An accurate and distraction-free vision-based structural displacement measurement method integrating Siamese network based tracker and correlation-based template matching

Yan Xu[a], Jian Zhang[a*], James Brownjohn[b]

[a] Jiangsu Key Laboratory of Engineering Mechanics, Department of Engineering Mechanics, Southeast University, Nanjing 210096, China.

[b] Vibration Engineering section, College of Engineering, Mathematics and Physical Sciences, University of Exeter, Exeter EX4 4QF, UK.

[*] corresponding author: jian@seu.edu.cn

ABSTRACT

Vision-based displacement measurement receives increasing attention on non-contact bridge monitoring while it faces challenges in long-time field applications due to the presence of environmental variations. To overcome this issue, this study proposes a novel distraction-free displacement measurement approach by integrating deep learning-based Siamese tracker with correlation-based template matching. The Siamese tracker used applies deep feature representations and learned similarity measures for image matching and also considers adaptive template update with time. Since the estimated bounding boxes by the Siamese tracker have size changes within frame sequences, a correction step is added to remove the centroid drifts between the template and the predicted target regions using correlation-based template matching. The proposed method is validated first in an indoor test and then implemented in monitoring tests on a short-span footbridge and a long-span road bridge, demonstrating its potential to handle challenging scenarios including partial occlusion, illumination changes, background variations and shade effects.

KEYWORDS

Displacement measurement, vision-based method, Siamese network, template matching, background variations.

1 INTRODUCTION

Bridge displacement is a significant metric for bridge condition assessment and of great interest to bridge owners. The displacement data collected during the normal operation reflect the bridge serviceability condition while the data collected during controlled vehicle load testing are useful for the estimation of load carrying capacity. For flexible bridges, the displacement data also carry short-time dynamic performance induced by the wind or traffic.

Vision-based measurement receives increasing attention in bridge displacement monitoring due to its advantages of non-contact, easy installation and cost effective, etc. Existing studies have demonstrated the potential application on structural condition assessment including system identification [1–3], finite element model calibration [4], damage detection [5] and bridge WIM system [6].

1.1 Review of existing vision-based displacement measurement methods

Vision-based displacement measurement is the process of localising target patterns in image sequences and converting the computed target motions in image plane to true structural displacement via a projection relationship. Thus, target region localisation is the key component with a few variants of methods available including correlation-based template matching, optical flow estimation and sparse keypoint matching, etc.

Correlation-based template matching is an area-based image matching method that works by searching in a new frame for an area most closely resembling a predefined template. The similarity measure is usually applied to image intensity values of grayscale images over a rectangle area with default resolution in pixel level. Interpolation schemes are added to refine the resolution to sub-pixel level. The method has been widely applied in structural monitoring from the earliest work on the Humber Bridge and Second Severn Crossing in 1990s [7,8] to recent displacement measurement applications on a railway bridge [9], a long-span bridge [10] and a high-rise building [11], etc. Instead of considering consistent two-dimensional rigid motion over the target area, the digital image correlation is an extension mostly used in experimental mechanics under large deformation defining a shape distortion function of the tracking area [12]. It was implemented in a short-span railway bridge monitoring exercise [13] but the large deformation assumption is usually unnecessary for bridge measurement purposes.

Correlation-based template matching method is sensitive to illumination variation, partial target occlusion, partial shading, and background disturbance, etc. and thus is often difficult to guarantee robust performance over a long time in outdoor field environmental conditions [14].

The classic optical flow estimation detects motions or flows of all pixels in an image resulting from brightness pattern shift [15]. The apparent velocity of movement is computed by variational approaches by minimising energy based on brightness constancy and spatial smoothness [16]. The measured results inherently contain sub-pixel resolution and the method was implemented for field monitoring tests on a footbridge [17] and bridge stay-cable vibration measurement [18,19]. Another popular variant is phase-based optical flow estimation based on local phase constancy assumption proposed by Fleet and Jepson [20]. The method mainly focuses on the application of system identification, i.e. extracting modal frequencies and mode shapes in laboratory tests [21,22] and identifying modal frequencies of high-rise tower buildings [23]. In one recent work, Dong etc. [24] proposed a deep learning based full field optical flow approach for displacement monitoring on a grandstand structure.

Sparse keypoint matching techniques apply the process of transforming an image into a collection of sparse feature representations and then finding their correspondences among image sequences using a suitable distance measure. The common descriptors for feature representations are scale-invariant feature transform [25], speeded up robust features [26] and Oriented FAST and Rotated BRIEF [27], etc. and the distance measures are usually the Euclidean distance in feature space [28] for float-point based descriptors and the Hamming distance [29] for binary descriptors. Khuc and Catbas [30,31] applied the FREAK and SIFT methods for deformation measurement in a stadium structure and a railway bridge, and Ehrhart and Lienhart [17,32] adopted the ORB method for deformation measurement in a short-span footbridge. Three review works [14,33,34] summarised challenges faced by vision-based displacement approaches, mainly concerning robustness with respect to environmental variations and camera mounting instability. These limitations could impose measurement uncertainty, especially for continuous long-time tests. Concerning tracking robustness under environmental variations (e.g. target pattern, illumination and background variations), deep learning (DL) based techniques could be a potential solution, learning patterns in visual inputs and improving prediction performance using big data and plentiful computing resources.

1.2 Review of deep learning based target tracking methods

Computer vision techniques have been widely used in structural inspection and monitoring applications, including surface defect detection [36], 3D reconstruction of structural geometry [37], strain [38] and displacement monitoring, etc. Vision-based methods for outdoor applications are highly susceptible to uncontrolled environmental conditions, such as lighting variations, shadows, atmospheric interference and wind gusts [39]. In the task of image-based defect detection, the research focus has recently moved from the earlier image processing methods (e.g. edge and boundary detection, background subtraction and thresholding, etc.) to DL techniques [39]. Most of them use a typical CNN or its variations to classify defeat images pre-trained on a large dataset [41] and successfully applied on images of different lighting conditions and viewing angles [42]. This could provide as a hint for the displacement monitoring task.

DL techniques employ multiple, deep layers of neurons that capture underlying pattern representations from a dataset, enabling them to learn richer abstractions of inputs. The classic feature matching by SIFT, SURF and ORB, etc., are describing the sparse and salient key-points by their local image gradient variables while the DL-based retrieval models for feature representation usually compute hierarchical layer-wise representations, capturing increasingly complex image characteristics. The DL-based feature retrieval models have validated to outperform SIFT-like detectors [35], particularly in cases where SIFT contains many outliers or cannot match a sufficient number of feature points. Besides the successful applications in automatic visual defect detection, DL techniques have been applied in structural health monitoring applications such as data anomaly detection in long-time monitoring data [43] and computer vision-based vibration measurement and modal frequency identification [44], etc.

Convolutional neutral networks (CNNs) are used primarily in computer vision applications. The task of target localization using the end-to-end learning framework is generalised as a classification problem where the decision boundary is obtained by online learning of a discriminative classifier using image patches from the target object and the background. One popular tracking framework is based on ‘Siamese networks’ following the template matching concept. A Siamese network consists of two-branch CNNs with tied parameters. It implicitly encodes the object template and the search region to deep feature representations in another space and then fuses them with a specific tensor to predicts their similarity. The earlier work by Bertinetto et al. [45] first proposed a Siamese tracker (SiamFC), followed by a few extension works. CFNet [46] adds a correlation filter to the template branch and makes the Siamese network shallower but more efficient. However, they are lack of bounding box regression requiring multi-scale test of high computation efforts. The SiamRPN tracker [47] introduces the region proposal network after the Siamese network and performs joint classification and regression for tracking. The DaSiamRPN tracker [48] further introduces a distractor-aware module and improves the discrimination power of the model. These trackers do not consider updating the template image which is inadequate for long-term tracking in presence of appearance changes, fast motion, or occlusion. Zhang et al. [49] proposed a convolutional neural network (CNN) architecture, UpdateNet to learn an adaptive target template update strategy given the initial template, the accumulated template and the template of the current frame. The UpdateNet architecture is general and can be integrated into all existing Siamese trackers. To adapt to the target’s scale and aspect ratio changes, the predicted bounding box is designed to have size changes among frame sequences instead of a fixed size as in traditional template matching. It has the advantage over the traditional template matching method to be robust to occlusion, lighting variation and pattern changes, etc. However, the predicted bounding box centroid might deviate from the template centre in the initial frame due to the predicted size changes. Therefore, it is infeasible to directly apply the method on the structural displacement measurement.

1.3 Purpose of this study

The field applications undergo environmental variations (e.g. illumination conditions, shadow, partial occlusion and other variations), making it challenging for a vision system to achieve a robust and accurate displacement measurement over a long time. To overcome these challenges, this study proposes a novel distraction-free target tracking approach by integrating deep learning-based Siamese tracker with traditional correlation-based template matching. There are a few variations of Siamese trackers for template matching and the DaSiamRPN tracker integrated with the UpdateNet for adaptive template updating is adopted in this work, which is robust in challenging scenarios over long-term monitoring. Different from the fixed target size setting in template matching, the Siamese tracker include a bounding-box regression layer to predict target localisation, which consists of four regression coefficients, two-directional position translation and size scaling of the bounding box. Since our task for structural displacement measurement is based on the quantification of image translations of the target region, a correction step is added to remove the centroid drifts between the template and the predicted bounding boxes due to image size changes using correlation-based template matching. The proposed method is validated first in a laboratory test and then implemented in monitoring tests on a short-span footbridge and a long-span road bridge, demonstrating its potential to handle challenging scenarios including occlusion, illumination and background changes.

To that end, section 2 introduces the basic principles for the DaSiamRPN tracker and the UpdateNet template update scheme as well as our proposed method for structural displacement measurement. Section 3 provides the information of an indoor validation test considering the scenarios of occlusion and illumination changes. Section 4 and 5 demonstrates two field monitoring tests on a short-span footbridge and a long-span road bridge in presence of background changes and shade influence, respectively.

2 PROPOSED METHOD

A vision-based system comprises camera devices, a computer with video processing software and accessories like a tripod. The video processing procedure could fit into a four-component framework in Figure 1, i.e. camera calibration, video streaming, target tracking and displacement calculation.

Camera calibration is to determine the projection transformation between the 2D image coordinate system and the 3D structural coordinate system. The projection transformation could be either scaling factor determined by the camera-to-target distance or planar homography calibrated based on a few planar point correspondences. Target tracking is critical in the video processing procedure to locate the target regions in the image plane through tracking methods. The structural displacement could be easily derived from the outputs of the previous two steps.

