Received June 15, 2020, accepted June 29, 2020, date of publication July 22, 2020, date of current version August 7, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.3011360

Stereo Vision-Based 3D Positioning and Tracking

ATIQUL ISLAM[1], (Graduate Student Member, IEEE), MD. ASIKUZZAMAN[2], (Member, IEEE), MOHAMMAD OMAR KHYAM[3], MD. NOOR-A-RAHIM[4], (Member, IEEE), AND MARK R. PICKERING[2], (Member, IEEE)

[1]ANU College of Engineering and Computer Science, The Australian National University, Canberra, ACT 2600, Australia
[2]School of Engineering and Information Technology, University of New South Wales, Canberra, ACT 2612, Australia
[3]School of Engineering and Technology, Central Queensland University, Melbourne, VIC 3000, Australia
[4]School of Computer Science and Information Technology, University College Cork, Cork 021, T12 YN60 Ireland

Corresponding author: Atiqul Islam (atiqul.islam@anu.edu.au)

ABSTRACT The evolution of technologies for the capture of human movement has been motivated by a number of potential applications across a wide variety of fields. However, capturing human motion in 3D is difficult in an outdoor environment when it is performed without controlled surroundings. In this paper, a stereo camera rig with an ultra-wide baseline distance and conventional cameras with fish-eye lenses is proposed. Its cameras provide a wide field of view (FOV) which increases the coverage area and also enables the baseline distance to be increased to cover the common area required for both cameras' views to perform as a stereo camera. We propose a passive marker-based approach to track the motion of the object. In this method, an adaptive thresholding method is applied to extract each small pink polyester marker from the video frames. As the cameras have fish-eye lenses, it is difficult to estimate the depth information using a pinhole camera model. We use a unique method to restore the 3D positions by developing a relationship between the pixel dimensions and distances in an image and real world coordinates. In this paper, occlusion detection is considered because, in the marker-based capturing of articulated human kinematics, the occlusion of a marker is one of the major challenges. The detection algorithm differentiates among types of occlusions and predicts any missing marker position where necessary. As this design is intended to be mounted on a moving carrier, such as a drone or car, a method for compensating the camera's ego-motion is proposed. The proposed 3D positioning and tracking system is tested in different situations to validate its applicability as a stereo camera rig as well as its performance for motion capture. The performance of the proposed system is compared with that of a standard motion capture system called Vicon and is shown to have the same order of accuracy while incurring less cost.

INDEX TERMS Motion capture, 3D positioning, stereo vision, motion tracking, high precision.

I. INTRODUCTION

MoCap, a popular nickname for motion capture, often conjures up images of a human motion capture system (MCS) that records the movements of a human and then uses the recorded data to analyze or animate them. Over the last decades, the human motion capture has drawn significant attention from both industry and academia because of the expectation that one day, they may be able to capture the human movement very precisely in a complex natural environment that does not require a dedicated control environment and sophisticated setup. Realizing such a capability could have a transformative impact on many applications domains including analysis, control, and surveillance. Attempts to develop an MCS to achieve this goal have generally relied on several technologies such as mechanical, magnetic, inertial, and optical [1]. Of these technologies, optical MCSs have been widely used because of their higher accuracy or precision. Although commercially available optical MCSs, such as Vicon [2], are very efficient at tracking the realistic motion of a subject, an expensive system setup with a large number of cameras inside a laboratory is required. Moreover, they are sensitive to lighting conditions, shadows and other factors that can interfere with light propagation. Thus, they are not suitable for outdoor environments where a large number of human activities take place. Therefore, as there is a need for low-cost and more convenient MCSs in environments in which Vicon fails, vision-based MCSs have

The associate editor coordinating the review of this manuscript and approving it for publication was Mu Zhou.

attracted the attention of many researchers, with their wide range of applications requiring different levels of accuracy. Therefore, to overcome these restrictions, lightweight stereo vision-based MCSs have been developed.

Stereo vision technology allows humans to perceive the depths of points on an object in 3D by capturing two images of the same scene from two different viewpoints using two different cameras separated by a certain distance, called the baseline distance, of a stereo system. Stereo vision methods discussed in [3]–[5] use one or more pairs of identical cameras mounted in accordance with their baseline distances and with their optical axes aligned. Reconstructing a scene in 3D is performed via epipolar rectification, feature detection, and the matching and triangulation of the correspondences of the scene from two camera viewpoints. This current study proposes to achieve motion capture using the described stereo vision technique. Although there are many techniques available, very few of them have an ultra-wide baseline distance between cameras. Despite significant efforts towards designing a capable yet lightweight and precise system for human pose estimation, the fundamental problem of obtaining comprehensive and reliable sensory information on complex environments remains to be solved.

In this paper, we use a very wide-baseline stereo camera system utilizing fish-eye lens cameras to obtain the stereo images. We use small pink polyester balls as passive markers placed on the subject body. The use of a marker reduces the computational complexity as it is not necessary to process the whole image. Creating a ground truth is an important contribution of this paper. The 3D construction of the marker's position from the ground truth also saves computational time. In marker-based motion capture, as occlusion of the marker is one of the most significant challenges, a robust algorithm is designed to overcome this problem. We also design a global motion compensation algorithm to track the local motion of an object on a moving platform. The proposed algorithm uses a reference object and rigid body transformations for the compensation of ego-motion of the cameras. The experimental results from the proposed system are compared with those from a commercially available optical MCS, Vicon, for both known and unknown trajectories. These results show that the proposed design has similar accuracy and precision to the Vicon system, but a much lower cost and structural complexity. Outdoor experiments show that there is no effect of light and shadows on the designed system. The experiment, where the target object and stereo camera rig both are moving on different platforms, shows that the ego-motion of the cameras is compensated with respect to the object's platform by the proposed algorithm. The main contributions of this paper are summarized below:

- We propose a 3D positioning and tracking system using a very wide-baseline stereo camera, which is suitable for both indoor and outdoor applications. We design the proposed system using low-cost fish-eye cameras with a very wide field of view (FOV).
- The proposed method extracts the markers from the stereo images captured by the two cameras. It is computationally efficient because we only process the extracted markers instead of the whole image.
- We develop a ground truth to establish the marker's disparity versus depth relationship, and then the depth versus pixel width and pixel height relationships to find the horizontal and vertical positions of the marker respectively.
- We design a robust algorithm to overcome the occlusion of the marker, which is a challenging problem in a marker-based motion capture system the aims to capture human motion.
- We propose an algorithm for compensating the camera motion with respect to the object's motion. When both the object and camera move, camera ego-motion compensation is utilized to provide the accurate local motion of the object.
- We perform several experiments for the validation of the proposed system in both indoor and outdoor environments, which reveals that our system achieves a similar accuracy with much lower cost and structural complexity when compared with the Vicon system.

This paper is organized as follows. Section II discusses the related studies of the proposed technique. The stereo camera calibration process is explained in Section III. Section IV describes the proposed 3D positioning and tracking method using stereo vision. The occlusion detection algorithm is given in Section V. Section VI presents compensation of the camera motion algorithm. Experimental results and analyses are given in Section VII. Finally, Section VIII summarizes and concludes this paper.

II. RELATED WORK

Many motion capture techniques have been proposed based on different methods. Human kinematic measurements are discussed in many literature surveys, such as [6], [7] while performing motion capture from a single image is proposed in [8]–[12]. As different human pose estimations cannot capture the same silhouette using a single camera, this led researchers to use multiple cameras. Moreover, the computational cost for stereo vision, coverage area, and accuracy of the depth map are also a few reasons to move from a single camera to a multi-camera system. The benefits of a multi-camera system are real-time transformation to 3D data from the image, already registered 3D image, passive 3D sensing, no projection light required and full field-of-view depth measurement to name a few.

A stereo vision-based system uses two cameras to recover the 3D pose of an object using 2D images [13], [14]. The extensive amount of literature published in the last few decades is an indication of the wide range of research that is happening in the field of stereo vision-based 3D motion tracking [15]–[18]. Recent research has shown that obtaining pose estimation and the reliable motion capture of a human in a natural scene using a stereoscopic method is achievable [3]. The automatic detection of action and behavior is also possible using an advanced algorithm. However, as this type of system is limited by the surface and visible structure of an image, achieving robustness and accuracy is difficult. Rather than relying solely on image features and structures, which makes a recognition task challenging, researchers have investigated using different types of marker-based systems. Based on the principles of stereo vision, finding marker locations in space has a variety of applications, one of which is for motion recognition [19]. Such a method consists of determining the 3D motion parameters of an object from a binocular image sequence, with its accuracy largely influenced by the cameras interior parameters and lens distortion. This kind of motion capture is limited to a closed laboratory environment rather than natural settings and depends on their algorithms being efficient in terms of robustness. Commercial MCSs provide accurate orientations and positions using active markers that emit light, typically in the infrared (IR) spectrum range. Although some emit visible light and can be estimated without processing, as they are too expensive and cumbersome for human subjects, passive markers are the preferred solution. Planar-coded markers are popular in robotic applications as they show projective behavior described by homographies [20]. They can be easily printed but, as they are usually placed on a planar object, are not visible from angular positions which make them less suitable for human subjects. On the other hand, small spherical markers may be attached to the limb of a human subject with less chance of occlusion from an angular view and as, from all projections, they exhibit circular shapes, it is easy to determine their centroids. With IR illumination, many researchers use passive retro-reflective markers in laboratory environments [21]–[23]. It has been demonstrated that passive marker-based MCSs are inexpensive and their accuracy higher than that of a magnetic system. Passive marker and stereo vision-based technologies are the best solutions for accurate and cost-effective motion capture.

Recent research in the computer vision community has focused on outdoor motion capture independent of controlled laboratory conditions. Capturing marker-based motion from a video in outdoors is considered very challenging. However, there are increasing demands for 3D models in the fields of communication, monitoring, virtual reality, surveillance, human-computer interactions, robot learning, etc. Therefore, the need for flexible and simple acquisition methods in uncontrolled environments as well as robust low-cost capturing is important for many new applications. This has motivated the use of popular consumer cameras and reconstructing human motion in 3D from multi-view video image sequences has recently become an important research field. Although multi-view human motion capture has been investigated for decades, only a few studies consider capturing human motion outdoors, with most methods using stationary camera scenarios. For motion capture in complex situations, prior knowledge of the object or scene is usually required [24]–[26]. However, this normally involves some assumptions, which eventually limit the applicability of tracking, particularly of human movements. Generally, human motion may appear very random and is difficult to model using classes of prior actions. This presents as the main challenge when markers are occluded in a marker-based method. While different approaches describe the process of switching between different motions [27], [28], even with strong prior knowledge, outdoor motion capture suffers from changes in illumination, occlusions, shadows and background clutter which conventional algorithms fail to handle. In this paper, attempts to solve these problems by means of a robust algorithm are discussed.

