School of Electrical Engineering, Computing and Mathematical Sciences

Computer Vision-Based Three-dimensional (3D) Vibration Displacement Measurement for Civil Structures

Yanda Shao
0000-0000-1928-6158

This thesis is presented for the Degree of
Master of Philosophy
of
Curtin University

September 2021

DECLARATION

I hereby declare that to the best of my knowledge and belief this thesis contains no material previously published by any other person except where due acknowledgement has been made.

This thesis contains no material which has been accepted for the award of any other degree or diploma in any university.

I have obtained permission from the copyright owners to use any third-party copyright material reproduced in the thesis, or to use any of my own published work in which the copyright is held by another party.

Signature: Date: 23. 11. 2021

ABSTRACT

Vibration displacement data is widely used in structural health monitoring (SHM) to analyse the health conditions of civil engineering structures. Traditionally, the displacement data is acquired by physical sensors such as linear variable differential transformers (LVDT) and laser displacement sensors (LDS), etc. When it is difficult to install displacement sensors, accelerometers are often used to indirectly derive the displacement from the accelerator measurements. In recent years, computer vision-based displacement measurements methods are receiving increasing attention, with inherent advantages in terms of cost-effectiveness and convenience. To date, most of the vison-based measurement methods are focused on two-dimensional (2D) measurements or in-plane displacement measurements, although civil structures are built in three-dimensional (3D) space and 3D displacements could provide more comprehensive information for SHM. The research of 3D vibration displacement measurements is in an early stage with many unsolved challenging tasks such as target-free 3D displacement measurement and tiny 3D displacement measurement, etc. This dissertation proposes an advanced binocular vision-based vibration displacement measurement system for target-free full-field 3D dynamic vibration displacement measurement of civil engineering structures. A state-of-the-art deep learning-based key point detection and matching algorithm is applied to achieve target-free measurement, which greatly improves the quantity and quality of the matching natural key points with low contrast. In most practical SHM scenarios, except under some extreme cases, the displacement response of civil structures is usually very small. Some of them can even be invisible to human eyes. The vision-based measurement methods or even some physical sensors would fail in such cases. Another contribution of this dissertation is a phase-based motion magnification algorithm which is employed to amplify the tiny displacement in videos to enable target-free 3D tiny vibration displacement measurement. The measurement accuracy of the tiny measurement system is achieved to the subpixel level or the submillimeter level in the engineering unit.

We conducted a series of experimental tests to evaluate the performance of the proposed vision measurement approaches based on deep learning. The performance of both the 3D and 3D tiny displacement measurement systems are evaluated through a steel cantilever beam vibration experiment conducted in a laboratory. A real field experiment is also conducted on a pedestrian

bridge on a university campus to evaluate the algorithm performance in practical applications. Displacements generated from the proposed vision system and the vibration frequencies converted from the displacements are compared with those measured by LVDTs and LDS, and/or acceleration responses measured by accelerometers. The results demonstrate that the performance of the proposed vision measurement system is on par with the traditional sensors, while it is much easier, cheaper, and safer to use. Accurate target-free and tiny 3D vibration displacement measurement can be achieved by the proposed vision based approach for civil engineering structural applications.

ACKNOWLEDGEMENTS

Foremost, I would like to express my sincere gratitude to my supervisors, Prof. Ling Li, Associate Prof. Jun Li, Dr. Senjian An and Prof. Wanquan Liu for their invaluable advice, continuous support, and patience during my MPhil study. It has been a wonderful and invaluable journey to be your students, and I will always be.

I would like to thank all my fellow research students and researchers (Antoni Liang, Nadith Pathirage, Qilin Li, Ruhua Wang, Huizhu Pan, Bradley Ezard, Yiwei Liu and Jie Liu) for many discussions and inspirations. Special thanks to Qilin and Ruhua for your company and mentorship.

I also wish to express my most profound appreciation to my family. Thanks to my parents, Mr. Changhua Shao and Mrs. Yan Tang, for their kind understanding and their warmest support. Thanks for believing in me all the time. It has been my source of strength along the way.

Last but not least, a special thanks goes to my dearest Miss Jiaorong Feng. Thanks for your constant company and efforts during such a long difficult period.

PUBLICATIONS

Yanda Shao, Ling Li, Jun Li, Senjian An, Hong Hao., 2021, “Computer vision based target-free 3D vibration displacement measurement of structures”, *Engineering Structures*, 246, 113040.

Yanda Shao, Ling Li, Jun Li, Senjian An, Hong Hao., “Deep learning assisted target-free 3D tiny structural vibration measurement”, Submitted to *Structural Control and Health Monitoring.*

STATEMENT OF CONTRIBUTION BY OTHERS

Chapter 3 and 4 of this thesis are based on work that have been published with joint-authorship. Details of authorship attribution statements can be found in Appendix.

Chapter 3 is based on the publication:

• **Yanda Shao.,** Ling Li., Jun Li., Senjian An., Hong Hao. (2021). Computer vision based target-free 3D vibration displacement measurement of structures, Engineering Structures, 246, 113040.

Chapter 4 is based on the publication:

• **Yanda Shao.,** Ling Li, Jun Li, Senjian An, Hong Hao., "Deep learning assisted target-free 3D tiny structural vibration measurement", Submitted to Structural Control and Health Monitoring.

We, the undersigned, endorse the above-stated contribution of work undertaken for each of the publications contributing to this thesis:

Candidate's Signature: Date:23. 11. 2021

Supervisor's Signature: Date:23. 11. 2021

Table of Contents

DECLARATION 2

ABSTRACT 3

ACKNOWLEDGEMENTS 5

PUBLICATIONS 6

LIST OF FIGURES 11

LIST OF TABLES 14

INTRODUCTION 15

1.1 Objective 16

1.1.1 Target-free Full-field 3D Vibration Displacement Measurement 17

1.1.2 3D Tiny Vibration Displacement Measurement 17

1.2 Structure of the Thesis 18

BACKGROUND 20

2.1 Multi-view 3D Reconstruction 20

2.2 Key Point Detection and Matching 22

2.3 Motion Magnification 26

2.4 Computer Vision for Displacement Measurement in SHM 27

2.5 Summary 30

TARGET-FREE THREE-DIMENSIONAL DISPLACEMENT MEASUREMENT 31

3.1 Introduction 31

3.2 Camera Calibration and Triangulation for 3D Reconstruction 32

3.2.1 Pinhole Camera Model 32

3.2.2 Two-view Camera Geometry 36

3.2.3 Camera Calibration 39

3.3 Key Point Detection, Matching and Tracking 44

3.3.1 Key Point Detection 44

3.3.2 Key Point Matching 49

3.3.2 Key Point Tracking 52

3.4 Experimental Studies on a Beam Structure 54

3.4.1 Experiment Setup 54

3.4.2 Results 56

3.5. Summary 64

3D TINY STRUCTURAL VIBRATION MEASUREMENT 65

4.1 Introduction 65

4.2 Motion Magnification 66

4.2.1 Complex Steerable Pyramid 68

4.2.2 Temporal Filtering and Denoising 70
4.2.3 Magnifying the Local Motion 71
4.2.4 The Limitation of Magnification 73
4.5 Experimental Validations 74
4.5.1 3D Vibration Tests of a Beam Structure 74
4.5.1.1 Experiment Setup 74
4.5.1.2 Results and Discussions 75
4.5.2 In-field Test on an Indoor Pedestrian Bridge 81
4.5.2.1 Experiment Setup 81
4.5.2.2 Results and Discussions 82
4.6 Sensitivity Investigations 86
4.7 Summary 88
CONCLUSION AND FUTURE WORKS 89
5.1 Conclusions 89
5.2 Future Works 90
BIBLIOGRAPHY 92
ATTRIBUTION STATEMENT 100

LIST OF FIGURES

Figure 1.1. An overview of the limitations of displacement measurement in SHM 16

Figure 2.1. two-view 3D point coordinate recovery 22

Figure 2.2. The flat area, edge area and corner area 23

Figure 3.1. The flowchart of the proposed pbinocular vision-based 3D vibration displacement measurement system 32

Figure 3.2. The pinhole model of a monocular camera 33

Figure 3.3. The architecture of epipolar geometry 37

Figure 3.4. The overview of the SuperPoint algorithm 45

Figure 3.5. The architecture of the MagicPoint network 45

Figure 3.6. The architecture of the VGG style encoder 46

Figure 3.7. Homographic adaptation 47

Figure 3.8. Joint training in SuperPoint 48

Figure 3.9. The architecture of SuperGlue network 50

Figure 3.10. The 3D vibration test setup 54

Figure 3.11. The installation of displacement sensors and setup of cameras 56

Figure 3.12. The performance of the proposed system. 57

Figure 3.13. Three direction displacement comparison of the 3D vibration test..... 59

Figure 3.14. Three direction displacement comparison for thirty key points of the 3D vibration test 60

Figure 3.15. Vibration frequencies obtained from the displacement measurement 63

Figure 4.1. The flowchart of deep learning assisted tiny 3D structural vibration measurement method..... 66

Figure 4.2. The flowchart of the phase-based motion magnification 67

Figure 4.3. The basis function of complex steerable pyramid..... 69

Figure 4.4. The waveform of 1D half-octave basis function 70

Figure 4.5. The flowchart of the phase-based motion magnification 72

Figure 4.6. A comparison of motion magnification with magnification factors inside and beyond the border 74

Figure 4.7. 3D tiny displacement measurement (original videos)..... 77

Figure 4.8. 3D tiny displacement measurement (magnified videos) 78

Figure 4.9. FFT spectrums of measured 3D vibrations in test..... 79

Figure 4.10. The similarity of 3D vibration trajectories of 30 key points in tiny movement test. 80

Figure 4.11. Experimental setup for indoor pedestrian bridge vibration tests..... 82

Figure 4.12. Vertical displacement near Accelerometer 5 obtained by the proposed vision method 83

Figure 4.13. Comparison of the vertical acceleration responses by the proposed vision approach and from Accelerometer 2. 84

Figure 4.14. Comparison of the vertical acceleration responses by the proposed vision approach and from Accelerometer 5. 84

Figure 4.15. FFT spectrums of acceleration time histories obtained by the proposed vision approach and recorded by Accelerometer 2..... 85

Figure 4.16. FFT spectrums of acceleration time histories obtained by the proposed vision approach and recorded by Accelerometer 5..... 86

LIST OF TABLES

Table 3.1. Sensors used in the 3D vibration test 56

Table 3.2. Displacement error analysis of the 3D vibration test 58

Table 3.3. Displacement relative errors of thirty key points 61

Table 3.4. Displacement correlation coefficients of thirty key points 62

Table 3.5. The relative errors of vibration frequencies and Fourier spectrum 64

Table 4.1. Displacement sensors used in the 3D vibration test 75

Table 4.2. Displacement error analysis of tiny vibration test 76

Table 4.3. Sensors installed on the indoor pedestrian bridge. 81

Table 4.4. Relative errors of obtained acceleration responses 85

Table 4.5. Sensitivity study of the vision system without and with motion magnification 87

CHAPTER 1

INTRODUCTION

From the early stage of human civilization, civil structures began to provide shelters, sacrificial places, cemeteries, etc., for human beings. Civil structure engineering that goes through development for thousands of years has become more complicated and it has already developed into a mainstream academic discipline. Modern people's life is relying on civil structures with more complex functions, for example, the demand for accommodation, transportation, and recreation. For civil structures, safety is always one of the basic requirements. Although this basic requirement has improved significantly, damages, even collapses, of civil structures are still a common problem. Civil engineering structures could be damaged in different ways. In extreme situations, after some accidents such as earthquakes, floods, and so forth, the structures could have some inner damages that may impact the further use, or the structure would directly collapse. On the other hand, the structure ages due to usage, and accidents could occur if the damage accumulates to a certain degree. Structural health monitoring (SHM) is a process of tracking the operational status, assessing the conditions, and detecting the damages of various types of structures, which provides an accessible approach that spends relative less effort to monitor the damages in the early stage to avoid severe economic losses and heavy casualties.

Computer vision is a branch of artificial intelligence (AI), which is used to understand the content of digital images using computers. Humans observe the world through both eyes, and the brain extracts and analyses features from the observed images. In computer vision, the images taken by cameras are processed by various algorithms to hopefully reach the equivalent performance of the human visual system or even beyond the human visual system. Recently, with neurobiology and deep learning rapidly developed over the last decade, the performance of computer vision on many applications has improved significantly. Today, computer vision has

been applied commonly in many real-world applications such as optical character recognition, machine inspection, surveillance, medical imaging, etc [1].

In the SHM community, with the development of cameras of low cost and high quality, and the remarkable progress of computer vision techniques, computer vision-based SHM has received increasing attention in recent years. A significant advantage of the computer vision methods is their easy setup and operation to extract information of civil structures from images or videos. Computer vision has been applied to crack detection, spalling detection, delamination detection, displacement measurement, vehicle load estimation, and so forth [2]. Displacement responses of civil structures, which directly reflects the structural all-inclusive stiffness, provide important signals for accurate assessments of structural conditions [3]. However, the acquirement of such important data is sometimes difficult using traditional physical sensors. Digital image sequences with their powerful characterization capabilities have an inherent advantage in displacement measurement.

1.1 Objective

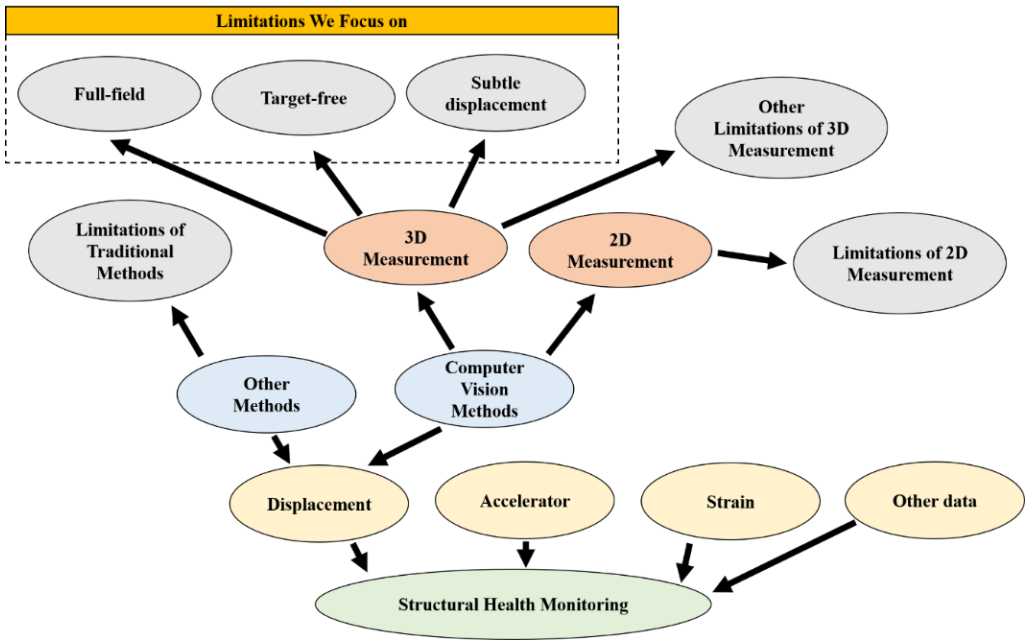


Figure 1.1. An overview of the limitations of displacement measurement in SHM

Figure 1.1 shows an overview of the challenges of the displacement measurement for SHM. The main goal of this thesis is to apply computer vision techniques for displacement measurement for structural health monitoring, especially the challenging 3D displacement measurement. Specifically, two algorithms are proposed to improve the 3D vision-based vibration measurement: one is for target-free full-field 3D vibration displacement measurement and the other is for 3D tiny vibration displacement measurement.

1.1.1 Target-free Full-field 3D Vibration Displacement Measurement

Natural key point detection and matching are essential but challenging processes for 3D reconstruction from 2D images. To assist key point detection and matching, high contrast artificial targets such as chessboard patterns are often utilized. However, with artificial targets, the installation that is time-consuming and labor-consuming prevents it from widely applying. Target-free measurement overcomes such difficulty hence is much widely applicable. Furthermore, measuring vibration displacements of every part of the structure provides more comprehensive information about the structure and its changes. Such kind of measurement is termed as full-field measurement in a vision displacement measurement system. A vision measurement system that can achieve target-free full-field 3D measurement rely on accurately detecting and matching a sufficiently large number of natural key points, which must distribute all over the whole structure. In this thesis, deep learning-based key point detection and matching algorithms are applied to realize the two-view matching of natural key points, and thus enable the target-free 3D displacement measurement. The numerous matched key points detected are distributed on every part of the measured structure. Thus, the vision system has full-field measurement capabilities. The vision-based displacement measurement system does not require the cumbersome installation of traditional sensors, nor any artificial targets to be fixed on the structure.

1.1.2 3D Tiny Vibration Displacement Measurement

Computer vision algorithms such as key point tracking algorithms usually cannot recognize tiny motions (typically smaller than one pixel). Such cases are quite common in practice when the recording cameras are placed far from the inspected structure. To measure movements that are in

sub-pixel level in recorded images/videos using computer vision methods, a motion magnification method is employed to magnify the tiny displacements in the videos. A binocular vision system as described in Sec.1.1.1 can then be applied to measure the 3D displacement from the videos. The tiny displacement measurement system is also target-free and full-field.

1.2 Structure of the Thesis

This thesis contains 5 chapters. The rest of the thesis are organized as below:

Chapter 2 introduces the preliminary knowledge related to this thesis. We begin with a brief introduction to the computer vision techniques relevant to this project. Some fundamental concepts of camera projection are first introduced. Classical key point detection algorithms in computer vision are then reviewed, followed by the key point matching algorithms which are important for 3D coordinate recovery. Next, we introduce a series of motion magnification algorithms, which are used to amplify tiny movements in videos. A literature review on the vision-based displacement measurement approaches in the field of structural health monitoring is also provided, with analysis on the advantages and disadvantages of the existing vision-based displacement measurement systems.

Chapter 3 proposes a binocular vision-based 3D displacement measurement system. We first introduce a calibration method based on epipolar geometry which does not require complicated manual measurement. Then a deep learning based key point detection and key point matching algorithm is introduced which is a key important component of the proposed vision system. The main contribution of this chapter is the development of a binocular vision system to achieve target-free 3D displacement measurement. This is the first study in the target-free full-field 3D vibration displacement measurement of civil engineering structures using a binocular camera system. The effectiveness of the binocular vision system is validated by a 3D vibration test on an iron beam in a laboratory.

Chapter 4 introduces a vision-based tiny displacement measurement method. Tiny movements captured by cameras are first amplified by a phase-based motion magnification algorithm to make them visible. The method is verified by a series of experimental tests. A 3D tiny vibration test in a laboratory is conducted to verify the system performance. The sensitivity of the vision system is also investigated. Next, a vibration test on an indoor pedestrian bridge is

conducted to investigate the in-field effectiveness of the proposed method. The main contribution of this chapter is the development of a vision-based binocular vision system for tiny 3D displacement measurement. Experimental results demonstrate that 3D subpixel level displacements, which are valuable data for structural health monitoring, can be effectively extracted by the proposed method.

Chapter 5 summarizes the contributions of the thesis. Potential directions of future research are also discussed.

CHAPTER 2

BACKGROUND

The goal of this chapter is to highlight the related background of the research objectives addressed in this thesis. As mentioned earlier, the focus of this thesis is to realize 3D target-free displacement measurement including tiny displacement measurement for civil structures. We will first go through the basics of multi-view 3D reconstruction. Next, we will review the fundamental concepts of the key point detection and key point matching algorithms. Some video motion magnification algorithms for realizing tiny movement measurement will be introduced. The computer vision-based displacement measurement techniques for SHM will also be covered.

2.1 Multi-view 3D Reconstruction

Single camera imaging projects an object in a 3D Euclidean space to a 2D projective space, hence images captured by a stationary single camera (monocular camera) only contain 2D in-plane information. The camera's image sensor converts the light signal into an electrical signal through photoelectric conversion and divides the photographed object into individual pixels. The photoelectric signal on the sensors is converted into a digital signal through an analog-digital converter and stored in the memory for further image processing [4]. The pinhole model is mainly designed for the cameras which contain charged coupled device (CCD) sensors. This model is widely used in the videogrammetry field to construct the mathematical model for the estimation of camera poses. Mathematically, the camera imaging process under the pinhole model can be expressed as $s \times m = P \times M$, in which m is a 2D pixel vector converted by a camera matrix P times a 3D real world point vector M, and s is an arbitrary scale factor. Continuously recording a scene can get a sequence of images, i.e., a video. For civil structure vibration displacement measurement, the scene that contains structural vibration is recorded by a video camera. Using computer vision techniques to extract the position of the structure in each image (frame) of a video, the dynamic vibration displacement responses of the structure can be derived.

The real-world structures always have 3D vibration and the vibration displacement in every direction is important data for SHM. However, the displacement on the depth direction is lost

during the camera projection due to the camera imaging procedure mapping the 3D scene to a 2D plane. Accurate reconstructing the 3D scene is an essential and challenging branch in computer vision area. There are many 3D reconstruction methods in the field of computer vision, such as multi-view depth estimation [5], structure-from-motion (SfM) [6, 7], monocular depth estimation [8, 9], and so on. The SFM techniques takes videos recorded by a movement camera as input, by analyzing the geometry relationship between different frames to extract the 3D scene. It works well on static scenes reconstruction, whereas measuring the displacement of a movement object in a scene could raise to large error. Estimating depth from a single image is an ill-posed problem as technically the depth for a single image has lost. The monocular depth estimation approaches adopt deep learning to mimic human learning behavior that makes prediction of the depth from a single RGB image become possible. It is similar as a cheaper Lidar with higher resolution which captures a scene and output a depth map. Although powerful, it is still a developing area, and the accuracy of its predicted depth is unsatisfied for SHM currently.

Multi-view depth estimation based on multi-view geometry is a mature method. It uses multiple cameras to take images of a scene from different perspectives and utilizes the correlation information between the images to recover the 3D information. A two-view camera system is shown in Figure 2.1, where M represents a 3D point in the real world, and m_1 and m_2 are its projective pixels in two-view images respectively. The camera matrix describes the projection transformation of a pinhole camera between 3D points in the real world and their 2D projections in an image. Without considering the projection error, if the matching between and m_1 and m_2 in two images can be established, the coordinate of a 3D point M can be obtained as the intersection in the 3D space of rays connecting the matched key points (representing the projections of M) and the optical centers of two cameras. 3D reconstruction of M can be achieved if matched key points m_1, m_2 and the camera matrices of camera 1 and camera 2 can be established. Such kind of method for 3D reconstruction is called triangulation. The fundamental requirement of the triangulation algorithm is the sufficiently large number of key points that need to be detected and matched.

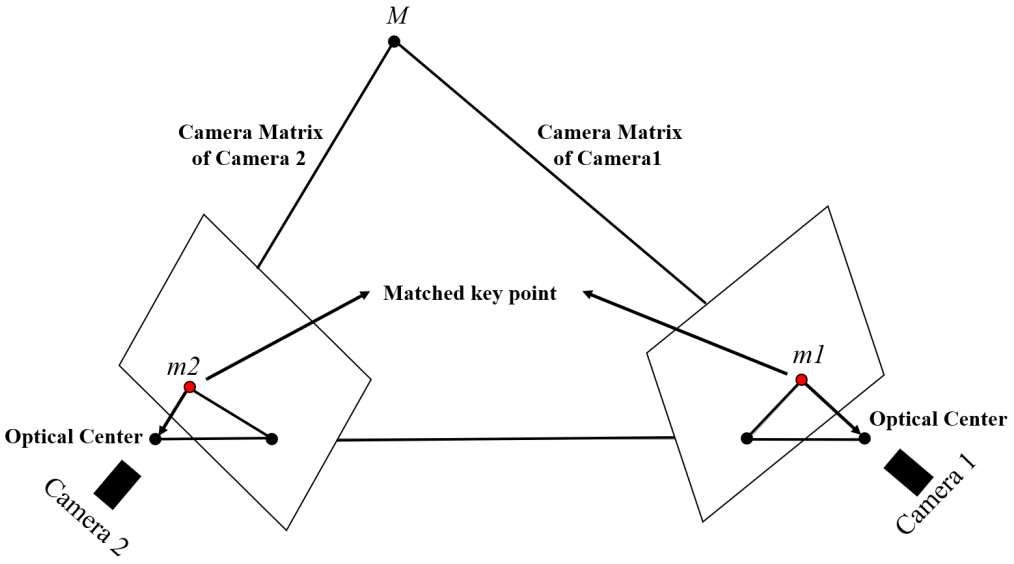


Figure 2.1. two-view 3D point coordinate recovery

2.2 Key Point Detection and Matching

A real world 3D object is projected into multiple different images when multiple cameras are employed. In the multi-view geometry, key points in different images are required to be detected and matched in order to solve tasks in geometric computer vision such as Simultaneous Localization and Mapping (SLAM) [10], Structure-from-Motion (SfM) [6, 7], and camera calibration [11]. In this context, key points are 2D pixel locations in images with two important properties: repeatability and well-description. The key points should be repeatable and stable from different view images. Besides the locations of key points, a unique identifier should be assigned to each key point for higher-level tasks such as key point matching [12]. This identifier is known as a descriptor, which is a unique descriptive vector endowed to each key point. Key points are distinguished from each other by these descriptors. There are certain properties that the key point descriptor should possess. For example, a descriptor of a key point should be make it highly distinctive from distant key points but similar to the key points in the vicinity. In order to facilitate key point matching, a key point should consist of two parts: a detector and a descriptor. The detector finds the location of key points; the descriptor sufficiently identifies the key points. Some key point detection algorithms have their designated feature description algorithms, while some others share description algorithms. In the field of computer vision, many studies have been

conducted to detect key points accurately and quickly. In the early stage of the key point detection research, corners are considered to be the good key points to detect, which can be traced back to 1977 [13, 14].

