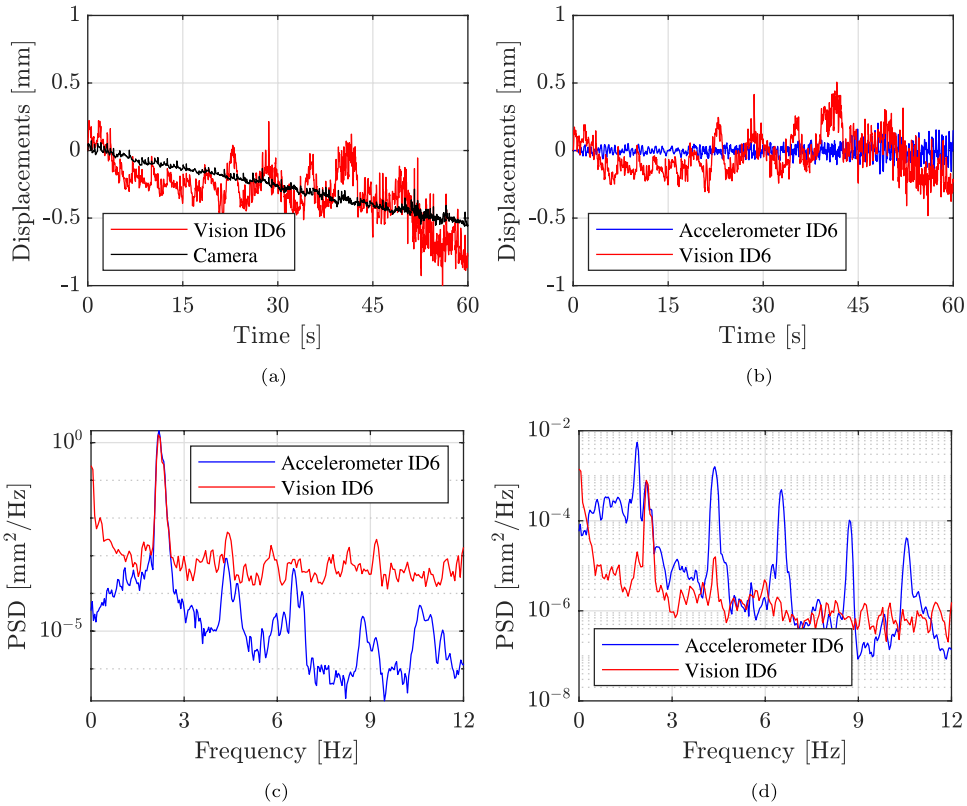

Fig. 7 Footbridge experiments: results of Test B. **a**, **b** Lateral displacement obtained from the vision-based approach (red) and from the post-processing of the measured acceleration (blue) at midspan (ID6). The black line in (**a**) indicates the average displacement estimated from the ground-based targets, representing the camera movement. In (**b**), the vision-based displacement has been filtered to remove camera motion, whereas in (**a**) it is shown unfiltered. (**c**, **d**) Power Spectral Density of the (**c**) vertical and (**d**) lateral displacement estimated from the vision-based approach (red) and the measured acceleration (blue)

