Measuring Human-Induced Vibrations of Civil Engineering Structures via Vision-Based Motion Tracking

Feng Zheng[a], Ling Shao[b,*], Vitomir Racic[c], James Brownjohn[d]

[a]*Department of Electronic and Electrical Engineering, University of Sheffield, United Kingdom.*
[b]*Department of Computer Science and Digital Technologies, Northumbria University, United Kingdom.*
[c]*Department of Civil and Structural Engineering, University of Sheffield, United Kingdom.*
[d]*College of Engineering, Mathematics and Physical Sciences, University of Exeter, United Kingdom.*

Abstract

We present a novel framework for measuring the body motion of multiple individuals in a group or crowd via a vision-based tracking algorithm, thus to enable studies of human-induced vibrations of civil engineering structures, such as floors and grandstands. To overcome the difficulties typically observed in this scenario, such as illumination change and object deformation, an online ensemble learning algorithm, which is adaptive to the non-stationary environment, is adopted. Incorporated with an easily carried and installed hardware, the system can capture the characteristics of displacements or accelerations for multiple individuals in a group of various sizes and in a real-world setting. To demonstrate the efficacy of the proposed system, measured displacements and calculated accelerations are compared to the simultaneous measurements obtained by two widely used motion tracking systems. Extensive experiments illustrate that the proposed system achieves equivalent performance as popular wireless inertial sensors and a marker-based optical system, but without limitations commonly associated with such traditional systems. The comparable experiments can also be used to guide

*Corresponding author. Tel: +44 7580591525.
Email address: ling.shao@ieee.org. (Ling Shao)

the application of our proposed system.

Keywords:
Object tracking, human induced vibration, ensemble learning, online learning.

1. Introduction

In civil engineering dynamics, there have been many problems related to vibrations of floors[1], footbridges [2], assembly structures (grandstands, spectator galleries, etc.), due to crowds or groups of human occupants walking, running, dancing and jumping. For example, the London Millennium Footbridge [3] opened on 10 June 2000 was closed almost immediately for nearly two years because of the unexpected movements occurred when a large crowd of pedestrians crossed the bridge. Just a year before, a similar vibration serviceability problem was observed on the newly built Solferino footbridge in Paris [4]. Also, in 2000 during a concert event, the cantilevers of the Cardiff Millennium stadium experienced excessive vibration amplitudes caused by people jumping so that the concert had to be stopped. Moreover, the modern structures have become more flexible and prone to human induced vibrations. Consequently, extensive research into the human-structure dynamic interaction phenomenon was launched. The research results were incorporated into two key design guidelines relevant to crowd loading of footbridges (France) [5] and grandstands (UK) [6] for civil engineers.

Human motion and the induced force have drawn much attentions of researchers from different areas for many years [7, 8]. Several reliable force models for active individuals [9, 10] are available. However, there is a lack of models describing dynamic loading of structures due to groups or crowds of people. How to use a model of individual loading to generate load models of multiple people still remains a challenge. It is unknown how people interact in groups of various sizes and what the level of synchronisation is between individuals under different circumstances, such as various visual and tactile stimuli. The main difficulty is to collect simultaneously the body motion data for multiple people in groups or crowds on real structures. Therefore, the key aim of this paper is to develop a new vision-based system which will enable robust collection of fundamental body data.

Although vision-based methods for human motion analysis have caught much attention of researchers and practitioners involved in gaming, security

and other related applications, the robustness of the systems is far from ideal. The key reasons for this are difficulties in setting up tracking targets and the environmental conditions. At present, these challenges can be partially solved using the robust object descriptors and adaptive appearance models [11, 12, 13, 14]. These methods can work well on data sets recorded under controlled conditions. However, due to the unpredictability of environmental changes, most existing methods cannot be applied directly in a real-world situation. In addition, they are usually unable to cope with the challenges appearing in a video sequence simultaneously. Thus, in this paper, a real-time system which contains a vision-based multiple object tracking algorithm [15] and a set of carefully selected hardware components is constructed to deal with the weaknesses of current systems.

The remainder of this paper is organized as follows. The background of measuring dynamic load and the contributions are given in Section 2. In Section 3, we describe the framework of the adopted object tracking algorithm. How to align the signals generated by different sensors is detailed in Section 4. Extensive experiments in comparison to classical sensors are presented in Section 5. We conclude this paper and discuss future work in Section 6.

2. Background and Contributions

2.1. Measuring dynamic load

Several researchers tried to adopt different systems to monitor activities of individual people and investigate the synchronization phenomenon of groups or crowds. Early attempts to measure human induced loading [16, 17, 18] were based on direct force identification using force plates and instrumented treadmills. However, their size places restrictions on studies of loading induced by multiple people [18]. An alternative approach is to measure the loading indirectly. According to [19], if the accelerations of body motion are known or measured, the ground reaction force (GRF) $\mathcal{F}$ [8] can be calculated indirectly using the basic principles of Newtonian mechanics, i.e., force is equal to mass times acceleration. Therefore, using the acceleration and mass of the individual, the GRF generated by a crowd can be computed by

$$\mathcal{F} = \sum_i m_i a_i - g \sum_i m_i, \tag{1}$$

where g is the static acceleration due to gravity, m_i is the body mass of the ith test subject and a_i is the dynamic acceleration due to body motion.

Generally, the body mass is supposed to be known, while acceleration of the body needs to be experimentally measured or estimated.

Experimental characterisation of the body motion is possible using optical marker-based motion tracking [10], wireless inertial sensors [20], video-based monitoring [21] or multichannel interacting model [22]. In [10], the accelerations of body segments were measured by tracking optical markers (Codamotion) stuck to the surface of the human body, and then used to generate force signals. However, due to interaction with daylight and the limitation of the number of markers, marker-based optical tracking systems are usually constrained to artificial laboratory environments. Alternative wireless inertial sensors [23] can be used in outdoor environments but are expensive and typically suffer from synchronising individual units in a wireless network. Moreover, the number of units within a wireless network is limited, which in turn restricts the number of monitored individuals within a crowd.

To overcome the limitations of conventional motion tracking sensors, a vision-based method can be considered. Video data captured by a camera (CCD or CMOS sensor) are becoming increasingly discussed as an innovative tool for measuring the motion of humans, structures or animals. Combined with the right video analysis algorithms used to detect the motion trajectory in the image space, vision-based methods have the potential to save time and money over conventional sensors. Research in vision-based motion tracking methods is topical [24], with a wide range of applications, such as surveillance [25], augmented reality, robotics and human-computer interaction. Compared with the conventional systems, vision-based methods for measuring human motion have the following advantages: (1) It is possible to measure people in outdoor environments rather than laboratory setting. This is because the system is less sensitive to illumination changes than marker-based sensors. (2) The number of tracking individuals is not limited. Due to the entire scenario being captured and no special tracking target (such as a Codamotion marker) being predefined, it is easy to track much more targets in the view at all times. (3) People are not aware of being recorded. No markers or inertial sensors need to be worn by participants. This will save time for preparation and lead to more natural captured body movement of test subjects, although there are ethical considerations to be addressed. (4) It is a cheap, remote and long-term monitoring system. The available commercial marker-based or wireless inertial systems are typically expensive and require external power.

Some research on digital image correlation (DIC) [26] methods to track

the movement of crowds does exist [21]. However, the suggested methods are built based on a strong assumption that the motion of each individual in a crowd is similar to the motion of surrounding people, i.e. everybody moves in the same direction. In reality, even when test subjects follow the same music, directions of their motion can be opposite. Moreover, each test subject has their own motion style or pattern, such as waving hands, nodding head and turning around, so occlusion often happens. All of these problems limit the application of DIC.