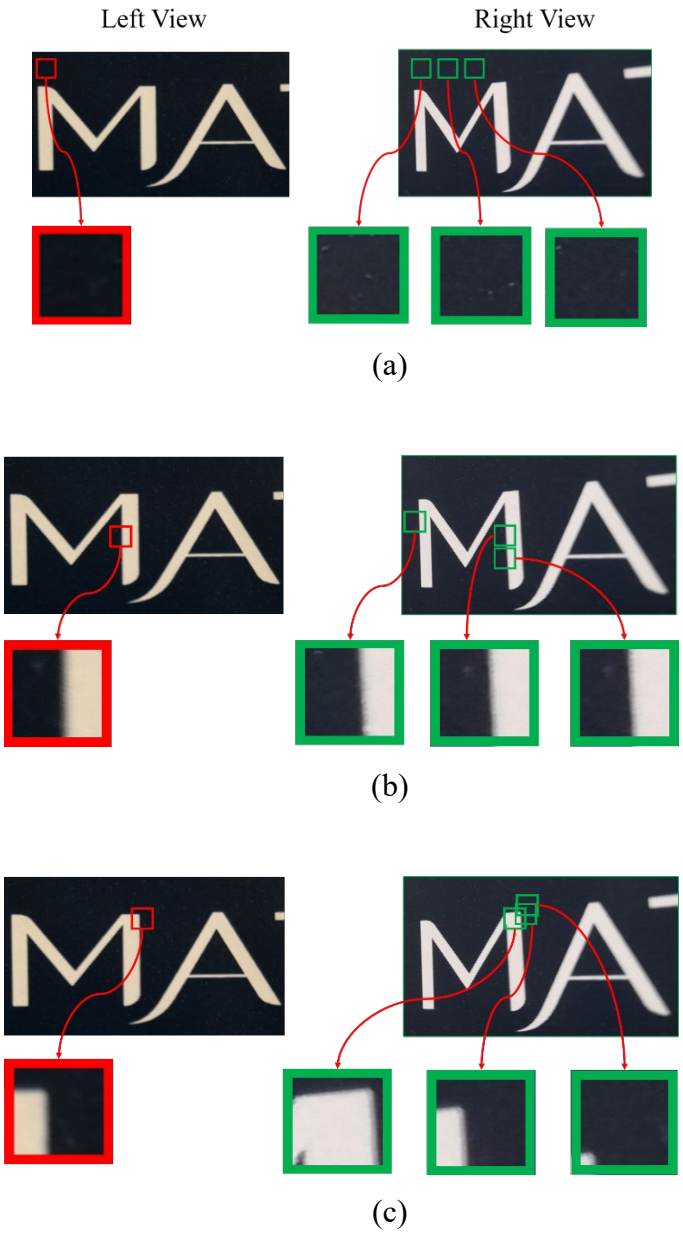


Figure 2.2. The flat area, edge area and corner area

The basic idea of the corner detection algorithm is to use a fixed sliding window (take a neighbourhood window of a certain pixel) to slide in any direction on the image. Comparing the two cases before and after sliding, if the pixel value in the window has a large change in any direction, then this window is considered as containing a corner point. In Figure 2.2 (a), pixels in a flat area have high similarities with each other. The distinctiveness of these pixels is lost; hence it is difficult to repeatedly detect them in another view. The uniqueness of key points on an edge is not obvious either. Sliding the window along the edge results in pixel value unchanged, as shown in Figure 2.2 (b). In contrast, the corner areas are highly distinctive. With slight movement of the sliding window, the pixel value in the area is changed dramatically, as shown in Figure 2.2 (c). Therefore, some studies try to find key points by detecting corners. For example, the famous Harris corner detector [15], Shi-Tomasi Corner Detector [16], Oriented FAST and Rotated BRIEF (ORB) [17] and Fast Retina Keypoints (FREAK) [18]. Meanwhile, corners have some disadvantages, e.g., they are sensitive to scale changing and image rotation [19]. Some other algorithms relied on blobs (regions that have constant properties for scaling, rotation, etc.) for key point detection, for example, the glorious Scale Invariant Feature Transform (SIFT) [19] method and Speeded Up Robust Features (SURF) [20] method. Both are based on Gaussian scale space analysis and the effective handling of the stability of the detection of key points on the scale changing or rotation of the images. Alcantarilla et al. proposed a key point detection algorithm based on computing the scale normalized determinant of Hessian Matrix on multiple scale levels, called KAZE features [21]. In recent years, some sparse local key points detection algorithms based on deep learning have been developed. LF-NET neural network [22] is introduced in 2018 by Yuki Ono et al, which is trainable end-to-end and does not require using a hand-crafted detector to generate training data. In 2019, Revaud et al. proposed an R2D2 neural network [23], which is a self-supervised neural network trained by a mixture of images with known transformations and point correspondences. The highlight of the method is that a style transfer method is applied to increase robustness against day-night illumination changes. Dusmanu et al. [24] proposed an approach called D2-Net that jointly train a key point descriptor and a detector. The algorithm postpones the key point detection to a later stage to produce more stable key points. Unfortunately, all these methods can only detect a small number of high quality key points, which are usually not enough for higher-level computer vision tasks.

Key point matching aims to establish the corresponding relationship between a set of key points in two different views. Traditional machine learning provides some solutions to achieve such tasks. Graph matching (GM) is one of the mainstream methods for key point matching. The key point matching task can be formulated as an optimal transportation problem and the corresponding key points can be found via maximizing the overall score unary correspondence. Zhou et al. [25] presented a method called Factorized Graph Matching (FGM), a graph matching algorithm that exploits the properties of the factorized affinity or graph matrix. This approach does not need to explicitly compute the affinity matrix and it provides a unified approach to frame several graph matching algorithms. Yu et al. [26] proposed a bunch of continuous approximations to the quadratic assignment problem (QAP). Based on the above theory, Generalized Graph Matcher (GGM) is designed for approximating discrete graph matching. By understanding the geometric properties of different regularization techniques and the gradient behaviour in the optimization process, Yu et al. [27] introduced a regularization technique for graph matching, which is based on the analysis of the determinant of the node matching matrix between two images. Different from graph matching, the point set registration (PSR) determines a global transformation. Resampling technique is (arguably) a prevalent paradigm of PSR and is represented by the classic RANSAC algorithm [28, 29] or many RANSAC related algorithms, such as Maximum Likelihood Estimation Sample Consensus (MLESAC) [30], Universal RANSAC (USAC) [31], Maximum Likelihood Estimation Sample Consensus (MLESAC) [32], etc. In recent years, some deep learning methods have been developed to improve upon the traditional key point matching algorithm. Zhang et al. [33] proposed the Order-Aware Network to learn two-view matching and geometry. The local context is acquired by the clustering of relevant nodes which is learned by the Differentiable Pooling (DiffPool) layer and Order-Aware Differentiable Unpooling (DiffUnpool) layer. Kwang et al. [34] introduced a single-shot technique, in which the matched points are classified into inliers and outliers by a deep neural network. They operate on sets of match points, estimated by NN search, thus the assignment structure is ignored, and visual information discarded.

2.3 Motion Magnification

In many real-life scenarios, the motions of an object can be very tiny. They can even be too small to be observed by naked eyes. Such kind of tiny motion is also difficult to be detected by computer vision algorithms. In some SHM cases, these sorts of tiny movements are very common and providing valuable information. Motion magnification algorithms are like microscopes to search for the invisible tiny motion in the video. These kinds of algorithms estimate the motion in the video and amplify them, and the invisible movement could become visible. Liu et al [35] presented the motion magnification technique in 2005. The algorithm groups the tiny motion of the input video into different motion layers and amplifies the motions of a layer selected by the user. “Holes” revealed by amplified motions are filled using texture synthesis methods. This study is the first attempt on such a difficult task to find the tiny motion in a video. It relies on accurate motion estimation, which is computationally expensive and difficult to make artefact-free, especially at regions of occlusion boundaries and with complicated motions. After seven years, in 2012, Wu et al. [36] introduced a Eulerian method for real-time video motion magnification. The authors believed that tiny changes in a dynamic environment can be revealed through the Eulerian Spatio-temporal processing of video sequences. The method does not perform feature tracking or optical flow computation but merely magnifies temporal colour changes using Spatio-temporal processing. However, as the amplification factor increases, the noise is also significantly amplified. In 2014, Wadhwa et al [37] described a new technology using the Riesz pyramid to decompose the images. The method allows for much faster and a real-time implementation of motion magnification. Although, the quality of the magnified videos in lots of scenarios of this method is better than the previous approach, the Riesz pyramid may has trouble at points where there is not a single dominant orientation. Tae-Hyun et al [38], introduced a deep learning based motion magnification approach. Different from other hand-design approaches, this method used a deep convolutional neuro network to train a spatial decomposition filter from a synthetic data set. The approach achieves better noise handling and has fewer edge artifacts. The very small motions sometime cannot be aware by the algorithm, which could lead to patchy magnification.

2.4 Computer Vision for Displacement Measurement in SHM

To analyse the state of a structure, it is essential to collect many data related to it, such as accelerator, displacement, strain and so on. Otherwise the SHM could become more arduous. The behaviours of structures can be described by the displacement responses which is one of the most important sorts of data for SHM [2]. The displacement response of structures in civil engineering applications is traditionally measured via physical sensors, which can be divided into two categories: contact-type sensors, e.g., linear variable differential transformer (LVDT), accelerometers, and noncontact-type sensors, e.g., laser displacement sensor (LDS), Global Navigation Satellite System (GNSS), and Lidar [39, 40, 41] etc. However, the use of these physical sensors is greatly restricted by accessibility. For example, the time/labour consuming and the associated cost for installing and maintaining of such measurement systems can be very high and in many cases can even be prohibitive, especially for large structures, such as long bridges or overpasses [42]. Another disadvantage of contact-type sensors is that a stationary platform must be used as a reference, which is often hard to find in practical scenarios [43]. The installation of noncontact-type sensors such as LDSs is easier and high measurement accuracy can be achieved. However, fixed platforms still need to be found for installation, and full-field measurement remains a big challenge for LDSs due to the installation and sparse beam density. GNSS systems are always used for large-scale structures, such as high-rise buildings, large bridges and so forth. Generally, the GNSS is integrated with some auxiliary means (Accelerometer [44], Robotic Total Station [45], etc.) to improve the measurement accuracy which can reach around 10-20 mm [46]. The accessibility of GNSS is the biggest problem for some medium- or small-scale structures. Lidars are widely used in self-driving vehicles. They can also be used to provide high-precision displacement measurement results for SHM. Nevertheless, Lidar systems have two main limitations: 1) for some low budget tasks, it is not worthwhile to spend a lot of money on an exorbitant Lidar system; 2) low resolution makes the radar systems unable to meet full-field measurement requirements.

Computer vision system, being convenience and economical, have been used extensively for measuring the displacement of civil structures since 1990s [47]. Vibrations of civil structures to be inspected are first recorded by one or multiple cameras. The images/videos are then processed by computer vision algorithms to extract vibrational displacements. One of the straightforward

advantages of the computer vision based approach is the simple setup and installation. Compared with the traditional contact-type displacement sensors which require a stationary and separate fixed platform for the installation, video cameras can be placed at a long distance in a non-contact manner, and are not affected by the structural vibrations to obtain the dynamic displacement responses of structures. Computer vision-based vibration displacement measurement approaches can be classified into two categories: two-dimensional (2D) displacement measurement and three-dimensional (3D) displacement measurement. The 2D or so-called in-plane vibration displacement can be extracted from single-view images as single-view images contain 2D information, but the depth direction information is lost over the camera imaging processing. At the present, for 2D displacement measurement, many vision-based methods such as target-free measurement, full-field measurement, and even tiny displacement measurement have been developed. They can be used for different SHM applications. 3D displacement measurement, on the other hand, attempts to acquire the out-of-plane displacement as well which is the displacement in the depth direction. Studies of out-of-plane vibration displacement measurement in SHM is still in the early stage.

To improve the accuracy and robustness of vision-based displacement measurement, it used to be a common practice to affix high-contrast artificial targets, such as chessboard patterns, spots and circle patterns, etc. on objects to assist the key point detection and tracking [47]. However, such artificial targets are not often available in real applications. It is almost impossible to affix artificial targets on every area of interest of a large civil structure to achieve full-field measurement. Due to the limited availability of artificial targets, target-free displacement measurement, which do not require any artificial targets to be installed, have aroused great research interests in recent years, which significantly increases the practicability of vision-based displacement measurement.

There have been numerous attempts to use vision-based methods for displacement measurement of civil structures. A 2D vision-based approach is presented by Cigada et al. [48] to measure the bridge response due to train pass-by using different image processing methods, and the results are compared with displacement sensors. Although both artificial target and target-free methods are tested, the approach highly relies on artificial targets. Feng et al. [49] developed a vision sensor system to measure the 2D displacement of civil structures based on template matching, in which both high-contrast artificial targets and natural key points were tested. Some

studies have also investigated the target-free 2D measurement based on key point detection and tracking. Kuddus et al. [50] presented a full-field target-free 2D tracking approach using BRISK key points and the KLT algorithm. Ji et al. [51] proposed a method to measure the 2D displacement of a target-free small cable by the optical flow algorithm. Bartilson et al. [52] presented a vision-based high precision 2D measurement system for traffic signal structure which can be rapidly set-up and taken-down to evaluate the structural stresses. Yoon et al. [53] proposed a 2D target-free vision measurement method tested on a six-story model, employing the Harris corner feature as the key points, and tracking them using the KLT optical flow algorithm. Morliler and Michon [54] introduced a practical vision-based 2D vibration measurement system with KLT trackers, which is used to estimate the displacement of a simple Oberst beam and the first two main modes of a helicopter blade.

Vision-based 3D displacement measurement of civil structures has also been studied. Park et al. [55] proposed a motion capture system with multiple cameras to measure 3D displacements with respect to a predetermined origin of the 3D coordinate system. The coordinates of artificial targets taken by each camera are used to calculate the 3D coordinates. Abdelbarr et al. [56] presented a method using an RGB-D sensor which is equipped with an RGB camera and an active depth sensor to measure multi-component displacement fields of flexible structures. Artificial markers are needed to install on the surface of the target structure. A target-free method is introduced by Gao et al. [57] to obtain the 3D dynamic responses of structural vibration of a double-tube five-layer structure. However, only the Y direction (vertical direction) result has been tested and verified. Limited studies investigated the feasibility and practicability of vision-based civil structural out-of-plane displacement measurement. The biggest limitation of the existing approaches is their dependence on artificial targets, due to the difficulty of natural key point detection and corresponding.

In some structural displacement measurement tasks, the displacement responses of civil engineering structures can be tiny, sometimes even invisible to human eyes. Sometimes the tiny displacement in videos can be caused by the distant positioning of the recording cameras. A structure could have large displacement on the engineering unit (mm, cm, etc.), but the motion is recorded as smaller than one pixel in the video. In such cases, accurately extracting such tiny displacement by computer vision algorithms is impossible. If the displacements are actually

smaller than a certain value of the engineering unit, even the traditional physical sensors could fail to effectively detect them. Motion magnification algorithms have been employed on videos with tiny motions for modal identification and 2D displacement extraction. Yang et al. [58] developed a blind source separation (BSS) based modal analysis algorithm in 2013, which utilized the phase-based motion amplification framework to acquire modal frequencies, damping ratios, and mode shapes of structures. Chen et al. [59] employed the phase-based motion magnification for 2D displacements extraction and demonstrated the algorithm's capability of qualitatively identifying the operational deflection shapes of a cantilever beam and a pipe from videos. Poozesh et al. [60] proposed a vision-based approach, which employed the phase-based motion magnification technique and stereo-photogrammetry to replace contact-type measurement sensors. However, artificial targets are required to be affixed on the structure to enhance the ability of the key point detection, and obtain the mode shapes of structures. Sarraf et al. [61] presented a method to extract the resonant frequencies and operational deflection shapes of a long wind turbine blade at the 2D level. The phase-based motion magnification method is applied to magnify the video of the tested structure to provide better visualization of structural modal characteristics.

2.5 Summary

In this chapter, we have reviewed the computer vision algorithms that are related with the 3D displacement measurement for SHM. We have also reviewed the existing computer vision applications on SHM and their shortcomings. The next Chapter will be focused on a target-free 3D displacement measurement system we designed for civil structure displacement measurement.

CHAPTER 3

TARGET-FREE THREE-DIMENSIONAL DISPLACEMENT MEASUREMENT

3.1 Introduction

One of the main objectives of this thesis is on the measurement of 3D displacements of civil structures. In this chapter such a measurement system is proposed which allows target-free full-field measurement. The video recording system consists of two independent monocular cameras, which recorded two simultaneous videos about the structure as the input of the system. The calibration of the binocular camera system is first implemented. Then natural key points are detected, matched, and tracked. The 3D displacement can be extracted from the matched key points in each frame. The performance of the proposed vision-based system is evaluated by a 3D vibration experimental test on a cantilever beam structure in the laboratory. In order to measure the accuracy of the proposed method, LVDTs and LDSs are used to measure the displacement responses of the tested structures at the same time. The results of the proposed vision-based approach are compared with the displacement responses measured by LVDTs and LDS. To the best of the authors' knowledge, this is the pioneer study to investigate the target-free full-field 3D vibration displacement measurement of civil engineering structures using a binocular camera system from video footages and deep learning techniques. The flowchart of the proposed binocular vision-based structure displacement measurement system is shown in Figure 3.1. Our contribution in this chapter is the design of a novel binocular camera system, which accomplish the requirement of 3D target-free vibration displacement measurement for SHM.

This chapter is reprinted, with permission, from [Shao, Y., Li, L., Li, J., An, S., & Hao, H. (2021). Computer vision based target-free 3D vibration displacement measurement of structures. Engineering Structures, 246, 113040. © 2021 Elsevier B.V. DOI: https://doi.org/10.1016/j.engstruct.2021.113040].

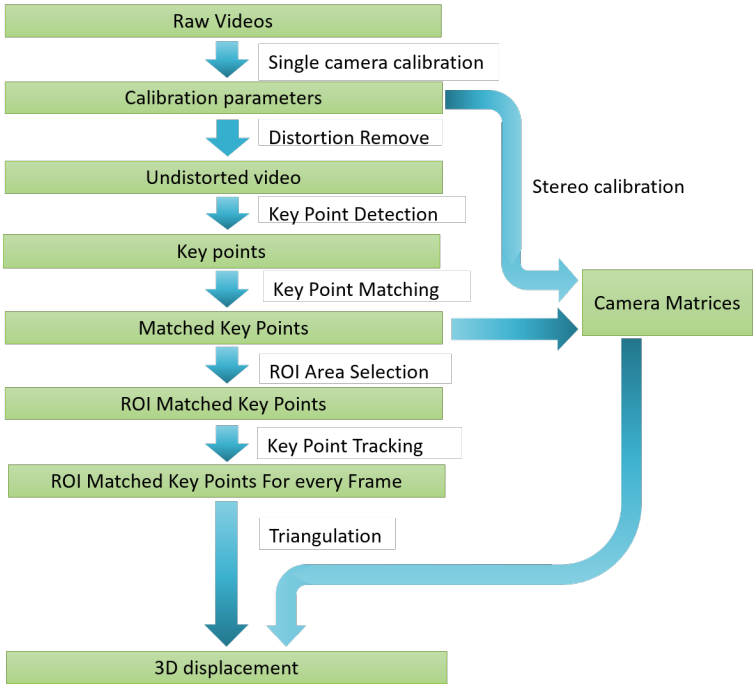


Figure 3.1. The flowchart of the proposed binocular vision-based 3D vibration displacement measurement system

3.2 Camera Calibration and Triangulation for 3D Reconstruction

3.2.1 Pinhole Camera Model

The pinhole camera model is employed to mathematically represent the geometric construction of the binocular camera system in this study. We first introduce the pinhole model of the monocular camera to pave the way for the whole process. The architecture of the pinhole model of a monocular camera is shown in Figure 3.2.

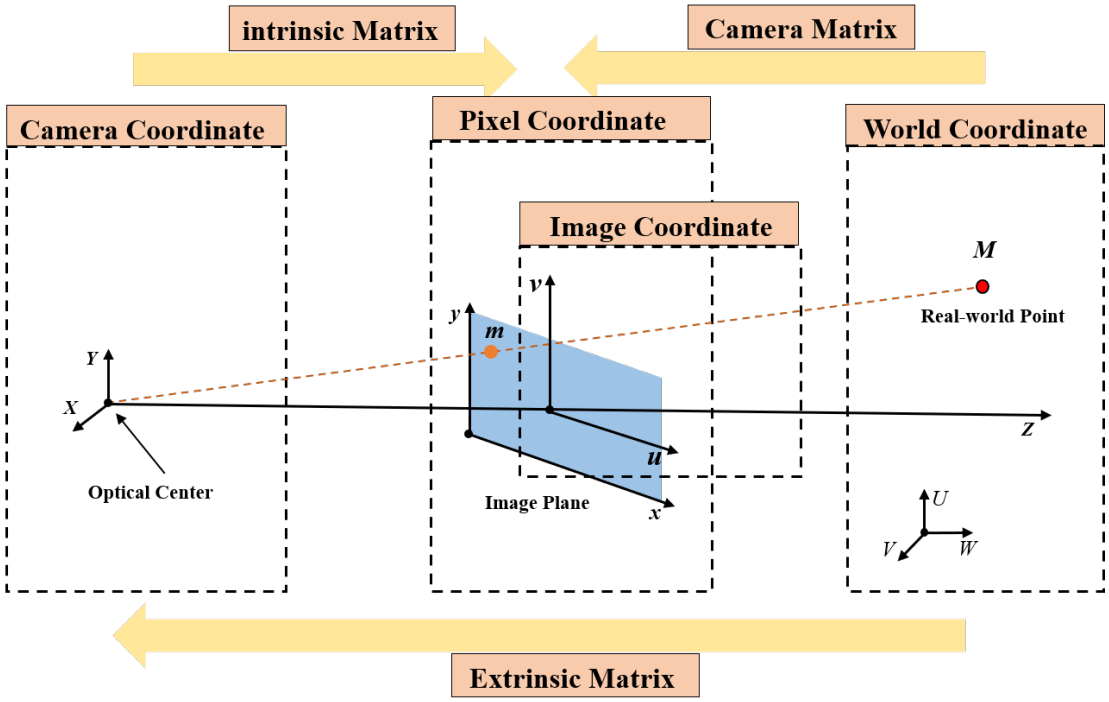


Figure 3.2. The pinhole model of a monocular camera

In the pinhole model theory, four coordinate systems are defined: World coordinate system, camera coordinate system, image coordinate system and Pixel coordinate system. The imaging process in Figure 3.2 is that an interest point *M* in the world coordinate system is projected to the pixel coordinate system as a pixel *m*. The world coordinate system is the reference system of the 3D real-world; we denote it by *(U, V, W)* in this thesis. The point *M* in the world coordinate system is first transferred to the camera coordinate system by a 3×3 extrinsic matrix, which is also a 3D coordinate system and the *Z* axes of it are perpendicularly cross the origin of the image coordinate system. It is denoted by *(X, Y, Z)*. The image coordinate system and pixel coordinate system are 2 dimensional. The interest point in the camera coordinate system is first transferred to the image coordinate system and then to the pixel coordinate system, denoted as *(x, y)* and *(u, v)* respectively, by a 3×4 intrinsic matrix. The camera matrix is a 3×4 matrix which is the product of the extrinsic matrix and intrinsic matrix. The interest point *M* can be directly transferred to pixel *m* by a camera matrix. The relationship between a 3D point in the world coordinate *M* and its 2D image *m* in the image coordinate for a monocular camera is given as:

$$s \times m = P \times M \tag{1}$$

where P is the camera matrix, s is a scale factor. The above equation can be expressed as:

$$S \times \begin{bmatrix} u \\ v \\ 1 \end{bmatrix} = A[R|t] \begin{bmatrix} X \\ Y \\ Z \\ 1 \end{bmatrix} \tag{2}$$

The extrinsic matrix is denoted as $[R|t]$, which is a rigid body transformation that consists of a rotation matrix and a translation matrix. A is the intrinsic matrix which consists of the principal point (u_0, v_0), the scale factors for the u and v axes of the image dX, dY, and the angle between the vertical and horizontal axes of the image sensor, namely θ, which is 90° if the axes are exactly perpendicular. f is the focal length. They can be expressed as:

$$[R|t] = \begin{bmatrix} r_{11} & r_{21} & r_{31} & t_1 \\ r_{12} & r_{22} & r_{32} & t_2 \\ r_{13} & r_{23} & r_{33} & t_3 \end{bmatrix} \tag{3}$$

$$A = \begin{bmatrix} \frac{1}{dX} & -\frac{cot\theta}{dX} & u_0 \\ 0 & \frac{1}{dYsin\theta} & v_0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} f & 0 & 0 & 0 \\ 0 & f & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix} = \begin{bmatrix} \frac{f}{dX} & -\frac{cot\theta}{dX} & u_0 \\ 0 & \frac{f}{dYsin\theta} & v_0 \\ 0 & 0 & 1 \end{bmatrix} \tag{4}$$

The error caused by the grinding in the manufacturing process of the camera lens will cause the radial distortion of the lens. This distortion makes the shape of the object in the video different from the actual object, which leads to errors in the measurement of the vision system. The tangential distortion, which is one of the two distortions, is ignored in our system since the modern manufacturing process has enough capability to reduce it to a negligible level. A solution is proposed in [11] to eliminate radial distortions. The redial distortion can be modelled as:

$$\begin{aligned} \breve{x} &= x_c(1 + k_1 r^2 + k_2 r^4) \\ \breve{y} &= y_c(1 + k_1 r^2 + k_2 r^4) \end{aligned} \tag{5}$$

where k_1 and k_2 are the radial distortion coefficients, (x_c, y_c) is defined as the normalized ideal image coordinate and $(\breve{x}, \breve{y})$ the distorted image coordinates, and

$$r = x_c^2 + y_c^2 \tag{6}$$

$$\begin{bmatrix} u_i \\ v_i \\ 1 \end{bmatrix} = \begin{bmatrix} \frac{1}{dX} & -\frac{cot\theta}{dX} & u_0 \\ 0 & \frac{1}{dY sin\theta} & v_0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} x_c \\ y_c \\ 1 \end{bmatrix}. \tag{7}$$

(u_i, v_i) is considered the ideal pixel coordinates. The angle, namely θ, between the vertical and horizontal axes of the image sensor is 90^{o}. Hence,

$$\begin{aligned} u_i &= \frac{x_c}{dX} + u_0 \\ v_i &= \frac{y_c}{dY} + v_0 \end{aligned} \tag{8}$$

and

$$\begin{aligned} \breve{u} &= \frac{\breve{x}}{dX} + u_0 \\ \breve{v} &= \frac{\breve{y}}{dY} + v_0 \end{aligned} \tag{9}$$

Substitute Eq. (8) into the formula above, we have

$$\begin{aligned} \breve{u} - u_0 &= (u_i - u_0)(1 + k_1 r^2 + k_2 r^4) \\ \breve{v} - v_0 &= (v_i - v_0)(1 + k_1 r^2 + k_2 r^4). \end{aligned} \tag{10}$$

So,

$$\begin{aligned} \breve{u} &= u_i + (u_i - u_0)(k_1 r^2 + k_2 r^4) \\ \breve{v} &= v_i + (v_i - v_0)(k_1 r^2 + k_2 r^4) \end{aligned} \tag{11}$$

or equivalently

$$\begin{bmatrix} (u_i - u_0)r^2 & (u_i - u_0)r^4 \\ (v_i - v_0)r^2 & (v_i - v_0)r^4 \end{bmatrix} \begin{bmatrix} k_1 \\ k_2 \end{bmatrix} = \begin{bmatrix} \breve{u} - u_i \\ \breve{v} - v_i \end{bmatrix}. \tag{12}$$

The radial distortion coefficients k_1 and k_2 can be obtained by the linear least-squares method to remove the distortion from the video images.

