13th International Conference on
Structural Health Monitoring of Intelligent Infrastructure
DOI: 10.3217/978-3-99161-057-1-144

A Comprehensive Approach for Vision-Based Dynamic Monitoring of Structures and Infrastructure

Federico Ponsi [1, 0000−0002−2808−4707], Ghita Eslami Varzaneh [1, 0000−0002−2985−9359], Giorgia Ghirelli [1],
Elisa Bassoli [1, 0000−0002−4919−1421], Loris Vincenzi [1, 0000−0003−2541−7104]
[1] University of Modena and Reggio Emilia, Department of Engineering "Enzo Ferrari", Modena, Italy
email: federico.ponsi@unimore.it, ghita.eslami@unimore.it, giorgia.ghirelli@unimore.it, elisa.bassoli@unimore.it, loris.vincenzi@unimore.it

ABSTRACT: Structural monitoring is crucial for extending the service life of civil structures. Vibration-based monitoring is widely employed across various applications, leveraging both traditional and innovative sensing technologies. Among these, video-based methods have emerged as a promising and cost-effective approach for evaluating structural displacements at critical points. This paper presents a novel vision-based procedure enabling accurate three-dimensional structural displacement measurement using only a single camera. The method applies to assessing dynamic effects on bridges subjected to dynamic loads. The algorithm extracts displacements by tracking predefined targets over time. Special attention is given to reconstructing small 3D displacements from videos that inherently capture two-dimensional projections of the scene. The procedure is validated through experiments on a steel frame in a controlled environment, comparing displacement time histories with imposed vibrations from a shaking table. The originality of this work lies in achieving accurate 3D measurements with minimal equipment, offering a practical and innovative solution for structural health monitoring.

KEY WORDS: Vision-based monitoring; Structural vibrations; Displacement tracking; Laboratory tests; Experimental validation.

1 INTRODUCTION

Structural Health Monitoring (SHM) is a crucial component of modern infrastructure management, offering valuable insights into the condition of structures and helping prevent catastrophic failures while extending their service life. SHM utilizes a combination of sensors, data analysis methods, and computational models to evaluate the performance and safety of civil infrastructure such as bridges, buildings, dams, and other critical structures. The main objective of SHM is to detect changes in the structural integrity or behaviour of a structure, often before visible damage occurs, ensuring continued safety, preventing collapses, and reducing maintenance and repair costs.

SHM systems typically employ a range of sensors to monitor structural responses, such as strain, displacement, and acceleration. These sensors, including accelerometers, strain gauges, displacement transducers, and fibre optic sensors, generate reliable data but they often offer limited spatial coverage and necessitate the installation of dense sensor networks, requiring the structure to be accessible. This can pose challenges during extreme events or when access is restricted, such as during periods of heavy traffic, in remote locations, or unsafe structures. A significant advancement in SHM has been the integration of contactless technologies, which enable the installation of sensors without the need for extensive cabling and, therefore, without interrupting the operation of the structure [1, 2, 3].

These contactless systems allow for easier deployment, even in hard-to-reach or remote areas. With that premise, non-contact monitoring has become increasingly popular. Contactless technologies for civil monitoring encompass a range of methods, including global navigation satellite systems (GNSS) [4, 5], satellite remote sensing [6, 7], terrestrial radar interferometry [8], and vision-based techniques [9]. Among these, vision-based techniques stand out as the only remote sensing approach that can reduce dependence on expensive industrial products [10]. Indeed, these methods have shown considerable promise even when using consumer-grade devices such as standard video cameras or smartphones [11, 12]. This progress is largely attributed to the development of low-cost technologies that provide high resolution and high frame rates, enabling accurate monitoring of large-scale structures in both static and dynamic fields.

These technologies use video feeds to track structural displacements, vibrations, and deformations, providing a flexible, cost-effective alternative to traditional methods and potentially eliminating the need for direct contact with the structure. The primary objective of such a system is to automatically and reliably transform video data into actionable insights. The fundamental concept behind vision-based monitoring is simple: a video of the structure being monitored is recorded, and the individual frames are analyzed, either in real-time or afterward, to extract motion data. This process generates displacement time histories, which can be further used to calculate strains, velocities, and accelerations. Vision-based methods offer several technical advantages, such as directly measuring displacements, which eliminates the need for the double integration of accelerations. Additionally, a single camera sensor can provide distributed monitoring, enabling the extraction of displacement data from multiple points on the structure within one video recording.

Beyond these technical benefits, the vision-based approach allows for substantial cost savings and significantly reduced setup efforts compared to traditional monitoring systems. Due to these advantages, vision-based techniques have garnered increasing attention in civil engineering research. Recent studies, including those by [13, 14, 15, 16], extensively review vision-based applications, including tests on bridges [17, 18, 19], and footbridges, [20, 21, 22].

A vision-based monitoring campaign and set up requires careful consideration and pre-planning, tailored to the specific structure being monitored. In civil engineering, two-dimensional measurements are typically favored because of their practicality and effectiveness. These measurements are

DOI: 10.3217/978-3-99161-057-1-144

commonly employed to monitor vertical and transverse vibrations of bridges, as well as the horizontal displacements of buildings and towers. In vision-based applications, a-priori estimating the expected displacements of the structure is crucial for selecting the right camera parameters and determining the optimal camera-to-structure distance, ensuring accurate detection of displacements. While a single camera is usually sufficient for detecting in-plane movements, capturing the full 3D motion of all relevant points can be challenging. Typically, this requires the use of multiple synchronized cameras, each focusing on different parts of the structure. Differently, the proposed procedure enables the reconstruction of 3D motion using just a single camera. This approach simplifies the installation, eliminates the need for video synchronization and the geometrical merging required in stereovision.