2.2. Contributions

The aim of this paper is to develop a vision-based motion tracking method to measure simultaneously the body motion of multiple individuals in a complex environment then enables the indirect measurement of human-induced loading [9, 10] and studies of synchronisation between individuals in groups or crowds in a real-world scenario. Thus, a camera system with high speed and resolution [27] is used for collecting the motion data. Aiming for addressing the challenges and abandoning the smooth motion assumption, a real-time robust object tracking algorithm, Learn++ [28], is designed to build the models of the tracker for each target. Moreover, due to the discrepancies of motion signals generated by the system and other classical sensors such as Codamotion [29] and Opal [30], an alignment method is proposed to measure the difference between the signals. The comprehensive comparison with conventional motion tracking technology can be used to guide the application of our proposed system. The added value of this research is that it will not only benefit structural engineering but will also benefit areas such as measurement of human movement in biomedicine, biomechanical rehabilitation, monitoring, performance optimisation and display of sport athletes, security surveillance and animation and virtual reality [31].

3. Algorithms

In this section, the challenges of vision-based object tracking are first introduced. Then, we detail the framework and every module of our adopted method. The adopted algorithm is composed of several modules: image patch representation, tracker training and tracker updating.

3.1. Challenges and algorithmic background

A perfectly robust vision-based system is far from being established, because many challenges induced by the target itself or the environment have

not been fully addressed. The object-induced challenges for object tracking include object deformation, in-plane or out-of-plane rotation, abrupt movement and moving out and in, while the environment-induced challenges include illumination change, motion of the camera or the background and partial occlusion. In this paper, we define "target" as the general object to be tracked. To address various challenges, different machine learning based methods are proposed. Generally, according to the type of samples used to train the model, the online adaptive algorithms can be divided into two groups: generative methods [12] which only use positive samples to infer the relationship between them, and discriminative methods [15] which use both positive and negative samples to train a classification hyperplane.

Our adopted method is based on discriminative online learning because of its separability and effectiveness. Thus, the discriminative online learning models are briefly reviewed. Discriminative methods generally consider object tracking as a classification problem. A classifier or a set of classifiers which are trained and updated online are used to make a decision for each sub-image patch. Due to the unpredictability of the object itself and the environment, different features would have different abilities for separating the object from the background. Choosing the most discriminative features will improve the robustness of object tracking methods. Some online feature selection mechanisms [32, 33] were proposed to improve tracking performance, by evaluating multiple features and adjusting the set of features. Two classical machine learning methods Support Vector Machine (SVM) [34] and AdaBoost [35] were introduced into object tracking by Aviden. After that, the online versions of AdaBoost [36] were used for feature selection in object tracking. Yan et al. [14] designed an ensemble framework for optimal selection of detectors and trackers to do multi-target tracking. Yoon et al. [37] used tracker selection and interaction for multiple feature fusion. Samples are the original information of the entire system of tracking. Weighting the samples changes the structure of the feature space so that an optimal classifier will be fast searched according to the desired feature space which is warped. Semi-supervised learning [13] and multiple instance learning [38] were adopted for sample selection.

3.2. Overview of the proposed method

The flowchart of our proposed system is shown in Fig. 1. Similar to most tracking-by-detection methods, in total, there are three main modules: tracker initiation in the first frame, target detection in the following frames

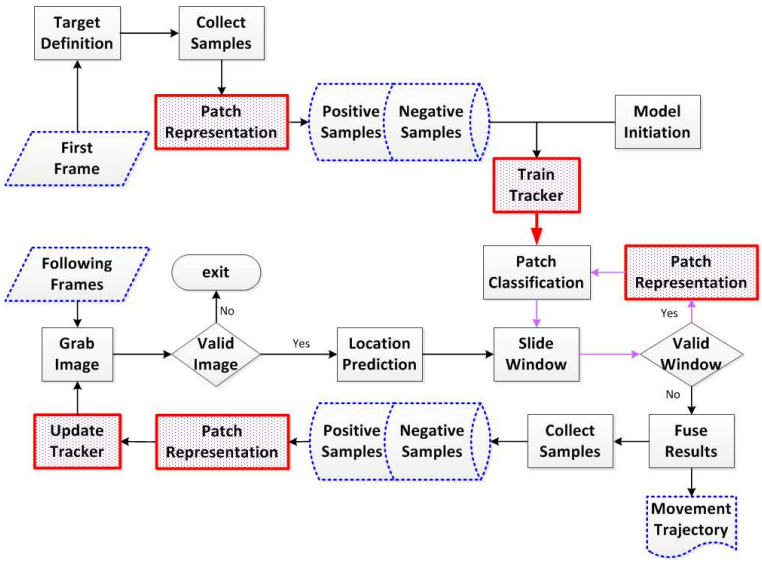


Figure 1: The flowchart of our proposed system. Red boxes denote the main modules.

and tracker update. The inputs of the system are the images captured by a high resolution and speed camera system. The outputs of the system are the motion trajectories of the multiple targets defined in the first frame. Once the trackers are generated and trained in the first frame, they can be used to detect the predefined objects in the following frames. In our implementation, one tracker is assigned to each defined object. Thus, for simplicity, we will consider the tracker for each object individually and just describe one tracker in the following sections.

The initiation module for a tracker includes four steps: target definition, collecting samples, parameters initiation and tracker training. Firstly, the target is generally defined by a set of pixels surrounded by a rectangle. In this paper, four parameters for the target are considered: (1) horizontal and vertical coordinates; (2) height and width of the rectangle. Secondly, after determining the location and size of the target in the first frame, two sets of samples will be collected. The positive samples are selected from the image patches which are sufficiently overlapped with the predefined rectangle (i.e., the intersection between a positive sample and the target divided by their union exceeds 0.75) while the negative samples come from other image patches randomly selected. Thirdly, each collected sample will be represented by a vector, which will be used in the detection module. How to represent the

image patches will be introduced in Section 3.3. Before the training, some parameters need to be initialised. Finally, based on the collected sample set, a discriminative tracker will be generated and trained based on the collected samples. How to train a tracker will be detailed in Section 3.4.

The object detection and model update modules are processed in one loop. At first, a new image will be grabbed speedily by a camera and transmitted to the memory of a workstation. Next, if the image is invalid, the system terminates at this point and outputs all the motion trajectories induced by the different subjects. Otherwise, a motion model $p(a_t|a_{t-1})$ will be used to predict the possible location in the present frame, where a_t denotes the state of object in the image space. Particle filter [39] or optical flow [40] methods can be adopted to achieve this step. According to the prediction, each location with high probability will be checked by the tracker. Same as in the model initiation module, the image patch is firstly represented by a vector. Then, the vector is considered as the input of the tracker and the output is the classification result for the image patch. The sliding window process in a sub-loop will not stop until all image patches with the high probability have been checked.

The classification results in the present frame will be used to update the tracker so that the adaptivity to the current environment can be improved. Same as in the first frame, the positive and negative samples represented by vectors are collected according to the detection result. Thus, the tracker can be updated by the information contained in the new data which represents the new environment. The details of the update module are presented in Section 3.5.

After all frames are processed, the displacement (motion trajectory) of the predefined object, which consists of a set of location points in the image space, can be produced. Next, through quadratic differential operation, the acceleration of motion is obtained. From the flowchart shown in Fig. 1, we can see that the three modules including patch representation, model training and model update are the three critical steps and will be elaborated in the following three sub-sections.

3.3. Image patch representation

Effective image patch representation is a significant step for achieving robust object tracking. In general, the rectangular patch can be converted to a vector with discriminative information extracted by patch representation. The most desirable property is the uniqueness so that each sample X_i, $X_i \in$

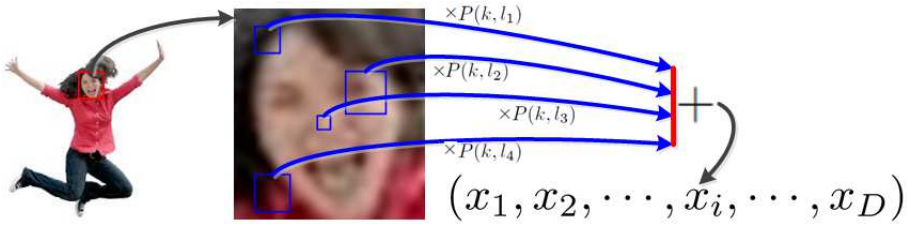


Figure 2: The pipeline of image patch representation. The enlarged patch containing a face is used to explicitly illustrate the intrinsic mechanism but need not to be processed in our system.