Target tracking method is the main factor to influence the measurement accuracy and robustness. Although there are a few variants of target tracking methods with field validations, it is still challenging for a long-term monitoring campaign in the presence of environmental variations. To overcome this limitation, a novel target tracking approach is proposed in this study by integrating deep learning-based Siamese tracker with traditional correlation-based template matching. The main flowchart is shown in Figure 1. With the region of interest (ROI) selected in the reference frame, a Siamese tracker (DaSiamRPN + UpdateNet) is employed to predict the bounding box on the current frame (Frame *i*). Since the predicted bounding box has size changes compared with the template, the output of the Siamese tracker is corrected using correlation-based template matching for target position refinement. The principals of the DaSiamRPN and UpdateNet are introduced in Section 2.1 and Section 2.2 provides the framework of our proposed method.

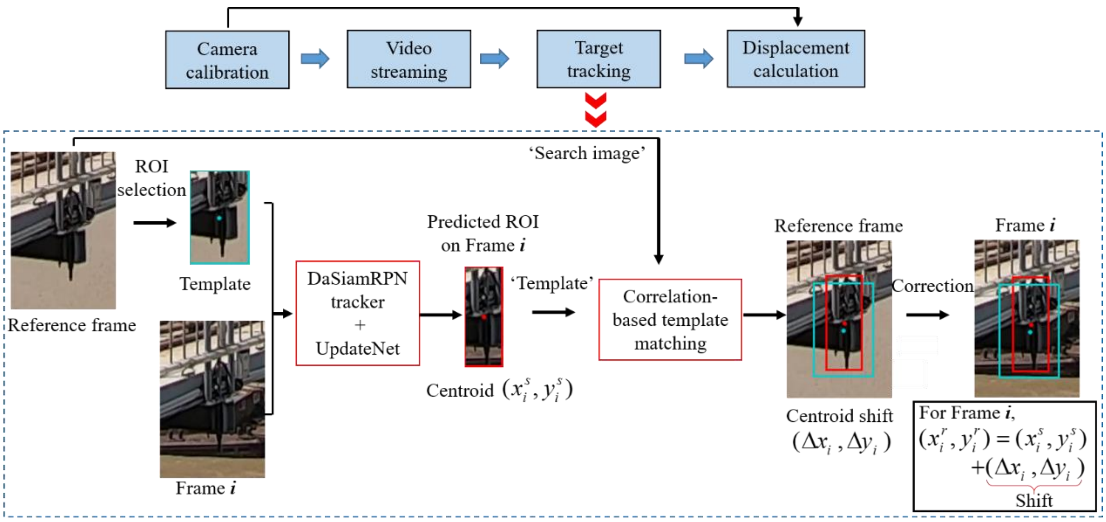


Figure 1 Flowchart of the proposed target tracking method.

2.1 Siamese tracker for object tracking

The Siamese tracker selected for initial template matching is the DaSiamRPN tracker integrated with the UpdateNet. This is because the DaSiamRPN architecture considers a distractor-aware learning module and a local-to-global search region strategy, making it robust in challenging scenarios like occlusion, illumination change and other variations. Also, the UpdateNet is integrated to learn an adaptive target template update strategy to meet the long-term monitoring requirement. The basic principles of the DaSiamRPN and the UpdateNet are briefed introduced here and more details could refer to [48,49].

DaSiamRPN

The DaSiamRPN for object tracking is based on the SiamRPNBIG architecture extended from the SiamRPN. The difference with the initial SiamRPN architecture [47] is that the image dimension of the search area in the current frame is expanded from 255×255 to 271×271 and the number of channels for CNN feature maps from 256 to 512 as shown in the Siamese branch. As shown in Figure 2, it consists of a Siamese subnetwork for feature extraction and a region proposal subnetwork for proposal extraction. Specifically, the Siamese network used is the modified AlexNet [50], where the groups from the second and fourth convolutional layers (conv2 and conv4) are removed. The Siamese feature extraction subnetwork consists of two branches, the template branch with the target patch in the template frame as input and the detection branch using the search region in the current frame as input. The two branches share parameters in CNN so that the two patches are implicitly encoded by the same transformation which is suitable for the subsequent tasks. The Region Proposal Network (RPN) subnetwork consists of two branches, the foreground-background classification branch and bounding box regression branch. The pair-wise correlation between the template feature map and current feature map is firstly computed on both branches with the template feature maps used as kernels. In the classification branch, the computed correlation features are passed through a softmax layer to derive the classification scores representing negative and positive activation of each anchor at corresponding locations on original map. In the regression branch, a linear regression layer is employed to predict four regression coefficients representing the position and size changes of the bounding box to refine the coordinates of the positive anchors. To train the RPN network, the training loss used is a multi-task loss combined by the cross-entropy loss for the classification branch and the smooth L1 loss with normalized coordinates for the regression branch, respectively.

Since high quality training data is crucial for the success of end-to-end representation learning, the DaSiamRPN framework includes a series of strategies to improve the generalization of the learned features and eliminate the imbalanced distribution of the training data. One is to expand the categories of positive pairs by introducing existing large-scale detection datasets and data augmentation techniques. Besides, diverse semantic negative pairs consisting of labelled targets both in the same and different categories are added in the training process.

In the online tracking process, a distractor-aware module is designed which can effectively transfer the general embedding to the current video domain. Distractors in context of the target are selected in each frame by the non-maximum suppression to generate a distractor set. Instead of directly using the cross correlation response between the template and the proposal with highest score in the embedding space as the similarity metric, this response is subtracted by the weighted sum of the cross correlation between the template and the distractor set to make full use of the negative label information.

To adapt the long-term tracking application which might include severe out-of-view or full occlusion, detection scores are taken as a metric indicating the tracking quality and an iterative local-to-global search strategy is designed to re-detect the target during failure cases.

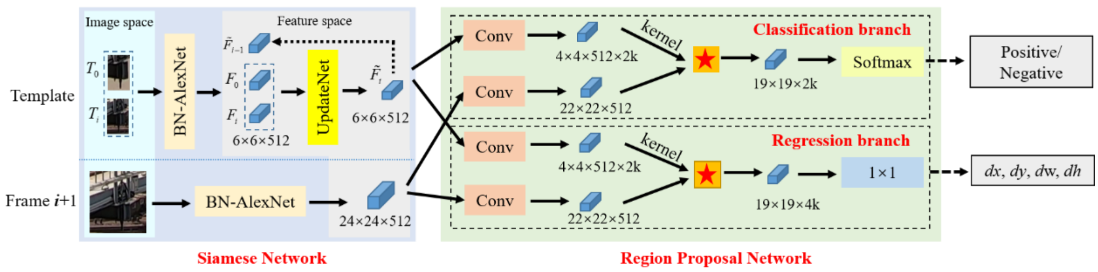


Figure 2 Structure of DaSiamRPN with the SiamRPNBIG architecture integrated with the UpdateNet for template updating.

UpdateNet

In the target tracking step, the template is the basis to find the best candidate region in the new frame and hence a good template is crucial for robust object tracking. However, the effectiveness of the template cannot be guaranteed for long-term object tracking on site in presence of illumination, occlusion and background variations. Therefore, the template must be updated iteratively to improve the matching efficiency for robust long-term object tracking.

The UpdateNet framework is a CNN which aims to estimate the optimal template $\tilde{F}_i$ for tracking the next frame (i+1) given the initial template (T_0),the template of the current frame

(T_i) and the last accumulated template in feature space ($\tilde{F}_{i-1}$). Specifically, the UpdateNet is a two-layer fully convolutional neural network: one $1\times1\times3{\cdot}\mathrm{C}\times96$ convolutional layer, followed by a ReLU and a second convolutional layer of dimensions $1\times1\times96\times\mathrm{C}$ where C = 512. As shown in Figure 2, the input of the UpdateNet is the initial template F_0, the last accumulated template $\tilde{F}_{i-1}$ and the template of the current frame F_i in the embedding feature space by a fixed fully convolutional network and the output is updated accumulated template of the current frame $\tilde{F}_i$. The training process is by minimizing the Euclidean distance between the updated template and the ground-truth template of the next frame. A multi-stage training approach is employed, and the first stage involves template updating using the standard linear update. In the posterior stages, the UpdateNet model trained in the previous stage is applied to get accumulated templates and for object location predictions. The UpdateNet which is compact can easily be integrated into existing Siamese trackers and here it is employed together with the DaSiamRPN.

The Siamese tracker is end-to-end offline trained with large-scale image pairs and then online-tracking as a local one-shot detection task. For the application, the pre-trained model used in the DaSiamRPN is the SiamRPNBIG model provided by the authors [51] trained on VID [52], Youtube-BB [53], COCO Detection [54] and ImageNet Detection [52] datasets with data augmentation. For the UpdateNet, the pre-trained model is derived using a three-stage training on the VOT2018 dataset [55] with the achieved expected average overlap of 0.403.

2.2 Proposed method

The Siamese tracker has the advantage of localising target regions in challenging scenarios like occlusion, illumination change and other variations. The output of bounding box predictions are scale and aspect ratio varied to adapt to the target pattern changes among frame sequences. Therefore, it is infeasible to directly employ the image coordinate changes of the predicted bounding box centroids for structural displacement measurement. To solve this issue, an additional step is supplemented to calculate the centroid drift between the initial template frame and the predicted bounding box regions due to region size changes.

Considering that the traditional correlation-based template matching achieves sub-pixel accuracy in ideal scenarios, it is adopted here to quantify the centroid drifts of the bounding boxes following the Siamese tracker. The correlation-based template matching is the process of searching in the current frame for an area most closely resembling a predefined template in the initial frame. A target region is selected as the template that is a subset image in the reference frame. A matching criterion is defined to evaluate the similarity degree between the template and the new frame and the criterion used is zero-mean normalised cross correlation coefficient (ZNCC). The target location in the new frame corresponds to the peak location in the similarity matrix that has resolution at pixel level. Subpixel interpolation schemes [9] are required to refine the tracking results to sub-pixel level and the interpolation method used in this study is zero-padding in frequency domain using the matrix multiplication form of discrete Fourier transform [56].

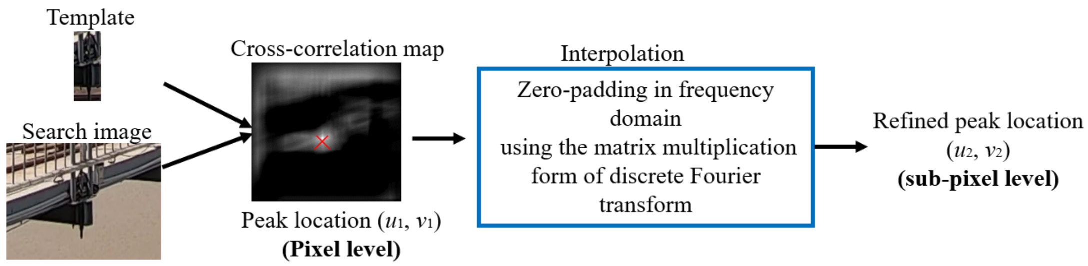


Figure 3 Framework of correlation-based template matching.