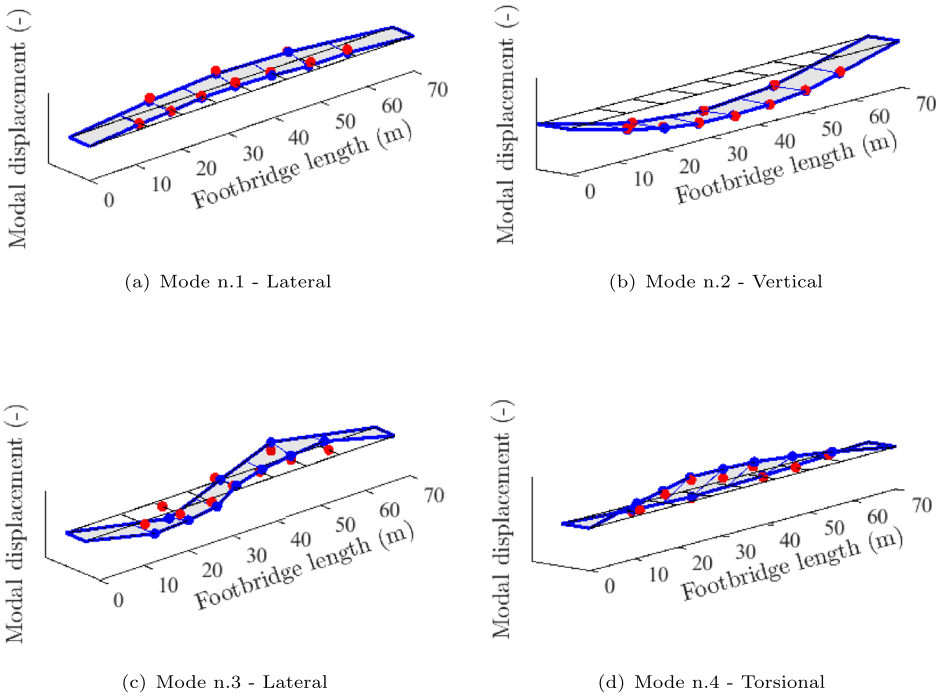

Fig. 8 Footbridge experiment: results of Test B. Mode shape components identified from the vision-based (red dots) and the accelerometer-based (blue dots) monitoring systems. Blue line and gray surface: deformed shape of the footbridge reconstructed from accelerometers. Black line: undeformed footbridge

Figure 7c and d present the Power Spectral Density (PSD) functions of vision-based and acceleration-based displacements in the vertical and lateral direction, respectively. As expected, the frequency content of the displacement estimated from the acceleration is more clearly defined. Nevertheless, the following results demonstrate that the displacement obtained via the vision-based approach still allows for a clear identification of the main natural modes of the footbridge.

Table 3 Footbridge experiments: results of dynamic identification. Natural frequencies obtained from the vision-based and accelerometer-based monitoring systems are denoted as f_{vis} and f_{acc}, respectively. The difference Δf represents the deviation of f_{acc} relative to f_{vis}, while mode shapes are compared in terms of MAC values

Mode n	Mode type	f_{vis}(Hz)	f_{acc}(Hz)	Δf (%)	MAC (%)
1	First lateral	1.871	1.870	0.05	96.7
2	First vertical	2.316	2.319	−0.13	99.6
3	Second lateral	4.297	4.303	−0.14	76.7
4	Torsional	4.429	4.419	0.23	94.3

Both vision-based and acceleration-based displacements are adopted to identify the footbridge dynamic parameters according to the COVariance-driven Stochastic Subspace Identification (SSI-COV) [52]. Modal properties identified in the two cases are summarized in Table 3 and Fig. 8. In particular, the first four natural modes are clearly identified, with frequencies below 5 Hz. The natural modes identified from the vision-based displacement strongly agree with those estimated from the acceleration-based displacement, with maximum differences in terms of natural frequency of 0.23%. Also the mode shapes are very close to each other with MAC (Modal Assurance Criterion) values higher than 90%, except for the third mode. Modes n.1 and 3 are mainly characterized by lateral motion, while Mode n.2 corresponds to a vertical bending mode. Mode n.4 exhibits the characteristic shape of a torsional mode.

5 Conclusions

This work presented and validated an automated vision-based procedure for the structural health monitoring of structures. The proposed methodology relies on the use of checkerboard targets, sub-pixel feature tracking, and P3P-based reconstruction to accurately extract 3D absolute displacements from video recordings using a single camera, even in the presence of camera motion and perspective distortions.