Motion tracking in a video captured by moving cameras usually requires motion compensation before an object is detected. The main challenge is propagating errors from motion compensation to tracking which produces further detection errors and is, eventually, computationally expensive. Interest in studying motion capture with moving cameras has increased with the development of cameras mounted on unmanned aerial vehicles (UAVs) [29], [30]. Although, because of their ever-changing FOVs, using moving cameras for motion capture is a very challenging problem, when both the object and cameras are moving almost parallel to each other, this problem is alleviated to some extent. More importantly, there is a chance of mixing the global camera and local object motions whereby the motion of the dominant camera, that is, the one that is always stationary with respect to the object, has to be compensated. Initially, researchers used graph matching or video registration to capture motion with moving cameras [31] and then background subtraction and the detection of a moving blob of interest. However, these methods suffer from difficulties such as video alignment problems, perspective distortions, and tracking and localization errors. A very basic problem of motion capture is feature extraction that can be greatly overcome using the marker-based approach in which representing the motion features is crucial. A motion trajectory is another well-known means of representation as it is compact, informative and spatio-temporally continuous [32], [33] for extracting tracking data. In a marker-based system, it is relatively easy to obtain the trajectory of each individual marker attached to the human body that represents a specific limb. Trajectory acquisition has been widely studied [34]. For example, multiple motion trajectories are tracked in [35] to analyze human activities. However, there are also many challenges associated with this approach, for example, the good features selected beforehand for tracking might be noisy, they sometimes do not represent the desired action and a discontinuity may occur in a trajectory due to a correspondence mismatch. In the method proposed in this paper, an added benefit is that no trajectory modeling is required as the position of each marker in each frame provides a clear trajectory of a human kinematic movement.

III. STEREO CAMERA CALIBRATION

In our proposed work, two low-cost Go-Pro Hero4 Black [36] cameras are attached to an aluminium bar with a baseline distance of one meter, as shown in Fig. 1. These are commercially available sport action cameras, each with a fixed focus and 12-megapixel ultra-wide FOV. In the proposed design, sport action cameras are used because of several benefits. The cost and robustness of these cameras, and small form factor are big reasons for selecting them while the fish-eye lens gives a very wide coverage of 170 degrees. These cameras also have high and varying frame rate, which is also a benefit for this work. Small pink polyester balls are placed on a subject's body to work as passive markers representing the torso for capture by the cameras. The cameras are synchronized using the camera remote, which can simultaneously turn on and off both cameras.

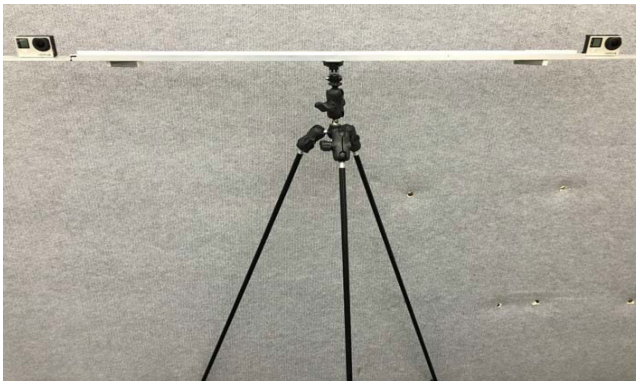


FIGURE 1. Stereo camera rig on a tripod with a baseline distance of one meter.

Camera calibration is a crucial part of a computer vision-based MCS. Most camera lenses impart a fixed quantity of distortion on a captured image, the rectification of which is essential for accurate transformation from the image coordinates to real-world coordinates. A calibration method delivers three significant pieces of information. First, the intrinsic parameters provide the focal length, a pixel skew and the principal point of the camera. Second, the extrinsic parameters represent the location and orientation of the camera on the world's axis. Third, the lens distortion coefficients provide artefacts of the camera's lens. Generally, distortions are either radially symmetric about the principal point or symmetric along a line passing through the principal point [37]. In barrel distortion, the points in an image are concentrated on the border and, in the center, spread in a radial direction whereas pincushion distortion exhibits contrary behavior as images shrink towards the center [20]. The cameras used in the proposed method show barrel distortion. Fig. 2(a) shows the image pair taken from two Go-Pro Hero4 Black cameras. The objects in the lab which are straight, such as shelves and pipes, appear curved in the image because of barrel distortion. The distortion increases with the distance from the center of the image. Distortion correction, or camera calibration, consists of approximating a camera's intrinsic parameters which are expressed mathematically by the following k matrix:

$$k = \begin{bmatrix} \alpha & 0 & u_0 \\ 0 & \beta & v_0 \\ 0 & 0 & I \end{bmatrix} \tag{1}$$

where u_0, v_0 is the image's center, called the principal point, α and β the focal lengths measured in the u and v pixel units respectively and I the identity matrix [38].

Camera calibration is well developed and has been widely studied in the literature, as mentioned in [37], [39], [40]. In our proposed work, the intrinsic and extrinsic parameters of each camera were determined using the method proposed by Zhang in [39]. Using multiple views of a planar calibration board is a flexible and convenient way of calibrating a stereo camera because, as each camera has a fixed focal length, its intrinsic parameters are always fixed, with only its extrinsic ones varying depending on the camera setup. A large square checkerboard calibration sheet with dimensions of 108 mm×108 mm is positioned in the FOV of both cameras at different positions and a synchronized video taken to obtain pairs of images for a range of perspective views of the pattern. Then, lens distortions are corrected and the images are rectified using the method described in [41], with an overall mean re-projection error of 0.50 pixels and the resulting images having epipoles at infinity. In order to reduce the correspondence search space between the left and right images in stereo vision processing, epipolar rectification is an important step which reduces computational efforts [20]. It ensures that each common marker observed by the two cameras is imaged on the same row. The image pairs before and after calibration are shown in Fig. 2. The figure shows that the image pair in Fig. 2(a) becomes vertically more aligned and the barrel distortion has been corrected in Fig. 2(b).

IV. PROPOSED 3D POSITIONING AND TRACKING SYSTEM

In this work, the two cameras are placed at a baseline distance of one meter. The cameras calibrated according to the method, as previously discussed in Section III, are used and arranged in a stereoscopic pattern to reconstruct the 3D coordinates of a markers' points, i.e., the centroids of the markers. The proposed 3D positioning and tracking system, which is based on the positions of the markers on the stereo image planes will be discussed in the following sub-sections.

A. PREPOSSESSING OF MARKERS

1) MARKER EXTRACTION

Marker extraction is an important part of a marker-based tracking method. The triangulation accuracy (i.e., tracking accuracy) is basically dependent on how accurately the locations of the markers are extracted from a scene. Reliably differentiating the markers of an image frame from the rest of the scene is the first step. While many proposed segmentation techniques are mainly application-orientated, we use color-based segmentation which, although requiring

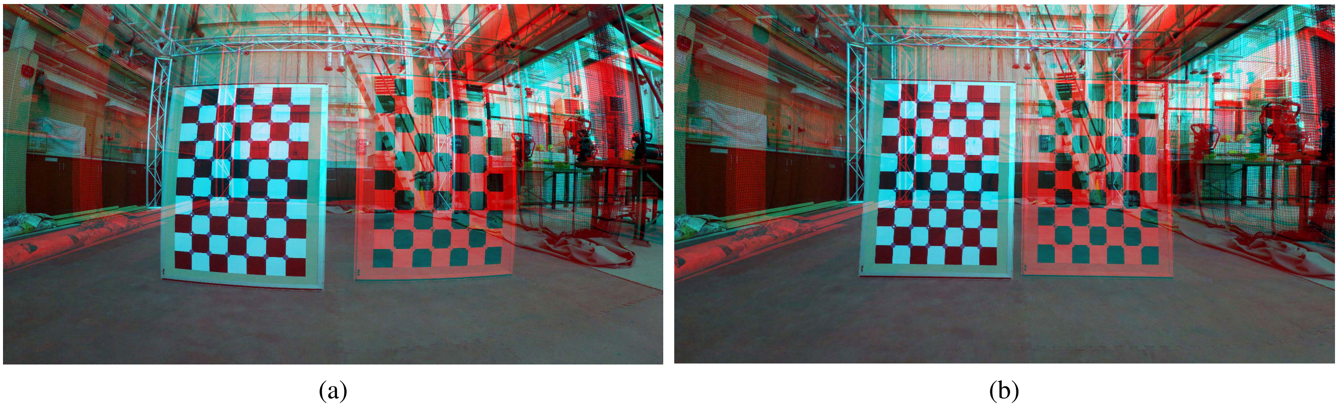


FIGURE 2. An example of the outcome of stereo camera calibration. Anaglyph stereo image pair (a) before and (b) after calibration. The barrel distortion in (a) has been corrected in (b) after calibration.

a controlled environment, works comparatively well if its computational speed is considered. In order for it to work properly, the markers must be of a distinctive color not seen anywhere else in the scene. Also, altering the lighting conditions may distort segmentation as a segmented blob changes according to the direction of the light. However, we use the centroid of a blob for our final calculations which partially solves this problem. Also, as the markers are spherical, the same shaped blob is projected on the image plane regardless of the location of any camera. These markers are made of very lightweight pink polyester balls of sizes that can be attached to the object and provide a strong illumination gradient. A block diagram of the marker extraction process is shown in Fig. 3. Generally, a camera captures a color image (RGB in this case) which is then pre-processed (color-space conversion) in order to perform segmentation, with the computational speed the main trade-off between accuracy and performance. As processing a whole frame at high resolution would be too computationally expensive, this is the main concern regarding practical real-time applications. Firstly, an input image is converted to the LAB color space using the transformation:

$$\begin{bmatrix} X \\ Y \\ Z \end{bmatrix} = \begin{bmatrix} 0.4124 & 0.3576 & 0.1805 \\ 0.2126 & 0.7152 & 0.0722 \\ 0.0193 & 0.1192 & 0.9505 \end{bmatrix} \begin{bmatrix} R \\ G \\ B \end{bmatrix} \tag{2}$$

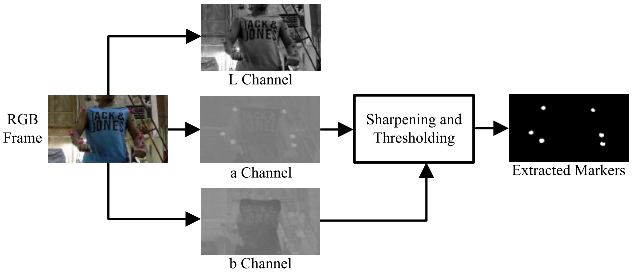


FIGURE 3. Block diagram of the marker extraction process. It takes an RGB image of the markers on the object as an input and produces a binary image containing only the markers.

$$\begin{aligned} L &= 116Y - 16 \\ a &= 500 \times (X - Y) \\ b &= 200 \times (Y - Z) \end{aligned} \tag{3}$$

Then, the markers are separated from the background image by applying an adaptive thresholding technique which offers good precision for an image with a strong illumination gradient. Next, the image is sharpened and the local thresholding value calculated. Finally, from the binary image for every extracted marker, the following geometric properties are calculated: area (A); number of identified pixels fitted; geometric centroid (C); and unit vector which defines the orientation (O) in degrees of its angle with respect to the image's coordinates.

2) MARKER CLASSIFICATION

After identifying the markers, a marker classification procedure, which aims to identify each marker's position and keep its label the same from the first to the last frame of the video, is applied. Both vertical and horizontal sorting of the centroid is applied for the first frame as it is considered the standard position of the subject. For subsequent frames, when the subject changes position, the algorithm chooses the correct marker when the following criteria are satisfied: E_l and E_r are minimum, where E_l and E_r are the Euclidean distances of each marker in the previous frame to all markers in the current frame. The algorithm propagates through the set of images and finds Euclidean distances between frames L_j and L_{j+1}, and R_j and R_{j+1}, where L and R denote the left and right camera frames and $j = 1, 2, \cdots, N$ the number of the frames. If E_l or E_r is greater than a certain threshold value, the marker is not correctly identified. The difference in the area of each marker between the previous and current frame should also be below a predefined threshold.