In addition to the camera(s), it is essential to identify the sections of the structure to be monitored. In this regard, it is possible to choose intrinsic notable elements of the structure itself [15, 23], such as prominent details, corners, holes, or bolts, or install artificial high-contrast targets on the sections of interest. The first option does not require access to the structure, avoiding traffic disruption, but the second option generally provides more accurate results.

While highly effective in many applications, vision-based monitoring is featured by critical aspects that cannot be overlooked. A key limitation is the sensitivity to environmental conditions, such as vibrations of the camera or its support due to user intervention or wind, non-uniform air refraction caused by temperature differences between the camera and the monitored object, ambient light condition, weather, and visibility, all of which can affect data accuracy. Literature on the assessment of environmental uncertainties in vision-based monitoring includes theoretical analyses and laboratory testing [16, 24], but outdoor experiments are still limited.

The accuracy of measurements relies not only on the camera technical specifications (hardware) but also on video post-processing (software), which includes challenging tasks such as camera calibration, target tracking and pixel-to-metric conversion. This paper presents a vision-based approach for accurately assessing the condition of civil structures and infrastructure, with particular focus on the transformation of image units into real-world units, which is crucial in large-scale civil constructions where perspective distortions can significantly affect measurement accuracy. Aiming to propose a reliable and validated vision-based method for real-world applications, this study evaluates the procedure in a controlled environment, focusing on detecting the dynamic displacement of a laboratory steel frame subjected to controlled shaking. For validation purposes, the vision-based results are compared to reference displacements, highlighting the potential of this method for accurate monitoring.

The paper is organized as follows: Section 2 outlines the procedure framework, detailing each step of the proposed method, from the setup of the monitoring campaign to the post-processing of the recorded video. Section 3 presents the experimental test, specifically designed to assess the performance of the procedure across different camera-to-structure distances. Finally, Section 4 addresses conclusions and future perspectives.

2 PROCEDURE FRAMEWORK

The proposed vision-based procedure aims to determine the actual dynamic displacement of a structure within its reference system, effectively filtering out camera vibrations and ensuring independence from the camera position and orientation.

The vision-based procedure relies on different transformation of coordinates. To provide clarity, the reference systems involved are described progressively as follows:

1. The 2D image reference system (π), which is related to camera sensor reference system by means of the focal length (f) and optical center (o) in a camera pinhole model. Specifically, the image reference system can be scaled and mirrored (with respect to the optical center) in order to obtain the sensor reference system. The image coordinates of this system are denoted as η and ξ.
2. The 3D real-world reference system (W) located in the optical centre, which represents millimeter displacements, derived from pixel displacements via a three-dimensional mapping process. One axis of the system points in the viewing direction of the camera, along the optical axis. The remaining axes define the plane orthogonal to the optical axis, representing the front side of the camera.
3. The structure reference system defined by the coordinates (x, y, z), which uniquely defines displacements along the main directions where structural motion occurs, ensuring that the results are independent of the camera pose.

The extraction of displacements within the image-plane (π) is straightforward and it is carried out by comparing image coordinates (η, ξ) across sequential frames. However, deriving the displacement time series in the structure reference system requires careful consideration of several key aspects.

These include the precise calibration of camera intrinsic parameters, accurate detection and frame-by-frame tracking of the target position within the image plane, establishing the correspondence between 2-D points in the image coordinate system (π) and their corresponding 3-D points in the real-world coordinate system (W) to account for potential perspective effects, roto-translating the results to align with the motion axes of the structure being analyzed, and filtering out unintended camera shaking to ensure measurement reliability.

To this end, a brief overview of the procedure is provided as follows:

- Stage 1: Monitoring set-up.
- Stage 2: Calibration of the camera.
- Stage 3: Post-processing of the recorded video, with the detection and the tracking of specific features.
- Stage 4: Perspective-3-Points method to establish the relationship between 2-D coordinates in the image-plane (π) and their corresponding 3-D points in the real-world (W).
- Stage 5: Transformation from real-world (W) to structure (x, y, z) reference system.
- Stage 6: Filtering of camera unintended vibrations.

2.1 Stage 1: Monitoring set-up

A vision-based monitoring system requires careful pre-planning based on the specific characteristics of the structure under observation. First and foremost, it is essential to identify the sections of the structure to be monitored, with each selected section being associated with distinguishing features to be

DOI: 10.3217/978-3-99161-057-1-144

tracked. These may either be intrinsic, notable characteristics of the structure itself or artificial high-contrast targets placed in areas of interest, yielding more accurate results at the expense of the need to directly access the structure. Specifically, the procedure is designed for high-contrast artificial targets featuring a checkerboard pattern.

Additionally, it is crucial to estimate the expected structural displacements, as the magnitude of these displacements helps determine the necessary level of accuracy. This, in turn, guides the selection of the camera specifications, including the camera-to-structure distance, which depends on factors such as obstacles or finding a stable vantage point; the optical lenses with an appropriate range of focal length (f) to ensure the desired field of view at that distance; and the frame rate, selected to adequately sample the expected vibration frequencies. For civil structures such as bridges and buildings, dominant modal frequencies typically lie below 10 Hz, indicating that frame rates of 30 frames per second (FPS) are generally adequate. This allows for the utilization of consumer-grade cameras, which is increasingly feasible thanks to recent technological advancements.