3.2.2 Two-view Camera Geometry

In the previous section, the monocular pinhole camera system is introduced. We use a binocular camera system which combines two monocular cameras to realize the 3D displacement measurement. Establishing a connection from the binocular camera system to the world coordinate of a measured object is important for a binocular camera system but using complicated manual measurement could result in a large measurement error. An Epipolar geometry based approach is applied to estimate the camera extrinsic matrices in this study. The epipolar geometry encodes the relationship between the two cameras. It is independent of scene structure and only depends on the intrinsic matrices and relative poses of the cameras.

The epipolar geometry is shown in Figure 3.3. A point *M* in 3D is projected to two pixels in the pixel coordinate systems, which are denoted as m and m' respectively. With the knowledge of camera origins and the image points, we can define an epipolar plane (π). The intersections of the epipolar plane and the two image planes are the epipolar lines l and l'. e and e' are epipoles. All the epipolar lines pass through their epipoles.

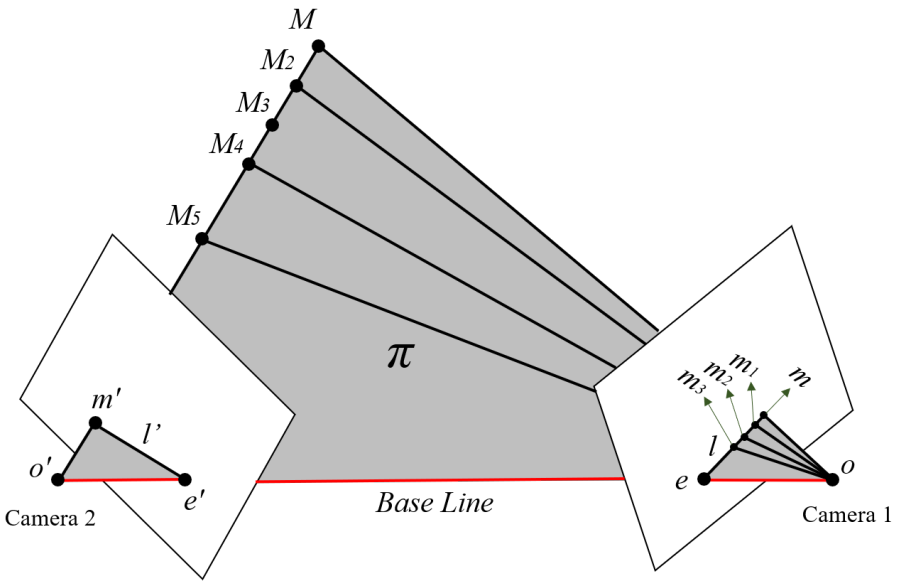


Figure 3.3. The architecture of epipolar geometry

The epipolar constraint is that the corresponding point of m can be found along the epipolar line l', and it must be somewhere on the line. A basic understanding of epipolar geometry allows us to create a strong constraint between image pairs without knowing the 3D structure of the scene. The epipolar geometry is encapsulated by an essential matrix or fundamental matrix.

Assume that we have canonical cameras, in which intrinsic matrices are I, the camera matrices of camera 1 and camera 2 can be written as:

$$P_1 = I[I \quad 0] \tag{13}$$

$$P_2 = I[R^T \quad -R^T t] \tag{14}$$

where R and t are the rotation and translation between two cameras. Assuming we have point m by Camera 1 and m' by Camera 2. In Camera 1's reference coordinate system, the location of m' is $Rm' + t$. The cross product of t and $(Rm' + t)$ generates a new vector perpendicular to the epipolar plane:

$$t \times (Rm' + t) = t \times Rm'. \tag{15}$$

Since m lies on the epipolar plane which is normal to $t \times (Rm')$, we have the following constraint:

$$m^T \cdot [t \times (Rm')] = m^T [t_\times] Rm' = \ 0. \tag{16}$$

The Essential matrix can be defined as:

$$E = [t_\times] R. \tag{17}$$

The Essential matrix has the following important properties: (1) the Essential matrix is a 3 × 3 matrix; (2) it contains 5 degrees of freedom; (3) it has rank of 2; (4) it is singular; and (5) a 3×3 matrix is an essential matrix if and only if two of its singular values are equal, and the third is zero.

If the points are already represented in one of the pixel coordinate systems, the camera matrix *P* can then be calculated via the following formula:

$$P_1 \ = \ A_1 [I \ \ 0] \tag{18}$$

$$P_2 \ = A_2 [R^T \ -R^T t] \tag{19}$$

where A_1 and A_2 are the intrinsic matrices of Camera 1 and Camera 2, respectively. R and t are the rotation and translation between the two cameras. The points in the pixel coordinate system can be converted to the camera coordinate system by the following formula:

$$m_c = \ A_1^{-1} m \tag{20}$$

$$m_c' = \ A_2^{-1} m' \tag{21}$$

Similar to (16), we have the constraint:

$$m_c^T [t_\times] R m_c' = m^T A_1^{-T} [t_\times] R A_2^{-1} m' = 0. \quad (22)$$

The Fundamental matrix F can be expressed as:

$$F = A_1^{-T} [t_\times] R A_2^{-1}. \quad (23)$$

3.2.3 Camera Calibration

Camera calibration is a common procedure in videogrammetry to obtain the geometrical projection information (camera matrix) of a camera system in order to recover the 3D information from images [11], and it is also crucial for removing geometric distortions [50].

A camera matrix is the product of the corresponding intrinsic matrix and extrinsic matrix. The method described in [11] is applied in this study to extract the intrinsic matrices A_1 and A_2. When the origin of the world coordinate is not clear, the 3D camera coordinate system of Camera 1 can be considered as the world coordinate system, hence the extrinsic matrix of Camera 1 is an identity matrix. Since the extrinsic matrices describe the conversion from the world coordinate system to the camera coordinate system, the transformation between the camera coordinate systems of Camera 1 and Camera 2 can be considered as the conversion between the world coordinate system and the Camera 2 coordinate system. Therefore, the rigid body transformation $[R|t]$ between two cameras is the extrinsic matrix of Camera 2.

Since images are in the pixel coordinate systems, the fundamental matrix F can be calculated by the celebrated eight-point algorithm [62]. For robustness, the eight-point algorithm actually requires more than eight pairs of matched key points as inputs, which can be provided by key point detection and key point matching, as described later in Section 3.3. In SHM, the unit of measurement in need is generally in engineering units (e.g., mm, cm, m) rather than in pixel units. However, the unit associated with a fundamental matrix is in pixels. Fortunately, the essential matrix is the specialization of the fundamental matrix to the case of normalized image coordinates, which are associated with engineering units. The normalized image coordinate of a point m is $m_c = A_1^{-1} m$. The essential matrix can be derived from the fundamental matrix by the following

formula:

$$E = A_1^T F A_2. \tag{24}$$

Once the essential matrix is known, the camera matrices can be retrieved. A method based on singular value decomposition (SVD) is employed to decompose the essential matrix to find the rigid body transformation $[R|t]$ between two cameras. We define the essential matrix E as:

$$E = [t]_\times R = SR \tag{25}$$

where S is a skew-symmetric matrix, which has the property $S = kUZU^T$. Here U is an orthogonal matrix, and Z is a block matrix $diag(s_1 D_1, \dots, s_n D_n, \dots, 0, \dots, 0)^T$. Followed by the property of skew-symmetric matrix, the determinant of S is 0, thus Z can be expressed as:

$$Z = \begin{bmatrix} 0 & 1 & 0 \\ -1 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} \tag{26}$$

In the situation without the scale factor k, the decomposition of the essential matrix can be expressed as $E = UZU^T R$. We also define an orthogonal matrix W:

$$W = \begin{bmatrix} 0 & -1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix} \tag{27}$$

The matrix W is a rotation and Z is skew symmetry. Furthermore, for these matrices we have:

$$ZW = diag(1,1,0) \tag{28}$$

$$ZW^T = -diag(1,1,0) \tag{29}$$

The SVD decompensation of $E = Udiag(1,1,0)V^T$ can be expressed by two situations:

$$E = UZU^T R = Udiag(1,1,0)(W^T U^T R) = Udiag(1,1,0)V_1^T \tag{30}$$

$$E = UZU^T R = Udiag(1,1,0)(-WU^T R) = Udiag(1,1,0)V_2^T \tag{31}$$

An important property of the essential matrix as mentioned before is that two of its singular values are equal, and the third one is zero. The decomposition of rotation matrix R can be written as:

$$R = UXV^T \tag{32}$$

where X is a rotation matrix. From Eq. (30) and Eq. (25), we have

$$Udiag(1,1,0)V^T = E = SR = (UZU^T)(UXV^T) = U(ZX)V^T \tag{33}$$

Hence, $ZX = diag(1,1,0)$, $X = W$ or $X = W^T$. Therefore, the rotation matrix has two situations:

$$\begin{aligned} R &= UWV^T \\ R &= UW^T V^T \end{aligned} \tag{34}$$

When a decomposition $E = SR$ have been determined, a translation vector t from S such that $[t]_\times = S$ needs to be computed. For such a t we have

$$St = [t]_\times t = UZU^T t = 0 \tag{35}$$

Since $St = 0$, it follows that $t = U(0,0,1)^T = u_3$, the last column of U. However, the sign of E, and consequently t, cannot be determined. Therefore, the translation vector has two possible choices:

$$t = \pm U(0,0,1)^T = \pm u_3 \tag{36}$$

The extrinsic matrix of Camera 2 is the combination of the rotation matrix and the translation vector. Therefore, there are four possible solutions for extracting the relative camera rotation R and translation t from the essential matrix. For the correct pair (R, t), the 3D triangulated point P exists in front of both cameras, which means that it has a positive z-coordinate with respect to both the camera coordinate systems.

Once the extrinsic matrix $[R|t]$ has been extracted, the camera matrices, namely P_1 for Camera 1 and P_2 for Camera 2, can then be simply derived from the intrinsic matrix and extrinsic matrix via the following formula:

$$P_1 = A_1[\mathrm{I}|0] \tag{37}$$

$$P_2 = A_2[R|t]. \tag{38}$$

The camera mapping procedure is fully described by camera matrices, which is one of the most essential inputs of the 3D triangulation for depth direction displacement recovery. A polynomial triangulation approach will be introduced in the next section.

3.2.4 Polynomial Triangulation for 3D Reconstruction

In this section, an introduction to the polynomial triangulation method is given based on [62] and [63]. Camera imaging is a process of projection from the 3D space to a 2D plane. If a camera is used for back-projection, a ray directed to infinity is employed. However, if two angled cameras are used for back-projection, two rays directed to infinity will intersect at a certain point. Triangulation is a method for the 3D reconstruction of image data using this method. However, due to the inaccuracy camera calibration and key point detection, errors on reconstructed 3D coordinates could happen. A nonlinear triangulation method, namely polynomial triangulation [62], is employed in our system to optimize the 3D coordinates of the key points.

With known camera matrices and corresponding points, two projection rays of the corresponding points can be calculated easily. However, the problem will be complicated when there are noises. Normally, two optical rays will not intersect perfectly under the influence of disturbance. In this study, the triangulation process is represented as a least-squares minimization

problem. The solution of the global minimum of the cost function is obtained using a non-iterative algorithm. The cost function is formulated as

$$\mathcal{C}(u, u') = d(u, \hat{u})^2 + d(u', \hat{u}')^2 \tag{39}$$

where u and u' are a pair of matched key points, $\hat{u}$ and $\hat{u}'$ are the estimated corresponding image points respectively, which should satisfy the epipolar constraint function $\hat{u}'^T F \hat{u} = 0$, with F being the fundamental matrix, which is calculated from the camera calibration process. The function $d(u, \hat{u})$ denotes the Euclidean distance between the two points u and $\hat{u}$. The corresponding 3D point can be recovered by any triangulation method, when $\hat{u}$ and $\hat{u}'$ are available. Since this method needs to solve a six-order polynomial, it is called the polynomial triangulation method.

Every pair of corresponding points satisfying the epipolar geometry must lie on a pair of epipolar lines, and the distance between u and $\hat{u}$ equals to the distance between u and the epipolar line l associated with $\hat{u}$, that is, $d(u, \hat{u}) = d(u, l)$. Similarly, $d(u', \hat{u}') = d(u, l')$, where l' is the epipolar line associated with $\hat{u}'$. Furthermore, l' can be computed from l based on the epipolar constraint. Therefore, Eq. (39) can be rewritten as

$$\mathcal{C}(u, u') = d(u, l)^2 + d(u', l')^2. \tag{40}$$

For both images, the epipolar lines l and l' can be parameterized by one variable t as $l(t)$ and $l'(t)$. Hence, the right side of Eq. (39) is a rational polynomial function of t [62]

$$\mathcal{C}(t) = d(u, l(t))^2 + d(u', l'(t))^2. \tag{41}$$

The optimal $\hat{u}$ and $\hat{u}'$ can be found by solving the following minimization problem

$$\min_t \mathcal{C}(t) = d(u, l(t))^2 + d(u', l'(t))^2. \tag{42}$$

The minimization of function $\mathcal{C}(t)$ can be solved by finding the real root of a six-degree polynomial [62]. The core of the polynomial triangulation method is to obtain the correct

corresponding points to perform back-projection for the optimal intersection.

3.3 Key Point Detection, Matching and Tracking

3.3.1 Key Point Detection

Artificial targets, such as fixed-sized high contrast planar patterns, have been commonly employed in most computer vision-based motion measurement and/or 3D reconstruction tasks. These patterns are treated as key points for subsequent processing according to task requirements. In recent years, target free methods have gained attention due to their wide applicability and advantages for practical applications. However, reliable detection of a sufficiently large number of matchable natural key points remains a challenging task. In this chapter, a 3D target-free approach is proposed by using natural key points to replace the artificial targets. These natural key points are obtained by a key point detection algorithm based on deep learning, rather than traditional methods.

The method used in our proposed system is inspired by a state-of-the-art deep learning-based algorithm called SuperPoint [64], which uses a self-supervised framework in training key point detectors and descriptors. The detection algorithm can be widely used in multi-view geometry problems in computer vision. The key point detector and fixed-length descriptors are generated in a single feed forward channel by applying the algorithm on a full-size image. The system bootstraps itself from many unlabeled images in a dataset (e.g., Microsoft-COCO), after a basic detector on synthetic data is pre-trained. Due to the lack of data for key point detection training, the unlabeled image is automatically labelled by a novel Homographic Adaptation procedure, which will be described in Section 3.1. A fully convolutional network that jointly derives key point detectors and descriptors from an image is trained using the generated labels. The overview of the algorithm is shown in Figure 3.4.

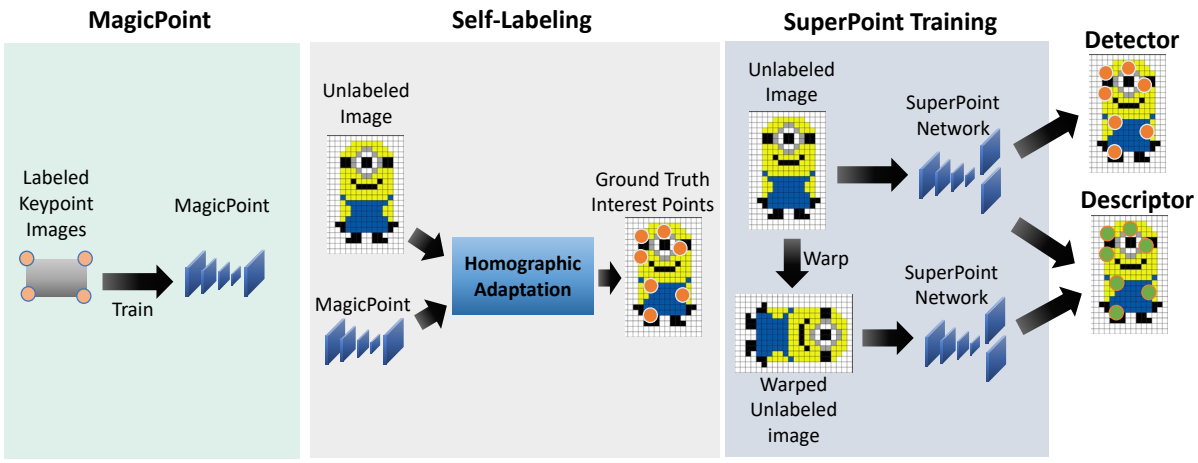


Figure 3.4. The overview of the SuperPoint algorithm

A basic detector called MagicPoint neural network [65] for dense prediction is designed, which operates on single grey scaled images and outputs the probability of each pixel that can be a key point. The architecture is shown in Figure 3.5.

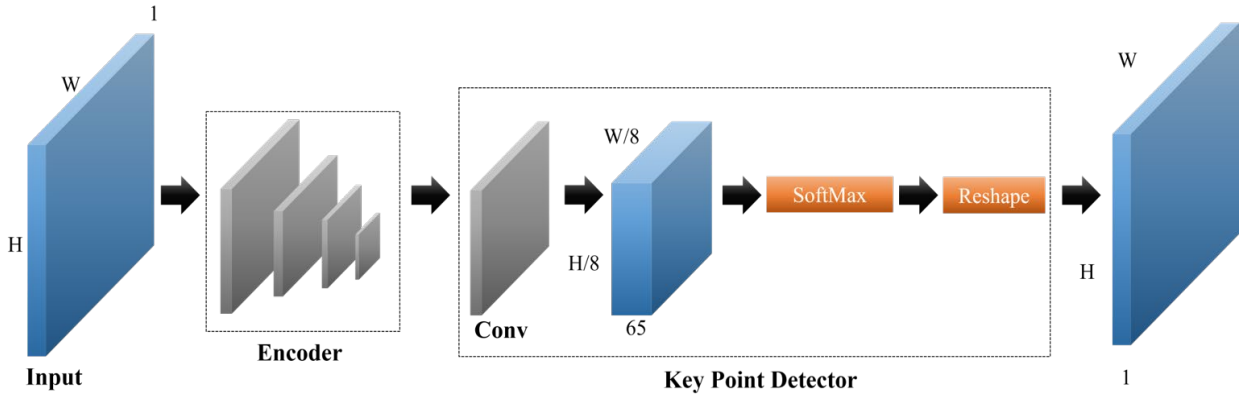


Figure 3.5. The architecture of the MagicPoint network

At the beginning of the MagicPoint convolutional neural network, a VGG-style encoder [66] is used to extract the features, which consists of convolutional layers, spatial downsampling via pooling, and nonlinear activation functions. The architecture is shown in Figure 3.6.

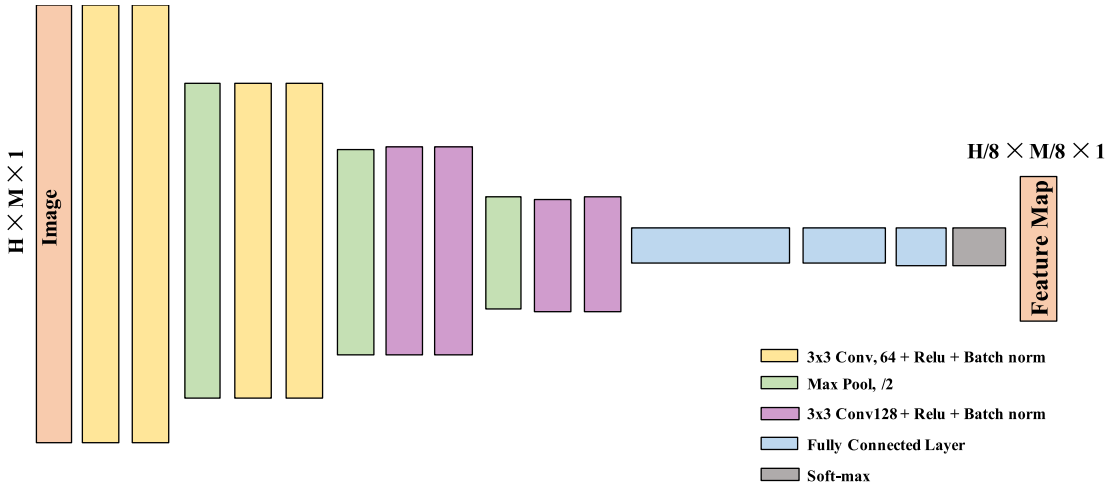


Figure 3.6. The architecture of the VGG style encoder

The encoder has four 3×3 convolutional layers with 64 channels and four 3×3 convolutional layers with 128 channels. The Rectified Linear Unit (ReLU) activation function and the batch normalization (Batch norm) are connected to every convolutional layer. Every two layers have a 2×2 max pool layer. Fully convolutional layers and a SoftMax layer are used at the end of the network. The dimension of the input image ($H \times W$) is reduced to a $H/8 \times W/8 \times 1$ cell grid by the encoder.

An explicit decoder based on Efficient Sub-Pixel Convolutional Neural Network (ESPCN) [67] is designed to upsample the feature map for resolution recovery. The output of the encoder is first fed into two convolution layers for feature extraction and a $H/8 \times W/8 \times 65$ feature map is outputted with 65 channels, which correspond to 8×8 grid regions of pixels plus a dustbin channel without key point detected. It is desirable to represent an object with key points evenly distributed on it rather than an uneven distribution (e.g., many key points are concentrated in a certain region, but very few in other places). Thus, a channel-wise Soft-max is used to compute the key points, which applies Soft-max on every 64 pixels region. To finally reshape the feature map to the original resolution, a periodic operator for spatial reshape processing is used to reshape the feature map to $H/8 \times W/8 \times 1$. Due to the lack of the key point training dataset, a large-scale synthetic dataset is created, which consists of simplified 2D geometries such as polygons, cubes, lines and ellipses, etc. The network is trained with this synthetic database using a standard cross-entropy as the loss function

$$\mathcal{L}_p(\mathcal{X}, Y) = \frac{1}{H_c W_c} \sum_{h=1;\omega=1}^{H_c, W_c} l_p(x_{h\omega}; y_{h\omega}) \tag{43}$$

where $x_{h\omega}$ is the predicted value from the MagicPoint network, $y_{h\omega}$ is the ground truth, $\mathcal{X}, Y$ are the sets of $x_{h\omega}$ and $y_{h\omega}$ respectively. $l_p(x_{h\omega}; y_{h\omega})$ is a standard cross-entropy loss. The MagicPoint performs reasonably well on images with real world objects, especially those with neat shapes, for example, cabinets, computers, windows and boxes. However, for images with more complicated shapes, the performance of MagicPoint is unsatisfactory, especially when the images are taken from various viewpoints.

The ability of the original MagicPoint for key point detection on the real image is improved by the homographic adaptation method which can provide quite accurate image-to-image transformation by manipulating the original images (rotating, scaling and translation etc.) to help the key point detector observe the scene from many different viewpoints and scales. The homographic adaptation is combined with the MagicPoint detector to improve the performance of the detector and generate the pseudo-ground truth key point for SuperPoint joint training.