3) CORRESPONDENCE OF MARKERS

For a marker-based tracking problem, correspondence matching is the most important part. In this paper, the epipolar geometry is used to search for correspondences.

It demonstrates the relationship between a pair of cameras observing the same scene, with the correspondence of a marker involving matching a marker in the left view with the same marker in the right view. Once a marker is identified in the left view, its depth could lie on any position of a line passing through the image plane. Using epipolar geometry, this line will map to a different line in the right camera's image plane.

B. ESTIMATION OF DEPTH FROM DISPARITY

The depth of a marker is determined from its disparity in the left and right images captured by the two cameras. A disparity versus depth relationship, which is very robust in terms of performance, very simple and saves computational time, is developed. Firstly, a process of generating a ground truth is conducted to convert 2D camera image coordinates to 3D world ones whereby a 140 mm×90 mm rectangular object is progressively moved away from the cameras' stereo setup. The reason behind this is to find disparity for every known depth. Its first position is 500 mm from the camera and then it is moved in regular intervals of 50 mm to 4450 mm away, with a video frame captured for each position and its dimensions in each frame in each position measured in pixels. With one-meter baseline distance, in the first few positions, the rectangular object is partially seen by the cameras thus ignored in the experimental result of ground truth. The experimental results of this experiment are used as the ground truth for depth estimation. An illustration of the experimental process is shown in Fig. 4.

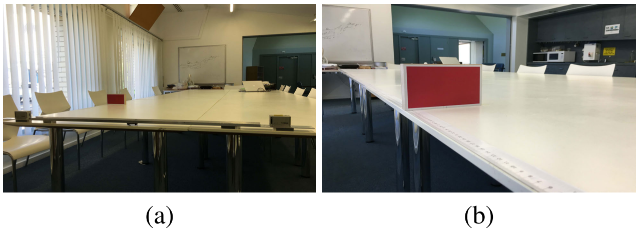


FIGURE 4. Experimental setup for estimating ground truth of depth measurement (a) setup and (b) rectangular object.

These experiments provide an indication of the proper image resolution to choose. As it is a conversion between an image's axis in pixels to a world axis in millimeters, a high image resolution gives more disparity and, consequently, clearer depth information. Fig. 5 shows the disparity and distance relationships for different resolutions, for example, 960p (960×1080), 1080p (1080×1920) and 4k (2160×3840) resolutions, with the curves indicating the coverage limit of the stereo rig. The maximum and minimum distances from a camera's FOV at which a subject can be captured are clear in the graph. With higher resolution, a minimum error can be achieved as the number of pixels increases and the pixel dimension decreases. Therefore, in this paper, the resolution of 4K is exploited for experimental purposes. Fig. 6 shows the relationship among the pixel sizes in millimeters in the x and

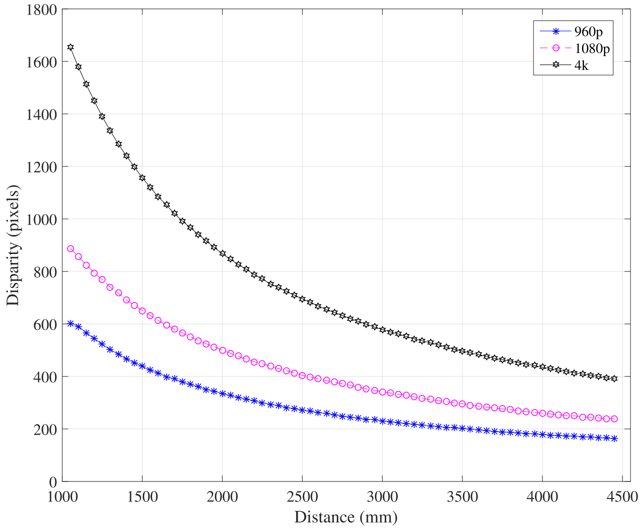


FIGURE 5. Distance (mm) versus disparity (pixel) at different resolutions.

y directions as well as the disparity in each position obtained from this experiment. In Fig. 6(a) the disparity versus distance relation is shown where a 6th degree polynomial is fitted on the original curve which is denoted by a black line and the fitted curve is presented by a red line. This 6th degree polynomial is given by:

$$d = a_1D^6 + a_2D^5 + a_3D^4 + a_4D^3 + a_5D^2 + a_6D + a_7 \tag{4}$$

where D is the disparity between the objects in the stereo images, in pixels, and d is the depth of the image point in world coordinates, in millimeters. Fig. 6(b) and Fig. 6(c) show the distance versus pixel width and height respectively. These curves are fitted with a quadratic equation as shown by the straight line. A pixel's height and width, in millimetres, at any distance, can be determined from the following equations using d obtained from the disparity vs distance relation in (4):

$$\begin{aligned} s_x &= b_1d^2 + b_2d + b_3 \\ s_y &= c_1d^2 + c_2d + c_3 \end{aligned} \tag{5}$$

where the multiplication factors s_x and s_y, in millimeters per pixel, are used to convert from image to world coordinates in the horizontal and vertical directions of the video frame respectively.

C. 3D RECONSTRUCTION

For the 3D reconstruction of marker positions using fish-eye lenses, a unique approach for transferring from an image to the world axis is used. Fig. 7 shows the overall block diagram of the 3D reconstruction process. Firstly, the left and right video sequences are taken as input. After the marker extraction and classification processes as described in Section IV-A, all the marker coordinates are stored for both left and right frames. A marker's coordinate position in the pixels in the left frame is considered (x_l, y_l) and (x_r, y_r) in the right frame.

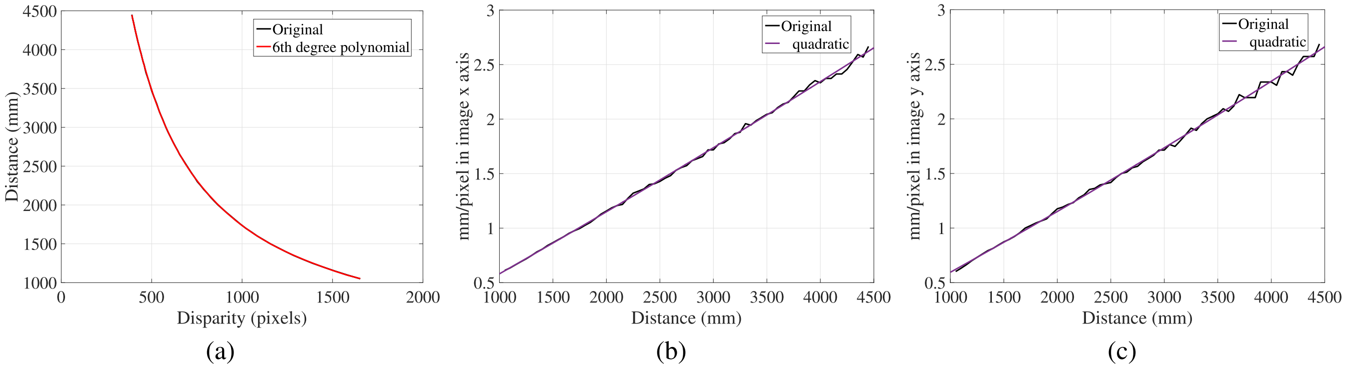


FIGURE 6. 2D to 3D relationships: (a) disparity (pixel) versus distance (mm); and pixel size (mm/pixel) at different distances (mm) in (b) horizontal and (c) vertical directions of a video frame.

It should be noted that $y_l \simeq y_r$, as the stereo images are rectified. For corresponding markers in the left and right frames, the disparity D is given by:

$$D_i = x_l(i) - x_r(i) \tag{6}$$

where $i = 1, 2, \cdots, m$ for the m markers in a frame. The depth d (i.e., the z coordinate in millimeters) is determined from the disparity D mentioned in (6) using (4), with the x and y coordinates, in millimetres, of a marker calculated using the following equations:

$$x_i = x_l(i) \times s_x \tag{7}$$

$$y_i = y_l(i) \times s_y \tag{8}$$

where

$$s_x, s_y \propto d \tag{9}$$

This operation is performed on each frame in the video sequence as shown in the block diagram in Fig. 7.

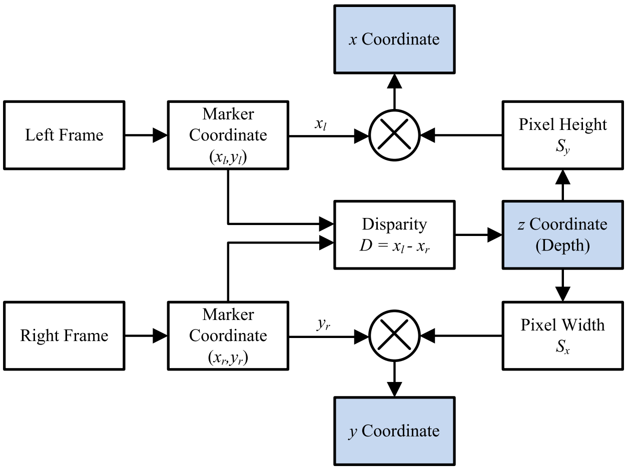


FIGURE 7. Block diagram of the proposed system for 3D reconstruction from video data.

V. OCCLUSION DETECTION

Occlusion is a major problem for articulated human motion capture in a marker-based approach. As occlusion can be caused by several scenarios, classifying the different types of occlusion makes determining a solution easier. Markers may disappear completely from a single frame or for an interval of frames and they may disappear from one or both camera images. For example, Fig. 8(a) shows the binary image of frame 140 in which all seven markers are present. In Fig. 8(b) an elbow marker is missing from frame 150 but then reappears in frame 160, as shown in Fig. 8(c). This situation can be defined as follows: For the set of markers M_i ($i = 1, 2, \cdots, m$, where m is the number of markers), any of these marker can be missing in the j^{th} frame (where $j = 1, 2, \cdots, N$, and N is the number of frames).

The above situation has several causes: the markers could be occluded; perfectly overlap each other; or be partially overlapping to appear as a whole. In this paper, only these three situations are considered. A block diagram showing the occlusion handling process is shown in Fig. 9. The first step is identifying the proper means for handling occlusions. Two different identification processes are used based on the 2D image of a marker blob and feedback from its 3D output location. For dealing with marker occlusions, in every experiment using the proposed system, the first frame is considered the reference one and assumed to have a standard position.