2.2 Stage 2: Calibration of camera parameters

Calibrating camera parameters is a crucial step for understanding how the sensor captures and processes visual data. In the current procedure, the calibration is performed according to the diffused approach proposed in [25]. This involves determining several parameters, including mm-to-pixel transformation factor (from sensor to image reference system), focal length (f), and lens distortion coefficients (which account for geometric distortions introduced by the lens). Calibration outcomes will be employed in Stage 4 for the derivation of the relationship between 2-D points in the image-plane (π) and their corresponding 3-D points in the real-world (W).

2.3 Stage 3: Post-processing of the recorded video

Video post-processing is composed of three basic steps: definition of the Regions of Interest (ROIs), feature detection and feature tracking.

ROIs are defined in the first frame of the video as areas surrounding specific targets located on the structure or on the ground. Targets on the structure are key points of interest for dynamic characterization, while targets on the ground are used for camera vibration filtering. In the application case study, checkerboard targets are adopted.

The definition of a ROI for each target allows to narrow the operational area within the video frames, where the features of the targets are detected, thus accelerating the automated analysis. The defined ROIs are managed as matrices of pixels, where each pixel is characterized by its 2-D coordinates (expressed in pixels relative to the frame upper-left corner) and a unique RGB intensity value.

The next step involves the detection of sparse feature points, also known as key points, which characterize the digital representation of each target. A key point is generally a small region of the image characterized by unique and invariant features, described by a matrix or a vector that encodes its characteristics. A wide variety of key point types have been proposed in the literature, along with specific algorithms for detecting and describing them [15]. In the presented procedure, the Harris-Stephens algorithm [26] is used to detect the internal corners of each checkboard target (see Figure 1 for an illustrative example). A corner represents the intersection of two edges, where an edge is characterized by a sharp change in image brightness. In addition to their distinctiveness due to RGB intensity contrast, corners exhibit geometric invariance properties, making them robust features for various applications. Their stability under transformations such as translation, rotation, and changes in scale or illumination enhances their suitability for tasks like detection and tracking.

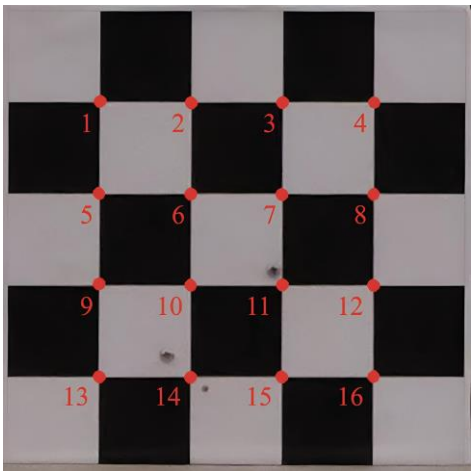

Figure 1. Feature detection via Harris function.

Once the checkboard corners are identified in the initial frames of the video, their locations in the subsequent frames are tracked with the Kanade-Lucas-Tomasi algorithm [27, 28], a well-established technique for visual tracking applications. In this process, the movement of the key points is determined through optical flow estimation. The output of this stage consists of a time series of coordinates in pixel units, for each corner of every checkboard target.

It should be carefully considered that the obtained displacement time series η and ξ solely represent the motion of the features within the image-plane (π). To determine the actual displacements of the structure, additional analysis is required, such as establishing the relationship between 2-D points in the image-plane (π) and their corresponding 3-D points in the real-world (W), mapping the movements to the real-world coordinate system and accounting for any unintended camera shaking.

2.4 Stage 4: Perspective-3-Points method to relate the 2-D image-plane coordinates to 3-D the real-world position

The mapping of the observed 2-D image-plane (π) coordinates into their actual 3-D real-world positions can be obtained by solving the so-called Perspective-n-Point (PnP) problem for the target corner coordinates at each frame. This implies to determine the 3D position and orientation of the camera based on a set of n 2D image points and their known corresponding 3D world coordinates. This is a fundamental problem that was first explored in the photogrammetry literature and later extended to the field of computer vision. The P3P method is a specific case of the PnP problem, where n = 3, namely the camera pose is computed according to the correspondence between 3 points.

DOI: 10.3217/978-3-99161-057-1-144

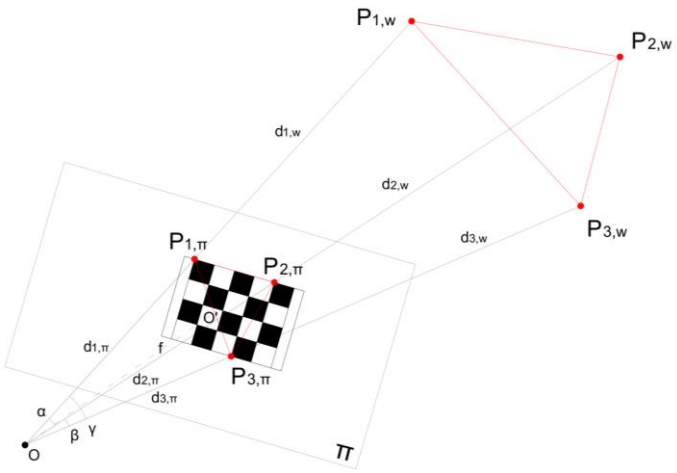

Figure 2. Geometric interpretation of the P3P method.