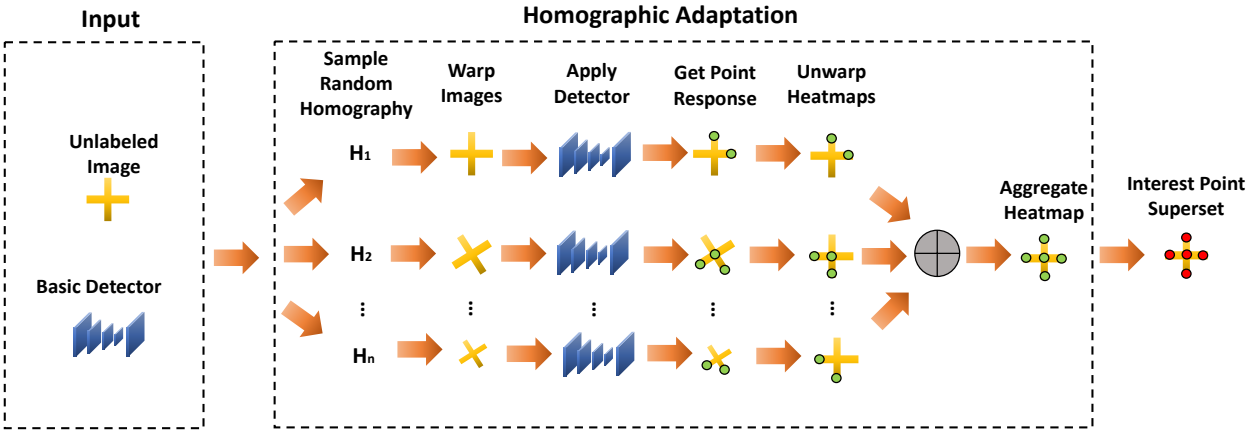


Figure 3.7. Homographic adaptation

It is believed that a good key point should be constantly detected on the original image and the modified image if an image is modified by homograph transformations. In other words, when

a scene is recorded from different views, the same key points should be detected from the images with different views. Therefore, the purpose of the Homographic Adaptation is like data augmentation in deep learning to generate more labelled key points which are repeatable in different views in natural images for SuperPoint training. As shown in Figure 3.7, the MagicPoint is first applied to an original natural image from Microsoft-COCO to detect some key points. Then the original image is modified by homographic adaptation to generate some warped images. The MagicPoint detector is applied to every warped image to generate more key points. Finally, all the key points detected from every modified image are combined with the original image. The clusters of key points generated from different images produce a new and improved detector, which can generate more repeatable and reliable labelled key points as the training data for joint training that is described below.

The last stage of the SuperPoint algorithm is joint training, which generates the SuperPoint detector and descriptor as shown in Figure 3.8.

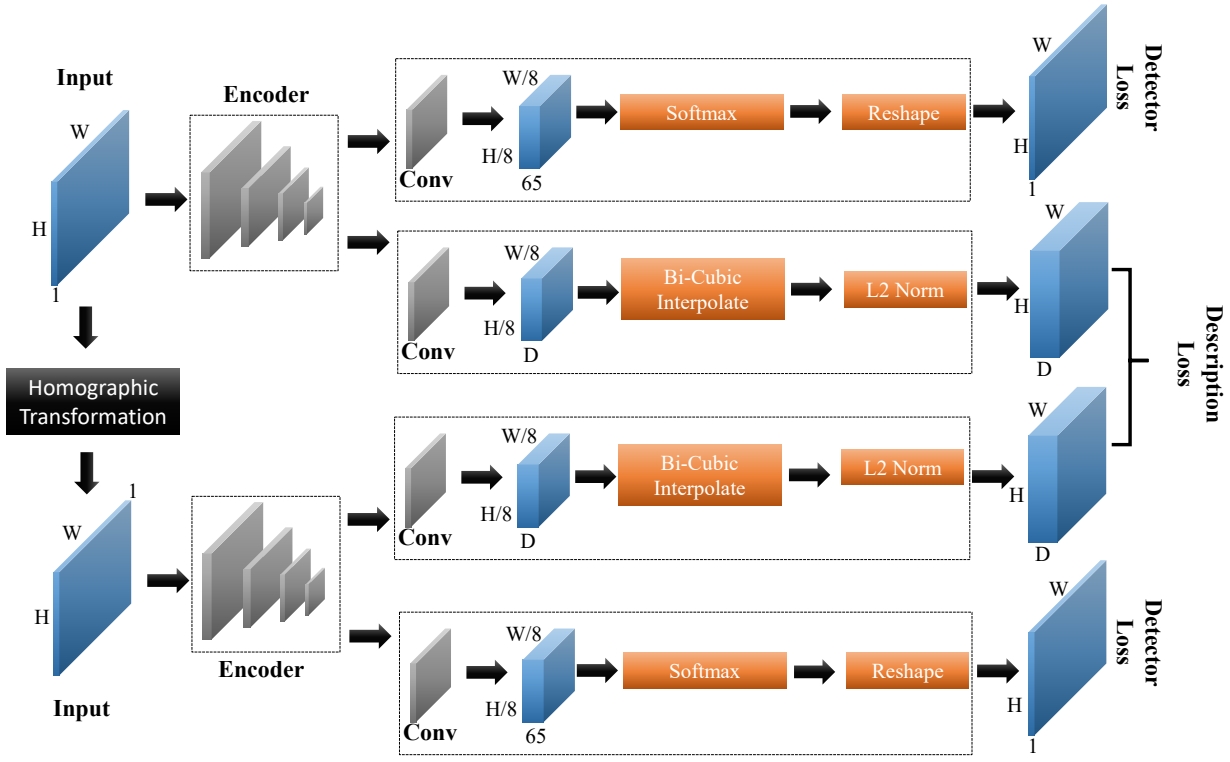


Figure 3.8. Joint training in SuperPoint

The SuperPoint network is trained by a Siamese training model. The original image is warped by a homography matrix. Two images are trained together in order to generate a descriptor. The encoder and detector of SuperPoint are exactly the same as those in the MagicPoint model. A fixed dimensional descriptor vector is commonly attached to each of these points for other tasks, e.g., key point matching. Traditionally, the key point descriptor is generated after the key point detection. The detector and descriptor are not directly connected in terms of computation and representation. The SuperPoint architecture is designed to share about 90% of computation between the detector and descriptor. An additional subnetwork that computes descriptors for points of interest is combined with the encoder, since the SuperPoint architecture is formed by a deep stack of convolutional layers that extract multi-scale features. In order to produce an L2 normalized fixed-length descriptor, a model similar to Universal Correspondence Network (UCN) [68] is applied first to generate a semi-dense grid of descriptors (e.g., one for every 8 pixels). The decoder then performs the bi-cubic interpolation of the descriptor and then normalizes the activations to be the unit size.

The final loss is a combination of the detector loss $\mathcal{L}_p$ and the descriptor loss $\mathcal{L}_d$:

$$\mathcal{L}(\mathcal{X},\mathcal{X}',\mathcal{D},\mathcal{D}',Y,Y',S) = \mathcal{L}_p(\mathcal{X},Y) + \mathcal{L}_p(\mathcal{X}',Y') + \lambda\mathcal{L}_d(\mathcal{D},\mathcal{D}',S) \tag{44}$$

where $\mathcal{L}_p(.,.)$ is the same as those in MagicPoint for the original image and the warped image, respectively. $\mathcal{L}_d(\mathcal{D},\mathcal{D}',S)$ is the loss function of the descriptor, $\mathcal{D}$ and $\mathcal{D}'$ are key point descriptors, we denote the entire set of correspondences for a pair of images with S.

For training the SuperPoint network, the MagicPoint must be pre-trained to act as a basic detector, using synthetic shapes for 200,000 iterations. SuperPoint is then trained on 80,000 wrapped images in MS-COCO dataset, and it is evaluated on the HPatches dataset [69] which has 116 scenes with 696 unique images. All training uses PyTorch [70] with mini-batch sizes of 32. Adam optimizer [71] is used during training with a default learning rate of 0.001. The exponential decay rate for the first moment estimates is 0.9 and for the second-moment estimates is 0.999.

3.3.2 Key Point Matching

In order to achieve 3D vibration displacement measurement, after the key points are detected

using the SuperPoint algorithm, the corresponding key points in the two camera images need to be matched. The matched key points play a vital role in solving the fundamental matrix and triangulation thus allowing the back projection from 2D to 3D. Without artificial targets, such kind of key point matching is a very challenging task in computer vision. A deep learning-based key point matching algorithm called SuperGlue [72] is recently proposed, which consists of two major parts, namely, the attentional graph neural network and the optimal matching layer. Partial point visibility and occlusion can be handled in the SuperGlue algorithm by solving an assignment optimization problem. The SuperGlue network is described in the following subsections and the architecture of the SuperGlue network is shown in Figure 3.9.

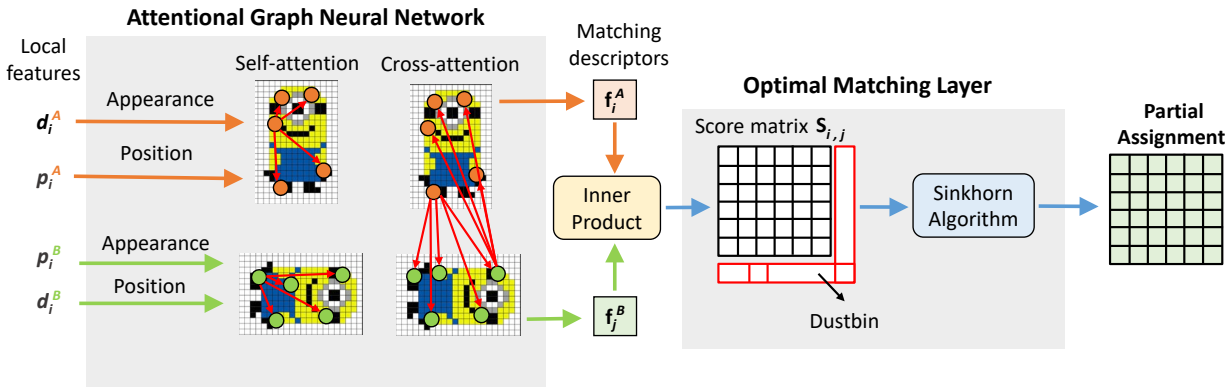


Figure 3.9. The architecture of the SuperGlue network

In the attentional graph neural network, a key point encoder is used to combine the positions p_i and the visual descriptors d_i for key point i through a Multilayer Perceptron (MLP). To make the features related to each other, self- and cross-attention layers are applied to create a new matching descriptor f_i. A new initial representation of key points on the first layer is generated as:

$$(0)_{\mathrm{x}_i} = d_i + MLP_{enc}(p_i). \tag{45}$$

To match key points, two images are observed to filter some key points and check back and forth. If they do not match, it is required to observe whether there are better key points around until the correct key points are determined [73]. In order to imitate the behaviour of human beings, an

attention graph neural network [74, 75] is designed to connect a key point with other key points in the same image (self-attention), and with key points in another image (cross-attention). Self-attention is like the process of human beings looking for similar key points in the same image. The cross-attention is to imitate the process of humans trying to match two images after finding good key points.

A message passing formulation [76, 77] is used to propagate information along with both types of attention. The residual message passing updates for all key points i in an image a is expressed as

$$(\ell + 1)_{\mathrm{x}_i^a} = (\ell)_{\mathrm{x}_i^a} + MLP([(\ell)_{\mathrm{x}_i^a} || m_{\varepsilon \to i}]) \tag{46}$$

where $[.||.]$ denotes concatenation; $(\ell)_{\mathrm{x}_i^a}$ is the intermediate representation of key point i in Image a at Layer ℓ; Message $m_{\varepsilon \to i}$ is the result of the aggregation from all key points, which is computed as a weighted average of the values. The final matching descriptors f_i of images are linear projections of the representation x_i in the last layer.

The optimal matching layer is responsible for generating a partial assignment matrix. Two assumptions are made for corresponding key points: 1) a key point is corresponding with at most a single key point in another image; 2) some key points will be discarded due to occlusion or detection failure. A confidence value can hence be defined for each possible correspondence between the corresponding key points derived from a partial assignment between two sets of key points. The goal of the optimal matching layer is to design a network to find the assignment H from two sets of local features.

The assignment matrix H can be obtained by maximizing the total score $\sum_{i,j} S_{i,j} H_{i,j}$, where S is the similarity score matrix: $S_{i,j} = < \mathrm{f}_i^a, \mathrm{f}_j^b >$. The optimization problem can be efficiently solved by the Sinkhorn algorithm [78, 79], whose entropy-regularized formulation can be easily plugged into a deep learning framework. SuperGlue is trained on ScanNet dataset [80] which has well-defined training, validation and test splits corresponding to different scenes. 230 million training data and 1500 test pairs are selected. Adam solver is applied to optimize the gradient descent. The constant learning rate is set as 0.0001 for the first 100k iterations, followed by an exponential decay of 0.999992 until the iteration number reaches 900k [72].

3.3.2 Key Point Tracking

In order to track the movement of the matched key points in in-plane directions (a.k.a., the X and Y directions) in a video sequence, the robust Kanade-Lucas-Tomasi (KLT) optical flow tracking algorithm [81, 82, 83] is applied to obtain their trajectories in the two directions. KLT retrieves the coordinates of a point in every frame by comparing the neighbours around the pixel to find the most similar one with the interested pixel. The algorithm is based on three assumptions: 1) Constant brightness; 2) Small displacement of pixels between consecutive frames; 3) Spatial consistency (Neighboring pixels move consistently).

Denoting the intensity of a point $U = (x, y)$ in the current image frame at time instant t as $I(x, y, t)$, the displacement of point U is $d = (\xi, \eta)$, and the intensity of point U at time instant $t + \tau$ is

$$I(x, y, t + \tau) = I(x - \xi, y - \eta, t) \tag{47}$$

with $J_U = I(x, y, t + \tau)$ and $I(U - d) = I(x - \xi, y - \eta, t)$. A small motion is represented as

$$J_U = I(U - d) + N(U) = I(U) - g \cdot d \tag{48}$$

where $N(U)$ is the noise (often assumed to be zero), g is the gradient vector, and d is the displacement vector of a point between two frames. The residual of the intensity changes for a tiny window can be denoted as

$$\epsilon = \int_W [I(U) - g \cdot d - J(U)]^2 \sigma dU = \int_W (h - g \cdot d)^2 \sigma dU \tag{49}$$

where $h = I(U) - J(U)$, and σ is a weighting function that could be one in the simplest case. To emphasize the window centre, σ can be set as a Gaussian function. The residual value is equal to zero when differentiating with respect to d

$$\int_W (h - g \cdot d) g \, \sigma dA = 0. \tag{50}$$

Since $(g \cdot d)g = (gg^T)d$ and d is assumed to be constant within the window $\mathcal{W}$, we have

$$\left(\int_{\mathcal{W}} gg^T \sigma dA\right) d = \int_{\mathcal{W}} hg\sigma dA \tag{51}$$

This is a system of two scalar equations with two unknowns. It can be rewritten as

$$Gd = e \tag{52}$$

where the coefficient matrix G is the symmetric 2×2 matrix:

$$G = \int_{\mathcal{W}} gg^T \sigma dA \tag{53}$$

and the right-hand side is the two-dimensional vector:

$$e = \int_{\mathcal{W}} (I - J) g \, \sigma dA. \tag{54}$$

The displacement vector d is calculated for each matched key point U in each frame of the video. The KLT algorithm is performed simultaneously in two synchronized videos, therefore the key points matched in the first frame are matched in each subsequent frame. These tracked key points are not only used to calculate the displacement in the *X* and *Y* directions but also used in the triangulation algorithm to extract the displacement in the *Z* direction.

The triangulation algorithm requires reasonably precise key point matching, so an outlier removing process is used on the key points tracked by the KLT algorithm. The outlier removing processing is commonly used in conjunction with key point matching and tracking to filter outlier points. In this study, an outlier removing algorithm [84] comparing forward-backward errors is adopted to improve the performance of the KLT algorithm. $V = (L_t, L_{t+1}, \ldots, L_{t+k})$ is defined as an image sequence (video) and u_t is a key point located on frame t. At first, the KLT tracker produces a trajectory by tracking the point forward in time. The trajectory can be represented as

$T_f^k = (u_t, u_{t+1}, \ldots, u_{t+k})$, where f indicates "forward" and k is the length of the video. Next the point in the last frame is tracked backwards to the first frame to obtain the trajectory $T_b^k = (\hat{u}_t, \hat{u}_{t+1}, \ldots, \hat{u}_{t+k})$, where $\hat{u}_{t+k} = u_{t+k}$. Various types of distances can be used to calculate the distance between T_f^k and T_b^k. In this study, Euclidean distance between u_t and $\hat{u}_t$ is employed for the estimation of the forward-backward error. In our experiments, when the bidirectional error is more than two pixels, the corresponding points will be considered invalid hence discarded.

3.4 Experimental Studies on a Beam Structure

3.4.1 Experiment Setup

To evaluate the performance of the proposed target-free vision-based approach for 3D vibration displacement measurement in structural engineering, a 3D vibration test on a steel cantilever beam (425 mm × 50 mm × 5 mm) are conducted in the laboratory. A bi-axial shake table and a uniaxial vibration shaker are used to produce the 3D vibrations of the testing structure. Figure 3.10 shows the experimental setup of the shake table, vibration shaker and testing specimen.

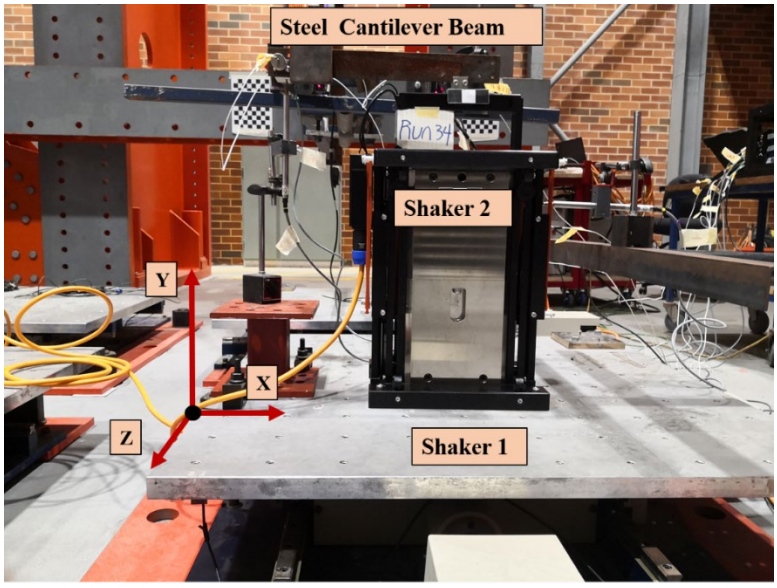
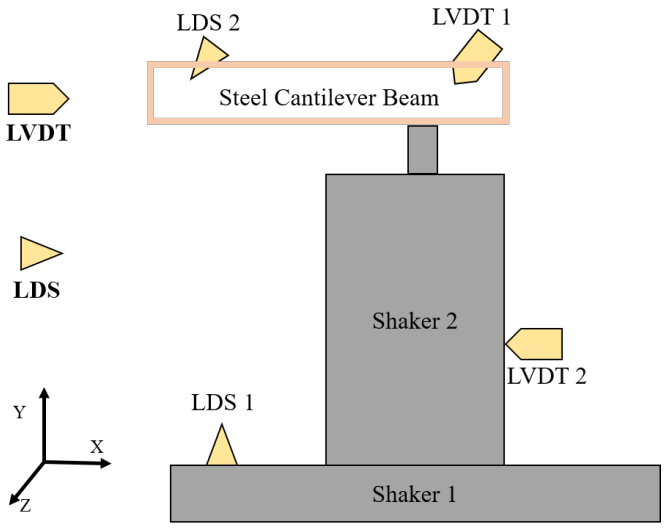


Figure 3.10. The 3D vibration test setup

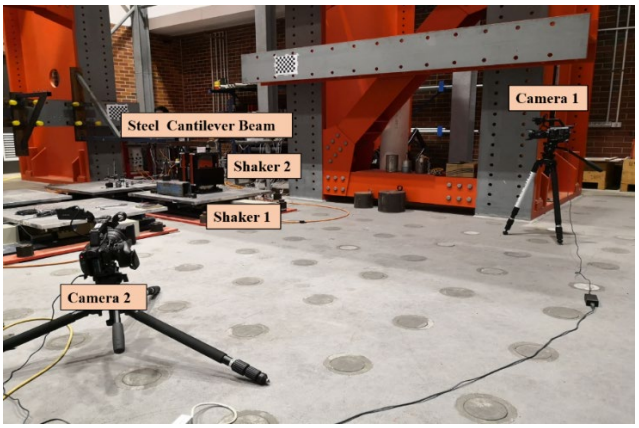
A bi-directional shake table (denoted as Shaker 1) is used to provide excitations in the *X* and *Z* directions. The excitation in the *Y* direction is provided by an APS 400 ELECTRO-SEIS long

stroke shaker (denoted as Shaker 2) that is mounted on the shake table. These two vibration shakers are used simultaneously during the vibration test to provide controllable vibration movements to the steel cantilever beam in three directions. To provide a comparison for the proposed vision-based approach for 3D vibration displacement measurement, two LVDTs, two LDSs are installed to measure the vibration displacement responses. Figures 3.11(a) shows the setup of the displacement sensors for the test. The synchronization between the vision system and the sensor system is manually adjusted.

The binocular vision system consists of two SONY video cameras (SONY PXW-FS5 4K XDCAM). A remote controller is used to remotely control these two cameras to start and finish recording synchronously. The frame size recorded by these two cameras is 1920 × 1080 and the frame rate is 50 fps (frame per second). The duration of each excitation of the vibrating shakers is 30 seconds. The experimental setup of these two cameras is shown in Figure 3.11(b). To simulate the realistic conditions, these cameras are setup at different heights with an angle between them.



(a)



(b)

Figure 3.11. The installation of displacement sensors and setup of cameras: (a) The sensor installation for the 3D vibration test; (b) The setting of the two cameras.

3.4.2 Results

In this experimental test, the sinusoidal wave excitations with an amplitude of 3 mm and a frequency of 3 Hz in X and Z directions are generated by Shaker 1. Shaker 2 is controlled by adjusting the voltage of the shaker controller, which cannot provide the exact value of vibration displacement amplitude and frequency. Multiple displacement sensors are installed to measure the vibration responses in three directions as shown in Figure 3.11(a). The details of these used sensors are shown in Table 3.1.

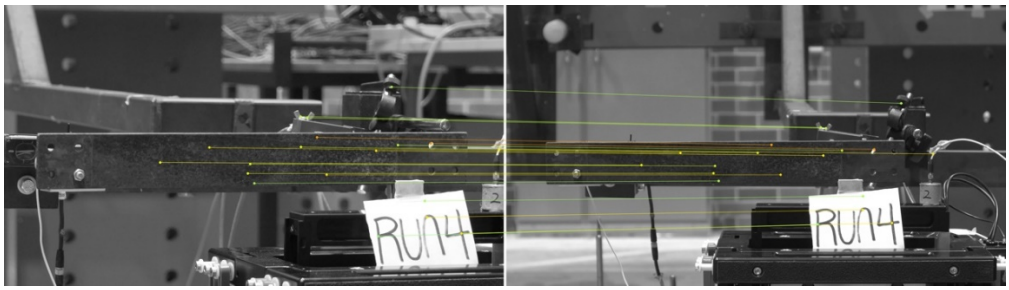
Table 3.1. Sensors used in the 3D vibration test

Sensor Name	Version	Measurement Direction
LDS 1	Keyence IL300	Y
LDS 2	Keyence IL300	Z
LVDT 1	HBM Displacement Transducer	Z
LVDT 2	HBM Displacement Transducer	X

The structural vibration is first recorded by two synchronized cameras, which means the image frames captured by the two cameras have one-to-one correspondence. The proposed algorithms as discussed in Section 3.3 are applied to detect and match natural key points for camera calibration. Figure 3.12(a) shows an example of the key points detected on and around the cantilever beam.

Figure 3.12(b) shows an example of the key point matching. Note that for viewing clarity Figure 3.12(b) only shows a small subset of the matched key points in an image pair. It is possible to detect and match many more key points than what are shown here using these algorithms. Numerically, in ScanNet indoor dataset, the precision of the proposed algorithms achieves 84.4% precision with a match score of 31%, whereas the SIFT with RANSAC only obtains 61.9% precision with a match score of 0.7% [72]. Here the precision is defined as the proportion of the correctly matched key points among all key points matched, and the match score is defined as the ratio of the number of correct matches key points over the total number of detected key points, matched or un-matched. From the matched key points, the camera parameters can be derived for subsequent tasks. Once the matched key points and camera parameters are accurately obtained, the 3D information can be recovered. The matched key points are then tracked by the KLT tracker frame by frame to obtain the displacement responses in the time domain.

(a)



(b)

Figure 3.12. The performance of the proposed system: (a) Some key points detected; (b) Corresponding points matched.

The vibration displacement responses measured by the proposed vision system against those obtained by physical sensors are shown in Figures 3.13 and 3.14. The cantilever is relatively stiff

and the excitation frequency applied in the test was small. The vibration mode of the cantilever beam was not excited, and the cantilever beam vibrated rigidly with the same displacement across the entire structure. Figure 3.13 shows the time domain 3D vibration displacement responses of an arbitrary key point on the middle of the beam structure measured by the proposed vision system, compared to those measured by various physical sensors. Figures 3.13(a), (c) and (e) show the complete time domain responses in the vibration duration of the test and Figures 3.13(b), (d) and (f) provide a clearer view by showing only part of the vibration displacements. It can be clearly observed that the vibration displacement measurements by the proposed vision approach match very well with those obtained by the physical sensors. It should be noted that the maximum displacements in the *X* and *Z* directions are only 3mm, and the maximum displacement in the *Y* direction is only around 1mm.