Different from general template matching applications, the ‘template’ used as the reference here is the image subset in the current frame predicted by the Siamese tracker as the target region. As shown in in Figure 1, the ‘search image’ which has a larger image size than the template is a subset image cropped from the initial reference frame. Regarding the two image subsets are the predicted resembling area with the highest similarity score by the Siamese tracker, the search range of the peak location in the correlation map could be limited to a small value (e.g. 10 pixels from the centroid). This constraint could effectively avoid the drift error due to apparent target pattern changes in challenging scenarios.

The output of target centroid location in the current frame is the Siamese output corrected by the estimated centroid drift by the correlation-based template matching.

To convert to the structural displacement, the projection relationship between the structural coordinate system and the image plane is pre-computed by the scaling factor using camera-to-target distance or by the planar homography matrix derived from a few planar point correspondences. It is noted that the tracked target is planar and coplanar with the computed 2D displacement.

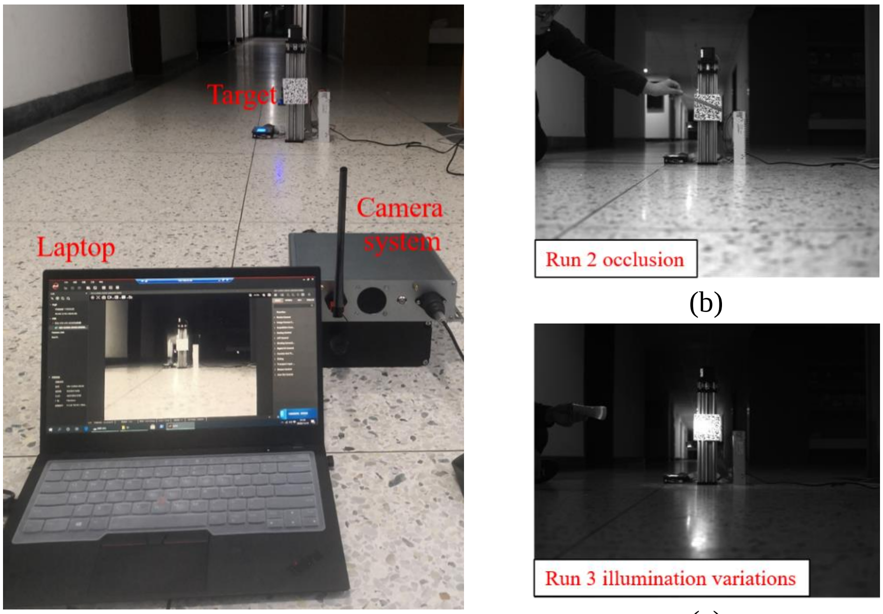
3 LABORATORY VALIDATION

To validate the effectiveness of the proposed method for structural displacement measurement on challenging scenarios, an indoor test of reciprocating motions triggered by a linear actuator was conducted considering two cases including partial occlusion and pattern variations due to illumination. Section 3.1 and 3.2 describe the test setup and the results obtained, respectively.

3.1 Test configuration

A linear actuator was implemented to generate reciprocating motion with the amplitude of 8 cm. The test configuration is shown in Figure 4. A target of speckle patterns with the dimension of 10 cm by 10 cm was attached to the centre of the cover tube.

The video acquisition device used is an industry camera (Hikvision MV-CA050-20UM) with the focal length of 16 mm which is arranged 1.77 m away from the target. The acquired images are grayscale. Three runs of tests were conducted and the differences of test conditions occurred in Run 2 and Run 3 are the presence of partially occlusion of the target and illumination changes caused by an adjacent lamp, respectively shown in Figure 4 (b) and (c).

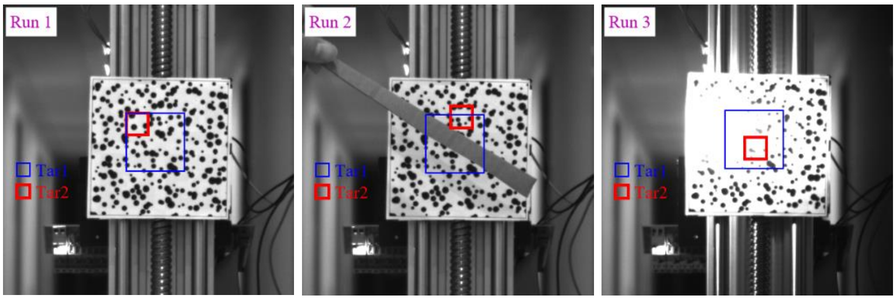


(a) (b) (c)

Figure 4 Test configuration (a) and sample frames captured by the camera in Run 2 (b) and Run 3 (c).

3.2 Test Results

To demonstrate the working performance of the traditional correlation-based template matching methods (denoted as CM in the following study), two targets on the speckle patterns are selected for analysis with the dimensions of 30 pixels by 30 pixels and 80 pixels by 80 pixels, respectively. Figure 5(a) shows the template frame captured in Run 1 and the image was cropped for better visualisation.



(a) (b) (c)

Figure 5 Selected target regions in the template frame in Run 1 (a), one sample frame at 14.2 s in Run 2 (b) and one frame at 19.5 s in Run 3(c).

Tracked motions by the CM in the three runs are indicated in Figure 6. The left column of the figure presents the tracked motions along the image width and height direction and the similarity scores between the predicted target region and the template image for the target region Tar1. The motion along the image width direction is expected to be zeros. For Tar1, the root mean squared errors (RMSE) for three runs are 0.03 pixel, 0.05 pixel and 0.10 pixel while the maximum deviations in three runs are 0.08 pixel, 0.17 pixel and 0.88 pixel. About the motion along the image height direction, the result measured in the ideal case (Run 1) is taken as the reference and the cross-correlation coefficients for the measurement in Run 2 and Run 3 both reach 99.9%. Due to the occurrence of partial occlusion in Run 2 and illumination variations in Run 3, the similarity scores experience sharp decreases in some time periods shown in Figure 6(c). Generally, the CM provides a reliable measurement for the larger target Tar1 in presence of partial target pattern changes.

The right column of Figure 6 presents the measured results for the smaller target Tar2. Taking the measurement in Run 1 as the reference, the measurement in Run 2 and Run 3 includes some sharp large deviations with the amplitude over 20 pixels in both the image width and height directions. Sample frames taken from Run 2 and Run 3 shown in Figure 5(b) and (c) indicate that these apparent measurement deviations correspond to tracking failures when large pattern changes occur on most of the target region.

Analysis results for both two targets demonstrate that the CM is not robust to severe target pattern changes over time and setting a threshold on similarity score measures is necessary to remove the measurement of low confidence or tracking failures. Also, it is better to select a large target region with stable patterns and consistent motions for tracking while it might be not available in field tests.

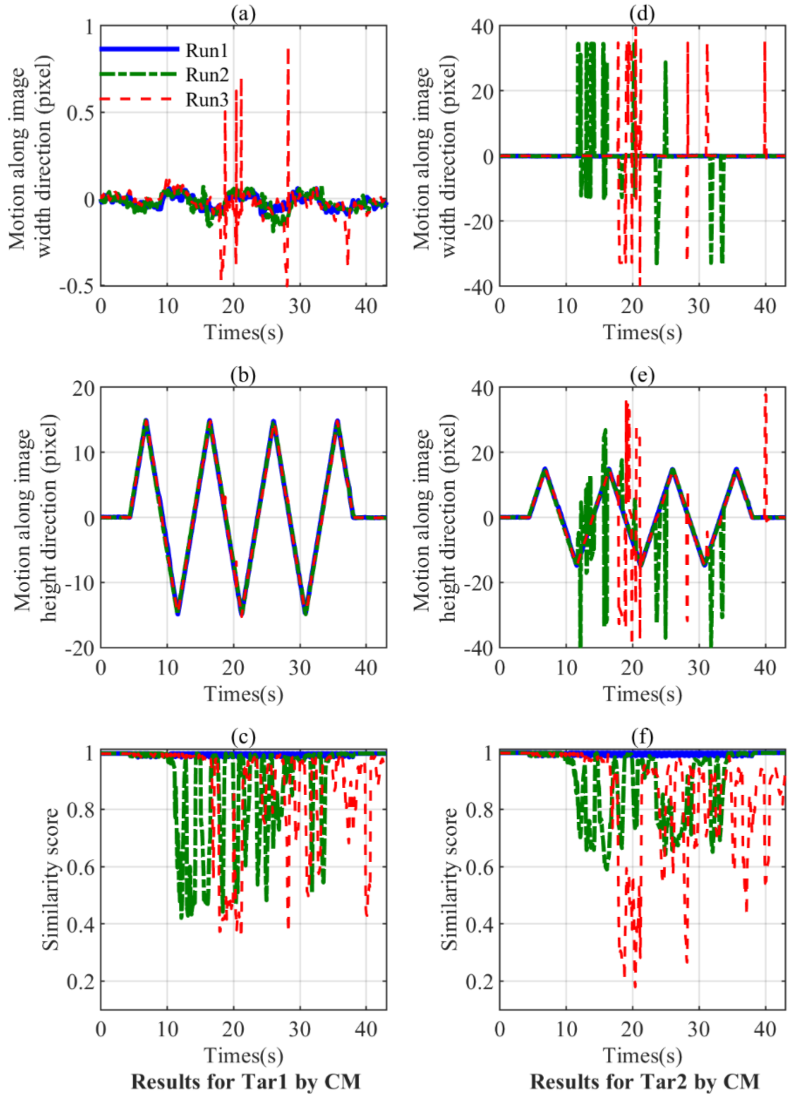


Figure 6 Image motions along two image directions and similarity scores using correlation-based template matching for the target Tar1 (left column, a-c) and the target Tar2 (right column, d-f).

As a comparison, two other classic target localisation methods, i.e. Lucas- Kanade (LK) optical flow and the SIFT matching were implemented to predict the motions of the larger target Tar1 in Run 3. For the LK optical flow estimation, the feature points in the template image was extracted first and then refined to sub-pixel level. Then the optical flow for the extracted sparse features was estimated in the subsequent frames using the iterative Lucas-Kanade method with pyramids. The outliers in feature point correspondences were filtered by apply RANSAC geometric transformation estimation. The remaining feature points are evaluated by the re-projection error of the 2D translation to further check the geometric consistency. The averaged image coordinate movement between feature point correspondences were taken as the image motion of the target region. For the SIFT matching, the key-points were detected for the computation of their descriptors in the template and the subsequent frames independently. The FLANN (Fast Library for Approximate Nearest Neighbours) match was implemented to find nearest neighbours in two sets of descriptors. The detected matches were sorted by their distance with the first 50% closest matches kept. The sorted key-points were post-processed for outlier removal similar to the process in the LK.

The results given in Figure 7 are the averaged image motions of sparse key-points after RANSAC geometric verification. It shows that the measurements contain many outliers due to insufficient number of matched key-points or large re-projection error after RANSAC geometric transformation. Thus, these two methods are not robust for patterns in severe varying lighting conditions.