R^D, will be taken as a point in the feature space and can be classified by a learned hyperplane f in this feature space. Normally, basic cues including intensity, colour, edge, gradient, texture and Harr-like low-level features are used to form a high-level representation. For example, the simplest way to describe a patch is to straighten the pixel intensity values of the image patch to a vector or to count the number of intensities. Recently, to build a robust feature representation, pairwise pixel comparisons attract much attention of computer vision researchers [41]. The advantages of features based on pixel comparisons include robustness to the illumination changes and minor deformation.

In our system, a pixel-comparison-based feature is used to represent the image patch. The framework of the representation is shown in Fig. 2. For each selected patch with a size of $W \times H$, e.g., the face region in Fig. 2, a set of smooth filters with different sizes is used. If the size of one filter is $w \times h$, then all the entries of the filter are defined by $\frac{1}{w \times h}$. Therefore, Integral Image [42] can be adopted to speedily calculate the convolution by multiplying a value $\frac{1}{w \times h}$. The filter sizes will be varied from 1×1 up to the image patch size $W \times H$. Thus, in total, $n_V = (W \times H)^2$ values ($V \in R^{n_V}$) are generated for each patch, where $W \times H$ values are generated by one filter. However, the curse of dimensionality is encountered because of the super-high dimension and too much redundant information. To avoid the curse of dimensionality, following [43], a set of random projections $P \in R^{D * n_V}$ is defined to embed the feature to a low dimensional space. This matrix is very easy to compute, as it only requires a uniform random number generator. By using the sparse projection, a low dimensional representation is obtained:

$$X_i(j) = \Upsilon(P(j, \cdot)V_i), \tag{2}$$

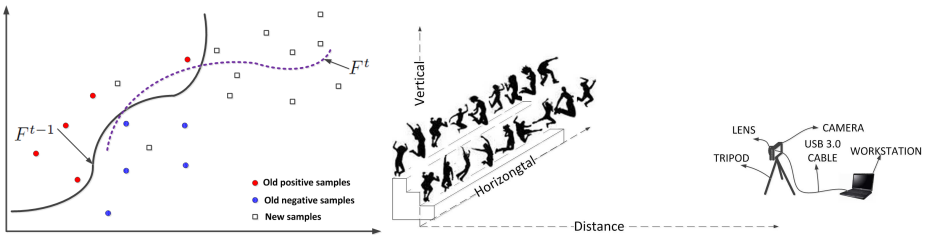
where Υ is the indicative function. Thus, $X_i(j) \in \{0, 1\}$ and the image patch has been binary coded. Due to the sparsity of projection (a small number of entries are non-zeros), the vast majority of the filters are not required to be computed so that the burden of computation is avoided.

3.4. Model training

A set of classifiers f_k are trained in the embedded feature space. Each classifier function will divide the space into two parts: one for the positive area corresponding to $sign(f_k(X)) > 0$ and the other for the negative area corresponding to $sign(f_k(X)) < 0$. The basic function can be defined by various types of formulation such as linear functions, kernel based methods, neural networks and density distribution. The classifiers are generally considered as the hyperplanes which are required to cross the low density area, maximise the maximum margin or preserve the manifold structure of samples.

In classical machine learning methods, the classifier parameters will be trained by using a fixed sample set assuming they are independent identically distributed. However, due to the non-stationary environment, this assumption is invalid in most cases of tracking problems. This is because the collected sample set in the first frame is just a set with a small number of samples and containing local information which cannot reflect the real density distribution. As a result, the classifiers trained in the previous set will suffer from the "concept drift" problems. Another difficulty in object tracking is that the various challenges frequently encountered in one scenario simultaneously, such as partial occlusion and rotation happening together. The classifiers used in the recent previous frames will be likely to fail in the new environment.

Learn++ [28], which is an ensemble of classifiers originally developed for incremental learning, can be adapted for solving the "concept drift" problem in the non-stationary environment or in data fusion applications. It specifically seeks the most discriminative information from each data set through sequentially generating an ensemble of classifiers. The classifiers trained on individual data sources are fine tuned for the given problem (concept drift). Learn++ can still achieve a statistically significant improvement by combining them, if the additional data sets carry complementary information. In this paper, assuming that the ensemble function set $\mathcal{E}^t$ and their corresponding weights w_l are available, the ensemble classifier F^t can be defined



(a) (b)

Figure 3: (a) The motivation of model updating. (b) The basic setting of the vision-based human-induced vibration detection system.

as

$$F^t(X_i) = \sum_{l:f_k \in \mathcal{E}^t} w_k f_k(X_i). \tag{3}$$

The details of how to calculate the weights of basic functions are referred to [28].

Each basic classifier f_k will correspond to n_B variables in the binary vector X_i. Δ_k denotes the index set of the variables used by f_k and $X_i^{\Delta_t}$ denotes the sub-vector of X_i corresponding to the index set Δ_k. The naive Bayesian is used as the basic classifier, which is defined as (assuming a uniform prior $p(y)$):

$$f_k(X_i) = \arg\max_y p(y|X_i^{\Delta_t}) = \arg\max_y \prod_{j \in \Delta_t} p(X_i(j)|y), \tag{4}$$

where the label $y \in \{-1, 1\}$. Given a data set $\mathbb{X}^t$ and its corresponding label set Y^t, the class distribution $p(\mathcal{X}_m|y)$ for each feature variable can be calculated according to the percentage of samples, where $\mathcal{X}_m$ denotes the mth variable of the representation. $N^t(y)$ is the total number of the samples belonging to class y in sample set $\mathbb{X}^t$ and $N^t(\mathcal{X}_m, y)$ is the number of these samples having a same code with $\mathcal{X}_m$. During training, the conditional distribution can be calculated by $p(\mathcal{X}_m|y) = N^t(\mathcal{X}_m, y)/N^t(y)$.

3.5. Model updating

Model updating is a critical step to increase the adaptivity of the proposed system. As shown in Fig. 3(a), F^{t-1} is an ensemble classifier used for the samples (circles) in previous frames but it cannot solve the problem in the current sample set (squares). However, the dotted line F^t seems to be the best classifier for the current environment. The aim of model updating is to

approximate the best classifier by incorporating the new sample set. In this case, the new sample set $(\mathbb{X}^t, Y^t)$ will be used to update the basic classifier f_k. $N_k^{t-1}(y)$ is the total number of the samples belonging to y used by f^k and $N_k^{t-1}(\mathcal{X}_m, y)$ is the number of these samples having a same code with $\mathcal{X}_m$ in the $t-1$ step. Thus, at the stage of updating, the conditional distribution can be updated by $p(\mathcal{X}_m|y) = (N_k^{t-1}(\mathcal{X}_m, y) + N^t(\mathcal{X}_m, y))/(N_k^{t-1}(y) + N^t(y))$. Meanwhile, the numbers will be updated as: $N_k^t(\mathcal{X}_m, y) \Leftarrow N_k^{t-1}(\mathcal{X}_m, y) + N^t(\mathcal{X}_m, y)$ and $N_k^t(y) \Leftarrow N^{t-1}(y) + N^t(y)$. We have $N_k^t(\mathcal{X}_m, y) = 0$ and $N_k^t(y) = 0$. Afterwards, the weights of basic classifiers will also be updated to construct the new tracker which is adaptive to the current environment. The details of this procedure are referred to [28]. By recomputing Eqn. 3, a new adaptive ensemble classifier F^t is obtained.

4. Aligning signals

To validate our proposed system, the signals should be compared with the ground truth signals. In this paper, we will consider the signals generated by Opal or Codamotion as the ground truth data. How to compare the signals generated by two different types of sensors will be introduced in this section. In general, there are four types of differences between the two signals including time domain translation, time domain scaling, amplitude translation and amplitude scaling.