Table 3.2 shows the relative errors (ϵ) in the measured time domain displacement responses in three directions and correlation coefficients (ρ) between the displacement responses obtained by physical sensors and the proposed vision approach in three directions. They are defined as:

$$\epsilon = \frac{||B_i - A_i||}{||B_i||} \times 100\% \tag{55}$$

$$\rho = \frac{1}{N-1}\sum_{i=1}^{N}\left(\frac{A_i - \mu_A}{\sigma_A}\right)\left(\frac{B_i - \mu_B}{\sigma_B}\right) \tag{56}$$

where the displacement trajectory A is measured by physical sensors (ground truth) and B is measured by the vision method, N represents the total sampling number, μ stands for the mean of a set of data, and σ is the standard deviation. The maximum relative error is less than 8% and the minimum correlation coefficient is larger than 0.9958, indicating that the time domain 3D vibration displacement responses are measured accurately by the proposed approach.

Table 3.2. Displacement error analysis of the 3D vibration test

Direction	**Error (ϵ)**	**Corr (ρ)**
X	4.09	0.9969
Y	4.41	0.9980
Z	7.20	0.9958

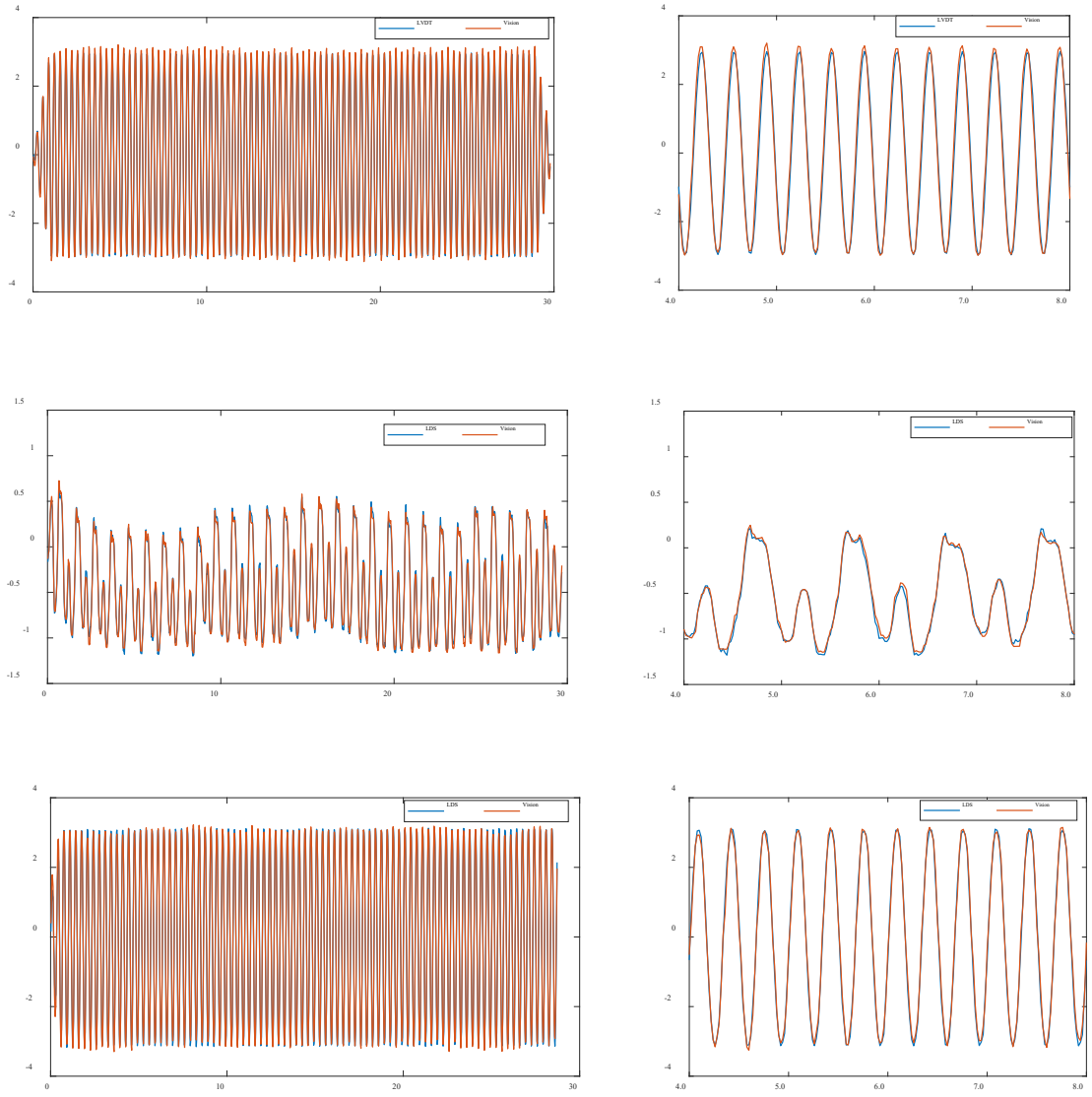


Figure 3.13. Three direction displacement comparison of the 3D vibration test: (a) Vision against LVDT 2 in X direction; (b) A zoomed-in view for (a); (c) Vision against LDS 1 in Y direction; (d) A zoomed-in view for (b); (e) Vision against LDS 2 in Z direction; (f) A zoomed-in view for (e).

Results in Figure 3.14 and Table 3.2 demonstrate the accuracy of the proposed vision system in measuring the 3D vibration displacement responses of a single key point. Since the cantilever

beam is a rigid body, the full-field measurement reliability can be checked by randomly taking a point on the cantilever, extracting the vibration displacement and comparing it with the physical sensor measurements. In Figure 3.14, the 3D displacement responses of thirty randomly chosen key points that cover the whole structure measured by the vision system are provided, in comparison with physical sensors.

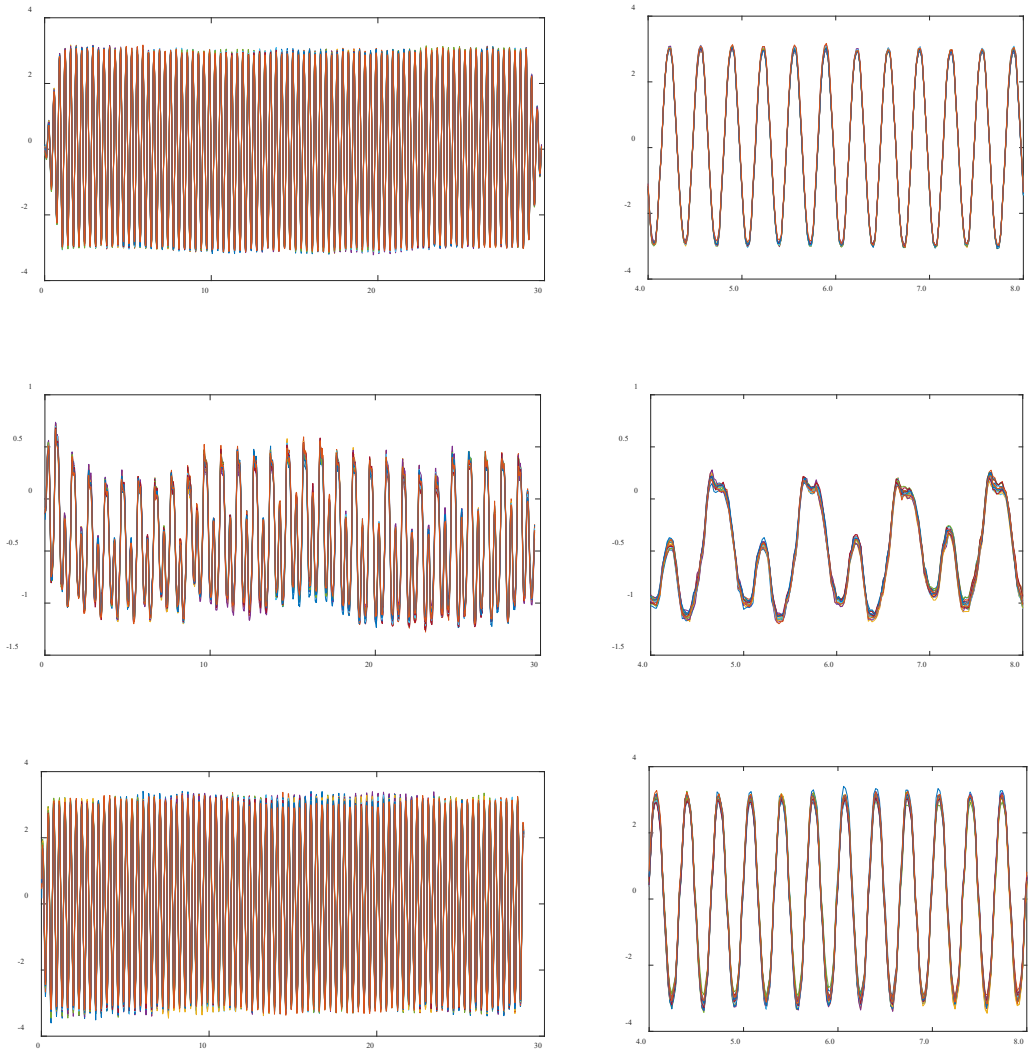


Figure 3.14. Three direction displacement comparison for thirty key points of the 3D vibration test: (a) Vision against LVDT 2 in X direction; (b) A zoomed-in view for (a); (c) Vision against LDS 1 in Y direction; (d) A zoomed-in view for (b); (e) Vision against LDS 2 in Z direction; (f) A zoomed-in view for (e).

Figure 3.14 shows the 3D displacement responses of multiple key points. Since the cantilever beam used is a rigid object, it is expected that the movements of different parts of the object are highly similar, and should match the measurement by physical sensors well. The measurement results by the proposed vision system shown in Figure 3.14 confirmed this. In fact, the trajectories of the 30 key points are so close that they can hardly be distinguished from each other in the plot. This experiment confirms that the proposed vision-based approach is able to provide full-field 3D displacement measurement for a structure. Tables 3.3 and 3.4 provide a numerical evaluation of the measurement in all 30 points, in terms of relative error and cross-correlation coefficients. The displacement responses measured by physical sensors are used as the ground truth which is denoted as G in the table, whose relative error is set as 0.00% and the correlation coefficient as 1.0000. The other rows in the two tables show the performance of the displacement measurement using the proposed vision method at every one of the 30 points. As shown in Table 3.3 and 3.4, the displacement measurement at all 30 key points have very high similarities with cross-correlation coefficients consistently over 0.99 in 3 directions, and the relative errors within 6% in the X, Y, and 10% in the Z direction.

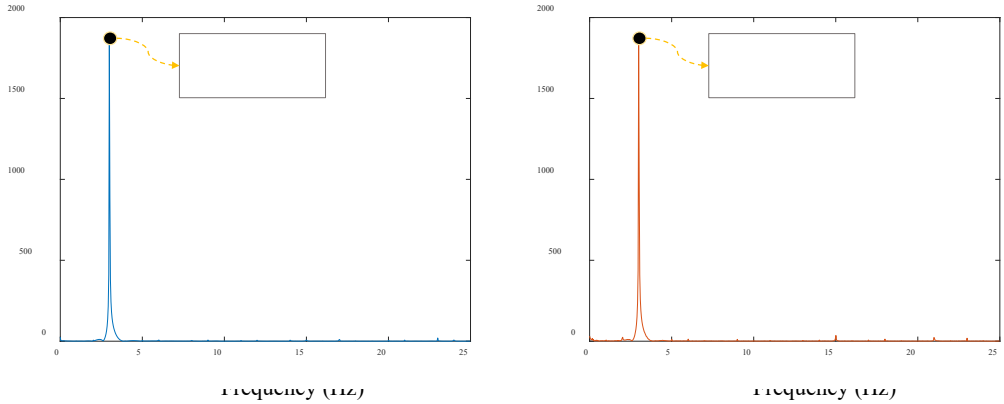
Table 3.3. Displacement relative errors of thirty key points

No.	Error (ϵ)			**No.**	Error (ϵ)			**No.**	Error (ϵ)		
	X	**Y**	**Z**		**X**	**Y**	**Z**		**X**	**Y**	**Z**
G	0.00	0.00	0.00	**11**	4.14	5.47	8.93	**22**	4.06	4.39	7.49
1	4.09	4.41	7.20	**12**	4.52	4.47	7.62	**23**	5.38	5.43	8.24
2	5.23	5.14	8.65	**13**	5.33	5.93	8.49	**24**	4.63	4.23	7.54
3	4.86	5.24	7.64	**14**	5.40	6.45	9.63	**25**	5.45	4.47	8.27
4	5.35	6.43	9.72	**15**	4.92	4.93	7.32	**26**	4.92	4.03	7.54
5	4.87	6.63	9.48	**16**	5.34	6.04	9.74	**27**	5.04	4.27	7.78
6	4.34	5.35	8.87	**17**	5.29	6.73	9.72	**28**	4.29	4.93	7.26
7	5.26	6.64	9.22	**18**	4.23	6.23	9.01	**29**	4.14	5.04	8.62
8	4.93	4.04	7.98	**19**	4.76	5.65	8.42	**30**	4.72	4.30	7.74
9	4.23	4.63	7.68	**20**	5.97	6.03	9.74	**-**	-	-	-
10	5.27	5.12	8.06	**21**	4.65	5.63	8.37	**-**	-	-	-

Table 3.4. Displacement correlation coefficients of thirty key points

No.	Corr. (ρ)			No.	Corr. (ρ)			No.	Corr. (ρ)		
	X	Y	Z		X	Y	Z		X	Y	Z
G	1.0000	1.0000	1.0000	**11**	0.9963	0.9974	0.9947	**22**	0.9963	0.9982	0.9957
1	0.9969	0.9980	0.9958	**12**	0.9956	0.9978	0.9953	**23**	0.9956	0.9980	0.9943
2	0.9959	0.9973	0.9943	**13**	0.9958	0.9982	0.9948	**24**	0.9978	0.9961	0.9952
3	0.9968	0.9972	0.9955	**14**	0.9962	0.9945	0.9933	**25**	0.9965	0.9969	0.9949
4	0.9957	0.9946	0.9935	**15**	0.9976	0.9965	0.9973	**26**	0.9963	0.9963	0.9953
5	0.9964	0.9945	0.9962	**16**	0.9954	0.9984	0.9957	**27**	0.9959	0.9963	0.9952
6	0.9962	0.9973	0.9948	**17**	0.9967	0.9963	0.9973	**28**	0.9968	0.9964	0.9958
7	0.9983	0.9964	0.9955	**18**	0.9987	0.9986	0.9932	**29**	0.9967	0.9958	0.9949
8	0.9943	0.9974	0.9975	**19**	0.9956	0.9973	0.9955	**30**	0.9963	0.9986	0.9954
19	0.9974	0.9963	0.9954	**20**	0.9956	0.9982	0.9957	**-**	-	-	-
10	0.9964	0.9953	0.9963	**21**	0.9976	0.9935	0.9943	**-**	-	-	-

The displacement responses measured from the presented binocular vision system and the physical displacement sensors are transferred to the frequency domain by Fast Fourier Transformation (FFT) to reveal the vibration frequencies of the structure. The results are shown in Figure 3.15. Table 3.5 shows the relative errors of the frequencies and Fourier spectrums. It is clearly observed that the displacement responses generated by the proposed vision system are able to obtain the vibration frequencies of the structure in three directions accurately, with the same accuracy and performance to measure vibration frequencies by the physical displacement sensors.



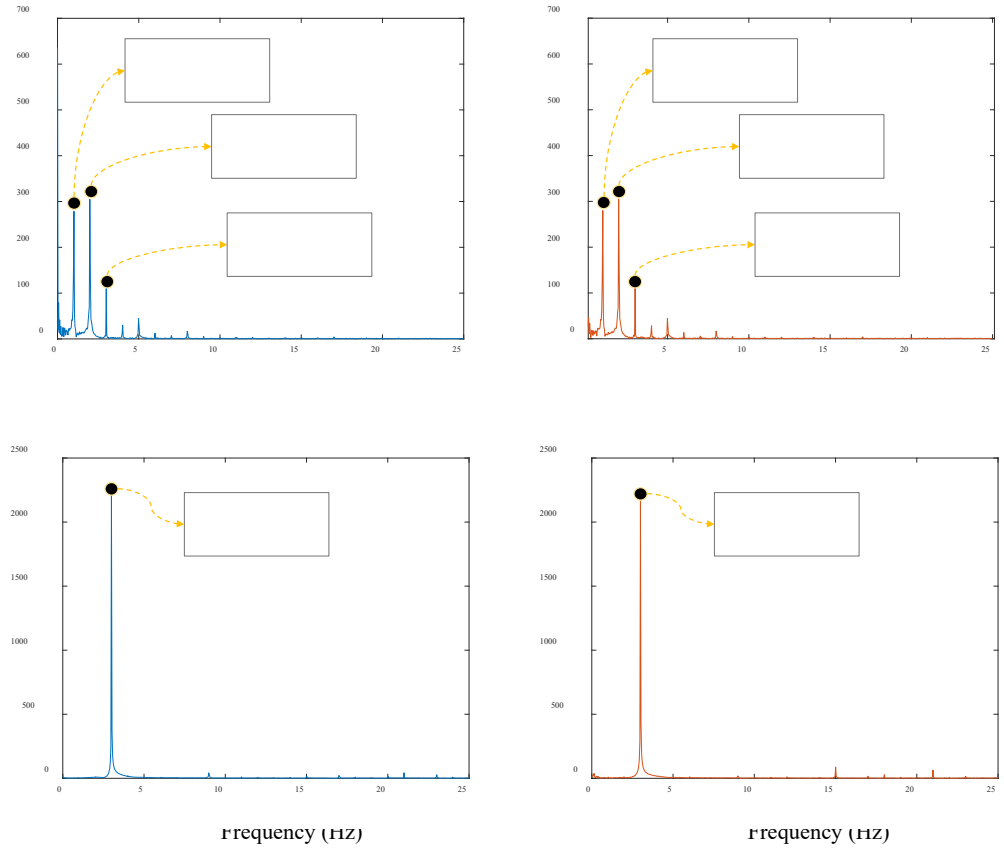


Figure 3.15. Vibration frequencies obtained from the displacement measurement: (a) in X direction obtained from LVDT 2; (b) in X direction obtained from the vision system; (c) in Y direction obtained from LDS 1; (d) in Y direction obtained from the vision system; (e) in Z direction obtained from LDS 2; (f) in Z direction obtained from the vision system.

Table 3.5 gives the numerical results of the obtained vibration frequencies comparing with those from physical sensors in three directions. The measured frequencies in X, Y and Z directions match perfectly with those from the physical sensors, while the measurement in amplitudes of Fourier Spectrum has relative errors less than 1.87%.

Table 3.5. The relative errors of vibration frequencies and Fourier spectrum

Direction	Peak	Relative error of Frequency (%)	Relative error of Fourier Spectrum (%)
X	First Peak	0.00	0.23
Y	First Peak	0.00	0.46
	Second Peak	0.00	0.16
	Third Peak	0.00	1.87
Z	First Peak	0.00	0.13

3.5. Summary

This chapter introduced a computer vision-based full-field 3D vibration displacement measurement approach without artificial targets. A binocular vision system with two video cameras is developed to measure the in-plane and out-of-plane vibration displacement responses of the target structure. Compared with the previous vision-based 3D displacement measurement methods, the proposed approach introduces state-of-art key point detection and key point matching algorithms based on deep learning to achieve highly accurate target-free measurement. Experimental results demonstrate that the proposed target-free vision system can accurately measure the vibration displacement responses and obtain the natural frequencies of structural vibrations in 3D.

In in-field measurements, the cameras are often located far away from the inspected structure. This could result in very tiny displacements recorded in the video, sometimes even invisible. There are also situations where the structure vibrations are indeed very tiny, but have significant impact in structural health monitoring. No matter whether displacement sensors or computer vision methods are used, accurate measurement of 3D tiny vibrations remains a very difficult task. In the next chapter, a vision-based tiny 3D displacement measurement system will be presented.

CHAPTER 4

3D TINY STRUCTURAL VIBRATION MEASUREMENT

4.1 Introduction

In this chapter, a deep learning-based binocular vision system for target-free full-field 3D tiny vibration measurement of civil engineering structures is introduced. A phase-based video motion magnification algorithm [85] is employed to achieve a high measurement accuracy of tiny vibrations at submillimeter level. This motion magnification method allows tiny movements in videos to be manipulated by analyzing the local phase at different orientations and scales. This processing does not involve the computation of optical flow, and it supports larger amplification factors and is significantly less sensitive to noise comparing to other methods [35, 36]. The advanced key point detection and matching algorithms based on deep learning techniques discussed in Chapter 3 are employed to achieve target-free displacement measurement. The accuracy and performance of the proposed approach are first evaluated through experimental tests on a steel cantilever beam in the laboratory. In-field experimental tests are then conducted on a pedestrian bridge on a university campus to further evaluate the performance of the proposed approach in practical applications. The tiny vibration measurements obtained from the proposed approach are compared with those measured by LVDTs and/or LDS, and the derived acceleration responses are compared with those measured from the installed accelerometers on the testing structures. The results demonstrate that 3D tiny vibration that are almost impossible to measure with existing non-physical sensor-based methods, are obtained with satisfactory accuracy by the proposed approach. The flowchart of the proposed system is shown in Figure 4.1.

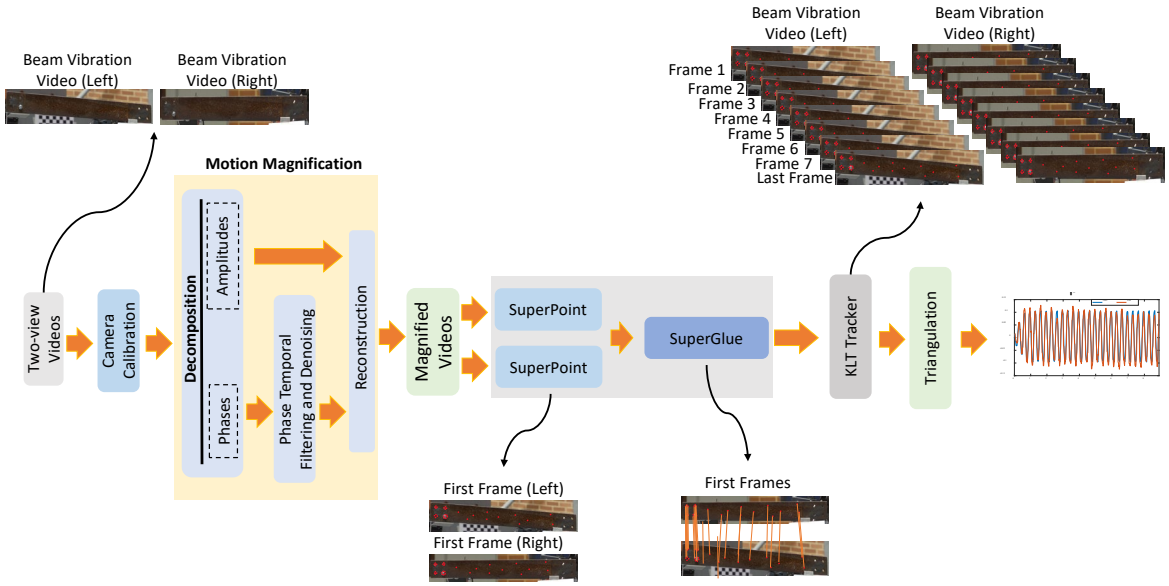


Figure 4.1. The flowchart of deep learning assisted tiny 3D structural vibration measurement method

4.2 Motion Magnification

Vibration displacement responses of civil engineering structures under some loading conditions, such as ambient conditions, could be very small, often not visible to human eyes. Under certain circumstances, measuring very small vibration displacement responses of civil engineering structures is necessary for SHM, for example, obtaining the small vibration displacement of long-span bridges and monitoring the settlement of bridge piers. For the in-field measurement tasks, it is impossible to guarantee that the camera setup can be very close to the structure. In those cases, the structural vibration in the video will be very tiny or even invisible to human eyes and it is possible that vision-based methods for measuring vibration displacement response would fail. It is hence necessary to perform motion magnification on the original video to enlarge the motion in the video so that the tiny displacement becomes visible for vision-based methods to detect, match and track key points and subsequently obtain the vibration displacement. In this study, a phase-based video motion magnification algorithm [85] is employed to magnify the motion in the vibration video and consequently enable the measurement of tiny displacement responses.

The local phase shift approximating the local motion has been well investigated in phase-based optical flow [86, 87] and it is a more robust representation than amplitude for motion field.

The phase-based motion magnification algorithm assumes that the motion is composed of multiple sinusoidal waves. The motions of pixels can hence be adjusted by changing the phase of these sinusoidal waves. However, by changing the global phase, the signal can only be processed globally, although the local movements in the video are often of more interest. Complex steerable pyramids [88, 89] are applied to decompose each image into different scales and directions. The phases at each location, orientation and scale are temporally extracted and amplified. The meaningless phase in the low amplitude area is blurred by an amplitude-weighted spatial Gaussian filter to improve the quality of the magnified videos. The amplified signals can then be reconstructed to form a motion-magnified video. The flowchart of the motion magnification technique is shown in Figure 4.2.