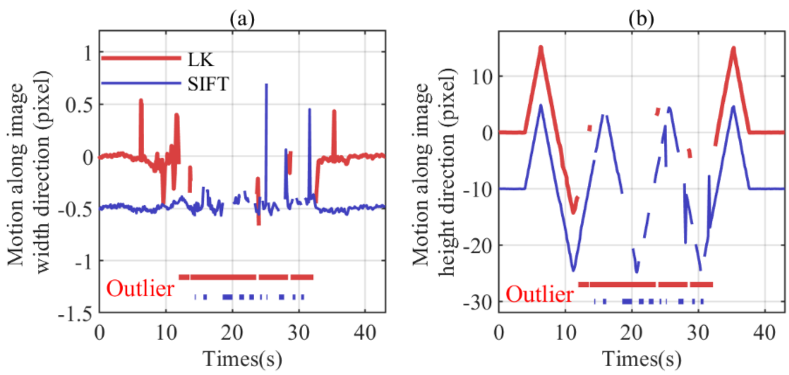


Figure 7 Image motions along image width (a) and height directions (b) for the target Tar1 in Run 3 measured by Lucas-Kanade (LK) optical flow and the SIFT matching.

The DaSiamRPN+UpdateNet and our proposed method (denoted as Siam and Siam+CM in the following study) were also implemented to analyse the video data by tracking the smaller target Tar2 and the results are shown in Figure 8. For the results measured by the Siam (left row), in Run 1, the measured motion in image width direction has a RMSE of 0.18 pixel and maximum deviation of 0.52 pixel while the motion in image height direction reaches a high similarity with results by the CM with the cross-correlation coefficient reaching 99.8%. In Run 2 and Run 3, the predicted bounding boxes deviate from the actual position with box size changes after 11s and 18 s when the target patterns experience apparent variations as shown in Figure 9. The predicted bounding boxes have apparent size changes and shift to adjacent resembling area with salient patterns.

For the results measured by the Siam+CM (right row), the measured motion in image width direction has the RMSEs of 0.04 pixel, 0.04 pixel and 0.06 pixel for the three runs while the maximum deviations are 0.10 pixel, 0.09 pixel and 0.46 pixel, respectively. Compared with the measurement by the CM in Run 1, the motions in image height direction in the three runs are of high similarity with the cross-correlation coefficient over 99.9%.

The 2D structural displacement for the target region Tar2 by Siam+CM is shown in Figure 10. The amplitudes in the vertical direction are 8.0 cm consistent with the test settings. The horizontal movement has a similar shape with the amplitude of 0.1 mm which is the leakage of vertical components caused by the error in the definition of the structural coordinate system.

The measurement speed of the proposed method is provided in Table 1. The Computer hardware specifications are the CPU, Intel Core i9-10900K (10 cores, 20 siblings) and the GPU, NVIDIA GeForce RTX 2080 Ti (11G). The Programming language is Python in Linux environment. By making use of the multiple processing cores, the measurement speed for CM could be increased from 26.4 FPS (frame per second) to 166.4 FPS. The proposed method Siam+CM which runs sequentially the Siam and CM process, reaches the tracking speed of 56.5 FPS which is sufficient for most real-time bridge monitoring applications.

Table 1 Measurement speed of the proposed method

Frame resolution (pixel)		2592×2048
Template dimension (pixel)		30×30
Search image dimension (pixel)		110×110
Measurement speed (FPS)	Siam	85.5
	CM	26.4
	CM (multi)	166.4
	Siam+CM (multi)	56.5

Observations indicate that the Siamese tracker (DaSiamRPN+UpdateNet) localises the target regions by size changes and shift to adjacent regions with high similarity in presence of target pattern changes. By supplementing the correlation-based template matching as a followed correction step, it reaches a robust and accurate measurement in challenging scenarios. The measurement results are evaluated by the actuator records and the root mean square (RMS) errors are summarised in Table 2. The maximum RMS errors in three runs are 0.05 mm in horizontal direction and 0.158 mm in vertical direction. The measurement accuracy in cases of pattern occlusion and varying illumination are similar to that in ideal case.

Table 2 Measurement error of the proposed method in three runs

Run No.	Horizontal direction	Vertical direction	
	RMS error (mm)	RMS error (mm)	Correlation coefficients
Run 1	0.048	0.138	99.95%
Run 2	0.044	0.089	99.98%
Run 3	0.050	0.158	99.92%

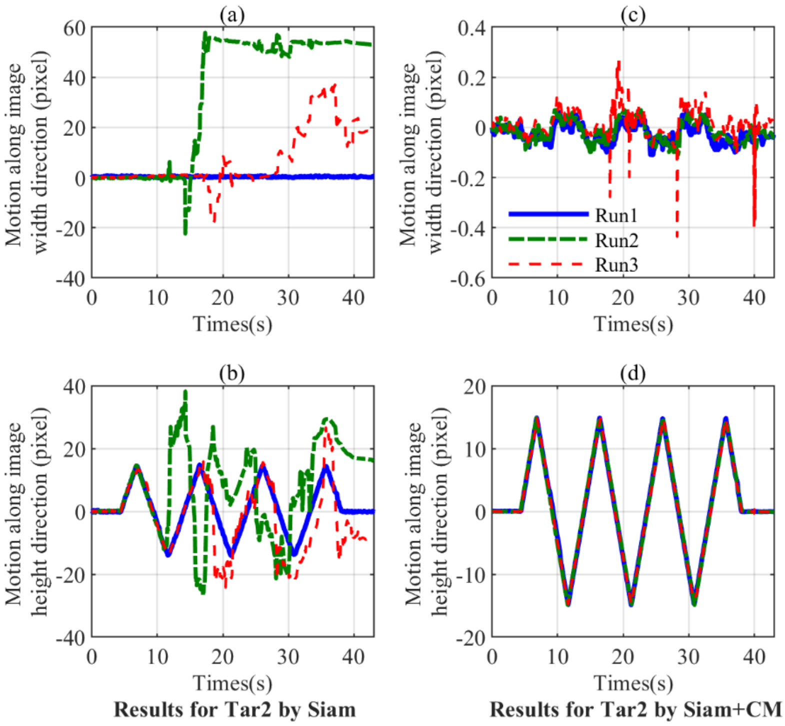
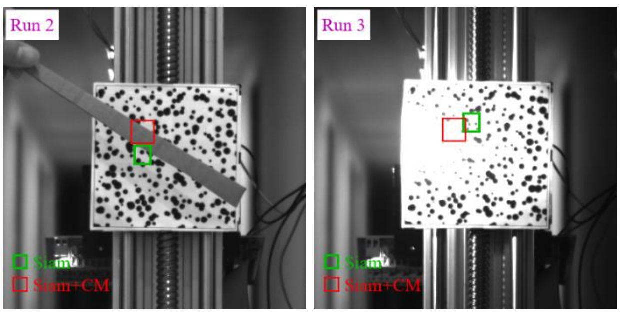


Figure 8 Image motions along two image directions for the target Tar2 using the DaSiamRPN+UpdateNet (left column, a-b) and our proposed method (right column, c-d).



(a) (b)

Figure 9 Demonstration of the predicted bounding boxes by the DaSiamRPN+UpdateNet and our proposed method on the target Tar2: (a) cropped frame in Run 2 at 14.2 s; and (b) cropped frame in Run 3 at 33.0 s.

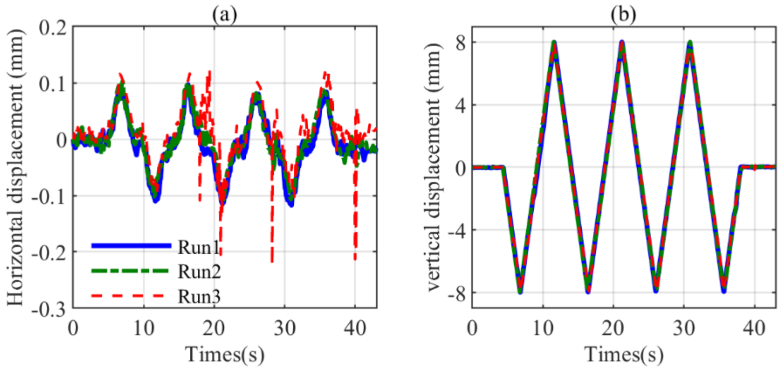


Figure 10 Displacement time histories along horizontal (a) and vertical directions (b) on the target Tar2 using our proposed method.

4 FIELD TEST ON A SHORT-SPAN FOOTBRIDGE

This section describes a case study of using the proposed target tracking method for measuring displacement of a suspension footbridge during bridge rehabilitation status. The recorded video streams include frequent background variations due to coal barge passage. Section 4.1 and 4.2 describe the test setup and the results obtained, respectively.

4.1 Bridge and test information

The tested footbridge as shown in Figure 11 (a) is a canal water overpass in Huai'an, China with the span length of 115.7 m. The camera used for video acquisition was a GoPro Hero7 which was mounted on a tripod in one platform of the east tower. The acquired videos were outputted for post-processing using the proposed method.

One sample frame is given in Figure 11 (b) and the tracked target is the deck area connecting the third vertical hanger, which is approximately 4 meters away from the east tower. The projection transformation used is the planar homography matrix which is calibrated using the six planar point correspondences (marked in Figure 11 (b)) between the image coordinates and the structural coordinates. The known dimensions from the design drawing is the distance between two adjacent vertical hangers (2.0 m) and the distance from the top parapet connectors to the bottom hanger connectors (1.41 m). Cargo barges frequently pass through the canal during the day time as shown in Figure 11 (c) which causes apparent background variations in the deck target region. The video frames were converted to grayscale before performing target tracking process.

(a)

(b) (c)

Figure 11 Pictures of the tested suspension footbridge (a), one sample frame with annotations of target regions and control points (b), and another video frame recorded when a cargo barge passed through (c).

4.2 Measurement results

To demonstrate the effectiveness of the proposed distraction-free displacement measurement method, a 400-second video stream is truncated for analysis. In the beginning 250 seconds, the target patterns experienced apparent background variations due to the passage of a series of cargo barges tied together with mooring cables.

Three classic target localisation methods (CM, LK and SIFT) were implemented to predict the motions with the results shown in Figure 12. For the CM, the tracked results with the similarity scores lower than 0.7 are taken as tracking failure and removed as outliers while there some still some local sharp peaks in the first 250 s. For the LK and SIFT, the outliers are removed by applying threshold (0.1 pixel) on re-projection error after RANSAC geometric transformation.

The measurements by the LK ad SIFT are of high similarity with the correlation coefficient of 97.1%. The CM results after 250 seconds have high consistency with the outputs by the other two methods. It shows that the LK and the SIFT are robust to the partial background distractions while the CM fails to acquire reliable results. This is because the CM is area-based matching which utilises the image intensity information of the whole rectangle target region while the LK and the SIFT only track the sparse and salient key-points within the rectangle target region.