Assume two discrete signals s_1 and s_2 which need to be aligned have different lengths: $l(s_1) \neq l(s_2)$, where $l(s_1)$ is the length of s_1. Five types of transformation, without changing the intrinsic properties of signals, are defined. (1) $s = T_S^H(s, \alpha)$: s has been transformed by a scaling factor α in time coordinate. (2) $s = T_T^H(s, \beta)$: s has been translated by a shifting step β in time coordinate. (3) $s = T_S^V(s, \gamma)$: s has been transformed by a scaling factor γ in amplitude coordinate. (4) $s = T_T^V(s, \delta)$: s has been translated by a shifting step δ in amplitude coordinate. (5) $s_1 = T_C(s_1, s_2)$: The latter part of s_1 has been cut off according to $l(s_2)$.

Besides the above five transformations, we also define three quantities to describe relationships of the two signals: (1) The energy difference of two signals: $e(s_1) - e(s_2)$, where $e(s_1) = \sqrt{\sum_i |s_1(i)|^2}$. (2) The correlation of two signals (requiring $l(s_1) = l(s_2)$):

$$c(s_1, s_2) = \frac{\sum_i (s_1(i) - \hat{s}_1)(s_2(i) - \hat{s}_2)}{\sqrt{\sum_i (s_1(i) - \hat{s}_1)^2} \sqrt{\sum_i (s_2(i) - \hat{s}_2)^2}}, \tag{5}$$

where $\hat{s}_1 = \sum_i s_1(i)/l(s_1)$. This quantity is not influenced by the energy difference of the two signals and is used to align the two signals in time coordinate. (3) The normalised distance from signal s_1 to signal s_2 (requiring $l(s_1) = l(s_2)$):

$$d(s_1, s_2) = \frac{e(s_1 - s_2)}{e(s_2)}. \tag{6}$$

Based on the aforementioned transformations and relationships, the two signals can be successfully aligned in both time and amplitude coordinates.

4.1. Time translation and scaling

Assume that s_c^k is the signal generated by the camera and s_o^k is the signal generated by Opal or Codamotion for the kth test subject. They are with different lengths: $l(s_c^k) \neq l(s_o^k)$. We will try to translate s_c^k and scale (down-sample) s_o^k so that the two signals can be matched in time coordinate. There are three operations: (1) Down-sample s_o^k: $T_S^H(s_o^k, \beta_o)$. (2) Shift s_c^k: $T_T^H(s_c^k, \alpha_c)$. (3) Cut the latter part of s_o^k according to the signal length of the camera: T_C.

Assume that the best translation step for s_c^k is $\hat{\alpha}_c$ and the best scaling factor for s_o^k is $\hat{\beta}_o$. The two best quantities can be found by optimizing the following objective function:

$$[\hat{\alpha}_c, \hat{\beta}_o] = \max_{\alpha_c, \beta_o} \sum_k c(T_T^H(s_c^k, \alpha_c), T_C(T_S^H(s_o^k, \beta_o), T_T^H(s_c^k, \alpha_c))).$$

The above objective is defined based on the following two facts in our experiment: (1) The length of signal s_o^k is larger than that of signal s_c^k. However, the latter part of s_o^k is meaningless because, at that moment, the test already ends. We need to cut the latter part of signal s_o^k so that the two signals have the same length. (2) For all test subjects, the sensors of Opal or markers of Codamotion belted on their bodies are synchronised. In addition, the whole view of all subjects is captured by one camera, thus the signals generated by the camera for different subjects can be considered synchronised. It means that, for all the subjects in one test, the matching points of two signals (start and end) are the same. As a result, the best matching points can be found according to the summation of correlations of all subjects.

4.2. Amplitude translation and scaling

Compared with the time coordinate where translation and scaling have same values for all test subjects, amplitude translation and scaling will be

different for different subjects. They depend on the initial values of the Opal sensors or the Codamotion markers and the positions of the test subjects in the camera view. Suppose that the two signals s_c and s_o have been matched in time coordinate according to the best values $\hat{\alpha}_c$ and $\hat{\beta}_o$. To match the two signals s_c and s_o in amplitude coordinate, two transformations are used: (1) Move signal s_o to around zero: $T_T^V(s_o, \delta_o)$. (2) Scale signal s_c: $T_S^V(s_c, \gamma_c)$. From the definition of correlation, we can see that the correlation of the two signals will not be changed by the above two operations. Thus, the objective function can be defined as:

$$[\hat{\delta}_o, \hat{\gamma}_c] = \min_{\delta_o, \gamma_c} |e(T_S^V(s_c, \gamma_c)) - e(T_T^V(s_o, \delta_o))|$$

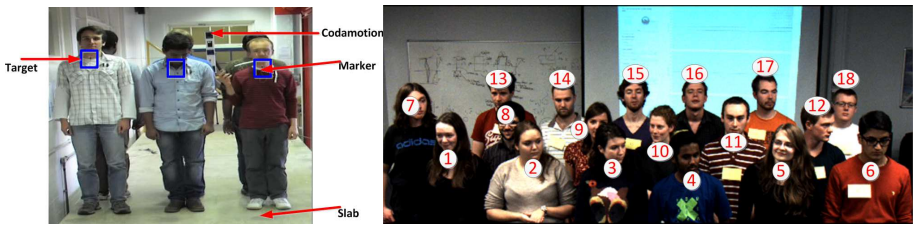
$$+\lambda d(T_S^V(s_c, \gamma_c), T_T^V(s_o, \delta_o)),$$

where λ is a regularisation parameter used to balance the two parts of the function. In this objective function, the energy function considers the global difference while the normalised distance considers subtle difference between two signals. By optimising this objective function, the two signals s_c and s_o will be aligned in both amplitude and time coordinates.

5. Experiments

To test the proposed vision-based system, two experiments are conducted in the Light Structures Laboratory (LSL) at the University of Sheffield, UK. The basic setting of the camera system is shown in Fig. 3(b). The only necessary consideration in this experimental setup was that the interest parts of moving participants (also called test subjects) should be in the camera field of view. During the studied activities, the bodies were moving predominantly in the vertical direction inducing vertical structural vibrations. Therefore, the vertical motion trajectories in the image space corresponding to the projected movement in the vertical direction were detected. To investigate the performance of the proposed system, motion signals generated by our system were compared against the marker-based Codamotion and/or Opal wireless inertial sensors both in time and frequency domains.

The hardware of the proposed system includes a Point Grey Flea USB3.0 CMOS sensor [27], a USB3.0 cable, a tripod and a portable workstation. The maximum resolution and FPS (frames per second) of the selected sensor are 2080×1552 and 60, respectively. A low distortion lens of $8mm$ made by NET



(a) (b)

Figure 4: (a) The setting of the first experiment: people performing on a slab. The blue rectangles denote the targets defined in our system. (b) The setting of the second experiment: people bouncing to music.

New Electronic Technology[1] is a C-mount lens and is attached to the camera. The sensor can be directly connected to a workstation by using a USB3.0 cable. To solve the speed problem of transmission from memory to hard drive, a portable solid state drive (SSD) is used. As a result, the integrated system is easy to be installed and convenient to be taken to wherever it is needed. More details of high efficiency and robustness of the tracking algorithm can be found in [15].

5.1. Experimental setting

5.1.1. People moving on a slab

In this experiment, six people were standing on and inducing vertical motion in a flexible slab strip structure in the LSL. The slab strip is a $2m$ wide, 15 tonne pre-stressed concrete slab spanning $11m$ between simple supports, and the six people were arranged in two rows. Three people in the front row were in view of the camera positioned at one end of the slab, as shown in Fig. 4(a). Two types of actions - bouncing and jumping - were investigated at a frequency from $2.0Hz$ to $2.5Hz$ which were synchronised by a metronome (held by the right subject in the first row). The Codamotion markers [29] were attached on the neck of every subject and two sets of Codamotion cameras were installed at the two ends of the slab. For each sequence, the test duration was set to 30 seconds. The sampling rates of Codamotion and the camera were set to $200Hz$ and $60Hz$, respectively.