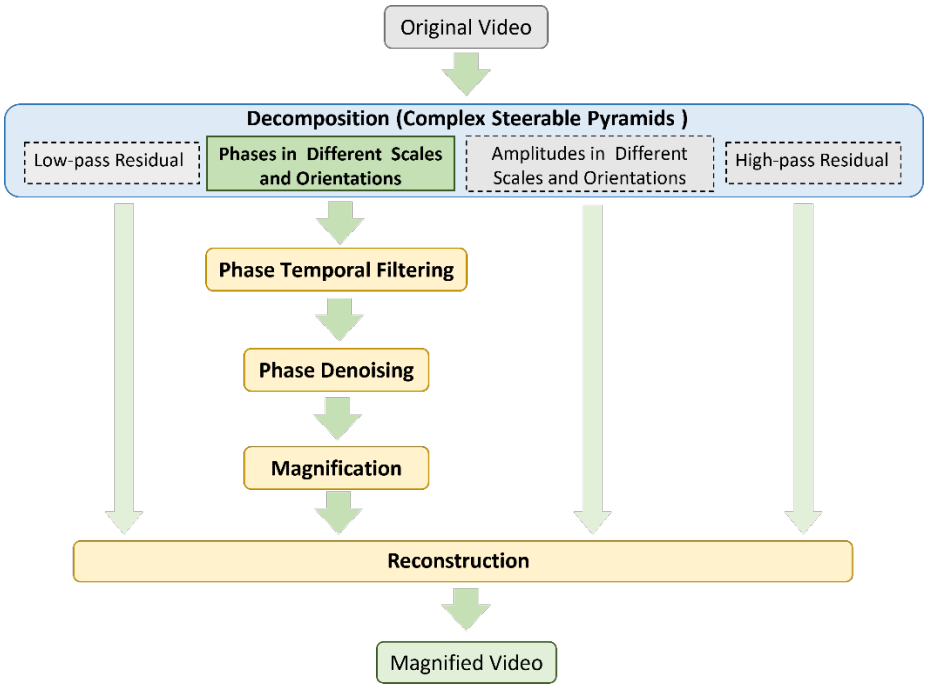


Figure 4.2. The flowchart of the phase-based motion magnification

An example for magnifying a one-dimensional (1D) translation movement is described below, in which the motion is magnified using the phase of global Fourier basis coefficients. A 1D image intensity is denoted by $I(u,t)$ at location u and time t. The image is then translated by a tiny motion function $\delta(t)$ to location $u-\delta(t)$ at time t, which can be express by $I(u,t) = f(u -$

$\delta(t)$). The image motion profile $f(u)$ at time 0 and $f(u-\delta(t))$ at time t can be written as a sum of complex coefficients times sinusoids corresponding to frequency ω by Fourier series decomposition

$$f(u) = \sum_{\omega} A_{\omega} e^{i\Phi_{\omega}} e^{-i\omega u} \tag{57}$$

$$f\big(u-\delta(t)\big) = \sum_{\omega} A_{\omega} e^{i\Phi_{\omega}} e^{-i\omega\left(u-\delta(t)\right)} = \sum_{\omega} A_{\omega} e^{i(\Phi_{\omega}+\omega\delta(t))} e^{-i\omega u}. \tag{58}$$

The phase difference $\Delta\Phi_{\omega}$ between time 0 and time t is

$$\Delta\Phi_{\omega} = \big(\Phi_{\omega} + \omega\delta(t)\big) - \Phi_{\omega} = \omega\delta(t). \tag{59}$$

Use a magnification factor φ to enlarge the phase and the image intensity $I(u,t)$ become

$$I(u,t) = f\big(u-(1+\varphi)\delta(t)\big) = \sum_{\omega} A_{\omega} e^{i\Phi_{\omega}+(1+\varphi)\omega\delta(t)} e^{-i\omega u} \tag{60}$$

It is clear that the global translation motion from time 0 to time t is amplified. Nevertheless, in most of the real scenarios, the motion in the video is local movement rather than global movement. In such cases, the Fourier transform cannot break the image into a representation consisting of exact sinusoids.

4.2.1 Complex Steerable Pyramid

To extract the local phase, the complex steerable pyramid is applied to decompose the images. The complex steerable pyramid decomposes the images with multi-scale, multi-oriented and aliasing-free subbands [88, 89]. The basis function of complex steerable pyramid is shown in Figure 4.3. Each basis function is complex therefore has a real part and an imaginary part. The

waveform is that a Gaussian envelop is covered on the sinusoid. The basis function decomposes the image to local amplitude and phase instead of the global decomposition of Fourier transformation.

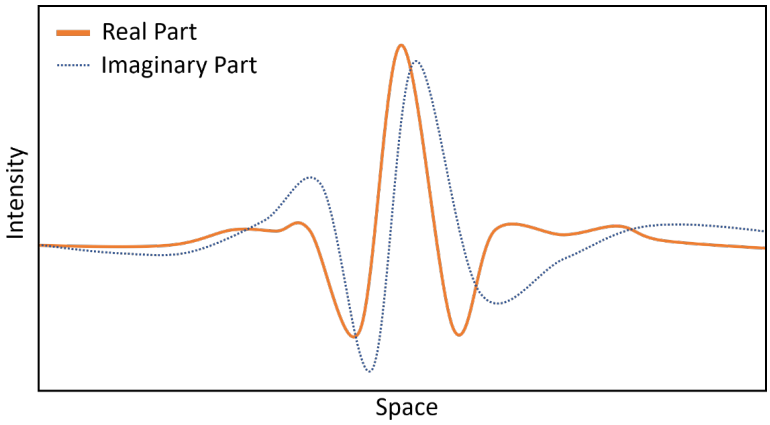


Figure 4.3. The basis function of complex steerable pyramid

The basis function of the complex steerable pyramid is convoluted with a 2D image with intensity $I(u, v, t)$ for extracting the local amplitudes and phases at each video frame. A set of transformed images in different scales and orientations can be expressed as

$$I_{\theta,\tau}(u, v, t) = I(u, v, t) * G_{\theta,\tau} = A_{\theta,\tau} e^{i\Phi_{\theta,\tau}(u,v,t)} . \tag{61}$$

where $G_{\theta,\tau}$ is the basis function corresponding to scale τ and orientation θ. Subtracting the phase at time 0, the phase shift between time 0 and time t can be obtained as

$$\Delta\Phi_{\theta,\tau}(u, v, t) = \Phi_{\theta,\tau}(u, v, t) - \Phi_{\theta,\tau}(u, v, 0). \tag{62}$$

The phase shift is then multiplied by a magnification factor φ to generate a new set of transformation $I_{A_{\theta,\tau}}(u, v, t)$ of the image sequence, in which the amplitudes are the same, but the phase shifts are magnified. The new video can be rebuilt by multiplying the transformed images by basis function and summing the scales and orientations to generate the video that the tiny motions are visualized

$$\sum_{\theta,\tau} I_{A_{\theta,\tau}}(u,v,t) * G_{\theta,\tau} = I_A(u,v,t), \tag{63}$$

The basis function introduced by [89] has 4-orientation, octave-bandwidth, which support real-time processing, but only allows the small motion-magnification. A 8-orientation half-octave pyramid that has two periods sinusoid under the Gaussian envelope supports larger amplification, but the processing time is longer [85]. A 1D half-octave basis function is shown in Figure 4.4. In our study, the half-octave pyramid is applied, since the vibration of civil engineering structures is normally tiny.

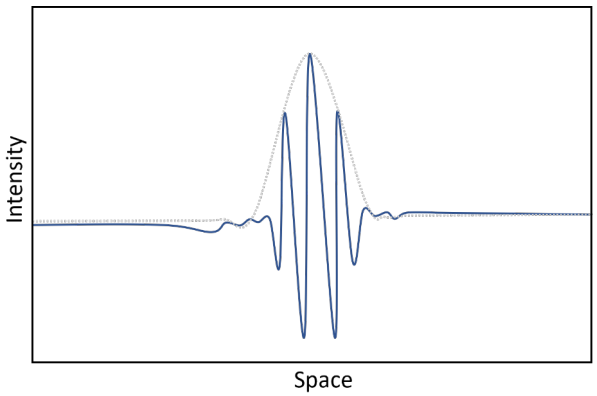


Figure 4.4. The waveform of 1D half-octave basis function

4.2.2 Temporal Filtering and Denoising

To manipulate the real signals, the signal-to-noise ratio (SNR) is a measure to compare the level of signal and noise. The performance of the output video can be improved by maximizing the SNR of the local phase changes. The components related to the noise are removed by temporally and spatially filtering, and the desired signal is preserved. Since different motions are occurred at different frequencies, the signals are also limited in a certain frequency range of interest by temporally filtering the signals. Simply applying narrowband linear filters [85], SNRs for motions occurring in a certain frequency range can be improved.

Although phase-based motion magnification has inherent noise characteristics, the noises in the input videos can also result in the noisy output magnified video that the noise signals are

amplified much more than deserved signals. Spatially Lowpassing the phase signal is an easy and efficient way for increasing the SNR of the input video [85]. For removing the meaningless phase-signals in low amplitude areas, an amplitude-weighted spatial Gaussian is applied to blur on the phases:

$$\frac{\big((\Delta\emptyset)A\big) * J_\rho}{A * J_\rho}. \tag{64}$$

where $J_\rho = e^{(-\frac{x^2+y^2}{2\rho^2})}$ is a Gaussian filter and the indices of amplitude A and phase $\emptyset$ have been suppressed for readability. The parameter ρ is equal to the widths of the spatial domain filter which is two pixels in our project. The results without this noise handling process can be avoided, because this step increases the computing speed, and the quality of the magnified videos is usually good without it.

4.2.3 Magnifying the Local Motion

A single basis function similar to the global phase-shift theorem of Fourier basis functions is demonstrated below to show the local phase shift approximates local translation. A basis function is modelled as a Gaussian window multiplying with a complex sinusoid

$$e^{(-\frac{x^2}{2\sigma^2})}e^{-i\omega x} \tag{65}$$

where σ is the standard deviation of the Gaussian function and ω is the frequency of the complex sinusoid. Due to the self-similar property of the basis function in the complex steerable pyramid, the ratio between σ and ω is fixed, which means the lower the frequency the higher the window. Multiplying the phase of the basis function by a complex coefficient $e^{-i\emptyset}$ results in

$$e^{(-\frac{x^2}{2\sigma^2})}e^{-i\omega x} \times e^{-i\emptyset} = e^{(-\frac{x^2}{2\sigma^2})}e^{-i\omega(x-\emptyset/\omega)} \tag{66}$$

The complex sinusoid under the window is translated, which is approximately a translation of the whole basis function by $\frac{\emptyset}{\omega}$. Conversely, the phase difference between two translated basis functions is proportional to translation. Specifically, supposing we have a basis element and its translation by δ, that is

$$e^{(-\frac{x^2}{2\sigma^2})}e^{-i\omega x}, e^{-\frac{(x-\delta)^2}{(2\sigma^2)}}e^{-i\omega(x-\delta)} \tag{67}$$

The local phase of each element only depends on the argument to the complex exponential and is $-\omega x$ in the first case and $-\omega(x-\delta)$ in the second. The phase difference is then $\omega\delta$ which is directly proportional to the translation. Local phase shift can be used both to analyze tiny translations and synthesize larger ones.

As same as the global translation example, the differences of the local phases extracted by the half-octave pyramid are multiplied with an amplification factor φ at each level of the steerable pyramid. These amplified phase differences are finally used to modify the phase of the pyramid to amplify the motion in the sequence. The magnified complex steerable pyramid is finally collapsed to synthesize the output motion magnified video. A 1D displacement example is shown in Figure 4.5. The local phase of complex steerable pyramid coefficients is used to amplify the motion of a translation movement.

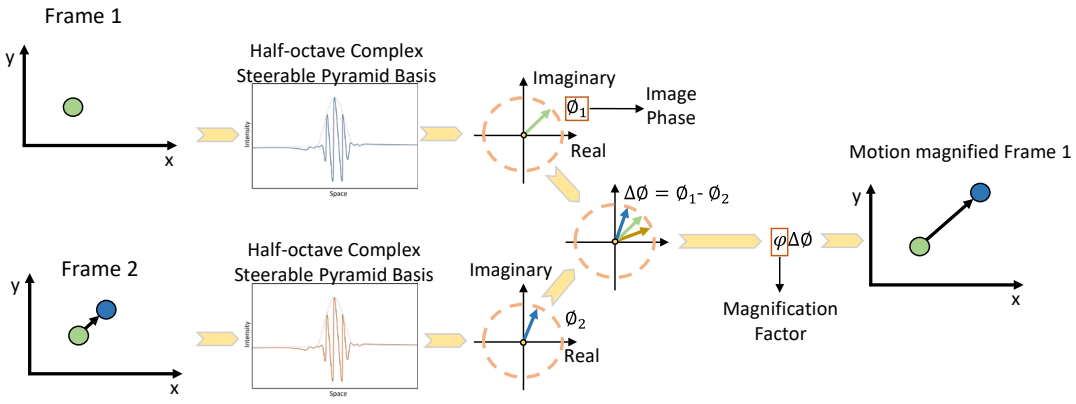


Figure 4.5. The flowchart of the phase-based motion magnification

Two frames from a video of a subtly translation is transformed to the complex steerable pyramid representation by projecting onto basis functions. The phase between the complex coefficients is computed and amplified. In Figure 4.5, only the coefficient corresponding to exactly one location and scale is shown. This processing is actually done on every pyramid coefficient. The new coefficients are used to shift the basis functions and a reconstructed video is produced in which the translation between the two frames is enlarged.

4.2.4 The Limitation of Magnification

Due to the limitation of the spatial support of the complex steerable pyramid and the essence of the motion magnification approach is shifting the image phase covered by the basis function, the phase will finally reach the border that it cannot be moved. Once the magnification factor is beyond this border, the approach cannot amplify the tiny motion according to the factor assigned and the images are blurred by noises. Figure 4.6 shows a frame from our cantilever beam vibration test, in which Figure 4.6(a) is the image without noise, in the contrast, Figure 4.6(b) is full of noise since the applied magnification factor is too large.

(a)

(b)

Figure 4.6. A comparison of motion magnification with magnification factors inside and beyond the border. (a) The magnification factor inside the border; (b) The magnification factor beyond the border.

In such a case, the displacement is impossible to be accurately extracted from the noisy video. It is important to ensure the factor chosen is always inside the border to avoid measurement errors. According to Ref.[85], the standard deviation of the Gaussian window is used as the border, which is described as

$$\varphi\delta(t) < \sigma\,. \tag{68}$$

where σ is the standard deviation of the half-octave handpass filter and φ is the magnification factor. Exceeding the limitation above can result in flaws or blur since some image pyramid components are not present in their proper ratios to reconstruct the desired motion.

4.5 Experimental Validations

4.5.1 3D Vibration Tests of a Beam Structure

4.5.1.1 Experiment Setup

Once the 3D tiny motions of civil structures magnified using the system described in the sections above, the vision-based target free full-filed displacement measurement system

introduced in Chapter 3 can be employed for their measurement. The performance of the 3D tiny vibration displacement measurement is first evaluated by conducting 3D vibration tests. The shaking table setup is the same as that described in Section 3.4.1. Two LDSs are installed for displacement measurement on Z direction to provide more accurate ground truth. The detail of the sensors is listed in Table 4.1.

Table 4.1. Displacement sensors used in the 3D vibration test

Sensor Name	Version	Measurement Direction
LDS 1	Keyence IL300	Z
LDS 2	Keyence IL300	Z
LVDT 1	HBM Displacement Transducer	Y
LVDT 2	HBM Displacement Transducer	X

4.5.1.2 Results and Discussions

The purpose of this test is to evaluate the performance of the proposed 3D vision system on measuring tiny vibration displacement. In this test, Shaker 1 is set to generate sinusoidal excitations with an amplitude of 0.1 mm and a frequency of 3 Hz in two directions (*X* and *Z*), and Shaker 2 is to generate a vertical vibration with the displacement within the range of 0.1mm.

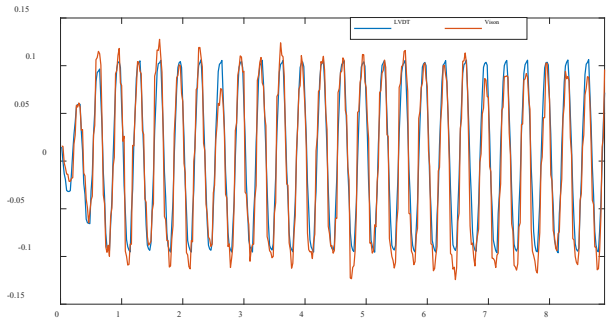
The vibration displacement measured by the proposed vision system without and with applying the motion magnification, compared with those obtained by physical displacement sensors are shown in Figures 4.7 and 4.8. The time-domain 3D vibration displacement responses of an arbitrary key point in the middle of the beam structure obtained from the original videos are shown in Figure 4.7. Figure 4.8 shows the 3D vibration displacement responses of the same point obtained from the magnified videos. Figures 4.7 and 4.8(a), (c) and (e) show the whole time series of the displacement vibration response of the complete test, and Figures 4.7 and 4.8(b), (d) and (f) provide a clearer view by zooming into a certain range of the time series. It can be observed that the displacement responses are generally tiny (on the scale of 0.1 mm with each pixel represents 0.3038mm in videos recorded by Camera 1 and 0.3213mm in those recorded by Camera 2). When the videos are not magnified, the accuracy of the measurement results from the vision-based method is unsatisfactory when compared with the physical sensors. When the phase motion magnification is applied and the magnified videos is used, the proposed vision-based approach is

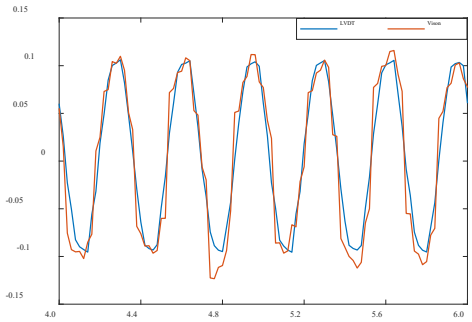
able to obtain vibration displacement measurements very similar to those obtained by the physical displacement sensors, as shown in Figure 4.8.

Table 4.2 shows the numerical analysis of the proposed methods. For video images without magnification, the relative errors are 30.29%, 55.14%, and 80.02% in X, Y, and *Z* directions, respectively, and the correlation coefficients are 0.9578, 0.8160, and 0.6001, respectively. These indicate a large error in displacement measurement. For the magnified video, the relative errors in the in-plane directions, namely *X* and *Y* directions, are less than 13%, and the relative error in the out-plane *Z* direction is about 37%. The correlation coefficient for the displacement in the *Z* direction is 0.9419, demonstrating much more accurate tiny displacement measurements than those without using motion magnification. Given that the vibration displacements in all three directions are in the magnitude of 0.1mm, which is extremely difficult to measure using the vision-based methods, the obtained results with motion magnification are highly satisfactory with the minimum correlation coefficient of 0.94.

Table 4.2. Displacement error analysis of tiny vibration test

Direction	**Original Videos**		**Magnified Videos**	
	Error (ϵ)	**Corr (ρ)**	**Error (ϵ)**	**Corr (ρ)**
X	30.29%	0.9578	9.02%	0.9949
Y	55.14%	0.8160	12.67%	0.9763
Z	80.02%	0.6001	37.64%	0.9419





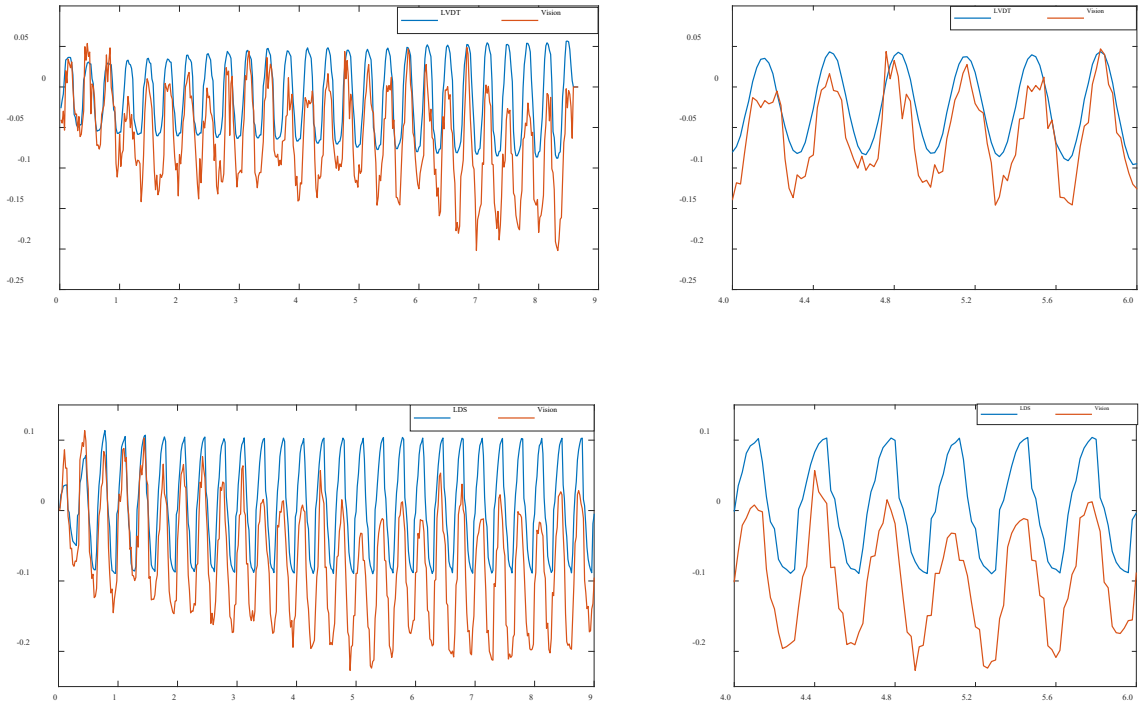
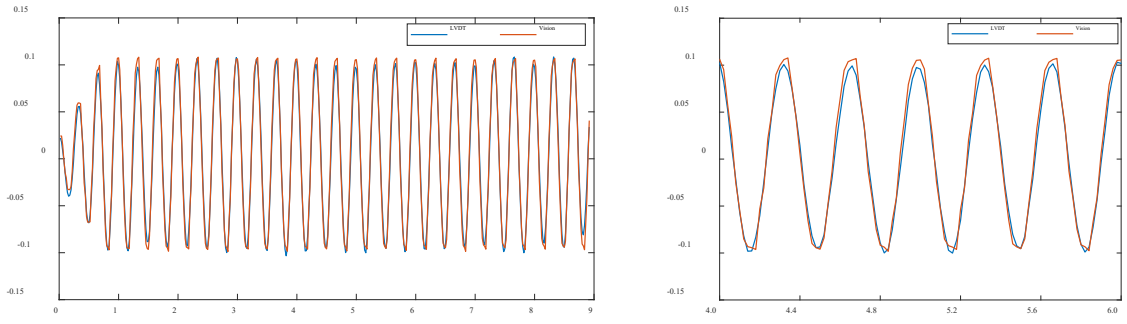


Figure 4.7. 3D tiny displacement measurement (original videos): (a) Vision vs. LVDT 2 in *X* direction; (b) A zoomed-in view of (a); (c) Vision vs. LVDT 1 in *Y* direction; (d) A zoomed-in view of (c); (e) Vision vs. LDS 2 in *Z* direction; (f) A zoomed-in view of (e).



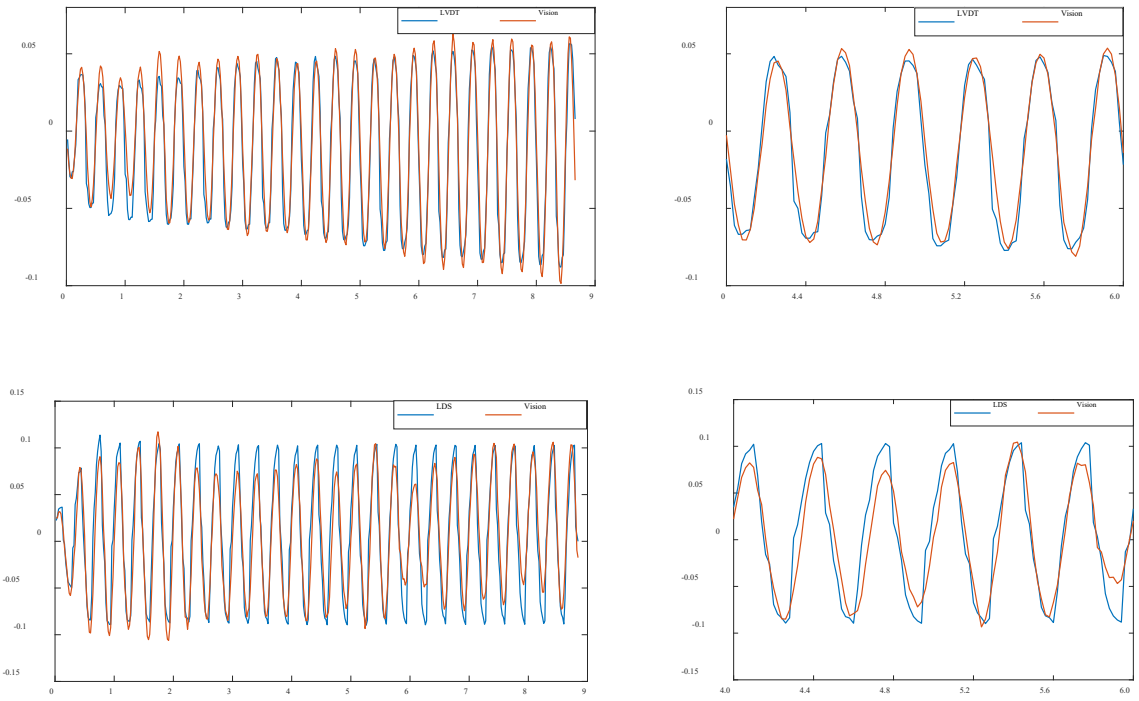


Figure 4.8. 3D tiny displacement measurement (magnified videos): (a) Vision vs. LVDT 2 in *X* direction; (b) A zoomed-in view of (a); (c) Vision vs. LVDT 1 in *Y* direction; (d) A zoomed-in view of (c); (e) Vision vs. LDS 2 in *Z* direction; (f) A zoomed-in view of (e).