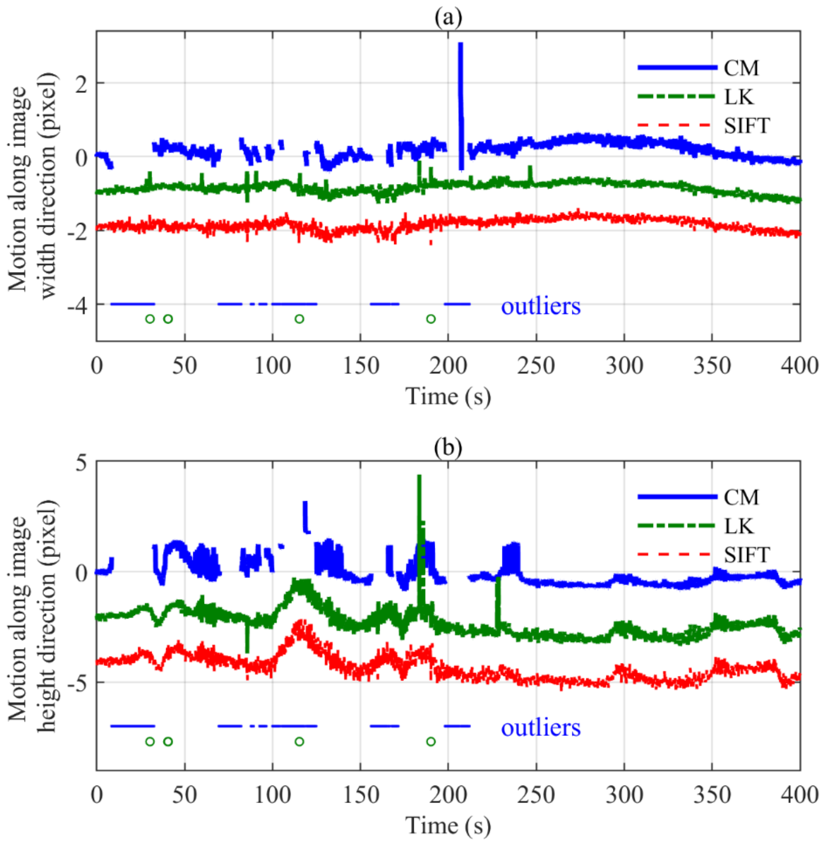


Figure 12 Image motions along two image directions (a-b) using the three classic methods (CM, LK and SIFT) for the deck target.

Besides the three classic methods, the Siam tracker and the proposed method (Siam+CM) were implemented to analyse the video data. The parameter settings are the same as that mentioned in section 3.2.

Figure 13 (a) and (b) shows the time histories of target motion along image width and height directions, and the similarity scores are given in Figure 13 (c). It is noted that the similarity scores in CM and Siam methods are calculated using different functions. In CM, the similarity score is measured by zero-mean normalised cross correlation coefficient of image intensity value between the template and the proposal region. In Siam, the similarity score is measured by the cross correlation of feature representation embedded by a modified AlexNet between the template and the proposal, subtracting the weighted sum of the cross correlation in embedding space between the template and the distractor set.

Four frames at the time steps of 4.6 s, 120.7 s, 204.0 s and 262.9 s are provided in Figure 14 for the demonstration of localisation results by the three methods.

For the Siam, the predicted bounding boxes in four sample frames are sensible even in presence of background distractions but the width and height for the bounding box is varied among video frames. Therefore, the predicted target motions which are taken from the centroid coordinates of the predicted bounding boxes are not accurate.

The proposed method (Siam+CM) is to refine the Siam output by correcting the centroid drift due to bounding box size changes. The tracked results are stable within 3 pixels in both directions. Compared with the measurements by the SIFT, the correlation coefficient reaches 96.0%. The tracked results by Siam+CM is reproduced independently in Figure 13 for better visualisation. Except the deck target, the tracked motions of a stationary target on the west tower is given together. It shows that the tracking resolution is within 0.2 pixel and that there is no apparent drift indicating a stable mounting condition of the camera system during the recording.

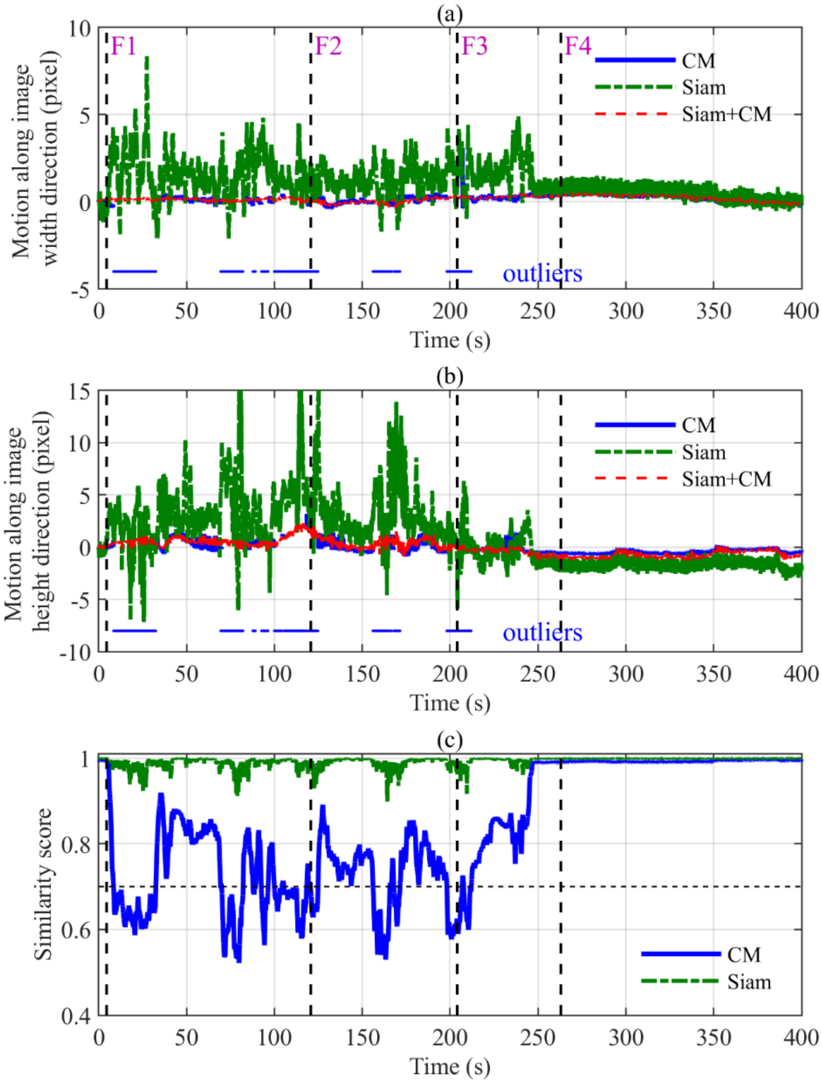


Figure 13 Image motions along two image directions (a-b) and similarity scores (c) using the three methods (CM, Siam, Siam+CM) for the deck target.

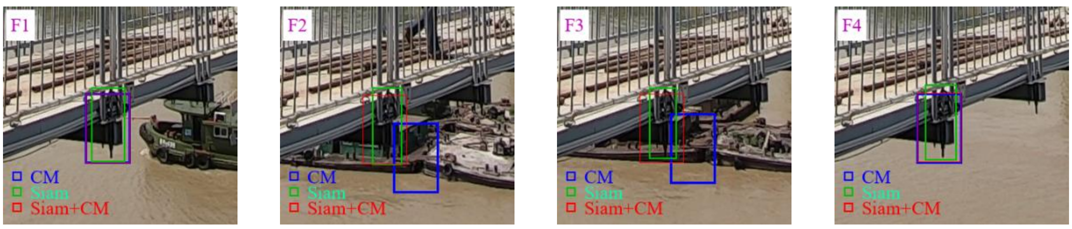


Figure 14 Predicted bounding boxes by the three methods in four frames F1~4.

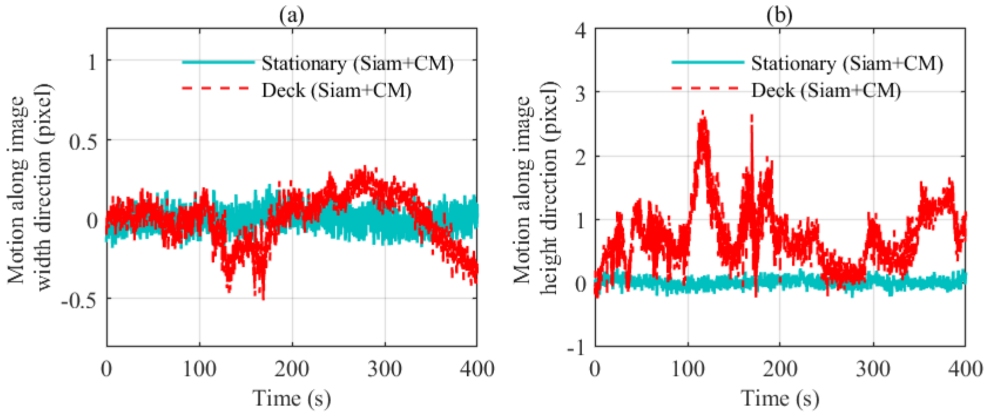


Figure 15 Image motions along two image directions (a-b) using the proposed method for the stationary and deck targets.

The 2D structural displacement at the deck target region is estimated by transforming the image coordinates of predicted target centroids using the pre-determined planar homography matrix. The results are shown in Figure 16 with the time history and frequency information. The maximum vertical displacement is approximately 1.6 cm compared with the initial state. It occurred when a group of maintenance workers passed through. The frequency components indicate more than four peaks below 2 Hz which means the bridge serviceability could be an issue of concern.

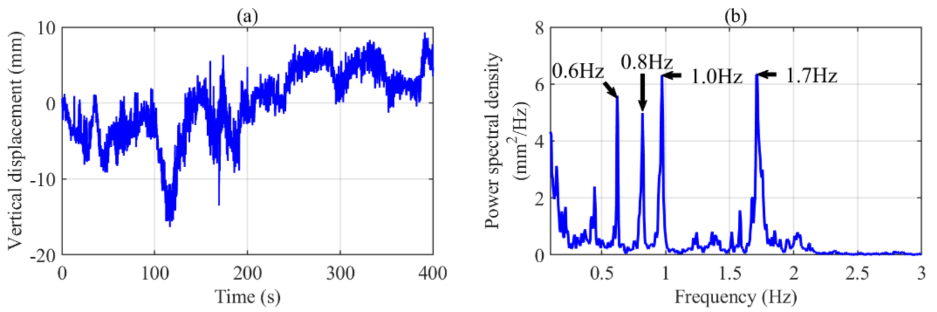


Figure 16 Measured displacement time history for the deck target by the proposed method (a) and the corresponding power spectral density (c).