The solution of the P3P framework used here has its origins in the studies of Grunert [29], and for nearly two centuries it has remained relevant in various research and applications. In this paper, the P*n*P method is used to define the position and orientation of both targets on a structure and the camera. The position of targets on the structure is determined by assuming the camera is not moving while the structure undergoes movement; the motion of the camera is obtained using a fixed target placed outside the shaking structure.

With reference to Figure 2, the intuitive procedure for solving the P3P problem is grounded in the resolution of the law of cosines, which is fundamental for calculating distances and angles in both the image-plane (π) and the real-world (W). The geometric interpretation of Figure 2 refers to a single triplet of checkboard corners, but it can be extended to every possible combination of corners. The law of cosines is first employed for the computation of the angles between the line of sight and the corners in the real-world system. In this phase, the results of Stage 2 and Stage 3 are exploited, namely the calibration outcomes and the coordinate time series of the corners. Then, the law of cosines is re-employed to compute the actual distances between the optical center and the corners of the physical target. This has been the subject of several studies due to the non-linear nature of the problem [30]. Here, the solution of Finsterwalder [31] is used for its high accuracy, as it does not involve any numerical approximations.

For each target, the P3P method is applied to every combination of triplets of corners, resulting in the distance between the optical center and each corner in real-world multiple times. Finally, the distance values related to the same corner are averaged to enhance the accuracy of the estimates. The procedure is repeated for each video frame, associated to a time instant through the frame rate, to obtain the target position at each time step. In this way, the reconstruction of the target displacement with respect to the camera optical center over time is carried out.

This approach is an alternative to the common approach relying on the simple scale factor for unit conversion. The latter only provides accurate results when the camera-to-target line of sight is perpendicular to the target plane. The proposed approach is more flexible, and it can adjust for perspective distortions caused by varying angles between the camera and multiple targets. This is particularly common when monitoring civil structures due to their large scale and/or the presence of unavoidable restrictions on camera positioning.

2.5 Stage 5: Coordinate transformation to the structure reference system

At this stage, a transformation of coordinates into the structure reference system is proposed to ensure a rapid and clear interpretation of the structural behaviour. If the target is positioned in such a way that it aligns as closely as possible with the directions of the main structural movements, two axes of the structure reference system are considered to be parallel to the target directions, and the third one is perpendicular to the target plane. Once this reference alignment is established, a matrix-based change of basis is applied for coordinate transformation.

This involves roto-translating the real-world reference system W into the structure reference frame. The process includes both rotation and translation operations to account for the differences in orientation and position between the two coordinate systems. The core of the transformation is the least-square fitting of a plane to the coordinates of the checkboard corner in the real-world reference system. This transformation allows the representation of displacements in the structure reference system, which is independent of the location and orientation of the camera.

2.6 Stage 6: Filtering of camera unintended vibrations

Up to this stage, the procedure provides the relative displacements between the camera and each target, expressed in the world coordinate system. However, these displacement time series can be affected by camera shaking, which can arise from external factors such as wind or unintended user interactions. To obtain accurate estimates of absolute displacements, it is essential to account for and eliminate these camera-induced contributions. This is achieved by using reference targets placed on the ground in stable positions. These targets are assumed to remain stationary throughout the observation period. Consequently, any apparent displacement they exhibit in the world coordinate system reflects movement of the camera, rather than motion of the targets themselves.

The same tracking procedure outlined in the previous stages can be applied to the ground-based targets to quantify their apparent displacements. The absolute displacements of the targets on the structure can therefore be obtained by subtracting the apparent displacements of the ground-based targets from the relative displacements of the targets on the structure.

In laboratory settings, the camera can often be stably fixed, and the environment is controlled (e.g., no wind), which reduces the need for this correction. In the field, maintaining comparable stability is more challenging, making this filtering step essential. However, since this study focuses specifically on laboratory conditions, field-related considerations are not addressed further.

3 PROCEDURE ASSESSMENT PERFORMED UNDER CONTROLLED CONDITIONS

In this Section, the validation of the procedure described in Section 2 is performed through a laboratory test. The aim is to evaluate the performance of the designed vision-based monitoring procedure and to assess its potential applicability in

DOI: 10.3217/978-3-99161-057-1-144

outdoor scenarios. The test involves the monitoring under controlled conditions of a scale steel frame subjected to excitation from a shaking table (see Figure 3). High-contrast artificial targets with a checkerboard pattern, measuring 250 mm by 250 mm, have been used. Two of them are connected to the base and the top floor of the frame, while a third one is located on the laboratory floor to identify potential camera movements.

The video-monitoring system consists of a Panasonic Lumix GH6 camera and Samsung S23 smartphone recording video in 4K and 8K resolution at 50 and 30 frames per second (FPS), respectively. The results presented below are based on videos captured by the camera, which has a lower resolution compared to the smartphone, making it more sensitive to the noise. Three different scenarios are considered by varying the location of the camera sensor. The distance between the camera and the steel frame for the three examined scenarios is listed in Table 1, measured using a laser meter. In both the scenarios, the angle of incidence between the line of sight and the target plane is nearly zero, implying an almost frontal view of the scene.