To solve the occlusion problem in a 2D image, the algorithm firstly identifies the proper circumstances based on the area A and orientation O of each marker. If there are less than the maximum number of markers and one marker exceeds an area and orientation threshold, a partial occlusion is deemed to have occurred. The large marker forms an ellipse and is then separated in the 2D image. If there are missing markers but no unusual blob size in the 2D image, an actual occlusion or full overlap is deemed to have occurred and the second identification process using feedback from the 3D locations of markers in the previous frame is performed. In the case of an actual missing marker, the inter-frame and inter-view correspondence algorithm assigns a new position to it based on knowledge of the previous frame. The predicted position

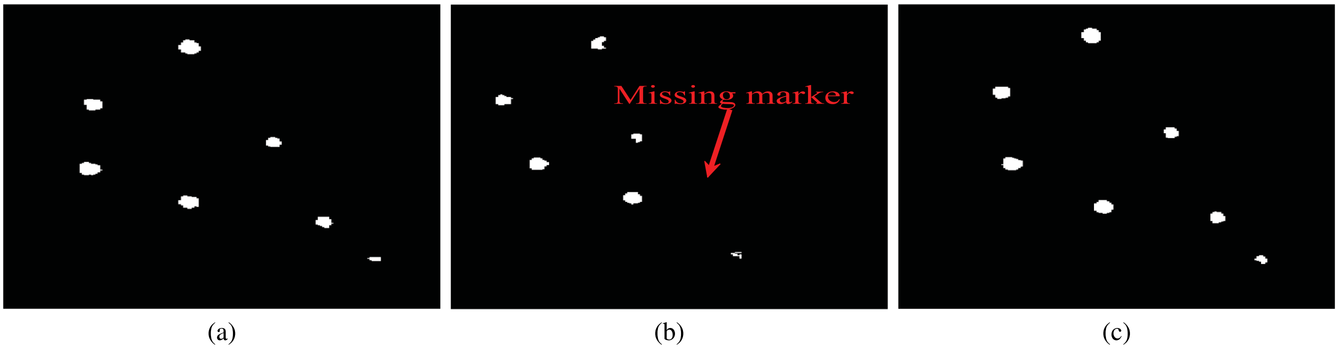


FIGURE 8. Positions of markers in different binary frames: (a) no occlusion in frame 140; (b) one marker occluded in frame 150; and (c) previously occluded marker reappearing in frame 160.

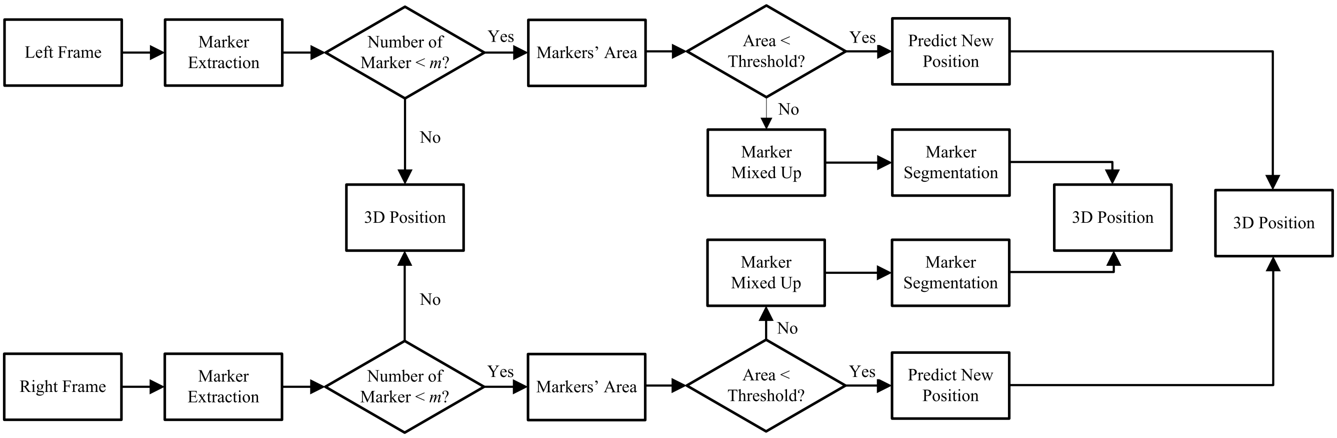


FIGURE 9. Flow chart of occlusion handling algorithm.

Algorithm 1 3D Position Synchronizing With Previous Frame

```
Require: $R$: Reference 3D position, $T$: Target 3D position,
    $m$: Number of marker, $N$: Number of frame
Initialize: $R = F(1)$ {Frame one as reference frame}
Initialize: $T = F(2)$ {Frame two as target frame}
Initialize: $T' = 0$ {New target position}
for $j := 2$ to $N - 1$ do
    for $i := 1$ to $m$ do
        for $k := 1$ to $m$ do
            $E(i, k) = ||R(i) - T(k)||$ {The Euclidean distance
            between $R$ and $T$}
        end for
        $I = \text{find}\,(E(i,:) = \min E(i,:))$
        $T'(i) = T(I)$
    end for
    $R = T'$ {Update reference}
    $T = F(j + 1)$ {Update target}
end for
```

then gives the 3D location of the marker. The predicted 3D location is validated by a feedback algorithm as shown in Algorithm 1. The distances of one marker in the current frame from each marker in the previous one are calculated. An accurately estimated marker in the current frame will show the minimum distance between themselves as the displacement of two consecutive frames is very low. If this situation continues for a long period, a linear interpolation is used to predict the trajectory of the marker. In the case of two markers being mixed up, they are separated in the 2D binary image based on their areas and orientation angles as they usually form an elliptical shape and the algorithm proceeds to obtain their 3D positions. As previously discussed, each 3D position is fed back to the previous frame and, if the minimum distance condition is satisfied, the algorithm proceeds to the next frame. If not, the marker is considered an actual missing one rather than a mixed one and the algorithm follows the rule for an actual missing marker.

VI. CAMERA MOTION COMPENSATION

An important challenge for motion capture using a moving camera is compensating for the camera motion, i.e., the ego-motion or camera motion compensation with respect to an object's motion. When an object is moving on another moving platform and the camera is also moving, it is difficult to capture the object's motion using this camera as there are two separate motions between the object and camera, that is, a global one between the platforms of the camera and object, and the object's local motion. Compensating the ego-motion of the camera will provide the accurate local motion of the subject for which a transformation-based system is proposed. In this method, a reference object is included in the subject's platform and its position tracked before the desired object's motion is tracked. A 'T'-shape composed of seven markers shown in Fig. 10 is considered as the reference object. The first position, i.e., the position of the object in the first frame is the reference's position to which the subsequent one (the target's position) is transformed by selecting proper rotation and translation parameters. A rigid-body transformation as shown in the block diagram in Fig. 11 is used. As the tracking is not dependent on the background, the problem of a moving background is eliminated. A moving object could be tracked by noting the reference attached to it while its velocity or background are not factors to be considered unless they were in the camera's FOV. The difference between two consecutive frames is compensated for by means of aligning the 'T'-shapes in the two frames. As they were merged by means of a rigid transformation and their displacement parameters fed back every time to the main algorithm for capturing the target object's motion, the cameras were effectively stationary with respect to the 'T'-shape in each frame. If a transformation T is applied from a set of points p to another set of points p', this can be mathematically expressed as:

$$p' = Rp + t \tag{10}$$

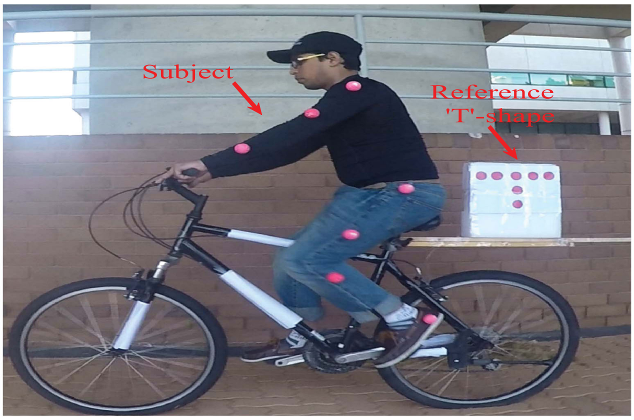


FIGURE 10. Subject with reference 'T'-shape in a single frame where both camera and subject are moving.

where R and t represent the rotations and translations respectively and are expressed as:

$$t = \begin{bmatrix} 1 & 0 & 0 & t_x \\ 0 & 1 & 0 & t_y \\ 0 & 0 & 1 & t_z \\ 0 & 0 & 0 & 1 \end{bmatrix} \tag{11}$$

where t_x, t_y and t_z are the translations in the x, y and z directions respectively, with the rotation R expressed as:

$$R = R_x R_y R_z \tag{12}$$

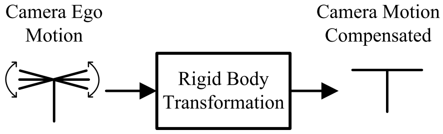


FIGURE 11. Block diagram of transformation technique for compensating camera motion.

where R_x, R_y and R_z are the rotations in the x, y and z directions respectively which can be defined as:

$$R_x = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & cos\theta & -sin\theta & 0 \\ 0 & sin\theta & cos\theta & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \tag{13}$$

$$R_y = \begin{bmatrix} cos\theta & 0 & sin\theta & 0 \\ 0 & 1 & 0 & 0 \\ -sin\theta & 0 & cos\theta & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \tag{14}$$

$$R_z = \begin{bmatrix} cos\theta & -sin\theta & 0 & 0 \\ sin\theta & cos\theta & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \tag{15}$$

As the numbers of points in the reference and target images are considered n, a n pair of matching points p and p', and the rotation R, the transformation matrices t will be [42]:

$$t = p'_0 - Rp_0 \tag{16}$$

where

$$p_0 = 1/n \sum_{i=1}^{n} p_i \tag{17}$$

and

$$p'_0 = 1/n \sum_{i=1}^{n} p'_i \tag{18}$$

denotes the centroid of the set of points and the rotation R which can be estimated by minimizing the error E_r as:

$$E_r = 1/n \sum_{i=1}^{n} |p'_i - Rp_i|^2 \tag{19}$$

The rotation and translation obtained in each frame are applied to the desired object to be tracked in each frame to achieve its actual position. Fig. 12 shows a block diagram of the proposed video stabilization process whereby, for each image in the input video sequence, features of the reference 'T'-shape are detected. The first frame is considered the reference one and assumed to be stabilized with respect to the camera, with the 'T'-shape in the next frame compared with it using rigid transformation. Then, starting from the second frame, each frame with the desired object is fed back with the parameters found from the reference shape to get the actual position of the desired object.

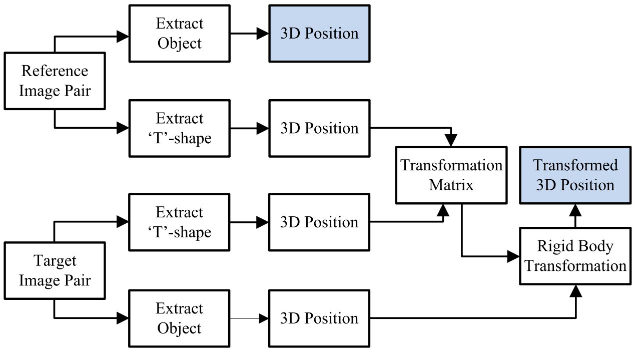


FIGURE 12. Block diagram of proposed camera motion compensation process.

VII. EXPERIMENTAL RESULTS AND ANALYSIS

In this section, the experimental data captured with the designed stereo camera rig and processed with the proposed algorithms were analyzed. Literature shows that many state-of-the-art 3D positioning and tracking methods have used known trajectories for evaluation purposes where tracking error was measured with respect to the known trajectories [43], [44]. However, in this paper, experiments were conducted to validate the accuracy of the proposed method by comparing with a Vicon MCS [2]. The Vicon system is considered as the 'gold standard' for motion capture and used as the ground truth because it offers sub-millimetric accuracy. It uses IR reflective markers and each of its Bonita cameras stream the accurate motions of the markers in a video sequence as well as optically whether placed on people, animals or machines. We conducted the experiments using the Vicon system with 14 Bonita cameras at the University of New South Wales (UNSW), Australia and the stereo camera rig described in Section III. The experimental setup was placed in the center of the Vicon's FOV so that a sufficient number of Vicon cameras could see the Vicon markers, which were placed on top of the camera markers. The data was collected at 30 frames per second (FPS) using the camera-based and Vicon MCSs, and the pair of cameras triggered by the remote. As the Vicon system provided coordinates from the origin and the camera system from the camera's center, an additional Vicon marker was set on the camera-based system for accurate transformation. To evaluate the proposed method, a set of experiments was conducted according to the procedures described in the following sub-sections.