5 FIELD TEST ON A LONG-SPAN BRIDGE

This section describes a case study using the proposed target tracking method for measuring deformation of a long-span road bridge in normal operation. The recorded video streams were approximately 26 minutes before the sunset and the selected target region in video frames experienced severe patterns changes due to moving shaded area. Section 5.1 and 5.2 describe the test setup and the results obtained, respectively.

5.1 Bridge and test information

The tested bridge is a long-span suspension bridge, the Humber Bridge in UK with a main span of 1410 m. A single day of field test using the vision-based monitoring system was performed to measure the displacement at mid-span of the bridge which has been reported in [57]. The study in [57] was about improving GPS measurement by a data fusion method with the vision-based data as data comparison. The focus here is different.

The vision-based monitoring system used in the test was a commercial system by Imetrum Limited, UK. The camera was equipped with the lens of 300 mm focal length mounted on a tripod at the base of the north tower as shown in Figure 17 (b). Essentially the camera is zoomed in on the artificial target of concentric rings which has been mounted in a 1 m x 1 m metal frame attached to the parapet on the east side of the bridge with on sample frame shown in Figure 17 (c). The acquired images are grayscale. The bridge long-term monitoring system includes two GPS rovers (Leica GMX902) mounted on the main cables at mid-span and a GPS base station at the bridge tower. The measured displacement by the GPS is used as reference in this study. Currently, the GPS is the common choice of displacement sensing in long-span bridges [58]. The accuracy level of GPS data is suggested to be up to 15 mm and 35 mm for horizontal and vertical measurements, respectively, at 98.5 percentile level without gross errors such as cycle slip or multipath [59].

The acquired videos were output for post-processing using the proposed method. The planar homography based on dimension correspondences in image plane and structural surface plane are adopted to transform the image coordinates (i.e. pixel) to structural coordinates (i.e. mm) and the known dimensions are from the width and height of the mounted metal frame. The output includes the two-dimensional translations along vertical and transverse directions of the bridge. A target region covering the main circular pattern on the artificial target shown in Figure 17 (c) is selected for analysis. The target pattern used as the template was from one frame in normal condition. The video streams for analysis were of 26 min recorded approximately one hour before the sunset and there are severe target pattern variations in the first 15 min.

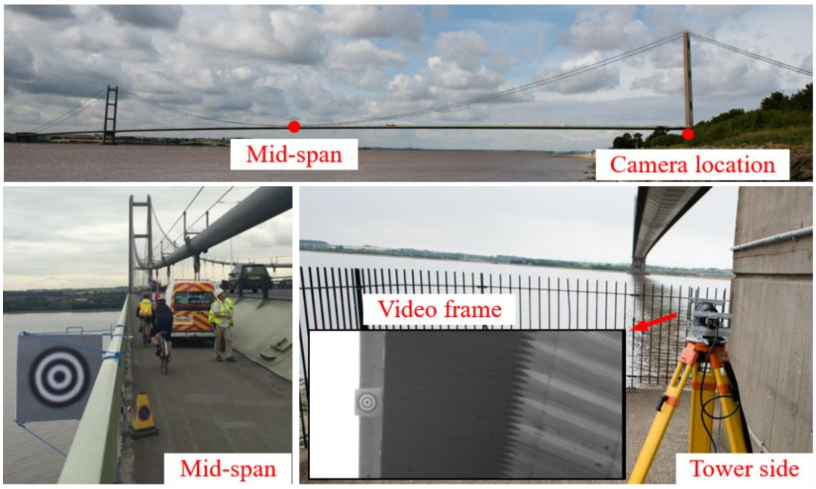
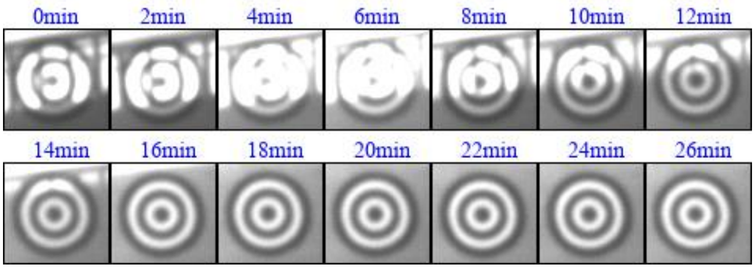
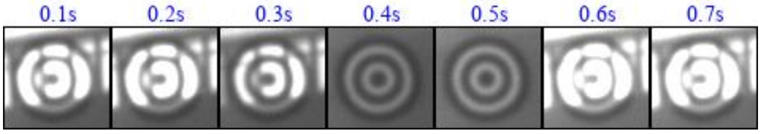


Figure 17 Test configration on the Humber bridge.

The video records include over 10 hours from a single day and most of them are under ideal environmental conditions (calm day, stable light conditions and stable tripod mounting). The chosen video stream for analysis includes apparent target patterns due to shade effect. A previous study [60] evaluated the working performance of three tracking methods (i.e. the template matching, LK optical flow, and SIFT matching) in this scenario. All of the three methods failed to acquire a robust measurement. In this study, the video records with shade effect is analysed by the proposed method (Siam+CM). In the first 16 min, due to the low sun elevation in the west, the target panel on the east side was partially in the shadow of the bridge railing and the target patterns are varied with time as shown in Figure 18 (a). Also, there was a sharp pattern change in less than one second during the vehicle passage. When one tall vehicle passed the mid-span of the bridge between the sun and the target, sunlight was completely blocked, making the whole target pattern visible in the image within 0.2 seconds as shown in Figure 18 (b). Then the target pattern recovered to the previous situation with parapet shades. In the end 10 min, there is no apparent target pattern change.



(a)



(b)

Figure 18 Target pattern variations in recorded videos in 26 mins due to shade effect (a) and sharp pattern changes in 0.7 second due to vehicle passage (b).

5.2 Measurement results

Three methods were implemented to analyse the video data including the CM, Siam and Siam+CM. The results are shown in Figure 19 and Figure 20.

The similarity scores by CM in the first 1000 seconds are mostly in a low status below 0.8 but also involve frequent jumps to a high value (over 0.9) due to vehicle passage. Some measurement outliers are observed at around 200 s as shown in Figure 20 (c) and it is observed that the circular patterns are less salient. The threshold of similarity scores for unreliable measurement evaluation is set as 0.5 through trials. The measurement after 400 s are highly similar with that by the proposed method. It indicates that from the CM is apparently not robust in presence of shading effect and apparent partial pattern variations.

For the Siam, the similarity score during the whole video records is always higher than 0.90 which indicates high confidence on measured results. As shown in Figure 20, the predicted bounding boxes in four sample frames are sensible but the bounding box is of varied dimensions. The predicted target motions taken from the centroid coordinates of the predicted bounding boxes are not accurate.

The tracked results by the proposed method (Siam+CM) are stable and the transformed vertical displacement data are presented in Figure 21 together with the GPS measurement. The cross correlation between two displacement signals is 94.7% and the RMS difference of the two measurements in 1600 seconds is 0.97 cm. The peak displacement under vehicle passage by vision-based method is 14.48 cm at 238 s and 14.72 cm at 1099 s, close to the GPS measurement (15.81 cm at 238 s and 14.60 cm at 1099 s). The previous study [57] compared the GPS and vision-based data by the CM method acquired at different time of the same day in ideal condition. The RMS difference of the two measurements is 0.75 cm, slightly smaller than the value (0.97 cm) in this study. Since the reference GPS displacement has limited accuracy (centimetre level [59]), the comparison study here could indicate that vision-based measurement integrating the proposed method could also reach centimetre-level accuracy in over 715 m camera-to-target distance when apparent shading and lighting changes occur. It could be a non-contact alternative to the GPS for measuring displacement data in long-span bridges.

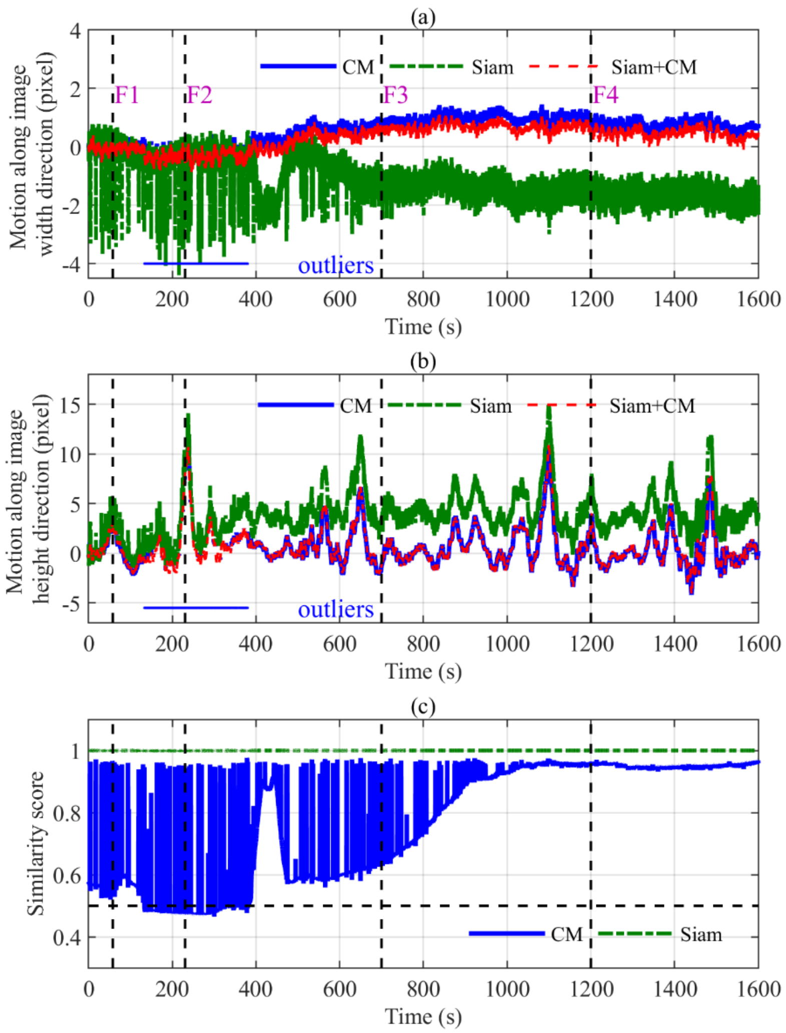


Figure 19 Image motions along two image directions (a-b) and similarity scores (c) using the three methods.