Several input excitations have been applied to the frame base during the tests. The results presented in the following refer to the Irpinia earthquake ground motion excitation [32], recorded on November 23, 1980, and reproduced by the shaking table along the *x*-axis of the structure, which is nearly horizontal to the recorded scene.

The accuracy of the vision-based monitoring system is assessed by comparing the estimated dynamic displacements with reference time histories. For the target at the base of the frame, the reference displacements are those imposed by the actuator of the shaking table. To validate the vision-based displacement for the target at the top of the frame, a Linear Variable Displacement Transducer (LVDT) is specifically positioned near the target for this purpose. The adopted LVDT measures displacements within the range [0, 100 mm], with sensitivity of 80 mV/V, excitation voltage equal to 10 V, and sampling frequency set at 200 Hz.

3.1 Results

The results of the monitoring conducted during the experimental test are discussed in this Section. Since the vision-based results for the target at the top of the frame exhibit similar accuracy, the following focuses on presenting the results for the target at the base of the frame.

The displacements of the frame base target in the 2-D image coordinate system, expressed in pixels and identified as detailed in Section 2.3, are shown in Figure 4 and Figure 5 for scenarios 1 and 2, respectively (similar conclusions can be drawn for scenario 3).

It can be observed that the vertical component of the motion, ξ, is approximately zero in both scenarios, since the imposed motion is horizontal to the structure and the camera is perpendicular to the target plane, implying no perspective effects. The difference between the amplitude of the horizontal displacement, η, in the two example scenarios is related to the distance between the camera and the frame, which is about 2 m for scenario 1 and 10 m for scenario 2, implying different pixel coverage on the examined target, as indicated in Table 1.

In this regard, it is specified that the target η-displacement time history in pixels (for example, Figure 5 for scenario 2) is

Figure 3. Laboratory experiment framework.

Table 1. Monitoring scenarios.

ID	Measured distance [m]	Target area [10^3 pxl^2]
1	1.84	186.75
2	10.90	93.02
3	25.21	18.22

calculated by averaging the results obtained by separately tracking the motion of the corners of the checkerboard target (see Figure 6, which shows the motion of four out of sixteen monitored corners, specifically the outer ones: points 1, 4, 13, 16 with numbering following Figure 1), a step that allows for an increase in the accuracy. Indeed, this approach minimizes errors from individual tracking by leveraging multiple data for a more reliable measurement.

Afterwards, following the procedure indicated in Section 2.4, Section 2.5, and Section 2.6, pixel displacements related to the image system are converted into 3D real-world displacements, projected into the structure reference system, and cleared from uncontrolled camera shaking, measured by evaluating the apparent motion of the fixed ground-based target. The vision-based dynamic displacement along the *x*-direction in the structure coordinate system, expressed in millimeters, is represented in Figure 7 and Figure 8 for scenarios 1 and 2, respectively, along with a comparison to the corresponding reference displacement. In this, the reference is the known displacement time history set by the shaking table, which demonstrates excellent validation of vision-based results for all the scenarios.

As discussed in Section 2.4, a key aspect in determining the three-dimensional mapping between image and real-world systems is the evaluation of the distance between the optical center and the monitored target, determined by means of the P3P method. This method allows the calculation of the distances between the camera and any triplet of checkerboard corners within each frame, after which the time-varying

DOI: 10.3217/978-3-99161-057-1-144

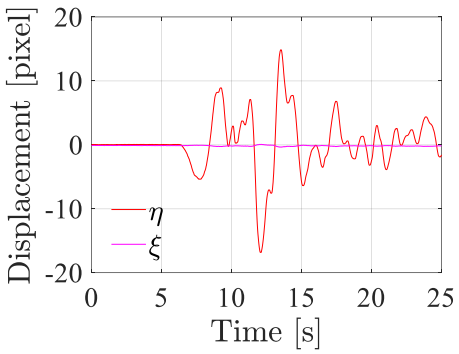

Figure 4. Scenario 1 - Image-plane horizontal and vertical displacements in pixel unit, η and ξ respectively.

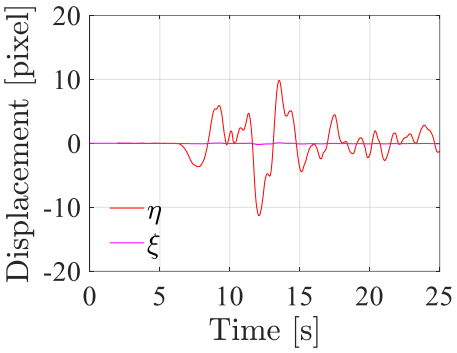

Figure 5. Scenario 2- Image-plane horizontal and vertical displacements in pixel unit, η and ξ respectively.

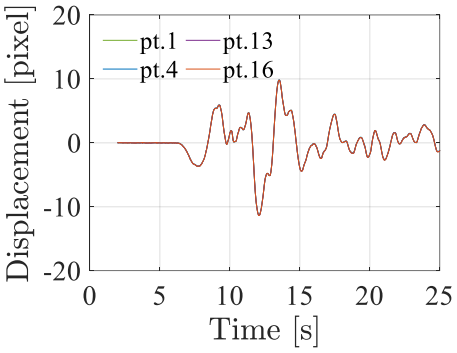

Figure 6. Scenario 2 - Image-plane horizontal displacement, η, obtained by tracking the four outer checkerboard key corners.

Table 2. Vision-based method accuracy.