[1] http://www.net-gmbh.com/

5.1.2. People bouncing to music

A group of 18 (S1-S18) persons took part in this test. Vertical movement of each individual was measured directly using miniature APDM Opal wireless accelerometers [30] attached to their bodies, while the whole group was simultaneously recorded by the video camera located $3m$ away from the group. Assuming a constant magnification matrix of the video image, this led to an approximate relation of $0.00131m$ per pixel. The Opal sensor has an acquisition frequency of $128Hz$, which results, according to Nyquist, in a resolve spectrum up to $64Hz$. The instrument has an accuracy of $0.0012m/s^2$ for accelerometer and maintains time-synchronisation of $\leq 1ms$ between sensors. Both the camera and the sensors were synchronised by a trigger signal measured by the Opal sensors and projected to the screen behind the participants to allow a rough determination of starting video frames. In each test, the participants were asked to bounce simultaneously to a given popular song for $40-50$ seconds. An example of the acquired images is shown in Fig. 4(b).

5.2. Problems Definition

To compare the displacement and acceleration generated by the proposed system to the signals generated by other two technologies, five problems can be identified and should be solved first.

5.2.1. Intrinsic noises

Firstly, it is difficult to select the same tracking targets on the body as the locations of the Coda markers or Opal sensors. For example, an Opal sensor is usually belted on the waist of a person, but motion of the waist is often impossible to track in video records due to frequent visual occlusions with other test subjects. Secondly, the tracking markers or sensors never represent the exact motion of a human body due to the relative movement between the clothes (e.g. belt) and the skin or the skin wobbling in case of overweight test subjects. As a result, these two kinds of noise bring some difficulties to the comparison between the signals generated by the camera and other motion tracking systems.

5.2.2. Time scale difference (Real sampling rate)

Sampling rates of Opal, Codamotion and the camera should be $128Hz$, $200Hz$ and $60Hz$ respectively, so the corresponding motion signals have to be resampled. In our experimental analysis, we found that either the real

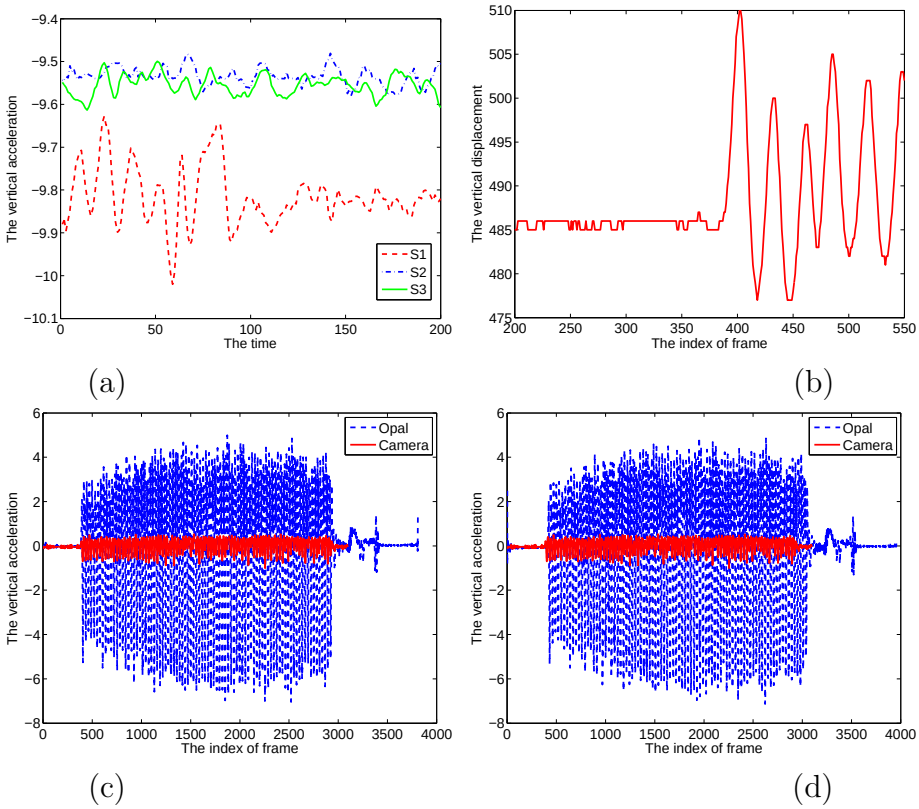


Figure 5: (a) The acceleration examples of Opal sensors: subjects 1, 2 and 3 in Trail 1. (b) The displacement example of camera. (c) The two accelerations are aligned in time coordinate when the signal generated by Opal is downsampled to 57.7. (d) Same with (c) but the signal generated by Opal is downsampled to 60.

sampling rate of the camera is not exactly $60Hz$ or the real sampling rate of Opal is not exactly $128Hz$, because of the loss of data or the system delay. Fig. 5(c) and Fig. 5(d) show two types of down-sampling rate: $57.7Hz$(Fig. 5(c)) and $60Hz$(Fig. 5(d)). We can see that the result of Fig. 5(c) is much better.

5.2.3. Time translation

In the first experiment, we can confirm the beginning frame according to the power lights of markers attached on necks of test subjects as shown in Fig. 4(a). In the second experiment, we set a screen in the camera view to indicate the start point of the Opal sensor. However, there is a delay of over 10 frames when the button becomes completely bright from dark. So, it is

Table 1: Comparison with Codamotion: the correlation of two displacements

ID	T1	T2	T3	T4	T5	Avg
S1	0.9896	0.9878	0.9754	0.9939	0.9719	0.9837
S2	0.9935	0.9926	0.9919	**0.9179**	0.9590	0.9710
S3	0.9917	**0.9960**	0.9938	0.9626	0.9553	0.9799
Avg	0.9916	0.9921	0.9870	0.9581	0.9621	

Table 2: Comparison with Codamotion: the normalised distance between the frequencies of two displacements

ID	T1	T2	T3	T4	T5	Avg
S1	0.0060	0.0089	0.0135	0.0018	0.0039	0.0068
S2	0.0044	0.0056	0.0026	0.0553	0.0288	0.0193
S3	0.0051	0.0035	**0.0020**	**0.1535**	0.0504	0.0429
Avg	0.0052	0.0060	0.0060	0.0702	0.0277	

hard to decide which frame is the best one to mark the start of recording, making time translation necessary to align the two signals.

5.2.4. Amplitude scale difference

The amplitude scale of signals generated by different sensors will be different because they are in different types of space. Moreover, the amplitude scale between the tracked trajectories in the image plane and the real motion of different test subjects will be also different as they were not standing in the same row. The precision per pixel strictly depends on the distance between the tracked target and the camera. A different distance means a different amplitude scale.

5.2.5. Amplitude translation

For different Opal sensors, the default initial values should be around the local intrinsic acceleration of gravity ($-9.8m/s^2$). However, in reality, they are a little different from each other. Three examples are given in Fig. 5(a) from the first point to the 200th point. For the camera, the means of all the accelerations are around 0. Thus, the best amplitude translation must be fixed to match the two signals in this coordinate.