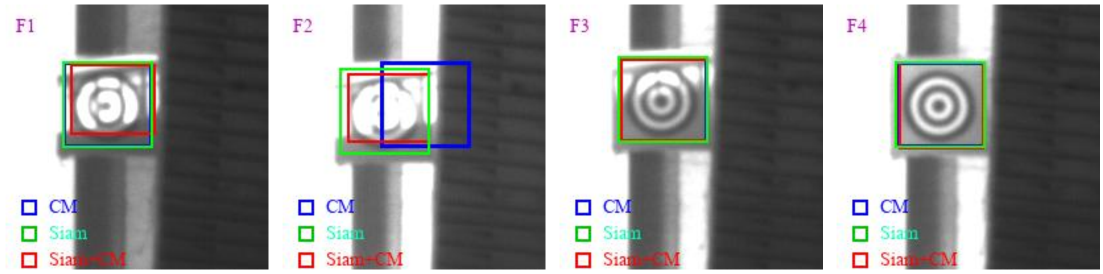


Figure 20 Predicted bounding boxes by three methods in four frames F1~4.

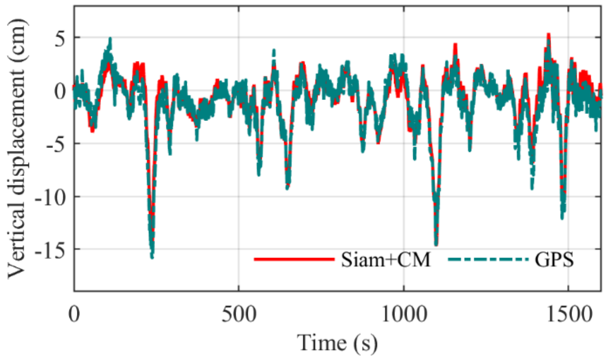


Figure 21 Measured displacement time history at bridge mid-span by the proposed vision-based method and the GPS.

6 CONCLUSIONS

This study proposes a novel distraction-free vision-based displacement measurement approach and provides indoor and field validation tests in challenging scenarios. The main conclusions are as follows:

1. The Siamese tracker (the DaSiamRPN combined with the UpdateNet) is an effective tool for coarsely localising the target regions in video frames. It uses deep feature representations and learned similarity measures for matching and also considers adaptive template update with time. It provides convincing tracking results under illumination variations. Also, it could adapt to the presence of severe target pattern changes through size changes of bounding boxes and local shift to cover the adjacent area. Thus, it is suitable for long-time continuous measurement.
2. The proposed method integrating the Siamese tracker with correlation-based template matching inherits the advantages of the Siamese tracker and also corrects the error in the Siamese tracker's output due to the image size changes of the estimated target regions. The method was validated in short-range and long-range monitoring campaigns considering

the scenarios with severe background variations, illumination changes and shade effect, providing stable and accurate displacement measurement results.

3. Existing applications of vision-based systems for bridge displacement measurement are usually limited to short-time monitoring tests. The proposed method resolves the problem of measurement robustness to environmental variations and could be potentially tied to a vision-based system stably fixed on a bridge component for long-time measurement.
4. The study evaluated the proposed method over three application cases and further study is necessary to evaluate the measurement accuracy and uncertainty for quality assurance of vision-based measurement system.

CREDIT AUTHOR STATEMENTS

Yan Xu: Conceptualization, Methodology, Software, Validation, Investigation, Data curation, Formal analysis, Methodology, Investigation, Visualization, Writing - original draft, Writing - Review & Editing, Funding acquisition. **Jian Zhang:** Conceptualization, Methodology, Data curation, Writing - original draft, Writing - Review & Editing. **James Brownjohn:** Conceptualization, Methodology, Writing- Reviewing and Editing.

DECLARATION OF COMPETING INTEREST

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

ACKNOWLEDGEMENT

This research was supported by the National Key R&D Program of China (2019YFC1511102) and the Jiangsu Natural Science Foundation (BK20190372). The authors express their sincere appreciation for their support. Finally, the authors would like to thank the four anonymous reviewers for their constructive comments.

REFERENCES

[1] H. Yoon, H. Elanwar, H. Choi, M. Golparvar-Fard, B.F. Spencer, Target-free approach for vision-based structural system identification using consumer-grade

cameras, Struct. Control Heal. Monit. 23 (2016) 1405–1416. https://doi.org/10.1002/stc.1850.

[2] E. Caetano, S. Silva, J. Bateira, Application of a vision system to the monitoring of cable structures, in: Seventh Int. Symp. Cable Dyn., 2007: pp. 225–236.

[3] B.K. Oh, J.W. Hwang, Y. Kim, T. Cho, H.S. Park, Vision-based system identification technique for building structures using a motion capture system, J. Sound Vib. 356 (2015) 72–85. https://doi.org/10.1016/j.jsv.2015.07.011.

[4] D. Feng, M.Q. Feng, Model Updating of Railway Bridge Using In Situ Dynamic Displacement Measurement under Trainloads, J. Bridg. Eng. 20 (2015) 1–12. https://doi.org/10.1061/(ASCE)BE.1943-5592.0000765.

[5] Y.-J. Cha, J.G. Chen, O. Büyüköztürk, Output-only computer vision based damage detection using phase-based optical flow and unscented Kalman filters, Eng. Struct. 132 (2017) 300–313. https://doi.org/10.1016/j.engstruct.2016.11.038.

[6] T. Ojio, C.H. Carey, E.J. OBrien, C. Doherty, S.E. Taylor, Contactless Bridge Weigh-in-Motion, J. Bridg. Eng. 21 (2016) 04016032. https://doi.org/10.1061/(ASCE)BE.1943-5592.0000776.

[7] G.A. Stephen, J.M.W. Brownjohn, C.A. Taylor, Measurements of static and dynamic displacement from visual monitoring of the Humber Bridge, Eng. Struct. 15 (1993) 197–208.

[8] J.H.G. Macdonald, E.L. Dagless, B.T. Thomas, C.A. Taylor, Dynamic measurements of the Second Severn Crossing, Proc. Inst. Civ. Eng. - Transp. 123 (1997) 241–248. https://doi.org/10.1680/itran.1997.29978.

[9] D. Feng, M. Feng, E. Ozer, Y. Fukuda, A Vision-Based Sensor for Noncontact Structural Displacement Measurement, Sensors. 15 (2015) 16557–16575. https://doi.org/10.3390/s150716557.

[10] X.W. Ye, Y.Q. Ni, T.T. Wai, K.Y. Wong, X.M. Zhang, F. Xu, A vision-based system for dynamic displacement measurement of long-span bridges: algorithm and verification, Smart Struct. Syst. 12 (2013) 363–379. https://doi.org/10.12989/sss.2013.12.3_4.363.

[11] W.Y. Liao, W.H. Chen, Y.Q. Ni, Y. Xia, Development of a vision-based real-time displacement measurement system for Guangzhou New TV Tower, in: F. Casciati, Giordano M. (Eds.), Proc. 5th Eur. Work. Struct. Heal. Monit., Sorrento, Naples, Italy, 2010: pp. 450–455.

[12] S. Baker, I. Matthews, Lucas-Kanade 20 Years On: A Unifying Framework, Int. J. Comput. Vis. 56 (2004) 221–255. https://doi.org/10.1023/B:VISI.0000011205.11775.fd.

[13] J. Guo, C. Zhu, Dynamic displacement measurement of large-scale structures based on the Lucas–Kanade template tracking algorithm, Mech. Syst. Signal Process. 66–67 (2016) 425–436. https://doi.org/10.1016/j.ymssp.2015.06.004.

[14] D. Feng, M.Q. Feng, Computer vision for SHM of civil infrastructure: From dynamic response measurement to damage detection – A review, Eng. Struct. 156 (2018) 105–117. https://doi.org/10.1016/j.engstruct.2017.11.018.

[15] S.S. Beauchemin, J.L. Barron, The computation of optical flow, ACM Comput. Surv. 27 (1995) 433–466. https://doi.org/10.1145/212094.212141.

[16] D. Sun, S. Roth, M.J. Black, Secrets of optical flow estimation and their principles, in: IEEE Conf. Comput. Vis. Pattern Recognit., 2010: pp. 2432–2439. https://doi.org/10.1109/CVPR.2010.5539939.

[17] M. Ehrhart, W. Lienhart, Development and evaluation of a long range image-based monitoring system for civil engineering structures, in: P.J. Shull (Ed.), Proc. SPIE Struct. Heal. Monit. Insp. Adv. Mater. Aerospace, Civ. Infrastruct., San Diego, California, United States, 2015: p. 94370K. https://doi.org/10.1117/12.2084221.

[18] Y.F. Ji, C.C. Chang, Nontarget Image-Based Technique for Small Cable Vibration Measurement, J. Bridg. Eng. 13 (2008) 34–42. https://doi.org/10.1061/(ASCE)1084-0702(2008)13:1(34).

[19] E. Caetano, S. Silva, J. Bateira, A vision system for vibration monitoring of civil engineering structures, Exp. Tech. 35 (2011) 74–82. https://doi.org/10.1111/j.1747-1567.2010.00653.x.

[20] D.J. Fleet, A.D. Jepson, Computation od Component Image Velocity from local phase information, Int. J. Comput. Vis. 5 (1990) 77–104.

[21] J.G. Chen, N. Wadhwa, Y.-J. Cha, F. Durand, W.T. Freeman, O. Buyukozturk, Modal identification of simple structures with high-speed video using motion magnification, J. Sound Vib. 345 (2015) 58–71. https://doi.org/10.1016/j.jsv.2015.01.024.

[22] Y. Yang, C. Dorn, T. Mancini, Z. Talken, G. Kenyon, C. Farrar, D. Mascareñas, Blind identification of full-field vibration modes from video measurements with phase-based video motion magnification, Mech. Syst. Signal Process. 85 (2017) 567–590. https://doi.org/10.1016/j.ymssp.2016.08.041.

[23] J.G. Chen, A. Davis, N. Wadhwa, F. Durand, W.T. Freeman, O. Buyukozturk, Video Camera-Based Vibration Measurement for Civil Infrastructure Applications, J. Infrastruct. Syst. 23 (2017) 11. https://doi.org/10.1061/(asce)is.1943-555x.0000348.

[24] C.Z. Dong, O. Celik, F.N. Catbas, E.J. O'Brien, S. Taylor, Structural displacement monitoring using deep learning-based full field optical flow methods, Struct. Infrastruct. Eng. 16 (2020) 51–71. https://doi.org/10.1080/15732479.2019.1650078.

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

[26] H. Bay, A. Ess, T. Tuytelaars, L. Van Gool, Speeded-up robust features (SURF), Comput. Vis. Image Underst. 110 (2008) 346–359.

[27] E. Rublee, V. Rabaud, K. Konolige, G. Bradski, ORB: An efficient alternative to SIFT or SURF, in: Proc. IEEE Int. Conf. Comput. Vis., IEEE, 2011: pp. 2564–2571. https://doi.org/10.1109/ICCV.2011.6126544.

[28] R. Szeliski, Computer Vision: Algorithms and Applications, Springer Science & Business Media, London, 2011. https://doi.org/10.1007/978-1-84882-935-0.