ID	Estimated distance [m]	$\sigma_{(0\text{-}5)s}$ [mm]	$\sigma_{(5\text{-}25)s}$ [mm]
1	1.85	0.0072	0.2220
2	10.94	0.1033	0.1382
3	25.27	0.1904	0.3807

Table 3. Peak values compared with the scale factor approach.

ID	Reference displacement [mm]	Estimated displacement [mm]	Scale factor displacement [mm]
1	9.48	9.30	9.74
2	9.48	9.17	9.28
3	9.48	9.37	9.07

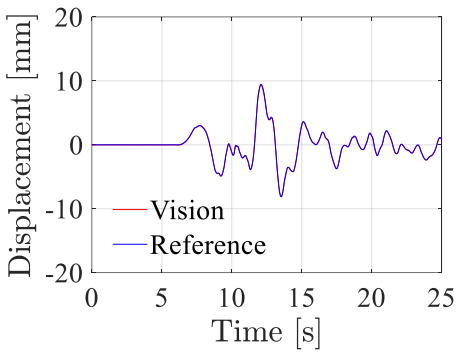

Figure 7. Scenario 1 - Horizontal displacement in the structure coordinate system.

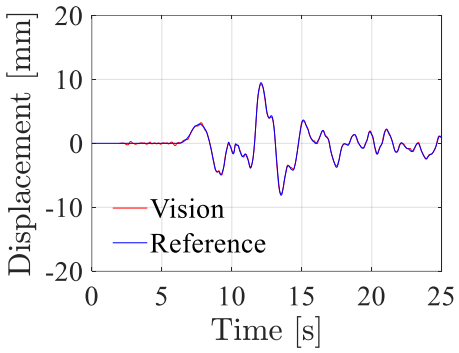

Figure 8. Scenario 2 - Horizontal displacement in the structure coordinate system.

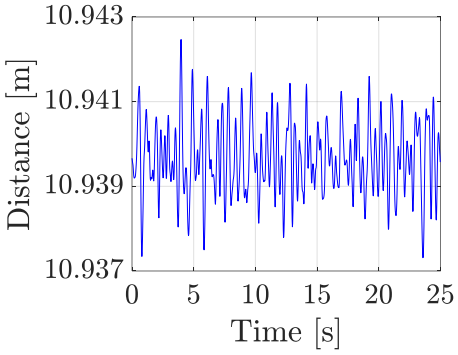

Figure 9. Scenario 2 - Variation of the estimated distance between the camera and the target centroid along time.

Table 4. Peak displacement relative error.

ID	Proposed method error [%]	Scale factor error [%]
1	1.90	2.95
2	3.27	2.11
3	1.05	4.32

distance between the target centroid and the camera is evaluated by averaging the triplet estimates.

The variation of the estimated camera-to-target distance along time is investigated in Figure 9, related to scenario 2. The estimation error is less than 1 mm, which is very small compared to the actual distance of 10.9 m. However, this error

DOI: 10.3217/978-3-99161-057-1-144

is larger than the small change in the real distance caused by the maximum allowed horizontal displacement (about 10 mm). This explains why accuracy decreases as the distance camera-to-scene increases. The camera-to-target distance from the P3P method, averaged over time, is shown in Table 2. P3P-based estimates (Table 2) are in close agreement with the distances measured on-site using a meter laser (Table 1), with relative errors of 0.1, 0.4, and 0.2 % for scenarios 1, 2, and 3.

As a metric in evaluating the reliability of the method, Table 2 also presents the standard deviation $\sigma_{(0\text{-}5)s}$ of the vision-based displacement time series in the first 5 seconds of the tests. During this interval, no excitation is transmitted by the shaking table, so the standard deviation reflects the signal noise and can be considered a measure of the accuracy. The obtained $\sigma_{(0\text{-}5)s}$ is of the order of one-hundredth of mm for scenario 1, denoting a very high accuracy. It reduces to about one-tenth of mm for scenarios 2 and 3, highlighting the impact of the target-to-camera distance on vision-based results.

To further quantify the noise, the standard deviation $\sigma_{(5\text{-}25)s}$ of the difference between reference- and vision-based displacement time histories is calculated for the time interval between 5 and 25 seconds, as reported in Table 2. The standard deviation of the error ranges between 0.1 and 0.4 mm including the three scenarios. Considering that the maximum displacement experienced by the examined (base) target is approximately 10 mm, the accuracy of the measurements is deemed satisfactorily high, with a standard deviation-to-amplitude ratio below 4%, even for the more distant case scenario, indicating reliable measurement accuracy.

Finally, a comparison of the results is presented to assess the accuracy of the P3P method in relation to the scale factor, a simpler and more widely used method in the literature. Scenario 1 is selected as an example, characterized by minimal uncertainty resulting from the (reduced) camera-to-target distance, allowing the error to be entirely attributed to the method employed. The scale factor is calculated as the ratio between the target side in metric units (250 mm) and its side in image units (432 pixel, based on the target area shown in Table 1), resulting in a value of 0.58 mm/pixel.