5.3. Experiment 1: displacements generated by Codamotion and Camera

In this subsection, the comparison of two vertical displacements generated by the camera and Codamotion are given. The tests are repeated five times and the five trials marked as T1, T2, T3, T4 and T5 are recorded. The

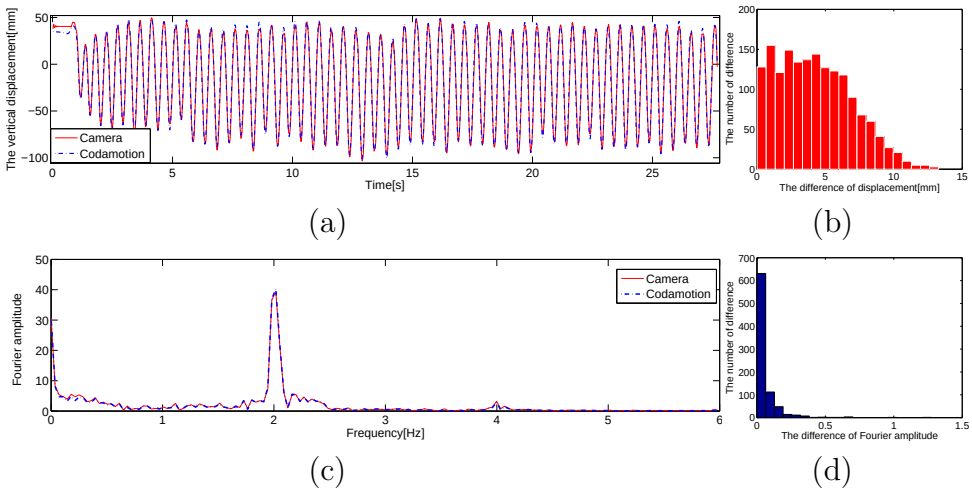


Figure 6: (a) The comparison of two vertical displacements by subject 3 in Trail 3. (b) The histogram of the displacement difference values between the camera and Codamotion. The difference values are the direct distances of two corresponding points in two signals and not normalised. (c) The corresponding Fourier amplitude comparison with frequency from 0 to 6. (d) The histogram of the Fourier amplitude difference values between the camera and Codamotion.

first three trials (T1, T2 and T3) correspond to the action bouncing while the last two trials (T4 and T5) correspond to the action jumping. After the translation and scaling operations in time and amplitude coordinates to the two displacements, the comparison results are shown in Table 1 and 2. A red bold number denotes the best match in all tests while the blue bold number denotes the worst one.

We can see that the correlation of any pair of signals is more than 0.95 with that of most pairs over 0.98. In the frequency domain, except for the motion of subject 3 in Trial 4, the normalised distance between any other pair of displacements is less than 0.06. Moreover, the displacements of bouncing are aligned better than the displacements of jumping, because the motion blur problem for jumping is more serious in the camera system. This can be solved by altering the lens focus for a specialised application if needed.

From the above analysis, we know the alignment of the two displacements by subject 3 in Trial 3 is the best. Fig. 6(a) shows the aligned displacements generated by camera and the Codamotion and Fig. 6(b) shows the corresponding frequency comparison from $0Hz$ to $6Hz$. It illustrates that the duration of this test is around 30 seconds and the main frequency of bounc-

ing is around $2Hz$. In fact, the test subjects were indeed bouncing with synchronisation by a metronome at $2Hz$. Hence, we conclude that the both two displacements reflect the intrinsic motion characteristics (dominant frequency). Next, Fig. 6(b) demonstrates that almost all the difference values between the two displacements are less than $10mm$ and Fig. 6(d) demonstrates the majority of frequency differences is less than $0.1Hz$. We can see that the two displacements are almost completely overlapped both in time and frequency domains. Therefore, we can conclude that the proposed vision-based system achieves similar results as Codamotion, but it possesses several advantages and has better extensibility.

Table 3: Comparison with Opal: the correlation of two accelerations

ID	T1	T2	T3	T4	T5	Avg
S1	0.17	0.14	0.19	0.17	0.22	0.18
S2	0.24	0.17	0.24	0.09	0.28	0.2
S3	**0.28**	0.19	0.28	0.19	0.25	0.24
S4	0.09	0.14	0.15	0.09	0.14	0.12
S5	0.26	0.22	0.27	0.09	**0.40**	**0.25**
S6	c.08	0.07	0.14	0.11	0.13	0.11
S7	0.08	0.1	**0.11**	0.12	0.19	0.12
S8	0.1	0.11	0.13	0.09	0.09	0.1
S9	0.23	0.11	0.18	0.08	0.2	0.16
S10	0.16	0.11	0.14	0.14	0.24	0.16
S11	0.22	0.17	0.26	**0.07**	0.17	0.18
S12	0.12	0.14	0.17	0.13	0.15	0.14
S13	**0.06**	**0.06**	0.12	**0.07**	**0.08**	**0.08**
S14	0.14	0.12	0.2	0.08	0.24	0.15
S15	0.09	0.2	0.21	**0.07**	0.15	0.15
S16	0.09	0.13	0.17	0.12	0.16	0.13
S17	0.09	0.12	0.12	0.12	0.15	0.12
S18	0.19	**0.27**	**0.39**	**0.20**	0.16	0.24
Avg	0.15	0.14	0.19	0.11	0.19	

5.4. Experiment 2: accelerations generated by Opal and Camera

The second test is also conducted five times using five pieces of music with different rhythms and the five trials marked as T1, T2, T3, T4 and T5 were recoded. For each trial, the Opal system was started first and stopped last so that it guarantees that the signals generated by the camera are within the duration of the signals generated by the Opal sensors. The 18 test subjects are denoted as S1 to S18.

Table 4: Comparison with Opal: the normalised distance of two accelerations

ID	T1	T2	T3	T4	T5	Avg
S1	0.29	0.27	0.36	0.27	0.33	0.3
S2	0.42	0.33	0.44	0.2	0.41	0.36
S3	0.46	0.36	0.5	0.28	0.42	0.4
S4	0.26	0.32	0.34	0.2	0.31	0.29
S5	**0.48**	0.39	0.49	**0.15**	**0.58**	0.42
S6	0.17	0.17	0.28	0.2	0.28	0.22
S7	0.16	0.19	0.26	0.2	0.29	0.22
S8	0.2	0.2	0.25	0.17	**0.17**	0.19
S9	0.36	0.23	0.35	**0.15**	0.35	0.29
S10	0.3	0.23	0.34	0.25	0.37	0.3
S11	0.35	0.35	0.51	0.17	0.37	0.35
S12	0.22	0.26	0.33	0.27	0.3	0.28
S13	**0.15**	**0.14**	0.3	0.16	**0.17**	0.18
S14	0.21	0.24	0.36	0.17	0.37	0.27
S15	0.19	0.36	0.36	**0.15**	0.32	0.28
S16	0.21	0.29	0.35	0.2	0.25	0.26
S17	0.2	0.22	**0.24**	0.2	0.31	0.23
S18	0.31	**0.42**	**0.58**	**0.31**	0.29	0.38
Avg	0.27	0.28	0.37	0.21	0.33	

5.4.1. Time domain analysis

Through the several operations (translation and scaling) to the two signals generated by the camera and Opal, we can find the best match of them. In this subsection, we will give quantitative analysis about the match. The correlation, normalised distance and their corresponding means of the correlation and distance are shown in Tables 3 and 4, separately, when the two signals are best aligned. A larger number in Table 3 means the two signals are aligned better whilst a larger number in Table 4 indicates the two signals are aligned worse. A red bold number denotes the best match in that trial while a blue bold number indicates the worst one. From Tables 3 and 4, we can see that signals of subject 13 are aligned the best four times. The signals of subject 18 are aligned worst three times and the signals of subject 5 are aligned worst twice.

The differences of two signals are from three aspects. The first one is the error from the vision-based tracking algorithm. This type of tracking method tries to find the location of a predefined object with highest probability in the image plane. However, due to limitations of the feature representation or the learning model, a small error to the location cannot be avoided. The

second aspect is the additional motion of different subjects. Even though the vision-based method gives the exact trajectories of movement, the final signals cannot be aligned completely because different subjects move with different ways. For example, subject 5 in our experiment always lowered her head, raised her hand to organise her hair, and turned her head to one side to talk with somebody else. These additional motions will bring some difficulties to align two signals of the same subject. Besides these types of additional motion, in fact, the Opal sensors cannot record the exact motion of a body because they were just fastened to the body or the clothes by a belt. The final aspect is the intrinsic difficulty caused by different body parts. Even though all test subjects strictly followed the rules of the experiment, there were still different motions for different parts of the body of the same test subject synchronised with the same music. For example, the head always nodded when the subject bounced at the lowest point. Also, for subjects with a large belly, the Opal sensor will record the motion of the belly.