The laboratory experiments on a steel frame confirmed the capability of the approach to detect very small displacements with a high level of accuracy, while the full-scale application to a steel footbridge demonstrated its effectiveness in identifying global modes, both bending and torsional. Comparisons with traditional accelerometer-based system highlighted the strong agreement between vision-based and reference measurements when the displacements are higher than the detection threshold, confirming the reliability of the proposed framework. Specifically, the approach achieves an accuracy of 0.08 pixels. The corresponding metric value depends on the specific application case. For instance, in the experimental case study, an accuracy of 0.1 mm at a 60 m camera-to-target distance is obtained, and 0.03 mm at 25 m. Beyond the influence of camera-to-structure distance, lighting conditions represent another critical factor. The results were obtained under constant illumination, while variable lighting could introduce additional uncertainty into the measurements.

The results emphasize the potential of vision-based monitoring as a cost-effective, contactless, and flexible tool for structural health monitoring, particularly in applications where conventional sensor networks are difficult or impractical to implement.

Acknowledgements The methodology adopted in the present research was partially developed in the frame of the FAR 2023 Project - FOMO line (Vision-based approaches for the structural health monitoring of existing bridges - VIS4SHM), contract E93C23002100007. The financial support of the University of Modena and Reggio Emilia and the “Fondazione di Modena" is gratefully acknowledged.

Data availability All data and code that support the findings of this study are available from the corresponding author on reasonable request.

Declarations

Conflict of interest The authors declare that they have no Conflict of interest.

References

1. Bassoli E, Vincenzi L, D’Altri AM, Miranda S, Forghieri M, Castellazzi G (2018) Ambient vibration-based finite element model updating of an earthquake-damaged masonry tower. Struct Control Health Monit 25(5):2150. https://doi.org/10.1002/stc.2150.e2150stc.2150
2. Quaranta G, Demartino C, Xiao Y (2019) Experimental dynamic characterization of a new composite glubam-steel truss structure. J Build Eng 25:100773. https://doi.org/10.1016/j.jobe.2019.100773
3. Ponsi F, Bassoli E, Vincenzi L (2022) Bayesian and deterministic surrogate-assisted approaches for model updating of historical masonry towers. J Civ Struct Heal Monit 12(6):1469–1492
4. Nicoletti V, Arezzo D, Carbonari S, Gara F (2022) Dynamic monitoring of buildings as a diagnostic tool during construction phases. J Build Eng 46:103764. https://doi.org/10.1016/j.jobe.2021.103764