[29] M. Calonder, V. Lepetit, C. Strecha, P. Fua, Brief: Binary robust independent elementary features, in: Eur. Conf. Comput. Vis., Crete, Greece, 2010: pp. 778–792.

[30] T. Khuc, F.N. Catbas, Completely contactless structural health monitoring of real-life structures using cameras and computer vision, Struct. Control Heal. Monit. 24 (2017) e1852. https://doi.org/10.1002/stc.1852.

[31] T. Khuc, F.N. Catbas, Computer vision-based displacement and vibration monitoring without using physical target on structures, Struct. Infrastruct. Eng. 13 (2017) 505–516. https://doi.org/10.1080/15732479.2016.1164729.

[32] M. Ehrhart, W. Lienhart, Monitoring of Civil Engineering Structures using a State-of-the-art Image Assisted Total Station, J. Appl. Geod. 9 (2015) 174–182. https://doi.org/10.1515/jag-2015-0005.

[33] Y. Xu, J.M.W. Brownjohn, Review of machine-vision based methodologies for displacement measurement in civil structures, J. Civ. Struct. Heal. Monit. 8 (2018) 91–110. https://doi.org/10.1007/s13349-017-0261-4.

[34] C.Z. Dong, F.N. Catbas, A review of computer vision–based structural health monitoring at local and global levels, Struct. Heal. Monit. (2020). https://doi.org/10.1177/1475921720935585.

[35] L. Zheng, Y. Yang, Q. Tian, SIFT Meets CNN: A Decade Survey of Instance Retrieval, IEEE Trans. Pattern Anal. Mach. Intell. 40 (2018) 1224–1244. https://doi.org/10.1109/TPAMI.2017.2709749.

[36] Y.J. Cha, W. Choi, O. Büyüköztürk, Deep Learning-Based Crack Damage Detection Using Convolutional Neural Networks, Comput. Civ. Infrastruct. Eng. 32 (2017) 361–378. https://doi.org/10.1111/mice.12263.

[37] F. Hu, J. Zhao, Y. Huang, H. Li, Structure-aware 3D reconstruction for cable-stayed bridges: A learning-based method, Comput. Civ. Infrastruct. Eng. 36 (2021) 89–108. https://doi.org/10.1111/mice.12568.

[38] M. Dutton, W.A. Take, N.A. Hoult, Curvature Monitoring of Beams Using Digital Image Correlation, J. Bridg. Eng. 19 (2014) 05013001. https://doi.org/10.1061/(asce)be.1943-5592.0000538.

[39] B.F. Spencer, V. Hoskere, Y. Narazaki, Advances in Computer Vision-Based Civil Infrastructure Inspection and Monitoring, Engineering. 5 (2019) 199–222. https://doi.org/10.1016/j.eng.2018.11.030.

[40] C. Koch, K. Georgieva, V. Kasireddy, B. Akinci, P. Fieguth, A review on computer vision based defect detection and condition assessment of concrete and asphalt civil

infrastructure, Adv. Eng. Informatics. 29 (2015) 196–210. https://doi.org/10.1016/j.aei.2015.01.008.

[41] C.V. Dung, L.D. Anh, Autonomous concrete crack detection using deep fully convolutional neural network, Autom. Constr. 99 (2019) 52–58. https://doi.org/10.1016/j.autcon.2018.11.028.

[42] F. Fang, L. Li, Y. Gu, H. Zhu, J.-H. Lim, A novel hybrid approach for crack detection, Pattern Recognit. 107 (2020) 107474. https://doi.org/10.1016/j.patcog.2020.107474.

[43] Y. Bao, Z. Tang, H. Li, Y. Zhang, Computer vision and deep learning–based data anomaly detection method for structural health monitoring, Struct. Heal. Monit. 18 (2019) 401–421. https://doi.org/10.1177/1475921718757405.

[44] R. Yang, S.K. Singh, M. Tavakkoli, N. Amiri, Y. Yang, M.A. Karami, R. Rai, CNN-LSTM deep learning architecture for computer vision-based modal frequency detection, Mech. Syst. Signal Process. 144 (2020). https://doi.org/10.1016/j.ymssp.2020.106885.

[45] L. Bertinetto, J. Valmadre, J.F. Henriques, A. Vedaldi, P.H.S. Torr, Fully-convolutional siamese networks for object tracking, Eur. Conf. Comput. Vis. (2016) 850–865. https://doi.org/10.1007/978-3-319-48881-3_56.

[46] J. Valmadre, L. Bertinetto, J. Henriques, A. Vedaldi, P.H.S. Torr, End-to-end representation learning for Correlation Filter based tracking, Proc. - 30th IEEE Conf. Comput. Vis. Pattern Recognition, CVPR 2017. 2017-Janua (2017) 5000–5008. https://doi.org/10.1109/CVPR.2017.531.

[47] B. Li, J. Yan, W. Wu, Z. Zhu, X. Hu, High Performance Visual Tracking with Siamese Region Proposal Network, Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (2018) 8971–8980. https://doi.org/10.1109/CVPR.2018.00935.

[48] Z. Zhu, Q. Wang, B. Li, W. Wu, J. Yan, W. Hu, Distractor-aware siamese networks for visual object tracking, in: Proc. Eur. Conf. Comput. Vis., 2018: pp. 101–117. https://doi.org/10.1007/978-3-030-01240-3_7.

[49] L. Zhang, A. Gonzalez-Garcia, J. van de Weijer, M. Danelljan, F.S. Khan, Learning the model update for siamese trackers, in: Proc. IEEE Int. Conf. Comput. Vis., 2019.

[50] A. Krizhevsky, I. Sutskever, G.E. Hinton, ImageNet classification with deep convolutional neural networks, in: Commun. ACM, 2017: pp. 84–90. https://doi.org/10.1145/3065386.

[51] https://github.com/foolwood/DaSiamRPN, (n.d.).

[52] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A.C. Berg, L. Fei-Fei, ImageNet Large Scale Visual Recognition Challenge, in: Int. J. Comput. Vis., 2015: pp. 211–252. https://doi.org/10.1007/s11263-015-0816-y.

[53] E. Real, J. Shlens, S. Mazzocchi, X. Pan, V. Vanhoucke, YouTube-BoundingBoxes: A large high-precision human-annotated data set for object detection in video, Proc. - 30th IEEE Conf. Comput. Vis. Pattern Recognition, CVPR 2017. 2017-Janua (2017) 7464–7473. https://doi.org/10.1109/CVPR.2017.789.

[54] T.-Y. Lin, M. Maire, S. Belongie, L. Bourdev, R. Girshick, J. Hays, P. Perona, D. Ramanan, C.L. Zitnick, P. Dollár, Microsoft COCO: Common Objects in Context, in: Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit., 2015: pp. 3686–3693. http://arxiv.org/abs/1405.0312.

[55] M. Kristan, A. Leonardis, J. Matas, M. Felsberg, R. Pflugfelder, L.Č. Zajc, T. Vojíř, G. Bhat, A. Lukežič, A. Eldesokey, G. Fernández, Á. García-Martín, Á. Iglesias-Arias, A.A. Alatan, A. González-García, A. Petrosino, A. Memarmoghadam, A. Vedaldi, A. Muhič, A. He, A. Smeulders, A.G. Perera, B. Li, B. Chen, C. Kim, C. Xu, C. Xiong, C. Tian, C. Luo, C. Sun, C. Hao, D. Kim, D. Mishra, D. Chen, D. Wang, D. Wee, E. Gavves, E. Gundogdu, E. Velasco-Salido, F.S. Khan, F. Yang, F. Zhao, F. Li, F. Battistone, G. De Ath, G.R.K.S. Subrahmanyam, G. Bastos, H. Ling, H.K. Galoogahi, H. Lee, H. Li, H. Zhao, H. Fan, H. Zhang, H. Possegger, H. Li, H. Lu, H. Zhi, H. Li, H. Lee, H.J. Chang, I. Drummond, J. Valmadre, J.S. Martin, J. Chahl, J.Y. Choi, J. Li, J. Wang, J. Qi, J. Sung, J. Johnander, J. Henriques, J. Choi, J. van de Weijer, J.R. Herranz, J.M. Martínez, J. Kittler, J. Zhuang, J. Gao, K. Grm, L. Zhang, L. Wang, L. Yang, L. Rout, L. Si, L. Bertinetto, L. Chu, M. Che, M.E. Maresca, M. Danelljan, M.H. Yang, M. Abdelpakey, M. Shehata, M. Kang, N. Lee, N. Wang, O. Miksik, P.

Moallem, P. Vicente-Moñivar, P. Senna, P. Li, P. Torr, P.M. Raju, Q. Ruihe, Q. Wang, Q. Zhou, Q. Guo, R. Martín-Nieto, R.K. Gorthi, R. Tao, R. Bowden, R. Everson, R. Wang, S. Yun, S. Choi, S. Vivas, S. Bai, S. Huang, S. Wu, S. Hadfield, S. Wang, S. Golodetz, T. Ming, T. Xu, T. Zhang, T. Fischer, V. Santopietro, V. Štruc, W. Wei, W. Zuo, W. Feng, W. Wu, W. Zou, W. Hu, W. Zhou, W. Zeng, X. Zhang, X. Wu, X.J. Wu, X. Tian, Y. Li, Y. Lu, Y.W. Law, Y. Wu, Y. Demiris, Y. Yang, Y. Jiao, Y. Li, Y. Zhang, Y. Sun, Z. Zhang, Z. Zhu, Z.H. Feng, Z. Wang, Z. He, The sixth visual object tracking VOT2018 challenge results, in: Eur. Conf. Comput. Vis., 2018. https://doi.org/10.1007/978-3-030-11009-3_1.

[56] M. Guizar-Sicairos, S.T. Thurman, J.R. Fienup, Efficient subpixel image registration algorithms, Opt. Lett. 33 (2008) 156–158. https://doi.org/10.1364/OL.33.000156.

[57] Y. Xu, J.M.W. Brownjohn, D. Hester, K.Y. Koo, Long-span bridges: Enhanced data fusion of GPS displacement and deck accelerations, Eng. Struct. 147 (2017) 639–651. https://doi.org/10.1016/j.engstruct.2017.06.018.

[58] J.M.W. Brownjohn, K.-Y. Koo, A. Scullion, D. List, Operational deformations in long-span bridges, Struct. Infrastruct. Eng. 11 (2015) 556–574. https://doi.org/10.1080/15732479.2014.951857.

[59] A. Nickitopoulou, K. Protopsalti, S. Stiros, Monitoring dynamic and quasi-static deformations of large flexible engineering structures with GPS: Accuracy, limitations and promises, Eng. Struct. 28 (2006) 1471–1482. https://doi.org/10.1016/j.engstruct.2006.02.001.

[60] Y. Xu, J.M.W. Brownjohn, Vision-based systems for structural deformation measurement: Case studies, Proc. Inst. Civ. Eng. Struct. Build. 171 (2018) 917–930. https://doi.org/10.1680/jstbu.17.00134.