A. EXPERIMENTAL ANALYSIS: STATIC CAMERA

We performed several experiments for the validation of the proposed system with respect to Vicon. The experiments in this section are classified into three groups. Firstly, the static camera with static object situation which was represented by a 3D grid recognition. Secondly, the static camera with a moving object but known trajectory situation which was represented by a pendulum tracking. Finally, the static camera with a moving object where the movement was unknown which was represented by tracking an articulated human subject. These experiments gave an idea about the accuracy of the measurement of 3D grid mesh, pendulum trajectory and articulated human movement by the proposed system. The trajectory of the pendulum and position of the 3D grid were measured by physical means. So, accurately measuring the trajectory and position indicates the accuracy of the proposed system.

1) 3D GRID PATTERN

The first situation is described in this section where the camera rig was static and the tracked object was also static. In this experiment, seven markers were placed in a vertical line on a rigid board and the board was stood upright vertically, with a similar setup for nine equidistant positions forming a 3D cube shape. The data were collected simultaneously by the proposed stereo camera rig and Vicon MCS, and the performance of the proposed system was evaluated by comparing their results.

Figure 13 shows a quantitative comparison of the markers' centers estimated by both systems. The system produced all three coordinates of each marker with respect to the world coordinates provided by the Vicon and arranged them in a 3D grid pattern. Fig. 13(a) and Fig. 13(b) show the errors after the 3D affine and rigid body transformations respectively. The proposed system was transformed to the Vicon system and the minimum root mean square errors (RMSEs) for the 3D affine and rigid body transformations were found to be 1.5576 mm and 8.3588 mm respectively. Table 1 shows the RMSEs of the proposed system relative to the Vicon system in x, y and z directions and overall. Because the error of a marker's position also depended on the proper placement of a Vicon marker in its centroid, this error may not have characterized the actual error of the system. Although, if the placement of a Vicon marker on a centroid were perfect, the error would be minimized, this couldn't be achieved by a bare eye measurement. This experiment provided an indication of any remaining distortion in the proposed system as it was clear that, as the distortions of a fish-eye lens were almost minimized by the algorithm, the results remained close to those obtained from the Vicon system. Therefore, with the proposed approach a static object was tracked with minimum distortion and errors. The next validation was to find out whether the proposed system can track a moving object properly.

TABLE 1. Root mean square errors of 3D grid pattern for the affine and rigid transformations from the proposed to Vicon system.

Transformations	RMSE (mm)			
	Directions			Overall
	x	z	y	
Affine	0.9169	2.2353	1.2004	1.5576
Rigid Body	9.1743	10.3023	4.3933	8.3588

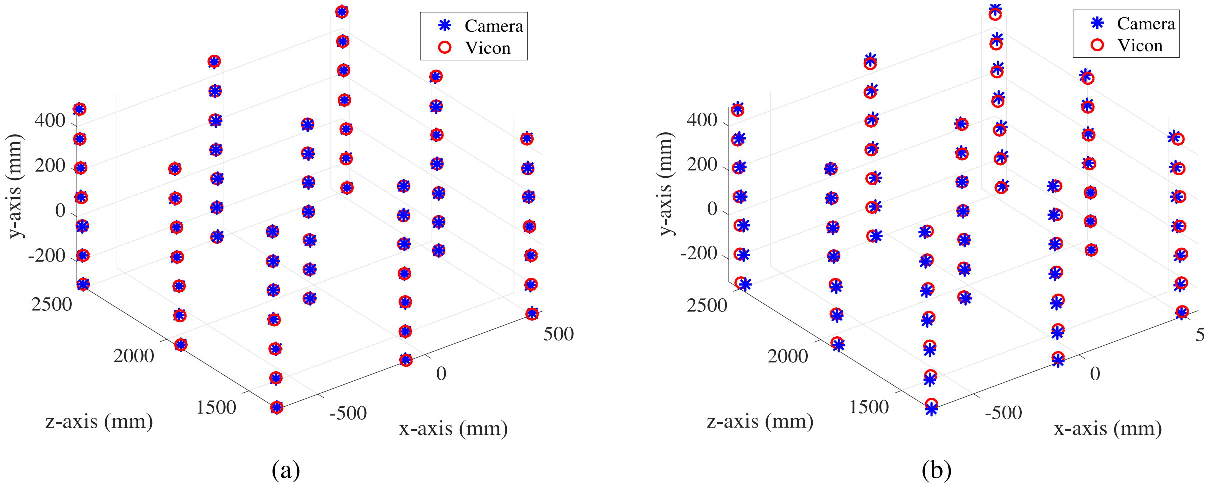


FIGURE 13. Performances comparison of the proposed and Vicon systems for 3D grid pattern where the 3D coordinates are transformed from the proposed camera-based to the Vicon system using (a) affine and (b) rigid body transformations.

2) PENDULUM TRAJECTORY TRACKING

For tracking a moving object when the proposed camera rig was static, first, a known movement was tracked which gave the idea about any deviation because of distortion or other errors. In this experiment, a known trajectory was considered to determine the accuracy of global estimations using the custom-made pendulum shown in Fig. 14 with one camera marker on it. The pendulum was manually pulled away from the bottom point and slowly released, with data collected using the camera-based and Vicon MCSs. Then as in the previous experiment, affine and rigid body transformations were conducted for the data obtained from the camera-based and Vicon systems where the proposed camera-based system was converted to Vicon coordinates, as shown in Fig. 15. These transformations give an idea about the variation of results from Vicon for the proposed system. Both of the transformations were done as the position of markers from Vicon and the proposed system were not the same centric due to manual placement. Table 2 shows the RMSEs of the proposed system with respect to Vicon for both affine and rigid transformations. The experimental results indicate that the proposed system produced a trajectory that was almost identical to that given by the Vicon MCS. To demonstrate the accuracy of the proposed system when tracking a moving object, the experiment was further conducted with multiple marker patterns, as discussed in the next section.

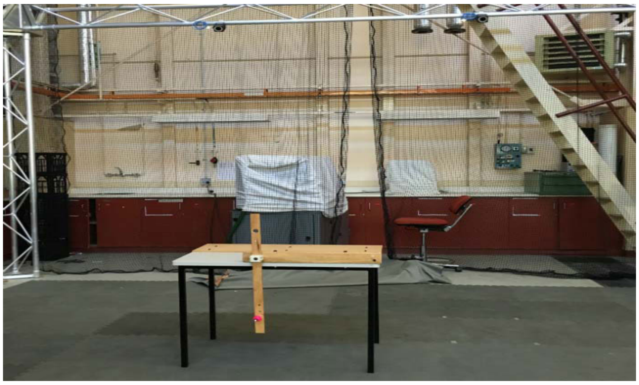


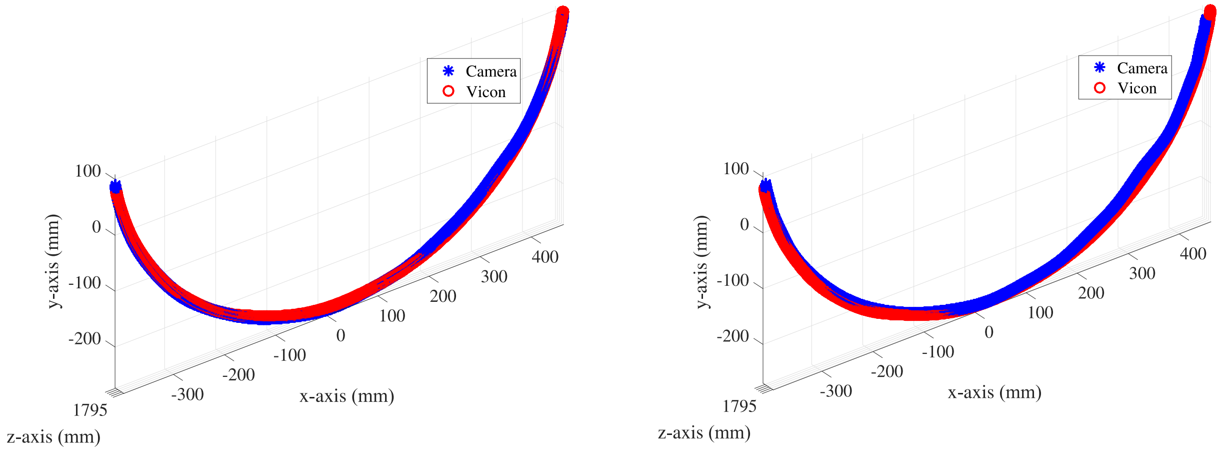
FIGURE 14. Pendulum trajectory tracking setup in the Vicon lab.

TABLE 2. Root mean square errors of pendulum trajectory tracking for the affine and rigid transformations from the proposed system to the Vicon system.

Transformations	RMSE (mm)			
	Directions			Overall
	x	z	y	
Affine	1.2847	0.2707	4.1270	2.5004
Rigid Body	5.9756	2.4459	5.4255	4.8692

3) HUMAN MOTION TRACKING

Prior to the use of our camera-based stereo motion capture system in an experimental context, in its developmental stage, substantial validation efforts were undertaken to assess its overall accuracy. At this stage, the third condition was validated where the camera was stationary but the object was moving in an unknown manner. A human subject was asked to perform kayaking movements in indoor Vicon environments in different situations, and simultaneous capturing performed using both the Vicon and camera-based systems. Seven markers were used to represent the subject's upper body limbs. One marker is placed on the subject's forehead and others on his/her left and right shoulders, elbows, and wrists. The calculated locations of each marker in each frame using the proposed system and the Vicon systems are shown in Fig. 16, and the RMSEs are shown in the Table 3. Regarding individual errors in the x, y and z directions, it can be seen that the error in the z-direction (depth) was greater than those in the x and y-directions. This was because the errors in



(a)

(b)

FIGURE 15. **Performances comparison of the proposed and Vicon systems for pendulum trajectory tracking where the 3D coordinates are transformed from the proposed camera-based to the Vicon system using (a) affine and (b) rigid body transformations.**

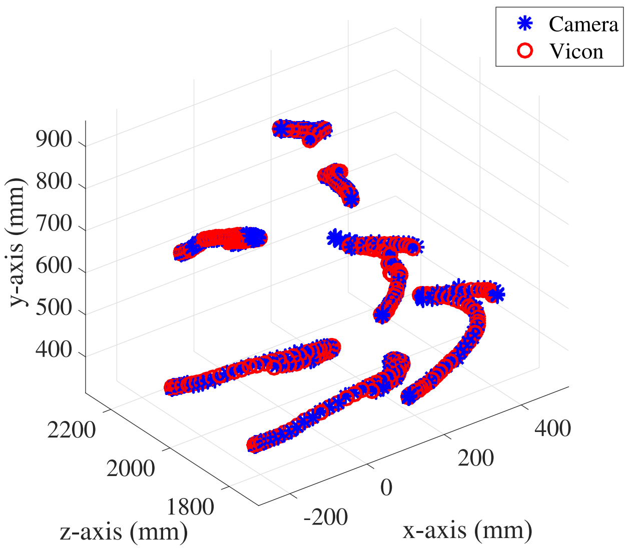


FIGURE 16. **Performance comparison of the proposed and Vicon systems for 3D reconstructions of marker trajectories obtained from one cycle of kayaking paddle strokes.**

TABLE 3. **Root mean square errors of marker trajectories obtained from one cycle of kayaking paddle strokes for the affine and rigid transformations from the proposed to Vicon system.**

Transformations	RMSE (mm)			
	Directions			Overall
	x	z	y	
Affine	2.8712	6.9809	1.6829	7.7336
Rigid Body	2.7214	6.9414	2.1202	7.7514

marker locations were magnified in the process for converting disparity to depth.