Using the scale factor method, the horizontal displacement η detected in the image system is simply scaled by the scale factor to obtain the physical displacement x. Thus, the peak absolute displacement of 16.83 pixels (see Figure 4) multiplied by the scale factor returns a physical value of 9.76 mm. As represented in Figure 7, the maximum absolute displacement detected by the P3P method is 9.30 mm, while the reference (i.e., imposed) value is 9.48 mm. This leads to relative errors with the reference peak displacement of 2.95 % for the scale factor approach and 1.90 % for the P3P method. Peak displacements obtained by the two methods for the other scenarios are shown in Table 3, leading to relative errors as indicated in Table 4. These results demonstrate a generally better performance of the P3P-based designed method compared to the scale factor approach. This indicates that the scale factor, which is designed for frontal views of small-scale objects, performs less effectively than P3P in its intended context. The discrepancy between the two methods is expected to become even more pronounced in the presence of inclinations in the line of sight relative to the structural displacement. These findings highlight the critical need for pixel-to-mm 3D mapping in real-world case studies, underscoring the significance of the present research.

4 CONCLUSIONS

This study proposes a vision-based approach for structural displacement monitoring, suitable for both dynamic and static conditions. It is a cost-effective, non-intrusive alternative to traditional sensing technologies. The approach involves the use of consumer-grade cameras and checkboard targets to be installed on the structure.

The reconstruction of the monitored target displacement is facilitated by computer vision algorithms, which detect the checkerboard corners in digital images and track their movement across consecutive frames. The proposed approach implements the Perspective-Three-Point (P3P) algorithm to establish a correspondence between the 2D image coordinates and the 3D world reference system coordinates. The flexibility of this approach makes it particularly suitable for a wide range of camera positions and orientations relative to the monitored structure. Additionally, unintended camera vibrations can be filtered out by tracking one or more targets placed externally to the structure, assumed to be stationary.

The methodology has been validated through a laboratory test on a steel frame excited by a shaking table. Specifically, performed tests focused on evaluating the impact of the camera-structure distance, or alternatively, the target area in the images to account for potential zoom variations, on the accuracy of the displacement estimates. The analysis considered two parameters: the standard deviation of the estimated displacements in the initial seconds of the test (where no excitation was applied) and the standard deviation of the difference between the estimated displacements and the reference values during the remaining part of the test. The first parameter represents the signal noise, whose order of magnitude increases from one-hundredth of a millimeter to one-tenth of a millimeter as the camera-target distance increases from 1.85 m to 25.21 m. Despite this variation, the second parameter, which measures the mean error of the displacement time series relative to the reference displacements, remains satisfactory in all the scenarios, with a standard deviation-to-amplitude ratio around 4 % in the greater structure-to-camera distance scenario.

In comparison, the scale factor approach, a simpler and more widely used method in the literature, was also considered. However, the P3P-based method demonstrated better accuracy in capturing displacement under laboratory conditions, suggesting that the scale factor may not be suitable for outdoor scenarios, which may involve non-frontal views and varying orientations of the targets. Additionally, the scale factor estimation requires user intervention, making it unsuitable for automated procedures.

These preliminary results underscore the potential of this vision-based approach for structural monitoring applications, paving the way for its broader adoption in civil engineering structures and infrastructure. Future research will focus on integrating these initial findings with further tests, particularly examining the effects of the inclination between the line of sight and the target, and refining the displacement accuracy.

DOI: 10.3217/978-3-99161-057-1-144

ACKNOWLEDGMENTS

This work was supported by the FAR 2023 Project - FOMO line (Vision-based approaches for the structural health monitoring of existing bridges, VIS4SHM). The financial support provided by the University of Modena and Reggio Emilia and the “Fondazione di Modena” is hereby gratefully acknowledged.

REFERENCES

[1] A. B. Noel, A. Abdaoui, T. Elfouly, M. H. Ahmed, A. Badawy and M. S. Shehata, Structural Health Monitoring Using Wireless Sensor Networks: A Comprehensive Survey, *IEEE Communications Surveys & Tutorials* 19(3): 1403-1423, 2017

[2] M. Abdulkarem, K. Samsudin, F.Z. Rokhani, and M.F. A Rasid, Wireless sensor network for structural health monitoring: A contemporary review of technologies, challenges, and future direction, *Structural Health Monitoring*, 19(3): 693-735, 2019.

[3] A. Sofi, J.J. Regita, B. Rane, and H.H. Lau, Structural health monitoring using wireless smart sensor network - an overview, *Mechanical Systems and Signal Processing*, 163: 108113, 2022.

[4] L. Poluzzi, M. Barbarella, L. Tavasci, S. Gandolfi, and N. Cenni, Monitoring of the Garisenda tower through GNSS using advanced approaches toward the frame of reference stations, *Journal of Cultural Heritage*, 38: 231-241, 2019.

[5] J. Yu, X. Meng, B. Yan, B. Xu, Q. Fan, and Y. Xie, Global navigation satellite system-based positioning technology for structural health monitoring: a review, *Structural Control and Health Monitoring*, 27(1): e2467, 2020.

[6] D.A. Talledo, A. Miano, M. Bonano, F. Di Carlo, R. Lanari, M. Manunta, A. Meda, A. Mele, A. Prota, A. Saetta, and A. Stella, Satellite radar interferometry: Potential and limitations for structural assessment and monitoring, *Journal of Building Engineering*, 46: 103756, 2022.

[7] E. Bassoli, L. Vincenzi, F. Grassi, and F. Mancini, A multi- temporal dinsar-based method for the assessment of the 3d rigid motion of buildings and corresponding uncertainties, *Journal of Building Engineering*, 73: 106738, 2023.