Fast Fourier Transformations (FFT) are performed to transfer the vibration responses measured by the binocular vision system with motion magnification and by the physical displacement sensors to the frequency domain in order to reveal the vibration frequencies of the structure. The results are shown in Figure 4.9. Although the 3D displacements presented in this test are very tiny, the proposed vision system can obtain a very good measurement of the displacements and subsequently vibration frequencies matched very well with those obtained by the physical sensors.

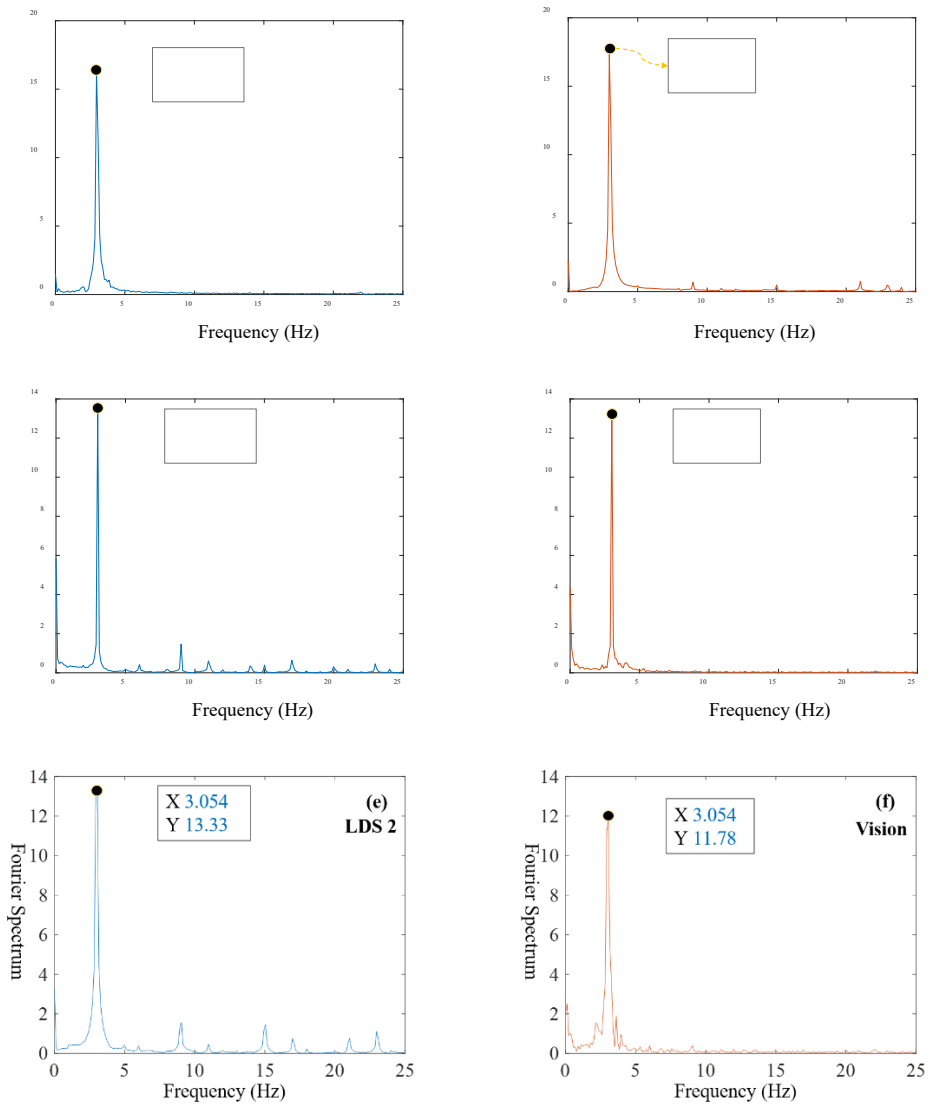


Figure 4.9. FFT spectrums of measured 3D vibrations in test: (a) *X* direction obtained from LVDT 2; (b) *X* direction obtained from vision method; (c) *Y* direction obtained from LVDT 1; (d) *Y* direction obtained from vision method; (e) *Z* direction obtained from LDS 2; (f) *Z* direction obtained from vision method.

Full-field displacement measurement capability of the proposed vision system for tiny vibration is again verified by randomly selecting multiple points all over the structure to check the measurement of their motions compared with physical sensors. If the vibration trajectories of the key points are highly similar, the full-field displacement measurement ability is validated. In Figure 4.10, the 3D displacement responses of thirty randomly chosen points that cover the whole

structure measured by the vision system are shown, in comparison with those measured by the physical sensors. It can be clearly seen that the obtained vibration displacement trajectories of the 30 key points are very close and can match accurately with those measured by the physical sensors, even when the vibrations in the three directions are all very tiny, with the magnitudes at 0.1mm in the *X* and *Z* directions, and 0.05 mm in the *Y* direction.

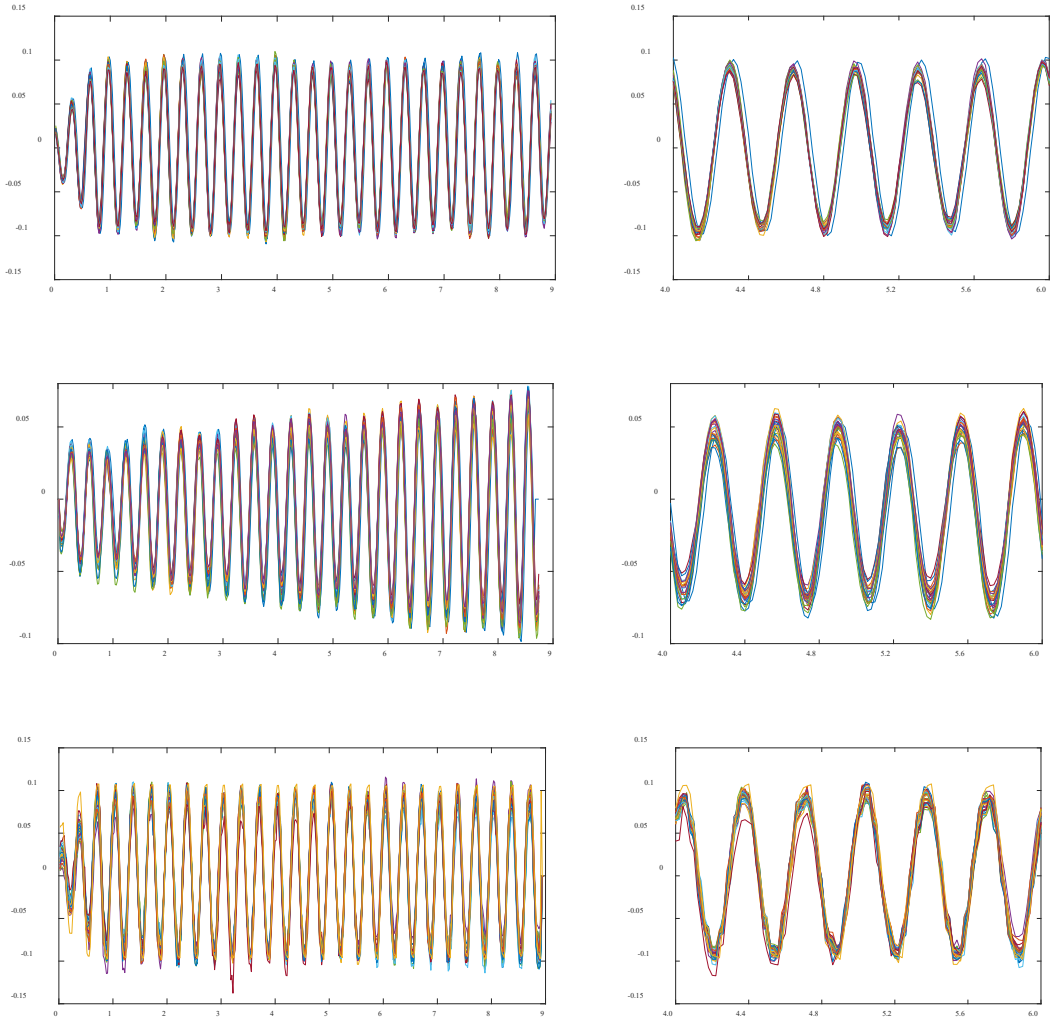


Figure 4.10. The similarity of 3D vibration trajectories of 30 key points in tiny movement test: (a) Vision approach vs. LVDT 2 in *X* direction; (b) A zoomed-in view of (a); (c) Vision approach vs. LVDT 1 in *Y* direction (d); A zoomed-in view of (c); (e) Vision approach vs. LDS 2 in *Z* direction; (f) A zoomed-in view of (e).

4.5.2 In-field Test on an Indoor Pedestrian Bridge

4.5.2.1 Experiment Setup

In order to verify the effectiveness of the proposed approach in measuring tiny vibration displacement responses for real world scenarios, vibration tests on an indoor pedestrian bridge were conducted, as shown in Figure 4.11(a). Since it is not feasible to install displacement sensors inside a building, six accelerometers were installed instead to measure acceleration responses in three dimensions at two locations, as shown in Figure 4.11(c). The two cameras as described in 3.4.1 were placed 6.5 meters away from the pedestrian bridge to capture its vibration, as shown in Figure 4.11(b). The details of the installed accelerometers are provided in Table 4.3.

Multiple shakers were initially intended to be used to vibrate the pedestrian bridge in three directions. However, due to safety concerns, only one APS 400 ELECTRO-SEIS long stroke shaker was eventually installed, generating small excitations in the vertical direction on the deck of the pedestrian bridge. Driven by the shaker, the pedestrian bridge produces tiny vibrations mainly in the vertical direction. Motion amplification on the original video is therefore necessary to enlarge the vibrations, which are almost invisible in the original videos. The proposed deep learning assisted target-free vision-based approach is used to analyse the magnified videos to perform the key point detection, matching and tracking, and subsequently obtain the 3D vibration displacement responses.

Table 4.3. Sensors installed on the indoor pedestrian bridge.

Sensor Name	Version	Measurement Direction
Accelerometer 1	PCB-393B04	Z
Accelerometer 2	PCB-393B04	Y
Accelerometer 3	PCB-393B04	X
Accelerometer 4	PCB-393B04	Z
Accelerometer 5	PCB-393B04	Y
Accelerometer 6	PCB-393B04	X

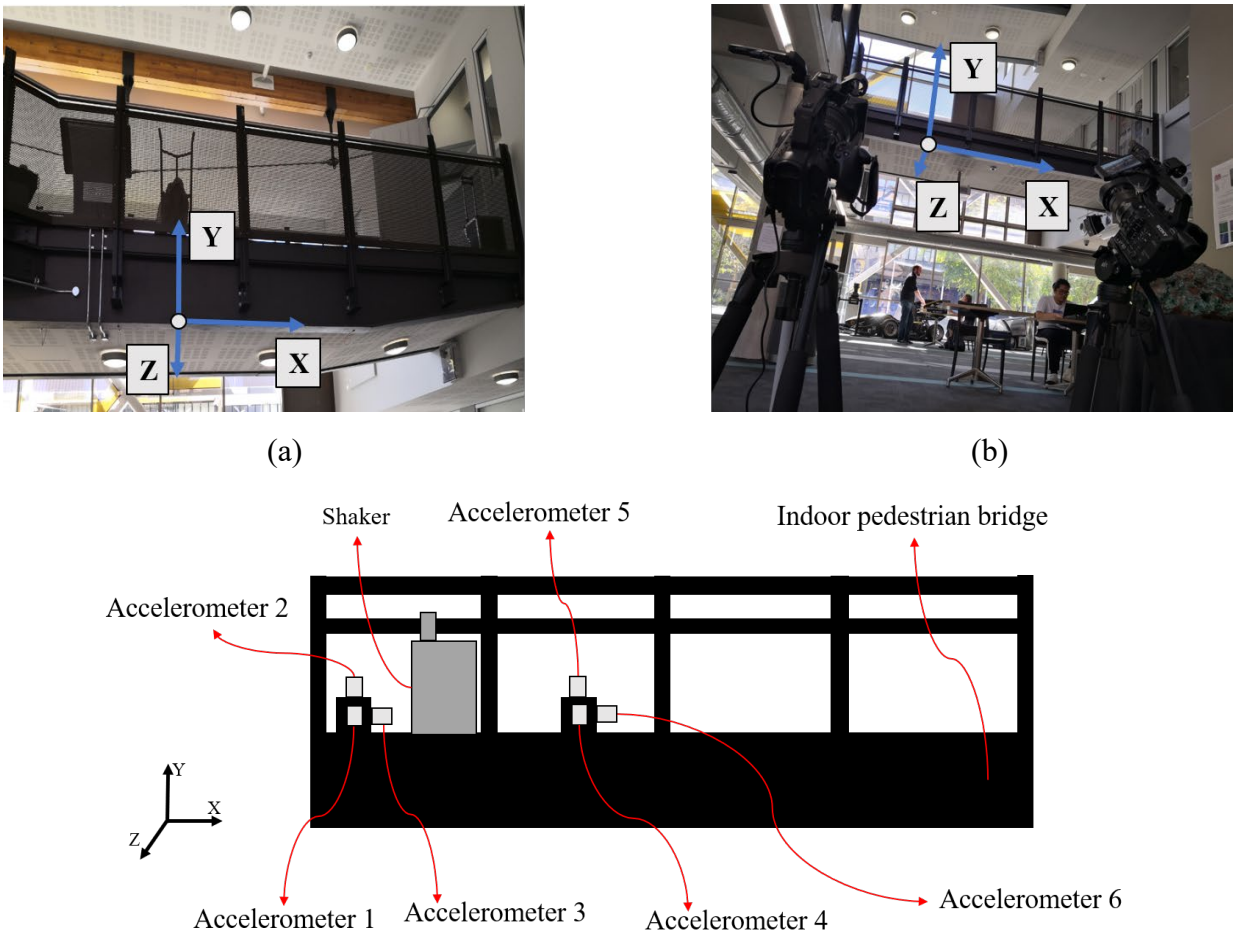


Figure 4.11. Experimental setup for indoor pedestrian bridge vibration tests: (a) The pedestrian bridge; (b) the Setup of two cameras; (c) Installed accelerometers and shaker.

4.5.2.2 Results and Discussions

Since only accelerometers are installed for this vibration test, displacement responses are not directly measured by the physical sensors. Vibration displacement responses are only obtained from the proposed vision measurement system. To validate the accuracy of the displacement measured and compare them with the acceleration responses measured by the accelerometers, acceleration responses are obtained by taking the second derivative of the displacement responses measured by the vision system. Since the sampling rate of the camera is 50 fps, and the unit of accelerometer data is m/s^2, the measured displacement is processed using the following formulas:

$$V(t) = 50 \times (y(t+1) - y(t)) \tag{67}$$

$$A(t) = 2500 \times (y(t+1) - y(t-1)) \tag{70}$$

where $y(t)$ is the displacement response at time instant t, $V(t)$ is the velocity response and $A(t)$ is the acceleration response.

Since the displacement in the *Y* direction is the focus in this test for accuracy validation, key points near Accelerometer 2 and Accelerometer 5, which are installed to measure the accelerations in the *Y* direction, are chosen to be evaluated. The vibration displacement responses of the key points near the accelerometers are defined by the average displacement of all the key points in the location of the installed sensor. Figure 4.12 shows the vertical displacement near Accelerometer 2 obtained by the proposed vision-based method, in which the displacement is less than 0.02 mm. Figure 4.13 shows the comparison of the acceleration responses obtained by Accelerometer 2 and the proposed vision-based approach. It can be clearly observed that the vibration acceleration responses measured by the proposed vision system match very well with those from the physical sensor (Accelerometer 2). An even better match is obtained for Accelerometer 5, as shown in Figure 4.14.

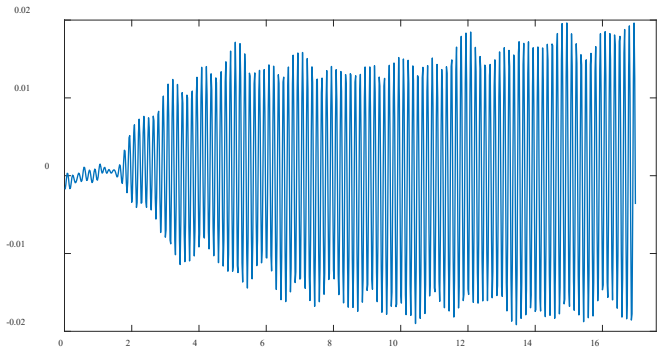


Figure 4.12. Vertical displacement near Accelerometer 5 obtained by the proposed vision method.

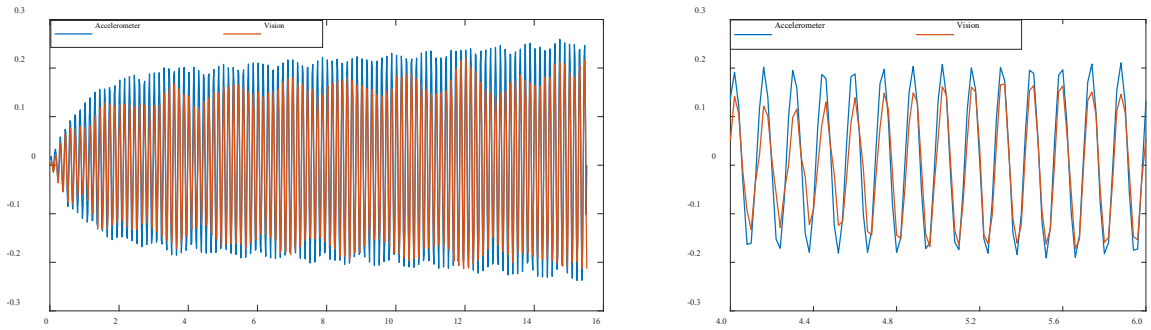


Figure 4.13. Comparison of the vertical acceleration responses by the proposed vision approach and from Accelerometer 2: (a) The complete time series; (b) A zoomed-in view of (a).

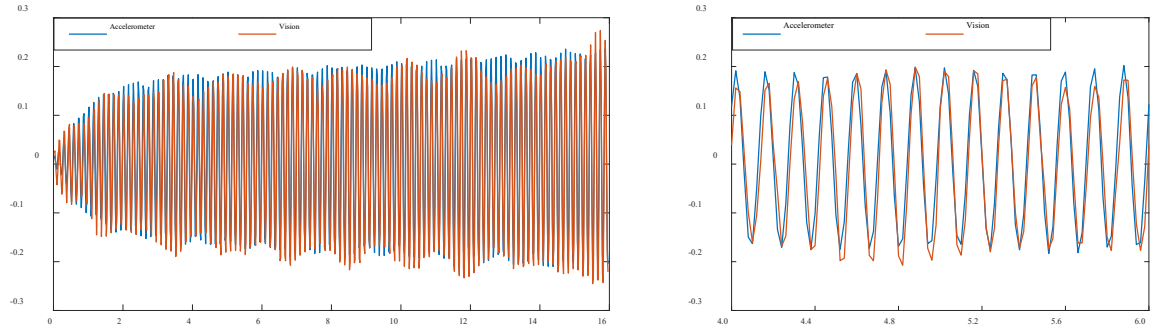


Figure 4.14. Comparison of the vertical acceleration responses by the proposed vision approach and from Accelerometer 5: (a) The complete time series; (b) A zoomed-in view of (a).

The relative errors in the acceleration responses obtained by the proposed vision approach against those measured from wired accelerometers are listed in Table 4.4. It is shown that the correlation coefficient is larger than 0.94, and the relative errors are less than 36%. The errors in the measured time domain acceleration responses could be from the inaccuracy of the numerical derivation process from displacement to acceleration and the uncertainty in the binocular vision system for measuring tine displacement.

Frequency domain analysis of the obtained acceleration responses by the proposed vision method and from the accelerometers are performed. The Fourier spectrum results for the acceleration responses near Accelerometer 2 and Accelerometer 5 are shown in Figures 4.15 and

4.16, respectively. The obtained natural frequencies by the proposed vision-based approach match very well with those from the installed accelerometers.

Table 4.4. Relative errors of obtained acceleration responses.

Position	Error (ϵ)	Corr (ρ)
Accelerometer 2	35.43%	0.9647
Accelerometer 5	33.24%	0.9462

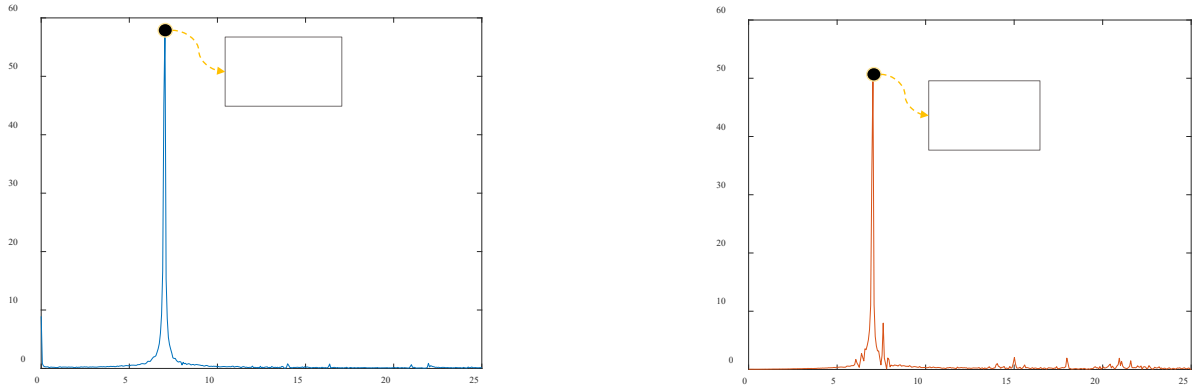


Figure 4.15. FFT spectrums of acceleration time histories obtained by the proposed vision approach and recorded by Accelerometer 2: (a) FFT spectrum of Y direction response obtained from Accelerometer 2; (b) FFT spectrum of Y direction response obtained by the proposed vision method.

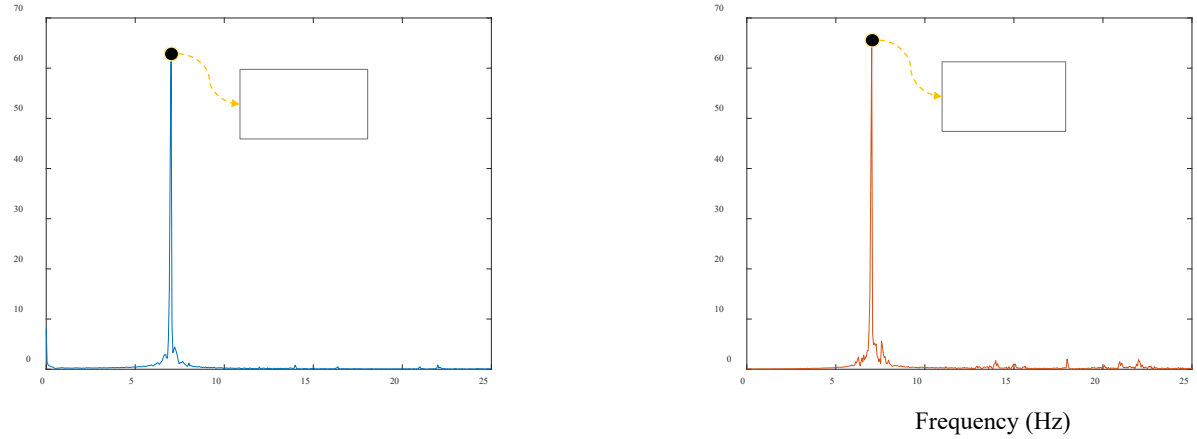


Figure 4.16. FFT spectrums of acceleration time histories obtained by the proposed vision approach and recorded by Accelerometer 5: (a) FFT spectrum of Y direction response obtained from Accelerometer 5; (b) FFT spectrum of Y direction response obtained by the proposed vision method.

4.6 Sensitivity Investigations

A binocular vision system is proposed for target-free 3D displacement measurement method for tiny vibration displacements in an order of less than 1 mm. Phase-based motion magnification is utilized to enlarge the movement of the vibration response videos. To examine the sensitivity of the proposed binocular vision system without and with motion magnifications, four sets of experiments are conducted. The accuracy of the vision system could be impacted by the position of the cameras, as well as the accuracy of key point matching and tracking. When conducting these four experiments, the cameras are positioned at the same locations as those for the 3D vibration tests of the steel beam in the laboratory as described in Section 3.4.1, which means that the camera matrices are not changed in the four experiments. Likewise, a pair of identical matched key points are used to limit the impact of the key point selection. The sensitivity analysis is conducted by using different magnitudes of vibration displacement to evaluate the performance of the proposed vision approach for tiny displacement measurement. For a fixed camera, it is easy to derive the real size (measured in the engineering units such as millimetres) represented by one pixel, which is determined by the distance from the camera to the recorded object. In our experimental setup, a pixel represents 0.3038mm in videos recorded by Camera 1 and 0.3213mm in those recorded by Camera 2.