The accuracy of the proposed system decreased due to a combination of calibration and rectification errors, the algorithm's performances for matching correspondences and deficiencies in the procedure for marker extraction. In the marker extraction process, the principal cause of the error was motion blur when the markers moved very fast. This made it difficult to estimate the proper shapes and centroid of the markers and, subsequently, find correspondences among frames. These difficulties can be overcome by a higher frame rate camera because with high frame rate motion blur can be minimized. The cameras used in this study are of high FPS but with the highest resolution, only 30 FPS can be used. With a higher frame rate than the one currently used, the error could be reduced even further. The proposed marker extraction and 3D reconstruction procedures can be used to reconstruct the 3D position of a subject in an outdoor environment.

4) COMPENSATION FOR OCCLUSION

The occlusion problem is present for motion capture of any articulated human movement using a marker-based system, as moving limbs can block the markers of other body parts. In this experiment, a movement was performed with an occlusion, as shown in Fig. 17(a) in which the left elbow marker is occluded by the right hand. The experimental result without considering the occluded marker is shown in Fig 17(b) and then the results after the occlusion detection algorithm was applied are shown with the Vicon data in Fig. 17(c). As shown in this figure, it is clear that occlusion was compensated as both sets of data follow a similar path. The predicted marker's position in the occluded area is compared with Vicon when considering Vicon as the true position of the occluded marker. Table 4 shows the RMSE in x, y and z directions and overall RMSE for the markers of plotted area for affine and rigid body transformation. The result shows lower RMSE than that achieved in Section VII-A3, this is because only a few positions of the marker near occlusion region are considered rather than the complete cycle of a back and forth movement. This reduces RMSE as both camera and Vicon in a straight

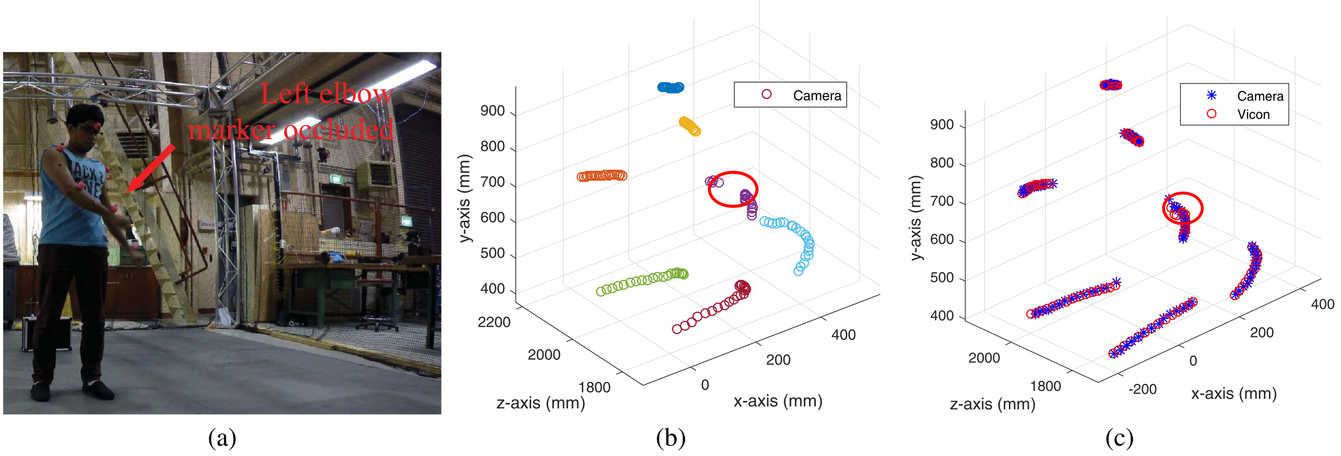


FIGURE 17. (a) Occlusion of marker in the frames of a video sequence, (b) 3D reconstructions of marker trajectories of kayaking paddle strokes with occlusion using the proposed method and (c) performance comparison of the proposed and Vicon systems for 3D reconstructions of marker trajectories of kayaking paddle strokes after occlusion correction.

TABLE 4. Root mean square errors of marker trajectories of kayaking paddle strokes after occlusion correction for the affine and rigid transformations from the proposed to the Vicon system.

Transformations	RMSE (mm)			
	Directions			Overall
	x	z	y	
Affine	1.5171	1.4299	1.0401	2.3298
Rigid Body	1.5573	1.5519	1.1195	2.4671

line of the trajectory without any reversing cycle of the back and forth motion.

B. EXPERIMENTAL ANALYSIS: MOVING CAMERA

For the compensation of camera ego-motion, the proposed stereo camera-based system is validated in two different situations: a moving camera with a static object and a moving camera with a moving object. In both cases, firstly, the motion tracking procedure described in Section IV was performed and then the camera motion compensation algorithm proposed in Section VI was applied.

1) MOVING CAMERA WITH STATIC OBJECT

To validate this case, in which the object was static but the camera moving, a board with a 'T'-shaped markers was stood upright and the camera mounted on a tripod. After starting to record the video sequence, the cameras were intentionally vibrated. For a better understanding, it was assumed that the camera rig observed a stationary scene while experiencing jitters. Fig. 18 shows the images in different frames from the video sequence of an indoor static scene in which it can be seen that their edges are not similar because of camera jitter. This was due to the 3D rotations and translations of the cameras. Fig. 19(a) illustrates the object's position due to jitter in the camera and Fig. 19(b) shows it retaining its original position after implementation of the camera motion compensation algorithm described in Section VI. It can be seen that the positions of the markers in different frames of a video sequence retained the 'T'-shape, i.e., the markers' positions are plotted exactly one over another. The obtained parameters from the camera balancing algorithm were fed back to the 'I'-shape in Fig. 19(c) and as the object was stationary, it got its desired stabilized position as shown in Fig. 19(d).

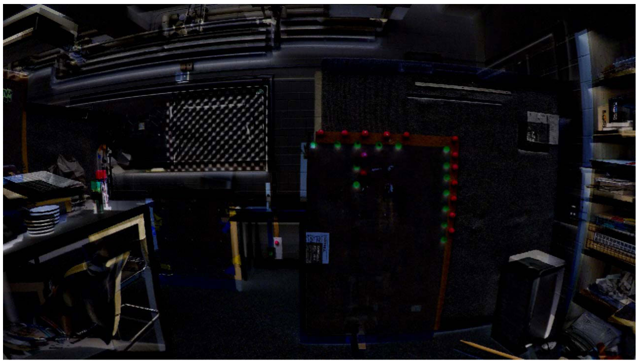
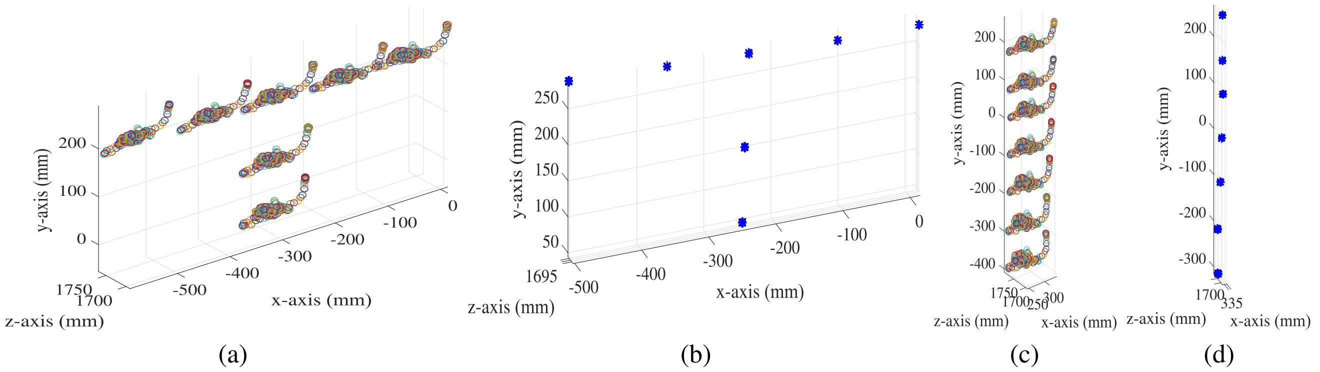


FIGURE 18. Difference between frame 1 and frame 150 in video sequence of stationary objects captured by moving camera.

TABLE 5. Root mean square errors of static 'T'-shape and 'I'-shape objects after applying camera motion compensation technique.

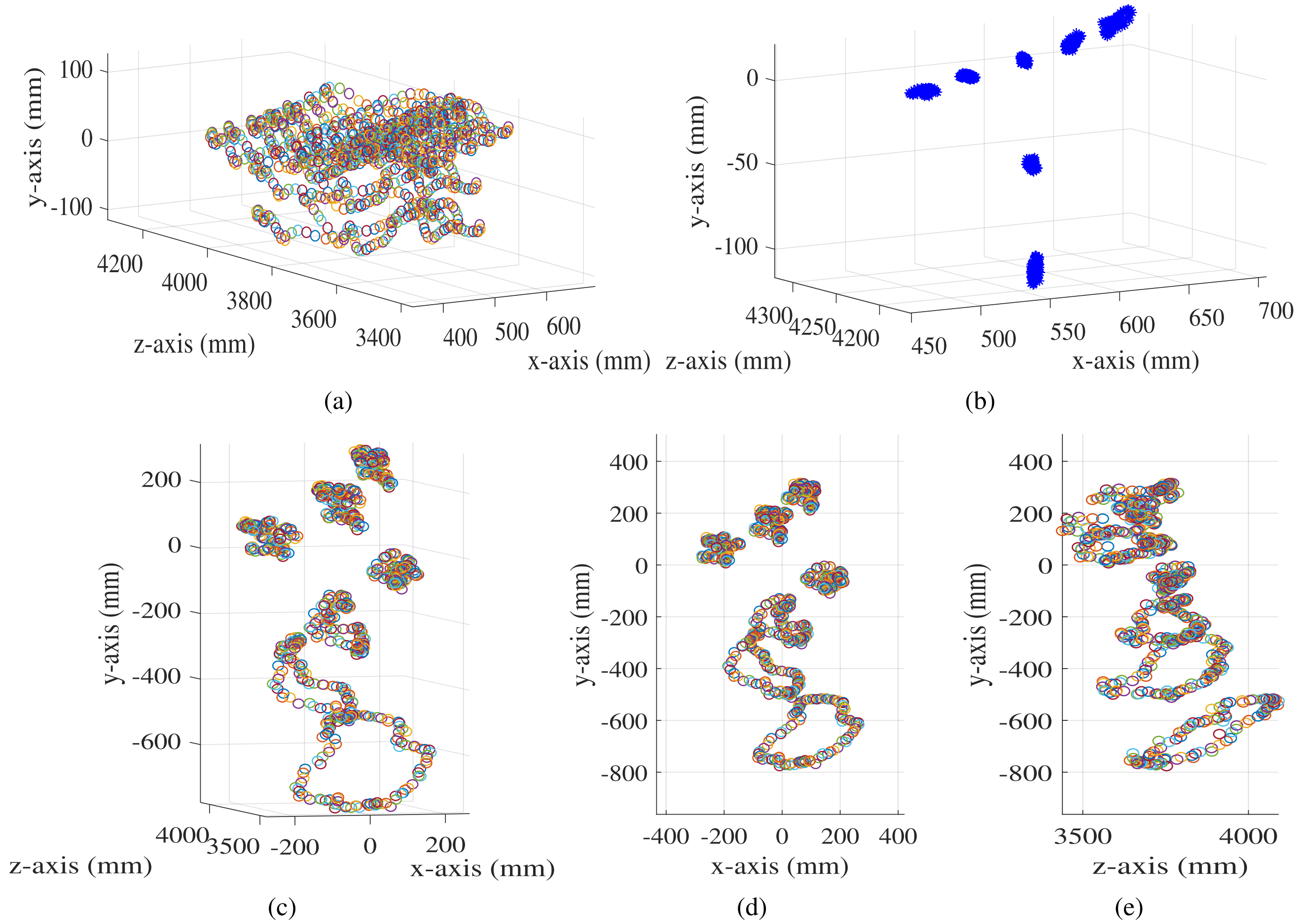
Objects	RMSE (mm)			
	Directions			Overall
	x	z	y	
'T'-shape	0.3021	0.3045	0.1684	0.2750
'I'-shape	0.4617	0.8492	0.3307	0.6244