Table 5: Comparison with Opal: the normalised distance of the frequencies of two displacements

ID	T1	T2	T3	T4	T5	Avg
S1	0.17	0.14	0.19	0.17	0.22	0.18
S2	0.24	0.17	0.24	0.09	0.28	0.2
S3	**0.28**	0.19	0.28	0.19	0.25	0.24
S4	0.09	0.14	0.15	0.09	0.14	0.12
S5	0.26	0.22	0.27	0.09	**0.40**	**0.25**
S6	c.08	0.07	0.14	0.11	0.13	0.11
S7	0.08	0.1	**0.11**	0.12	0.19	0.12
S8	0.1	0.11	0.13	0.09	0.09	0.1
S9	0.23	0.11	0.18	0.08	0.2	0.16
S10	0.16	0.11	0.14	0.14	0.24	0.16
S11	0.22	0.17	0.26	**0.07**	0.17	0.18
S12	0.12	0.14	0.17	0.13	0.15	0.14
S13	**0.06**	**0.06**	0.12	**0.07**	**0.08**	**0.08**
S14	0.14	0.12	0.2	0.08	0.24	0.15
S15	0.09	0.2	0.21	**0.07**	0.15	0.15
S16	0.09	0.13	0.17	0.12	0.16	0.13
S17	0.09	0.12	0.12	0.12	0.15	0.12
S18	0.19	**0.27**	**0.39**	**0.20**	0.16	0.24
Avg	0.15	0.14	0.19	0.11	0.19	

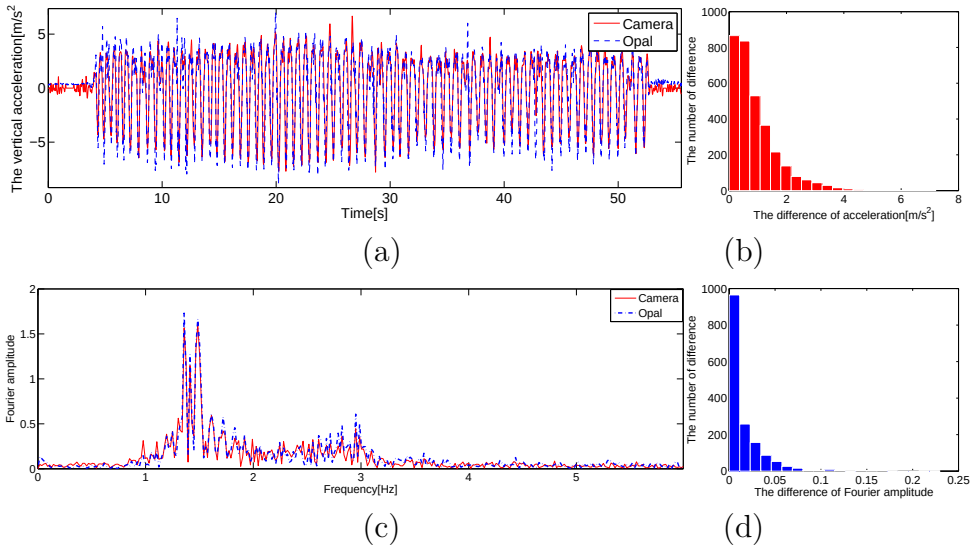


Figure 7: (a) The comparison between the two accelerations of Subject 13 in T2 in the time domain. (b) The histogram of the displacement difference values between the camera and Opal. (c) The corresponding Fourier amplitude comparison with frequency from 0 to 6. (d) The histogram of the Fourier amplitude difference values between the camera and Opal.

5.4.2. Fourier spectral analysis

Besides the time domain analysis, we also investigate the frequency difference for all the test subjects. The results are given in Table 5. The most remarkable point is that the normalised distance in the frequency domain is less than that in the time domain. This means that the translations in time or the vertical direction probably contribute to the most difference between the two accelerations, but the most dominant frequencies have also been captured by the camera system. In fact, the frequency analysis of acceleration induced by human is the most important step to monitor the fitness of a structure. The second point is that the results in Table 5 have the same trend as in Table 4. The pair of signals which has a lower comparison score in the time domain still has a smaller normalised distance between them. For example, the two signals of subject 13 in Trial 2 are the most similar pair in Tables 3 and 4 and they are also the most similar pair in Table 5.

Table 6: The percentage of points satisfying the conditions

Condition (m/s^2)	> 3	> 2	> 1	> 0.7	> 0.5
Percentage	0.031	0.096	0.337	0.483	0.617

5.4.3. The best match of subject 13 of Trial 2

According to the comparison results in time and frequency domains, we realise that the two signals generated by the camera and Opal on subject 13 in Trial 2 (T2) is aligned best. The correlation of the aligned signals is up to 0.93. Also, the normalised distance of the two signals is 0.14. Fig. 7(a) shows that there are around 50 seconds activities corresponding to the music played for 50 seconds. The generated accelerations by the two systems are both between $7m/s^2$ and $-9m/s^2$ and the two signals are similar with each other. However, it is worth to mention that the two signals come from motions of different parts of the body. The red signal is generated by the camera and the head is considered as the target whilst the blue dashed one is generated by the Opal sensor and the belly or the waist is considered as the target. As a result, it is obvious that there are some differences between the movements of the parts. We also investigate difference values of the two accelerations when they are aligned. From Table 6, firstly, we can see that the difference values of most points are less than $0.7m/s^2$. Next, we know that difference values of about 32% of the points are less than $0.5m/s^2$ and that of only 3% of the points is larger than $3m/s^2$. Moreover, the histogram is given in Fig. 7(b). It illustrates that the difference values of most of the points are with a low value. And the mean and variance of the difference values are $0.909m/s^2$ and $0.696m/s^2$, respectively. c Same as the analysis in the time domain, the Fourier spectral comparisons are shown in Fig. 7(c) and Fig. 7(d). We can see that there are two peaks of the frequencies around $0.98Hz$ and $1.96Hz$ in the two signals and most of the frequencies are similar with each other. Fig. 7(c) just shows the comparison with the frequencies from $0Hz$ to $6Hz$ and the higher frequencies of the two accelerations both tend to be zero. Generally, the frequency difference between the two accelerations is proportional to the amplitude. However, the Fourier amplitudes are almost the same around the dominant frequency of the activities as shown in Fig. 7(c). It demonstrates that both systems capture the main motion characteristics of the activities. Moreover, Fig. 7(d) illustrates that almost all the frequency differences are less than $0.06Hz$. As a result, we can conclude that the proposed camera system can capture almost all the characteristics of human induced motion

as the inertial sensor Opal.

6. Conclusion and Discussions

Based on the extensive investigation of the two comparative experiments with two motion tracking systems Codamotion and Opal, the following conclusions can be safely drawn.

First, the proposed vision-based system shows good performance in the field of human-induced vibrations. Compared with the classical sensors, such as marker-based and inertial sensors, this system can be easily installed and used in various environments. Due to the ability of remotely capturing the whole view of the scene, there is no limitation on the number of cameras. Moreover, no markers or other instruments are attached to the human body, so body motion can be natural (i.e. not restricted by hardware) and test subjects are not necessarily aware of being recorded. In addition, the adopted Learn++ based object tracking algorithm overcomes the difficulties encountered in realistic scenarios, such as moving out of view and partial occlusion.

Next, the possible errors are from the following three sources. (1) The motions of the selected different body parts will slightly differ so that the measured movements will be also different. This contributes to the main differences between the two types of systems. However, the error can be avoided by selecting the parts of the body with little uncontrollable movement, such as chest or neck. (2) The loss of data in the process of transmission will also lead to error in the trajectories in the time coordinate. At present, we use a low-end portable workstation and an improved hardware system will potentially improve the overall performance. (3) The non-vertical imaging plane will also lead to different scales of vertical displacements or accelerations for different test subjects.