5. Lu T, Liu J, Guo T, Zhang L, Xia Y (2025) Traffic-induced fatigue damage evaluation of long-span suspension bridge integrating 27-year monitoring data and multi-scale finite element analysis. J Civ Struct Heal Monit 15:2299–2319. https://doi.org/10.1007/s13349-025-00936-8
6. Abu Dabous S, Feroz S (2020) Condition monitoring of bridges with non-contact testing technologies. Autom Constr 116:103224
7. Wen W, Zhang C, Zhai C, Guo J, Hu J (2025) A method for automatic monitoring structural earthquake response using surveillance video. J Build Eng 112:113737. https://doi.org/10.1016/j.jobe.2025.113737
8. Poluzzi L, Barbarella M, Tavasci L, Gandolfi S, Cenni N (2019) Monitoring of the garisenda tower through gnss using advanced approaches toward the frame of reference stations. J Cult Herit 38:231–241
9. Yu J, Meng X, Yan B, Xu B, Fan Q, Xie Y (2020) Global navigation satellite system-based positioning technology for structural health monitoring: a review. Struct Control Health Monit 27(1):2467
10. Yu L, Cao H, Lian J, Li Y, Fan L, Liu H, Liu Y (2025) Gnss-ppp for dynamic deformation monitoring of large-scale engineering structures and a case study. J Civ Struct Heal Monit 15:2213–2227. https://doi.org/10.1007/s13349-025-00945-7
11. Talledo DA, Miano A, Bonano M, Di Carlo F, Lanari R, Manunta M, Meda A, Mele A, Prota A, Saetta A, Stella A (2022) Satellite radar interferometry: potential and limitations for structural assessment and monitoring. J Build Eng 46:103756
12. Bassoli E, Vincenzi L, Grassi F, Mancini F (2023) A multi-temporal dinsar-based method for the assessment of the 3d rigid motion of buildings and corresponding uncertainties. J Build Eng 73:106738
13. Castagnetti C, Bassoli E, Vincenzi L, Mancini F (2019) Dynamic assessment of masonry towers based on terrestrial radar interferometer and accelerometers. Sensors 19(6):1319
14. Fradelos Y, Thalla O, Biliani I, Stiros S (2020) Study of lateral displacements and the natural frequency of a pedestrian bridge using low-cost cameras. Sensors 20(11):3217
15. Wang M, Zhu Z, Koo K, Brownjohn J (2025) Gnss time-synchronised wireless vision sensor network for structural health monitoring. J Civ Struct Heal Monit 15:2725–2747. https://doi.org/10.1007/s13349-025-00953-7
16. Xu Y, Brownjohn JMW (2018) Review of machine-vision based methodologies for displacement measurement in civil structures. J Civ Struct Heal Monit 8:91–110
17. Yoon H, Elanwar H, Choi H, Golparvar-Fard M, Spencer BF Jr (2016) Target-free approach for vision-based structural system identification using consumer-grade cameras. Struct Control Health Monit 23(12):1405–1416
18. Zhao X, Ri K, Wang N (2017) 2017 Experimental verification for cable force estimation using handheld shooting of smartphones. J Sensors 2017(1):5625396
19. Kim S-Y, Lee S-J, Choi K-K (2025) Dynamic characteristics of multi-degree-of-freedom frame systems derived from vision-based displacement measurement using consumer-grade camera video. J Civ Struct Heal Monit 15:2661–2677. https://doi.org/10.1007/s13349-025-00959-1
20. Spencer BF Jr, Hoskere V, Narazaki Y (2019) Advances in computer vision-based civil infrastructure inspection and monitoring. Engineering 5(2):199–222
21. Zona A (2020) Vision-based vibration monitoring of structures and infrastructures: an overview of recent applications. Infrastructures 6(1):4
22. Dong C-Z, Catbas FN (2021) A review of computer vision-based structural health monitoring at local and global levels. Struct Health Monit 20(2):692–743
23. Feng D, Feng MQ (2017) Experimental validation of cost-effective vision-based structural health monitoring. Mech Syst Signal Process 88:199–211
24. Chen JG, Adams TM, Sun H, Bell ES, Büyüköztürk O (2018) Camera-based vibration measurement of the world war i memorial bridge in portsmouth, new hampshire. J Struct Eng 144(11):04018207
25. Xu Y, Brownjohn J, Kong D (2018) A non-contact vision-based system for multipoint displacement monitoring in a cable-stayed footbridge. Struct Control Health Monit 25(5):2155
26. Lydon D, Lydon M, Taylor S, Del Rincon JM, Hester D, Brownjohn J (2019) Development and field testing of a vision-based displacement system using a low cost wireless action camera. Mech Syst Signal Process 121:343–358
27. Dong C-Z, Bas S, Catbas FN (2020) Investigation of vibration serviceability of a footbridge using computer vision-based methods. Eng Struct 224:111224
28. Ye X, Yi T-H, Dong C, Liu T (2016) Vision-based structural displacement measurement: system performance evaluation and influence factor analysis. Measurement 88:372–384
29. Eltouny K, Gomaa M, Liang X (2023) Unsupervised learning methods for data-driven vibration-based structural health monitoring: a review. Sensors 23(6):3290. https://doi.org/10.3390/s23063290
30. Pan X, Yang T, Xiao Y, Yao H, Adeli H (2023) Vision-based real-time structural vibration measurement through deep-learning-based detection and tracking methods. Eng Struct 281:115676. https://doi.org/10.1016/j.engstruct.2023.115676
31. Xin C, Wang C, Xu Z, Qin M, He M (2022) Marker-free vision-based method for vibration measurements of rc structure under seismic vibration. Earthq Eng Struct Dyn 51(8):1773–1793. https://doi.org/10.1002/eqe.3637
32. Han Y, Wu G, Feng D (2022) Vision-based displacement measurement using an unmanned aerial vehicle. Struct Control Health Monit 29(10):3025. https://doi.org/10.1002/stc.3025
33. Grunert JA (1841) Das pothenotische problem in erweiterter gestalt nebst ûber seine anwendungen in der geodâsie. Grunerts archiv fûr mathematik und physik 1:238–248
34. Shao Y, Li L, Li J, An S, Hao H (2021) Computer vision based target-free 3d vibration displacement measurement of structures. Eng Struct 246:113040. https://doi.org/10.1016/j.engstruct.2021.113040
35. Zhang S, Ni P, Wen J, Han Q, Du X, Xu K (2024) Automated vision-based multi-plane bridge displacement monitoring. Autom Constr 166:105619. https://doi.org/10.1016/j.autcon.2024.105619
36. Tan D, Li J, Hao H, Nie Z (2023) Target-free vision-based approach for modal identification of a simply-supported bridge. Eng Struct 279:115586
37. Wang M, Ao WK, Brownjohn J, Xu F (2022) Completely non-contact modal testing of full-scale bridge in challenging conditions using vision sensing systems. Eng Struct 272:114994
38. Geiger A, Moosmann F, Car Ö, Schuster B (2012) Automatic camera and range sensor calibration using a single shot. In: 2012 International Conference on Robotics and Automation, pp. 3936–3943 IEEE
39. Mostafa K, Hegazy T (2021) Review of image-based analysis and applications in construction. Autom Constr 122:103516
40. Paneru S, Jeelani I (2021) Computer vision applications in construction: current state, opportunities and challenges. Autom Constr 132:103940
41. Lucas BD, Kanade T (1981) An iterative image registration technique with an application to stereo vision. In: Proceedings of the 7th International Joint Conference on Artificial Intelligence - Volume 2. IJCAI'81, pp. 674–679. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA

42. Tomasi C (1991) Detection and tracking of point features. 9:137–154
43. Harris C, Stephens M (1988) A combined corner and edge detector. Alvey Vision Conference 15:10–5244
44. Li D, Cheng B, Wang K (2024) Self-calibrating technique for 3d displacement measurement using monocular vision and planar marker. Autom Constr 159:105263. https://doi.org/10.1016/j.autcon.2023.105263
45. Fischler MA, Bolles RC (1981) Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. Commun ACM 24(6):381–395
46. Lu X (2018) A review of solutions for perspective-n-point problem in camera pose estimation. J Phys: Conf Ser 1087:052009. https://doi.org/10.1088/1742-6596/1087/5/052009
47. Wang B, Hu H, Zhang C (2020) Geometric interpretation of the multi-solution phenomenon in the p3p problem. J Math Imaging Vis 62(9):1214–1226. https://doi.org/10.1007/s10851-020-00982-5
48. Haralick RM, Lee C-N, Ottenberg K, Nölle M (1994) Review and analysis of solutions of the three point perspective pose estimation problem. Int J Comput Vis 13:331–356. https://doi.org/10.1007/BF02028352
49. Luzi L, Puglia R, Russo E, D'Amico M, Lanzano G, Pacor F, Felicetta C (2017) Engineering strong-motion database: a gateway to access european strong motion data. In: 16th World Conference on Earthquake Engineering
50. Guidorzi R, Diversi R, Vincenzi L, Simioli V (2010) (2010) Mems-based sensing for health monitoring of buildings. In: Fifth European Workshop on Struct Heal Monit 1:901–906 (**DEStech Publications, Inc**)
51. Lee HS, Hong YH, Park HW (2010) Design of an fir filter for the displacement reconstruction using measured acceleration in low-frequency dominant structures. Int J Numer Meth Eng 82(4):403–434
52. Peeters B, De Roeck G (1999) Reference-based stochastic subspace identification for output-only modal analysis. Mech Syst Signal Process 13(6):855–878