It should be noted that the first frame is considered as the reference in our proposed camera motion compensation technique. Therefore, the RMSEs of both the T'-shape and 'I'-shape objects ware calculated with respect to this frame. Table 5 summarizes the average RMSEs of the proposed



(a) (b) (c) (d)

FIGURE 19. Capturing static object using moving stereo camera: (a) original position of 'T'-shape; (b) after applying camera motion compensation technique; (c) 'I'-shape in its original position; (d) after compensation of camera ego-motion with respect to reference 'T'-shape.



(a) (b)

(c) (d) (e)

FIGURE 20. Capturing a moving object using a moving stereo camera system: (a) original position of 'T'-shape and (b) 'T'-shape after applying camera motion compensation technique; (c) 3D positions, (d) xy-axes view and (e) zy-axes view of markers' trajectories during a pedalling motion while a reference 'T'-shape in (b) is used to balance its global motion with respect to the camera.

camera motion compensation system. The RMSE values indicate that the proposed system can balance the camera motion with very good accuracy.

2) MOVING CAMERA WITH MOVING OBJECT

For this experiment, a cyclist was tracked when both he/she and the stereo camera rig were moving. The cyclist was cycling slowly in an outdoor environment while a person, holding the camera rig facing him/her, was walking in parallel with the bicycle. In this experimental setup, a reference 'T'-shape was also used on the tracked object's platform to balance its global motion with respect to the camera so that the camera rig was always stationary with respect to the object as described in Section VI. In Fig. 20(a), it can be observed that the 'T'-shapes of the consecutive frames were not similar. Their variations were very large, especially in

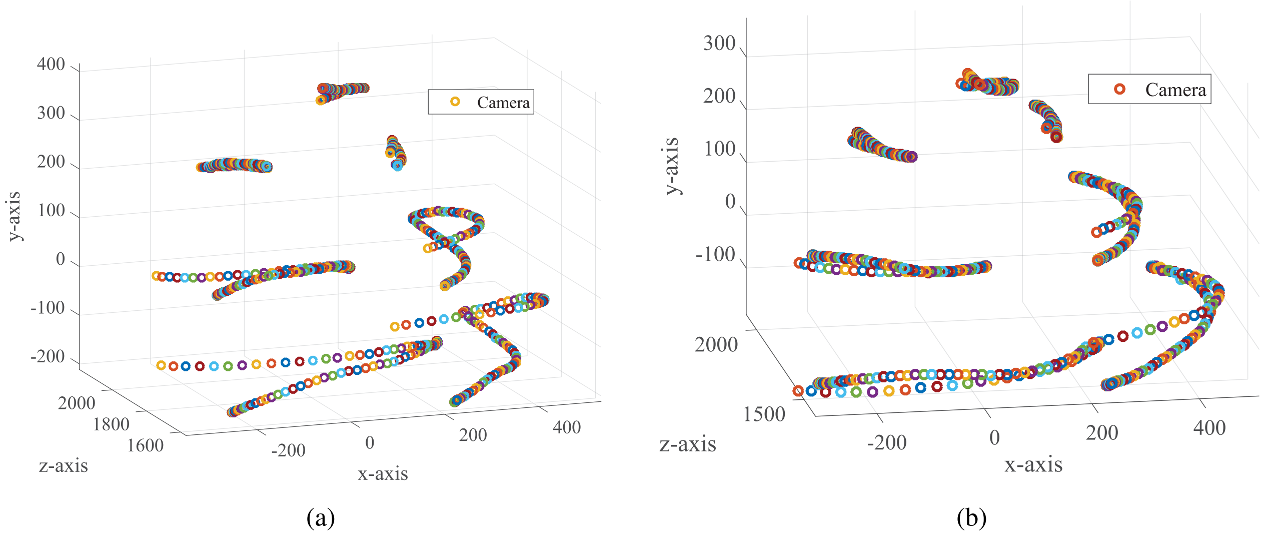


FIGURE 21. Experimental outcomes from kayaking paddle strokes obtained using a static stereo camera setup in an outdoor environment under conditions of (a) cloudy daylight and (b) sunny daylight.

the z-direction. Therefore, we applied the proposed camera motion compensation algorithm to retain their original positions, as shown in Fig. 20(b). The RMSE values in Table 6 show that the proposed algorithm compensated the camera motion accurately with a small error when objects were also moving.

TABLE 6. Root mean square errors of moving 'T'-shape object after applying camera motion compensation technique.

Object	RMSE (mm)			
	Directions			Overall
	x	z	y	
'T'-shape	6.368	5.018	3.064	5.073

As the 'T'-shapes were merged by means of a rigid transformation and their displacement parameters fed back every time to the main algorithm for capturing the cyclist's motion, the cameras were stationary with respect to the cyclist in each frame. Therefore, every marker's position in 3D was tracked for the cyclist overcoming the global motion caused by camera jitters. As a result, the only 3D motion-captured was that of the subject's limb, that is, his/her local motion, and was not affected by the global motion between the moving camera and moving platform he/she was riding. In this experiment, a side view of the pedalling motion of lower limbs of the body was tracked. Fig. 20(c) indicates the 3D trajectories obtained for the rider on a moving bicycle, as shown in Fig. 10. The xy-axes and zy-axes views of Fig. 20(c) are shown in Fig. 20(d) and Fig. 20(e) respectively. The shape at the bottom, which is a pedalling motion of the cyclist's ankle, should be a circle if the ankle's marker always stays at the same zy-plane. However, the markers on the cyclist's body were changing their positions in the z-direction as shown in Fig. 20(e). It should be noted that the proposed system was compared with a Vicon one in indoor experiments. However, the comparison was not possible in outdoor as it is not a suitable environment for the Vicon system.

C. EXPERIMENTAL ANALYSIS: ILLUMINATION CHANGE

In this part of our experimental analyses, the stereo camera rig was mounted on a tripod in different places in an outdoor environment with a diverse background and lighting conditions. In the area of the camera's view, a human with seven markers attached, performed the same kayaking activities as described in Section VII-A3. In these experiments, the stereo camera rig did not experience any jitter as it was strongly mounted on flat ground by means of a suction tripod. The experimental results obtained using the proposed method with two different lighting conditions on different days, one cloudy and the other sunny, are shown in Fig. 21. It should be noted that the experiment with a bicycle, as discussed in Section VII-B2, was also conducted with a diverse background and illumination conditions consistent with an outdoor setting. The experimental outcome obtained for the rider on a moving bicycle using the moving stereo camera is shown in Fig. 20. Although an environment prone to being sensitive to different lighting conditions, Fig. 20 and Fig. 21 show that the robustness of the proposed algorithms and designed stereo camera rig are capable of overcoming any relevant issues.

VIII. CONCLUSION

This paper proposed a new way of measuring human kinematics in both indoor and outdoor environments using a stereo vision-based technique. The proposed scheme demonstrates a method for using fish-eye lens cameras with wide baseline distances to capture human kinematic movements. Two commercially available low-cost sport action cameras with small form factors were used in a marker-based approach. The design of a wide-baseline distance stereo camera rig produced both a large disparity and large coverage area. Although this encountered a problem of losing a scene's common area, this was overcome using a selection of two cameras with ultra-wide angle fish-eye lenses. This enabled a large area to be seen by both cameras rather than only a long-baseline distance between the two and also provided a large FOV as a by-product. The fish-eye cameras were calibrated using Zhang's planar calibration method to eliminate the mismatch in their vertical positions. As a fish-eye lens shows non-linearity, we developed a ground truth for the depth versus disparity relationship which produced a depth map from disparity data. Since this is a marker-based system, a motion trajectory was obtained from each marker's 3D position in each frame which eliminated any need for 3D modeling of the trajectory and hence reduced the computational requirements of the system. To validate the proposed system, experiments were conducted in both indoor and outdoor environments for different scenarios. During the experiments, highly visible markers were placed on a subject's torso and the 3D locations of their centroids detected using the proposed technique. To estimate the errors in the system, these locations were also calculated using a Vicon MCS. In comparison with the Vicon system, the proposed system has the same order of accuracy while incurring less cost and structural complexity. However, unlike the Vicon system, the proposed system requires extra post-processing time to calculate the 3D positions of the markers. Moreover, this system was not implemented in real-time. The video data were first recorded using the designed camera rig and then further processed in the computer. This procedure could be converted into a full real-time MCS by incorporating real-time communications with the cameras. In the Internet of Things (IoT) era, such a feature could enable many real-time intelligent applications, where decision-making capability and the accuracy of the decisions being made have to be accomplished locally. Despite the inherent limitations of the proposed system, its overall RMSE was still in the range of few millimeters, which is an acceptable level of precision for many applications that require high accuracy. Its major advantage is that it can be applied in an outdoor real-world environment with any lighting conditions.

REFERENCES

[1] G. Welch and E. Foxlin, "Motion tracking: No silver bullet, but a respectable arsenal," *IEEE Comput. Graph. Appl.*, vol. 22, no. 6, pp. 24–38, Nov. 2002.

[2] *Vicon Optical Motion Capture System*. Accessed: Feb. 3, 2020. [Online]. Available: http://www.vicon.com

[3] G. Seguin, K. Alahari, J. Sivic, and I. Laptev, "Pose estimation and segmentation of multiple people in stereoscopic movies," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 8, pp. 1643–1655, Aug. 2015.

[4] F. AlQahtani, J. Banks, V. Chandran, and J. Zhang, "Three-dimensional head pose estimation using a stereo camera arrangement," in *Proc. Int. Conf. Mach. Vis. Appl. (ICMVA)*, 2018, pp. 28–35.

[5] M. J. Domínguez-Morales, Á. Jiménez-Fernández, G. Jiménez-Moreno, C. Conde, E. Cabello, and A. Linares-Barranco, "Bio-inspired stereo vision calibration for dynamic vision sensors," *IEEE Access*, vol. 7, pp. 138415–138425, 2019.