In future work, we will investigate the influence to the reconstructed ground reaction force by different body parts and confirm which part or parts will be the best to use. Also, it is worth to adopt a calibration step to measure the distance between the camera and the target.

Acknowledgement

This work was supported by the project "Synchronisation in dynamic loading due to multiple pedestrians and occupants of vibration-sensitive structures" funded by EPSRC [Reference: EP/I029567/1].

[1] H. Hashim, Z. Ibrahim, H. A. Razak, Dynamic characteristics and model updating of damaged slab from ambient vibration measurements, Measurement 46 (4) (2013) 371–1378.

[2] J. Skeivalasa, M. Jureviciusb, A. Kilikeviciusb, V. Turlac, An analysis of footbridge vibration parameters, Measurement 66 (2015) 222–228.

[3] P. Dallard, T. Fitzpatrick, A. Flint, S. L. Bourva, A. Low, R. Ridsdill, M. Willford, The london millennium footbridge, The Structural Engineer 79 (22) (2001) 17–33.

[4] A. N. Blekherman, Autoparametric resonance in a pedestrian steel arch bridge: Solferino bridge, paris, Journal of Bridge Engineering 12 (6) (2007) 669–676.

[5] French association of civil engineering, Technical guide footbridges: Assessment of vibrational behaviour of footbridges under pedestrian loading, Tech. rep., Paris, France: Setra (2006).

[6] Joint Working Group (JWG), Dynamic performance requirements for permanent grandstands subject to crowd action: recommendations for management, design and assessment, Tech. rep., The Institution of Structural Engineers, London, UK (2008).

[7] E. C.-Y. Yang, M.-H. Mao, 3d analysis system for estimating intersegmental forces and moments exerted on human lower limbs during walking motion, Measurement 73 (2015) 171–179.

[8] J. Garza-Ulloa, H. Yu, T. Sarkodie-Gyan, A mathematical model for the validation of the ground reaction force sensor in human gait analysis, Measurement 45 (4) (2012) 755–762.

[9] V. Racic, A. Pavic, J. M. W. Brownjohn, Experimental identification and analytical modelling of human walking forces: Literature review, Journal of Sound and Vibration 326 (2009) 1–49.

[10] V. Racic, J. M. W. Brownjohn, A. Pavic, Reproduction and application of human bouncing and jumping forces from visual marker data, Journal of Sound and Vibration 329 (2010) 3397–3416.

[11] D. D. Doyle, A. L. Jennings, J. T. Black, Optical flow background estimation for real-time pan/tilt camera object tracking, Measurement 48 (2014) 195–207.

[12] X. Mei, H. Ling, Robust visual tracking using l1 minimization, in: Proc. ICCV, 2009.

[13] G. Li, L. Qin, Q. Huang, J. Pang, S. Jiang, Treat samples differently: Object tracking with semi-supervised online covboost, in: Proc. ICCV, 2011.

[14] X. Yan, X. Wu, I. A. Kakadiaris, S. K. Shah, To track or to detect? an ensemble framework for optimal selection, in: Proc. ECCV, 2012.

[15] F. Zheng, L. Shao, J. Brownjohn, V. Racic, Learn++ for robust object tracking, in: Proc. BMVC, 2014.

[16] S. Kerr, Human induced loading on staircases, Ph.D. thesis, Mechanical Engineering Department, University of London, London (1998).

[17] A. Belli, P. Bui, A. Berger, A. Geyssant, J. Lacour, A treadmill ergometer for three-dimensional ground reaction forces measurement during walking, Journal of Biomechanics 34 (1) (2001) 105–112.

[18] V. Racic, A. Pavic, J. M. W. Brownjohn, Number of successive cycles necessary to achieve stability of selected ground reaction force variables during continuous jumping, Journal of Sports Science and Medicine 8 (2009) 639–647.

[19] M. Bobbert, H. Schamhardt, B. Nigg, Calculation of vertical ground reaction force estimates during running from positional data, Journal of Biomechanics 24 (24) (1991) 1095–105.

[20] M. Jina, J. Zhaoa, J. Jinb, G. Yuc, W. Li, The adaptive kalman filter based on fuzzy logic for inertial motion capture system, Measurement 49 (2014) 196–204.

[21] A. Caprioli, S. Manzoni, E. Zappa, People-induced vibrations of civil-structures: Image-based measurement of crowd motion, Experimental Techniques 35 (3) (2011) 71–79.

[22] S. J. Lee, Y. Motai, H. Choi, Tracking human motion with multichannel interacting multiple model, IEEE Transactions on Industrial Informatics 9 (3) (2013) 1751–1763.

[23] E. Palermoa, S. Rossib, F. Marinid, F. Patana, P. Cappaa, Experimental evaluation of accuracy and repeatability of a novel body-to-sensor calibration procedure for inertial sensor-based gait analysis, Measurement 52 (2014) 145–155.

[24] H. Yang, L. Shao, F. Zheng, L. Wang, Z. Song, Recent advances and trends in visual tracking: A review, Neurocomputing 74 (18) (2011) 3823–3831.

[25] C. Cheng, K. Kawaguchi, A preliminary study on the response of steel structures using surveillance camera image with vision-based method during the great east japan earthquake, Measurement 62 (2015) 142–148.

[26] H. Schreier, J.-J. Orteu, M. A. Sutton, Image Correlation for Shape, Motion and Deformation Measurements, Springer US, 2009.

[27] Point Grey Research Incorporation, Flea3 usb 3.0 getting started manual, http://www.ptgrey.com, Richmond, BC, Canada, 2012.

[28] R. Elwell, R. Polikar, Incremental learning of concept drift in nonstationary environments, IEEE Transactions on Neural Networks 22 (10) (2011) 1517–1531.

[29] Charnwood Dynamics Ltd., Codamotion user manuals, in: http://www.charndyn.com/, Leicestershire,UK, 2009.

[30] APDM Inc., Apdm sensor whitepaper technical specifications, http://apdm.com/, Portland USA, 2013.

[31] A. Yilmaz, O. Javed, M. Shah, Object tracking: A survey, ACM Computing Surveys 38 (4) (2006) 1–45.

[32] R. T. Collins, Y. Liu, M. Leordeanu, Online selection of discriminative tracking features, IEEE Transactions on PAMI 27 (10) (2005) 1631–1643.

[33] J. Wang, X. Chen, W. Gao, Online selecting discriminative tracking features using particle filter, in: Proc. CVPR, 2005.

[34] S. Avidan, Support vector tracking, IEEE Transactions on PAMI 26 (8) (2004) 1064–1072.

[35] S. Avidan, Ensemble tracking, IEEE Transactions on PAMI 29 (2) (2007) 261–271.

[36] H. Grabner, C. Leistner, H. Bischof, Semi-supervised on-line boosting for robust tracking, in: Proc. ECCV, 2008.

[37] J. H. Yoon, D. Y. Kim, K.-J. Yoon, Visual tracking via adaptive tracker selection with multiple features, in: Proc. ECCV, 2012.

[38] B. Babenko, M.-H. Yang, S. Belongie, Robust object tracking with online multiple instance learning, IEEE Transactions on PAMI 33 (8) (2011) 1619–1632.

[39] M. Isard, A. Blake, Icondensation: Unifying low-level and high-level tracking in a stochastic framework, in: Proc. ECCV, 1998.

[40] B. D. Lucas, T. Kanade, An iterative image registration technique with an application to stereo vision, in: Proc. IJCAI, 1981.

[41] M. Ozuysal, P. Fua, V. Lepetit, Fast keypoint recognition in ten lines of code, in: Proc. CVPR, 2007.

[42] P. Viola, M. Jones, Rapid object detection using a boosted cascade of simple features, in: Proc. CVPR, 2001.

[43] P. Li, T. J. Hastie, K. W. Church, Very sparse random projections, in: Proc. KDD, 2006.