[8] C. Castagnetti, E. Bassoli, L. Vincenzi, and F. Mancini, Dynamic assessment of masonry towers based on terrestrial radar interferometer and accelerometers, *Sensors*, 19(6): 1319, 2019.

[9] Y. Fradelos, O. Thalla, I. Biliani, and S. Stiros, Study of lateral displacements and the natural frequency of a pedestrian bridge using low-cost cameras, *Sensors*, 20(11): 3217, 2020.

[10] Y. Xu, J.M.W. Brownjohn, and F. Huseynov, Accurate deformation monitoring on bridge structures using a cost-effective sensing system combined with a camera and accelerometers: Case study, *Journal of Bridge Engineering*, 24(1): 05018014, 2019.

[11] H. Yoon, H. Elanwar, H. Choi, M. Golparvar-Fard, and B.F. Spencer Jr, Target-free approach for vision-based structural system identification using consumer-grade cameras, *Structural Control and Health Monitoring*, 23(12): 1405-1416, 2016.

[12] X. Zhao, K. Ri, and N. Wang, Experimental verification for cable force estimation using handheld shooting of smartphones, *Journal of Sensors*, 2017(1): 5625396, 2017.

[13] Y. Xu and J.M.W. Brownjohn, Review of machine-vision based methodologies for displacement measurement in civil structures, *Journal of Civil Structural Health Monitoring*, 8: 91-110, 2018.

[14] B.F. Spencer Jr, V. Hoskere, and Y. Narazaki, Advances in computer vision-based civil infrastructure inspection and monitoring, *Engineering*, 5(2): 199-222, 2019.

[15] C.Z. Dong and F.N. Catbas, A review of computer vision-based structural health monitoring at local and global levels, *Structural Health Monitoring*, 20(2): 692-743, 2021.

[16] A. Zona, Vision-based vibration monitoring of structures and infrastructures: An overview of recent applications, *Infrastructures*, 6(1): 4, 2020.

[17] D. Feng and M.Q. Feng, Experimental validation of cost-effective vision-based structural health monitoring, *Mechanical Systems and Signal Processing*, 88: 199-211, 2017.

[18] M. Wang, W.K. Ao, J.M.W. Brownjohn, and F. Xu, Completely non-contact modal testing of full-scale bridge in challenging conditions using vision sensing systems, *Engineering Structures*, 272: 114994, 2022.

[19] J.G. Chen, T.M. Adams, H. Sun, E. Santini Bell, and O. Büyüköztürk. Camera-based vibration measurement of the world war I memorial bridge in Portsmouth, New Hampshire. *Journal of Structural Engineering*, 144(11): 04018207, 2018.

[20] Y. Xu, J.M.W. Brownjohn, and D. Kong, A non-contact vision-based system for multipoint displacement monitoring in a cable-stayed footbridge, *Structural Control and Health Monitoring*, 25(5): e2155, 2018.

[21] C.Z. Dong, S. Bas, and F.N. Catbas, Investigation of vibration serviceability of a footbridge using computer vision-based methods, *Engineering Structures*, 224:111224, 2020.

[22] D. Lydon, M. Lydon, S. Taylor, J. Martinez Del Rincon, D. Hester, and J.M.W Brownjohn, Development and field testing of a vision-based displacement system using a low cost wireless action camera, *Mechanical Systems and Signal Processing*, 121: 343-358, 2019.

[23] D. Tan, J. Li, H. Hao, and Z. Nie, Target-free vision- based approach for modal identification of a simply-supported bridge, *Engineering Structures*, 279: 115586, 2023.

[24] X.W. Ye, T.H. Yi, C.Z. Dong, and T. Liu, Vision-based structural displacement measurement: System performance evaluation and influence factor analysis, *Measurement*, 88: 372-384, 2016.

[25] A. Geiger, F. Moosmann, Ö. Car, and B. Schuster: Automatic camera and range sensor calibration using a single shot, *IEEE International Conference on Robotics and Automation*, pp. 3936-3943, 2012.

[26] C. Harris and M. Stephens, A Combined Corner and Edge Detector, *Proceedings of the 4th Alvey Vision Conference*, pp. 147-151, 1988.

[27] B.D. Lucas and T. Kanade. An Iterative Image Registration Technique with an Application to Stereo Vision. *International Joint Conference on Artificial Intelligence*, pp. 674-679, 1981.

[28] C. Tomasi and T. Kanade. Detection and Tracking of Point Features. Carnegie Mellon University Technical Report CMU-CS-91-132, 1991.

[29] J. A. Grunert. Das pothenotische problem in erweiterter gestalt nebst bber seine anwendungen in der geodasie, *Grunerts Archiv fur Mathematik und Physik*, 238–248, 1841 (in German)

[30] B. Wang, H. Hu, and C. Zhang. Geometric Interpretation of the Multi-solution Phenomenon in the P3P Problem, *Journal of Mathematical Imaging and Vision* 62: 1214–1226, 2020.

[31] R.M. Haralick, C.N. Lee, K. Ottenberg, M. Nölle. Review and Analysis of Solutions of the Three Point Perspective Pose Estimation Problem, *International Journal of Computer Vision*, 13(3): 331-356, 1994.

[32] L. Luzi, R. Puglia, E. Russo, M. D’Amico, G. Lanzano, F. Pacor, and C. Felicetta. Engineering strong-motion database: a gateway to access European strong motion data. *In 16th World Conference on Earthquake Engineering*, 2017.