[6] N. Sarafianos, B. Boteanu, B. Ionescu, and I. A. Kakadiaris, "3D human pose estimation: A review of the literature and analysis of covariates," *Comput. Vis. Image Understand.*, vol. 152, pp. 1–20, Nov. 2016.

[7] M. B. Holte, C. Tran, M. M. Trivedi, and T. B. Moeslund, "Human pose estimation and activity recognition from multi-view videos: Comparative explorations of recent developments," *IEEE J. Sel. Topics Signal Process.*, vol. 6, no. 5, pp. 538–552, Sep. 2012.

[8] P. F. Felzenszwalb and D. P. Huttenlocher, "Pictorial structures for object recognition," *Int. J. Comput. Vis.*, vol. 61, no. 1, pp. 55–79, Jan. 2005.

[9] Y. Yang and D. Ramanan, "Articulated human detection with flexible mixtures of parts," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 12, pp. 2878–2890, Dec. 2013.

[10] L. Pishchulin, M. Andriluka, P. Gehler, and B. Schiele, "Poselet conditioned pictorial structures," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2013, pp. 588–595.

[11] C. Wang, Y. Wang, Z. Lin, and A. L. Yuille, "Robust 3D human pose estimation from single images or video sequences," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 5, pp. 1227–1241, May 2019.

[12] X. Zhao, L. Lyu, J. Zhang, and C. Lyu, "An image-constrained particle filter for 3D human motion tracking," *IEEE Access*, vol. 7, pp. 10294–10307, 2019.

[13] T. B. Moeslund and E. Granum, "3D human pose estimation using 2D-data and an alternative phase space representation," in *Proc. Workshop Hum. Modeling, Anal. Synth.*, 2000, p. 1.

[14] T. B. Moeslund and E. Granum, "A survey of computer vision-based human motion capture," *Comput. Vis. Image Understand.*, vol. 81, no. 3, pp. 231–268, Mar. 2001.

[15] J. K. Aggarwal and L. Xia, "Human activity recognition from 3D data: A review," *Pattern Recognit. Lett.*, vol. 48, pp. 70–80, Oct. 2014.

[16] T. B. Moeslund, A. Hilton, and V. Krüger, "A survey of advances in vision-based human motion capture and analysis," *Comput. Vis. Image Understand.*, vol. 104, nos. 2–3, pp. 90–126, Nov. 2006.

[17] N. Lazaros, G. C. Sirakoulis, and A. Gasteratos, "Review of stereo vision algorithms: From software to hardware," *Int. J. Optomechatronics*, vol. 2, no. 4, pp. 435–462, Nov. 2008.

[18] E. van der Kruk and M. M. Reijne, "Accuracy of human motion capture systems for sport applications; state-of-the-art review," *Eur. J. Sport Sci.*, vol. 18, no. 6, pp. 806–819, Jul. 2018.

[19] L. P. Maletsky, J. Sun, and N. A. Morton, "Accuracy of an optical active-marker system to track the relative motion of rigid bodies," *J. Biomech.*, vol. 40, no. 3, pp. 682–685, Jan. 2007.

[20] R. Hartley and A. Zisserman, *Multiple View Geometry in Computer Vision*. Cambridge, U.K.: Cambridge Univ. Press, 2003.

[21] A. P. Shon, J. J. Storz, and R. P. N. Rao, "Towards a real-time Bayesian imitation system for a humanoid robot," in *Proc. IEEE Int. Conf. Robot. Autom.*, Apr. 2007, pp. 2847–2852.

[22] M. Loper, N. Mahmood, and M. J. Black, "MoSh: Motion and shape capture from sparse markers," *ACM Trans. Graph.*, vol. 33, no. 6, p. 220, 2014.

[23] A. Chatzitofis, D. Zarpalas, S. Kollias, and P. Daras, "DeepMoCap: Deep optical motion capture using multiple depth sensors and retro-reflectors," *Sensors*, vol. 19, no. 2, p. 282, Jan. 2019.

[24] C.-S. Lee and A. Elgammal, "Coupled visual and kinematic manifold models for tracking," *Int. J. Comput. Vis.*, vol. 87, nos. 1–2, pp. 118–139, Mar. 2010.

[25] A. Zanfir, E. Marinoiu, and C. Sminchisescu, "Monocular 3D pose and shape estimation of multiple people in natural scenes: The importance of multiple scene constraints," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 2148–2157.

[26] S. Liu, Y. Li, and G. Hua, "Human pose estimation in video via structured space learning and halfway temporal evaluation," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 29, no. 7, pp. 2029–2038, Jul. 2019.

[27] J. Chen, M. Kim, Y. Wang, and Q. Ji, "Switching Gaussian process dynamic models for simultaneous composite motion tracking and recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 2655–2662.

[28] J. Gall, A. Yao, and L. Van Gool, "2D action recognition serves 3D human pose estimation," in *Proc. Eur. Conf. Comput. Vis.*, 2010, pp. 425–438.

[29] L. Xu, Y. Liu, W. Cheng, K. Guo, G. Zhou, Q. Dai, and L. Fang, "FlyCap: Markerless motion capture using multiple autonomous flying cameras," *IEEE Trans. Vis. Comput. Graphics*, vol. 24, no. 8, pp. 2284–2297, Aug. 2018.

[30] B. Zhao, Y. Tang, C. Wu, and W. Du, "Vision-based tracking control of quadrotor with backstepping sliding mode control," *IEEE Access*, vol. 6, pp. 72439–72448, 2018.

[31] J. Xiao, H. Cheng, H. Sawhney, and F. Han, "Vehicle detection and tracking in wide field-of-view aerial video," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, Jun. 2010, pp. 679–684.

[32] C. Rao, A. Yilmaz, and M. Shah, "View-invariant representation and recognition of actions," *Int. J. Comput. Vis.*, vol. 50, no. 2, pp. 203–226, 2002.

[33] S. Wu and Y. F. Li, "Flexible signature descriptions for adaptive motion trajectory representation, perception and recognition," *Pattern Recognit.*, vol. 42, no. 1, pp. 194–214, Jan. 2009.

[34] A. Yilmaz, O. Javed, and M. Shah, "Object tracking: A survey," *ACM Comput. Surv.*, vol. 38, no. 4, p. 13, 2006.

[35] J. Min and R. Kasturi, "Activity recognition based on multiple motion trajectories," in *Proc. 17th Int. Conf. Pattern Recognit. (ICPR)*, vol. 4, 2004, pp. 199–202.

[36] *Go-Pro*. Accessed: Feb. 3, 2020. [Online]. Available: http://www.gopro.com

[37] R. Tsai, "A versatile camera calibration technique for high-accuracy 3D machine vision metrology using off-the-shelf TV cameras and lenses," *IEEE J. Robot. Autom.*, vol. RA-3, no. 4, pp. 323–344, Aug. 1987.

[38] J. Heikkila and O. Silven, "A four-step camera calibration procedure with implicit image correction," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, Jun. 1997, pp. 1106–1112.

[39] Z. Zhang, "A flexible new technique for camera calibration," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 11, pp. 1330–1334, Nov. 2000.

[40] J. Weng, P. Cohen, and M. Herniou, "Camera calibration with distortion models and accuracy evaluation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 14, no. 10, pp. 965–980, Oct. 1992.

[41] A. Fusiello, E. Trucco, and A. Verri, "A compact algorithm for rectification of stereo pairs," *Mach. Vis. Appl.*, vol. 12, no. 1, pp. 16–22, Jul. 2000.

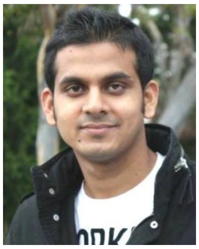
[42] A. David and P. Jean, *Computer Vision: A Modern Approach*. Upper Saddle River, NJ, USA: Prentice-Hall, 2002, pp. 654–659.

[43] A. Kumar and P. Ben-Tzvi, "Spatial object tracking system based on linear optical sensor arrays," *IEEE Sensors J.*, vol. 16, no. 22, pp. 7933–7940, Nov. 2016.

[44] J. Zhang, Z. Liu, Y. Gao, and G. Zhang, "Robust method for measuring the position and orientation of drogue based on stereo vision," *IEEE Trans. Ind. Electron.*, early access, Mar. 25, 2020, doi: 10.1109/TIE.2020.2982089.

ATIQUL ISLAM (Graduate Student Member, IEEE) received the B.Sc. degree in electrical and electronic engineering from the Khulna University of Engineering & Technology, Khulna, Bangladesh, in 2012, and the M.S. degree in electrical engineering from the University of New South Wales, Canberra, ACT, Australia, in 2017. He is currently pursuing the Ph.D. degree in human–computer interaction with The Australian National University under a very competitive Australian Government Research Training Program (AGRTP) scholarship. His research interests include signal and image processing, human–computer interaction, medical imaging, and video coding.

MD. ASIKUZZAMAN (Member, IEEE) received the B.Sc. degree in electronics and telecommunication engineering from the Rajshahi University of Engineering & Technology, Rajshahi, Bangladesh, in 2010, and the Ph.D. degree in electrical engineering from the University of New South Wales, Canberra, ACT, Australia, in 2015, under a very competitive University International Postgraduate Award Scholarship. He was a Research Associate with the School of Engineering and Information Technology, University of New South Wales, from 2015 to 2019, where he is currently a Senior Research Associate. His current research interests include 2D and 3D video watermarking, privacy preservation, 3D modeling, deep learning, medical imaging, and video coding. He was the Technical Program Chair for the 2018 International Conference on Digital Image Computing: Techniques and Applications. He is also serving as an Associate Editor for IEEE ACCESS.



MOHAMMAD OMAR KHYAM received the Ph.D. degree in electrical engineering from the University of New South Wales, Australia, in 2015. From 2016 to 2017, he was a Postdoctoral Research Fellow with the National University of Singapore. From 2018 to 2019, he was also a Postdoctoral Research Fellow with Virginia Tech, USA. He is currently a Lecturer with Central Queensland University, Australia. His current research interests include signal processing, wireless communication, deep learning, and robotics.

MD. NOOR-A-RAHIM (Member, IEEE) received the Ph.D. degree from the Institute for Telecommunications Research, University of South Australia, Adelaide, SA, Australia, in 2015. He was a Postdoctoral Research Fellow with the Centre for Infocomm Technology, Nanyang Technological University, Singapore. He is currently a Senior Postdoctoral Researcher (MSCA Fellow) with the School of Computer Science and IT, University College Cork, Cork, Ireland. His research interests include control over wireless networks, intelligent transportation systems, information theory, signal processing, and DNA-based data storage. He was a recipient of the Michael Miller Medal from the Institute for Telecommunications Research, University of South Australia, for the most Outstanding Ph.D. Thesis, in 2015.

MARK R. PICKERING (Member, IEEE) was born in Biloela, Australia, in 1966. He received the B.Eng. degree from the Capricornia Institute of Advanced Education, Rockhampton, QLD, Australia, in 1988, and the M.Eng. and Ph.D. degrees from the University of New South Wales, Canberra, ACT, Australia, in 1991 and 1995, respectively, all in electrical engineering. He was a Lecturer from 1996 to 1999, a Senior Lecturer from 2000 to 2009, and an Associate Professor from 2010 to 2017 with the School of Electrical Engineering and Information Technology, University of New South Wales, where he is currently a Professor. His research interests include video and audio coding, medical imaging, data compression, information security, data networks, and error-resilient data transmission.