Table 4.5 shows the 3D displacement magnitudes in each test, and the resulted relative errors in the time domain responses from the original videos and the magnified videos, respectively. The *X* direction displacements are larger than one pixel in Tests 1 and 2, and are smaller than one pixel in Test 3 and Test 4. In the *Y* direction, only the displacement in Test 1 is larger than one pixel. Displacements in the *Z* direction are triangulated by those in *X* and *Y* directions, therefore the accuracy of the displacement measurement in *Z* is affected by the displacement measurements in both the *X* and *Y* directions. It can be observed from Table 4.5 that the measurement errors using the original video in all three directions are acceptable when the displacements in both the *X* and

Y directions are larger than one pixel. However, when the displacement in either *X* or *Y* direction is less than one pixel, the accuracies of the vibration measurements with the original videos in the direction with less-than-1-pixel displacement and in the *Z* direction decrease dramatically, as shown in the *Y* and *Z* directions in Test 2. In Tests 3 and 4 where the displacement responses are very small in both the *X* and *Y* directions, the relative errors with the original videos are very large. This clearly shows the necessity of performing motion magnification when the vibration displacements are small, especially if they are invisible to human eyes. When motion magnification is applied, as demonstrated in Tests 2, 3 and 4, much more accurate displacement measurements in all three directions are achieved with smaller relative errors and larger correlation coefficients. From the sensitivity study, the proposed approach can obtain reasonable displacement measurements even when the displacements are at the scale of 0.05 mm as shown in Test 4, which is less than 1/6 of a pixel, and is hardly visible to naked eyes.

Table 4.5. Sensitivity study of the vision system without and with motion magnification

Test	**Applied Peak Displacement**	**Original Video**		**Magnified Video**	
		Error (ϵ)	**Corr. (ρ)**	**Error (ϵ)**	**Corr. (ρ)**
1	X: 0.6 mm	8.33%	0.9905	N/A	N/A
	Y: 0.4-0.6 mm	12.72%	0.9723	N/A	N/A
	Z: 0.6 mm	10.63%	0.9837	N/A	N/A
2	X: 0.4 mm	14.45%	0.9697	15.24%	0.9795
	Y: 0.2-0.32 mm	31.92%	0.9263	12.34%	0.9863
	Z: 0.4 mm	40.34%	0.9171	20.03%	0.9723
3	X: 0.2 mm	30.29%	0.9538	9.02%	0.9949
	Y: 0.1-0.2 mm	55.14%	0.8660	12.67%	0.9763
	Z: 0.2 mm	80.02%	0.6001	37.64%	0.9419
4	X: 0.2 mm	30.43%	0.9591	9.34%	0.9908
	Y: 0.05-0.06 mm	115.03%	0.1858	37.64%	0.9419
	Z: 0.2 mm	110.76%	0.3055	51.04%	0.8976

From the above sensitivity study, it is recommended that when the displacement magnitude in either *X* or *Y* direction is smaller than one pixel, motion magnification should be applied first before using the proposed vision system to obtain the target-free vibration displacement measurement. When movements are larger than one pixel, motion magnification might not be necessary. In fact further errors could be introduced due to the process of amplifying the original motion. For example, in the X direction of Test 2 shown in Table 4.5, the vibration displacement is larger than

one pixel. The relative error from the original videos is 14.45%, whereas the displacement obtained from the magnified videos has a relative error of 15.24%.

4.7 Summary

A deep learning assisted vision-based approach for target-free 3D tiny vibration displacement measurements of structures is proposed. To measure tiny vibration displacement responses, a phase-based video motion magnification algorithm is used to amplify the motion of objects in the video to achieve 3D measurement of tiny vibrations. Both the laboratory and in-field experimental results demonstrate that the proposed target-free vision system can accurately measure tiny vibrations of structures and obtain the natural frequencies of structural vibrations in 3D space accurately.

CHAPTER 5

CONCLUSION AND FUTURE WORKS

5.1 Conclusions

Following the worldwide construction of infrastructures, more and more civil structures are built to improve people's life quality. Especially in developing countries or areas, a large number of infrastructure facilities have been built intensively within a short period of time. Monitoring the health conditions is essential for these rapidly built structures, since accidents of civil structures may cause significant economic loss, even cost lives. However, compared to the efforts and funds invested in the design and construction of civil structures, the funds for SHM are limited. Moreover, in SHM, the investment and return cannot be balanced in a short time since the accidents of civil structures are very rare. Thus, economic and effective methods for SHM are keenly required. Traditional SHM approaches are costly and sometimes unaffordable for structure management organizations.

In this thesis, we propose vision-based displacement measurement approaches which can be easily accessible and significantly reduce the costs of SHM. The proposed methods overcome the limitations of traditional SHM methods and enable target-free, full-field and tiny movement measurements. The proposed vision system realizes 3D target-free displacement measurement for SHM. A two-view 3D reconstruction approach using a binocular camera system is used for depth direction displacement extraction. The proposed system uses natural key points instead of artificial targets, which avoids complicated pattern installations to save cost. Sophisticated deep learning-based key point detection and matching algorithms are used. The quality and quantity of the detected natural key points are significantly better than traditional key point detection and matching algorithms. The KLT optical flow tracker is used for accurately retrieving 2D displacements frame by frame. Vision-based measurements of tiny displacements are also enabled using the proposed system, which was traditional impossible to perform. The tiny motion in the video is amplified by a phase-based motion magnification algorithm, which visualizes invisible

motions in videos and enables computer vision algorithms to handle and process the subtle changes in videos. The vision system is validated by a series of experimental vibration tests in the lab and the measurement results are directly compared with various kinds of physical sensors, achieving highly accurate results. The proposed vision systems are also evaluated by in-field tests on a pedestrian bridge on a university campus. The motion magnification is applied in the tests since the vibration displacement of the pedestrian bridge is tiny. Since the bridge is not allowed to install displacement sensors, the results are indirectly compared with accelerators measured by multiple accelerometers, which confirms that the proposed method performs comparably with physical sensors.

In summary, this thesis addresses problems in regard with 3D vibration displacement measurement of civil engineering structures. The contributions in this thesis are listed as follows: 1) A binocular camera measurement system that applied novel computer vision algorithms is proposed to realize the target-free 3D displacement measurement; 2) The system is then extended to tiny 3D vibration displacement measurement that is feasible and more general for in-field measurement of tiny vibration displacement responses.

5.2 Future Works

The proposed approaches in this thesis focus on laboratory and in-door structural applications. The environment in such a situation is controllable that allows the systems to run stably. In the contrast, the outdoor environment is more unpredictable, such as in the raining or snowing weather, the raindrop or snowflakes appearing in the scene, which could impact the capability of the vision system. The proposed system will further consider testing on the outdoor structures under different weather conditions to improve the system robustness for the outdoor environment.

Moreover, our system can be installed on a UAV to measure the structures that are inaccessible for people. For instance, large scale bridge and tunnel structures always need to be monitored but it is very hard to install the vision system. A drone or a UAV equipped with vision measurement systems could help in this situation. Drones are not static platforms, novel visual odometry techniques are necessary to be designed to eliminate the drone’s motion for accurately measuring the displacement of the structures.

It is interesting to extend our works to using single cameras for 3D displacement measurement, which can improve the flexibility of the vision measurement system and further cutting the budget of the measurement system. The single image 3D reconstruction is a growing and energetic area in computer vision. A popular strategy at the moment is using self-supervised learning to train a model with a single image and a depth map as input and output, respectively. The algorithms developed based on this strategy need to be modified for the application in SHM in order to improve the measurement precision.

BIBLIOGRAPHY

[1] Granlund, G. H., & Knutsson, H. (2013). Signal processing for computer vision. *Springer Science & Business Media*.

[2] Dong, C. Z., Celik, O., Catbas, F. N., O'Brien, E. J., & Taylor, S. (2020). Structural displacement monitoring using deep learning-based full-field optical flow methods. *Structure and Infrastructure Engineering*, *16*(1), 51-71.

[3] Feng, D., & Feng, M. Q. (2015). Model updating of railway bridge using in situ dynamic displacement measurement under trainloads. *Journal of Bridge Engineering*, *20*(12), 04015019.

[4] Johnson, D., & Johnson, D. (2005). How to do everything with your digital camera. *McGraw-Hill/Osborne*.

[5] Andrew, A. M. (2001). Multiple view geometry in computer vision. *Kybernetes*.

[6] Wang, J., Zhong, Y., Dai, Y., Birchfield, S., Zhang, K., Smolyanskiy, N., & Li, H. (2021). Deep Two-View Structure-from-Motion Revisited. *In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8953-8962.

[7] Luo, X., Huang, J. B., Szeliski, R., Matzen, K., & Kopf, J. (2020). Consistent video depth estimation. *ACM Transactions on Graphics (TOG)*, *39*(4), 71-1.

[8] Godard, C., Mac Aodha, O., Firman, M., & Brostow, G. J. (2019). Digging into self-supervised monocular depth estimation. *In Proceedings of the IEEE/CVF International Conference on Computer Vision*, 3828-3838.

[9] Ji, P., Li, R., Bhanu, B., & Xu, Y. (2021). MonoIndoor: Towards Good Practice of Self-Supervised Monocular Depth Estimation for Indoor Environments. *arXiv preprint arXiv:*2107.12429.

[10] Mur-Artal, R., Montiel, J. M. M., & Tardos, J. D. (2015). ORB-SLAM: a versatile and accurate monocular SLAM system. *IEEE transactions on robotics*, *31*(5), 1147-1163.

[11] Zhang, Z. (2000). A flexible new technique for camera calibration. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *22*(11), 1330-1334.

[12] Yi, K. M., Trulls, E., Ono, Y., Lepetit, V., Salzmann, M., & Fua, P. (2018). Learning to find good correspondences. *In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2666-2674.

[13] Moravec, H. P. (1977). Techniques towards automatic visual obstacle avoidance.

[14] Moravec, H. P. (1979). Visual mapping by a robot rover.

[15] Harris, C. G., & Stephens, M. (1988). A combined corner and edge detector. *In Alvey Vision Conference*, *15*(50), 10-5244.

[16] Shi, J. (1994). Good features to track. *In 1994 Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 593-600.

[17] Rublee, E., Rabaud, V., Konolige, K., & Bradski, G. (2011). ORB: An efficient alternative to SIFT or SURF. *In 2011 International Conference on Computer Vision*, 2564-2571.

[18] Alahi, A., Ortiz, R., & Vandergheynst, P. (2012). Freak: Fast retina keypoint. *In 2012 IEEE Conference on Computer Vision and Pattern Recognition*, 510-517.

[19] Lowe, D. G. (2004). Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, *60*(2), 91-110.

[20] Bay, H., Tuytelaars, T., & Van Gool, L. (2006). Surf: Speeded up robust features. *In European Conference on Computer Vision.* Springer, Berlin, Heidelberg, 404-417.

[21] Alcantarilla, P. F., Bartoli, A., & Davison, A. J. (2012, October). KAZE features. *In European Conference on Computer Vision.* Springer, Berlin, Heidelberg, 214-227.

[22] Ono, Y., Trulls, E., Fua, P., & Yi, K. M. (2018). LF-Net: Learning local features from images. *arXiv preprint arXiv:*1805.09662.

[23] Revaud, J., Weinzaepfel, P., De Souza, C., Pion, N., Csurka, G., Cabon, Y., & Humenberger, M. (2019). R2D2: repeatable and reliable detector and descriptor. *arXiv preprint arXiv:*1906.06195.

[24] Dusmanu, M., Rocco, I., Pajdla, T., Pollefeys, M., Sivic, J., Torii, A., & Sattler, T. (2019). D2-net: A trainable cnn for joint description and detection of local features. *In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8092-8101.

[25] Zhou, F., & De la Torre, F. (2015). Factorized graph matching. *IEEE transactions on pattern analysis and machine intelligence*, *38*(9), 1774-1789.

[26] Yu, T., Yan, J., Wang, Y., Liu, W., & Li, B. (2018, December). Generalizing graph matching beyond quadratic assignment model. *In Proceedings of the 32nd International Conference on Neural Information Processing Systems*, 861-871.

[27] Yu, T., Yan, J., & Li, B. (2020). Determinant regularization for gradient-efficient graph matching. *In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 7123-7132.

[28] Fischler, M. A., & Bolles, R. C. (1981). Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, *24*(6), 381-395.

[29] Choi, S., Kim, T., & Yu, W. (1997). Performance evaluation of RANSAC family. *Journal of Computer Vision*, *24*(3), 271-300.

[30] Torr, P. H., & Zisserman, A. (2000). MLESAC: A new robust estimator with application to estimating image geometry. *Computer Vision and Image Understanding*, *78*(1), 138-156.

[31] Raguram, R., Chum, O., Pollefeys, M., Matas, J., & Frahm, J. M. (2012). USAC: a universal framework for random sample consensus. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *35*(8), 2022-2038.

[32] Torr, P. H., & Zisserman, A. (2000). MLESAC: A new robust estimator with application to estimating image geometry. *Computer Vision and Image Understanding*, *78*(1), 138-156.

[33] Zhang, J., Sun, D., Luo, Z., Yao, A., Zhou, L., Shen, T., & Liao, H. (2019). Learning two-view correspondences and geometry using order-aware network. *In Proceedings of the IEEE/CVF International Conference on Computer Vision*, 5845-5854.

[34] Yi, K. M., Trulls, E., Ono, Y., Lepetit, V., Salzmann, M., & Fua, P. (2018). Learning to find good correspondences. *In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2666-2674.

[35] Liu, C., Torralba, A., Freeman, W. T., Durand, F., & Adelson, E. H. (2005). Motion magnification. *ACM Transactions on Graphics (TOG)*, *24*(3), 519-526.

[36] Wu, H. Y., Rubinstein, M., Shih, E., Guttag, J., Durand, F., & Freeman, W. (2012). Eulerian video magnification for revealing subtle changes in the world. *ACM Transactions on Graphics (TOG)*, *31*(4), 1-8.

[37] Wadhwa, N., Rubinstein, M., Durand, F., & Freeman, W. T. (2014, May). Riesz pyramids for fast phase-based video magnification. *In 2014 IEEE International Conference on Computational Photography (ICCP)*, 1-10. IEEE.

[38] Oh, T. H., Jaroensri, R., Kim, C., Elgharib, M., Durand, F. E., Freeman, W. T., & Matusik, W. (2018). Learning-based video motion magnification. *In Proceedings of the European Conference on Computer Vision (ECCV)*, 633-648.

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

[40] Abdelbarr, M., Chen, Y. L., Jahanshahi, M. R., Masri, S. F., Shen, W. M., & Qidwai, U. A. (2017). 3D dynamic displacement-field measurement for structural health monitoring using inexpensive RGB-D based sensor. *Smart Materials and Structures*, *26*(12), 125016.

[41] Sony, S., Laventure, S., & Sadhu, A. (2019). A literature review of next‐generation smart sensing technology in structural health monitoring. *Structural Control and Health Monitoring, 26*(3), e2321.

[42] Fukuda, Y., Feng, M. Q., & Shinozuka, M. (2010). Cost-effective vision-based system for monitoring dynamic response of civil engineering structures. *Structural Control and Health Monitoring*, *17*(8), 918-936.

[43] Ribeiro, D., Calçada, R., Ferreira, J., & Martins, T. (2014). Non-contact measurement of the dynamic displacement of railway bridges using an advanced video-based system. *Engineering Structures*, *75*, 164-180.

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

[45] Moschas, F., Psimoulis, P. A., & Stiros, S. C. (2013). GPS/RTS data fusion to overcome signal deficiencies in certain bridge dynamic monitoring projects. *Smart Structures and Systems*, *12*(3-4), 251-269.

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

[47] Feng, D., & Feng, M. Q. (2018). Computer vision for SHM of civil infrastructure: From dynamic response measurement to damage detection–A review. *Engineering Structures*, *156*, 105-117.

[48] Cigada, A., Mazzoleni, P., & Zappa, E. (2014). Vibration monitoring of multiple bridge points by means of a unique vision-based measuring system. *Experimental Mechanics*, *54*(2), 255-271.

[49] Feng, D., & Feng, M. Q. (2016). Vision-based multipoint displacement measurement for structural health monitoring. *Structural Control and Health Monitoring*, *23*(5), 876-890.

[50] Kuddus, M. A., Li, J., Hao, H., Li, C., & Bi, K. (2019). Target-free vision-based technique for vibration measurements of structures subjected to out-of-plane movements. *Engineering Structures*, *190*, 210-222.

[51] Ji, Y. F., & Chang, C. C. (2008). Nontarget image-based technique for small cable vibration measurement. *Journal of Bridge Engineering*, *13*(1), 34-42.

[52] Bartilson, D. T., Wieghaus, K. T., & Hurlebaus, S. (2015). Target-less computer vision for traffic signal structure vibration studies. *Mechanical Systems and Signal Processing*, *60*, 571-582.

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

[54] Morlier, J., & Michon, G. (2010). Virtual vibration measurement using KLT motion tracking algorithm. *Journal of Dynamic Systems, Measurement and Control*, *132*(1): 011003.

[55] Park, S. W., Park, H. S., Kim, J. H., & Adeli, H. (2015). 3D displacement measurement model for health monitoring of structures using a motion capture system. *Measurement*, *59*, 352-362.

[56] Abdelbarr, M., Chen, Y. L., Jahanshahi, M. R., Masri, S. F., Shen, W. M., & Qidwai, U. A. (2017). 3D dynamic displacement-field measurement for structural health monitoring using inexpensive RGB-D based sensor. *Smart Materials and Structures*, *26*(12), 125016.

[57] Gao, S., Ye, Z., Wei, C., Liu, X., & Tong, X. (2019). Development of a high-speed videogrammetric measurement system with application in large-scale shaking table test. *ISPRS Annals of Photogrammetry, Remote Sensing & Spatial Information Sciences, IV-2/W7*, 33-38.

[58] Yang, Y., Dorn, C., Mancini, T., Talken, Z., Kenyon, G., Farrar, C., & Mascareñas, D. (2017). Blind identification of full-field vibration modes from video measurements with phase-based video motion magnification. *Mechanical Systems and Signal Processing*, *85*, 567-590.

[59] Chen, J. G., Wadhwa, N., Cha, Y. J., Durand, F., Freeman, W. T., & Buyukozturk, O. (2015). Modal identification of simple structures with high-speed video using motion magnification. *Journal of Sound and Vibration*, *345*, 58-71.

[60] Poozesh, P., Sarrafi, A., Mao, Z., Avitabile, P., & Niezrecki, C. (2017). Feasibility of extracting operating shapes using phase-based motion magnification technique and stereo-photogrammetry. *Journal of Sound and Vibration*, *407*, 350-366.

[61] Sarrafi, A., Mao, Z., Niezrecki, C., & Poozesh, P. (2018). Vibration-based damage detection in wind turbine blades using Phase-based Motion Estimation and motion magnification. *Journal of Sound and Vibration*, *421*, 300-318.

[62] Hartley, Richard, and Andrew Zisserman. (2003) Multiple view geometry in computer vision. *Cambridge University Press*.

[63] Hartley, R. I., & Sturm, P. (1997). Triangulation. *Computer Vision and Image Understanding*, *68*(2), 146-157.

[64] DeTone, D., Malisiewicz, T., & Rabinovich, A. (2018). Superpoint: Self-supervised interest point detection and description. *In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 224-236.

[65] DeTone, D., Malisiewicz, T., & Rabinovich, A. (2017). Toward geometric deep slam. *arXiv preprint arXiv:*1707.07410.

[66] Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:*1409.1556.

[67] Shi, W., Caballero, J., Huszár, F., Totz, J., Aitken, A. P., Bishop, R., ... & Wang, Z. (2016). Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network. *In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 1874-1883.

[68] Choy, C. B., Gwak, J., Savarese, S., & Chandraker, M. (2016). Universal correspondence network. *arXiv preprint arXiv:*1606.03558.

[69] Balntas, V., Lenc, K., Vedaldi, A., & Mikolajczyk, K. (2017). HPatches: A benchmark and evaluation of handcrafted and learned local descriptors. *In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 5173-5182.

[70] A. Paszke, S. Gross, S. Chintala, and G. Chanan. PyTorch. https://github.com/pytorch/pytorch.

[71] Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:*1412.6980.

[72] Sarlin, P. E., DeTone, D., Malisiewicz, T., & Rabinovich, A. (2020). Superglue: Learning feature matching with graph neural networks. *In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4938-4947.

[73] Chun, M. M. (2000). Contextual cueing of visual attention. *Trends in Cognitive Sciences*, *4*(5), 170-178.

[74] Mucha, P. J., Richardson, T., Macon, K., Porter, M. A., & Onnela, J. P. (2010). Community structure in time-dependent, multiscale, and multiplex networks. *Science*, *328*(5980), 876-878.

[75] Nicosia, V., Bianconi, G., Latora, V., & Barthelemy, M. (2013). Growing multiplex networks. *Physical Review Letters*, *111*(5), 058701.

[76] Gilmer, J., Schoenholz, S. S., Riley, P. F., Vinyals, O., & Dahl, G. E. (2017). Neural message passing for quantum chemistry. *In International Conference on Machine Learning*, 1263-1272. PMLR.

[77] Battaglia, P. W., Hamrick, J. B., Bapst, V., Sanchez-Gonzalez, A., Zambaldi, V., Malinowski, M., ... & Pascanu, R. (2018). Relational inductive biases, deep learning, and graph networks. *arXiv preprint arXiv*:1806.01261.

[78] Sinkhorn, R., & Knopp, P. (1967). Concerning nonnegative matrices and doubly stochastic matrices. *Pacific Journal of Mathematics*, *21*(2), 343-348.

[79] Cuturi, M. (2013). Sinkhorn distances: Lightspeed computation of optimal transport. *Advances in Neural Information Processing Systems*, *26*, 2292-2300.

[80] Dai, A., Chang, A. X., Savva, M., Halber, M., Funkhouser, T., & Nießner, M. (2017). Scannet: Richly-annotated 3d reconstructions of indoor scenes. *In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 5828-5839.

[81] Lucas, B. D., & Kanade, T. (1981). An iterative image registration technique. *In Proceedings of Imaging Understanding Workshop*.

[82] Tomasi, C., & Kanade, T. (1991). Detection and tracking of point features. *Technical Report CMU-CS-91-132*, Carnegie, Mellon University.

[83] Shi, J. (1994). Good features to track. *In 1994 Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 593-600.

[84] Kalal, Z., Mikolajczyk, K., & Matas, J. (2010). Forward-backward error: Automatic detection of tracking failures. *In 2010 20th International Conference on Pattern Recognition*, 2756-2759.

[85] Wadhwa, N., Rubinstein, M., Durand, F., & Freeman, W. T. (2013). Phase-based video motion processing. *ACM Transactions on Graphics (TOG)*, *32*(4), 1-10.

[86] Fleet, D. J., & Jepson, A. D. (1990). Computation of component image velocity from local phase information. *International journal of computer vision*, *5*(1), 77-104.

[87] Gautama, T., & Van Hulle, M. A. (2002). A phase-based approach to the estimation of the optical flow field using spatial filtering. *IEEE transactions on neural networks*, *13*(5), 1127-1136.

[88] Simoncelli, E. P., & Freeman, W. T. (1995, October). The steerable pyramid: A flexible architecture for multi-scale derivative computation. *In Proceedings of International Conference on Image Processing*, *3*, 444-447.

[89] Portilla, J., & Simoncelli, E. P. (2000). A parametric texture model based on joint statistics of complex wavelet coefficients. *International Journal of Computer Vision*, *40*(1), 49-70.

ATTRIBUTION STATEMENT

Chapter 3 of this thesis is based on work that have been published with joint-authorship. We hereby make an authorship statement to clarify the contribution of individual authors. Chapter 3 is based on the publication:

- **Yanda Shao.,** Ling Li., Jun Li., Senjian An., Hong Hao. (2021). Computer vision based target-free 3D vibration displacement measurement of structures, *Engineering Structures*, 246, 113040.

	Conception and Design	Data Acquisition and Manipulation	Programming	Experiments	Interpretation and discussion	Manuscript Writing and Revision	Total Contribution
Co-author 1 (Yanda Shao)	√	√	√	√	√	√	50%
Co-author 1 acknowledgement: I acknowledge that these represent my contribution to the above research output Signed:							
Co-author 2 (Ling Li)	√	√			√	√	15%
Co-author 2 acknowledgement: I acknowledge that these represent my contribution to the above research output Signed:							
Co-author 3 (Jun Li)	√			√	√	√	15%
Co-author 3 acknowledgement: I acknowledge that these represent my contribution to the above research output Signed:							
Co-author 4 (Senjian An)	√				√	√	10%
Co-author 4 acknowledgement: I acknowledge that these represent my contribution to the above research output Signed:							
Co-author 5 (Hong Hao)	√				√	√	10%
Co-author 5 acknowledgement: I acknowledge that these represent my contribution to the above research output Signed:							

Chapter 4 of this thesis is based on work that have been submitted to a journal. We hereby make an authorship statement to clarify the contribution of individual authors. Chapter 4 is based on the submitted paper:

- **Yanda Shao.,** Ling Li, Jun Li, Senjian An, Hong Hao., "Deep learning assisted target-free 3D tiny structural vibration measurement", Submitted to *Structural Control and Health Monitoring.*

	Conception and Design	Data Acquisition and Manipulation	Programming	Experiments	Interpretation and discussion	Manuscript Writing and Revision	Total Contribution
Co-author 1 (Yanda Shao)	√	√	√	√	√	√	50%
Co-author 1 acknowledgement: I acknowledge that these represent my contribution to the above research output Signed:							
Co-author 2 (Ling Li)	√	√			√	√	15%
Co-author 2 acknowledgement: I acknowledge that these represent my contribution to the above research output Signed:							
Co-author 3 (Jun Li)	√			√	√	√	15%
Co-author 3 acknowledgement: I acknowledge that these represent my contribution to the above research output Signed:							
Co-author 4 (Senjian An)	√				√	√	10%
Co-author 4 acknowledgement: I acknowledge that these represent my contribution to the above research output Signed:							
Co-author 5 (Hong Hao)	√				√	√	10%
Co-author 5 acknowledgement: I acknowledge that these represent my contribution to the above research output Signed:							