A Real-Time Photogrammetric System for Acquisition and Monitoring of Three-Dimensional Human Body Kinematics

Long Chen, Bo Wu, Yao Zhao, and Yuan Li

Abstract

Real-time acquisition and analysis of three-dimensional (3D) human body kinematics are essential in many applications. In this paper, we present a real-time photogrammetric system consisting of a stereo pair of red-green-blue (RGB) cameras. The system incorporates a multi-threaded and graphics processing unit (GPU)-accelerated solution for real-time extraction of 3D human kinematics. A deep learning approach is adopted to automatically extract two-dimensional (2D) human body features, which are then converted to 3D features based on photogrammetric processing, including dense image matching and triangulation. The multi-threading scheme and GPU-acceleration enable real-time acquisition and monitoring of 3D human body kinematics. Experimental analysis verified that the system processing rate reached ~18 frames per second. The effective detection distance reached 15 m, with a geometric accuracy of better than 1% of the distance within a range of 12 m. The real-time measurement accuracy for human body kinematics ranged from 0.8% to 7.5%. The results suggest that the proposed system is capable of real-time acquisition and monitoring of 3D human kinematics with favorable performance, showing great potential for various applications.

Introduction

Real-time capture and response for human locomotion at a large scale is of great importance for various applications, such as monitoring actions of patients in physical rehabilitation (Karunarathne *et al.* 2014), enhancing safe conditions of workers in industrial robotics (Seo *et al.* 2015), analyzing the movements of athletes (Gholami *et al.* 2019), and human-computer interaction in virtual reality (Jaimes and Sebe 2007). Because such applications benefit from accurate extraction and analysis of 3D human body kinematics, real-time photogrammetric systems capable of these types of measurements have been extensively researched in recent years.

With the advances of computer processing capabilities, human pose recognition has been shifted from single image-based (Agarwal and Triggs 2006; Shotton *et al.* 2011) to image sequences (Zhou *et al.* 2016), and further evolved to dynamic human pose recognition from video sequences (Wang *et al.* 2019). Nevertheless, the complexity of these algorithms prohibited the real-time processing of human post recognition. The recent development of smart cameras (Carraro *et al.* 2016) and red-green-blue-depth (RGB-D) sensors (Tang *et al.* 2020; Wu *et al.* 2019; Tang *et al.* 2016), which can directly capture 3D information in a given scenario, has enabled cost-efficient estimation and tracking of 3D human body posture in real time. However, the sensors are limited by the workable distance, field-of-view (FOV), and reliability. In contrast, ordinary RGB cameras are capable of capturing large-scale scenarios with a large FOV. Previous studies using RGB cameras only recognized human posture in 2D (Shotton *et al.* 2011; Jalal and Kim 2014), resulting in the loss of vital information in the depth dimension. Multi-view stereo techniques (Seitz *et al.* 2018) enabled retrieving 3D information from 2D images using photogrammetric approaches, and 3D human body kinematics can be subsequently extracted from the retrieved 3D information. However, 3D reconstruction of large-scale scenes using dense image matching (Haala 2013) is time-consuming, especially for systems without hardware acceleration. The complexity of dense image matching algorithms requires balancing the efficiency and quality of the matching results, which impedes the possibility of real-time processing.

The advent of the central processing unit (CPU) with multi-threaded capabilities and the development of the GPU-acceleration technologies make real-time computations possible. This paper presents a cost-effective photogrammetric system consisting of a stereo pair of RGB cameras. The system utilizes multi-threading and GPU-acceleration techniques as well as deep learning to extract and measure 3D human body features at a large scale in real-time, which enable further analysis of 3D human kinematics, such as step length, moving speed, arm angle, knee angle, etc., from the video sequence.

The remainder of this paper is organized into four sections. Section "Related Works" consists of reviews of related works, and the section "System Development" provides detailed descriptions of the developed system. The experimental evaluations are presented in the section "System Implementation and Evaluation", and the section "Conclusions and Discussion" consists of concluding remarks and suggestions for future work.

Related Works

2D Human Body Feature Extraction

A good algorithm for human body feature extraction can improve the efficiency and accuracy of human body tracking

Long Chen, Bo Wu and Yao Zhao are with the Department of Land Surveying and Geo-Informatics, The Hong Kong Polytechnic University, Hung Hom, Kowloon, Hong Kong (bo.wu@polyu.edu.hk).

Yuan Li was with the Department of Land Surveying and Geo-Informatics, The Hong Kong Polytechnic University, and is now with the School of Geospatial Engineering and Science, Sun Yat-Sen University, China.

Contributed by Long Chen, August 24, 2021 (sent for review August 26, 2020; reviewed by Nadine Rüegg, Christian Zimmermann, and Max Scharz).

Photogrammetric Engineering & Remote Sensing
Vol. 87, No. 5, May 2021, pp. 363–373.
0099-1112/21/363–373

doi: 10.14358/PERS.87.5.363

and monitoring system. Early studies of human-computer interaction (Jaimes and Sebe 2007) introduced variable computer vision algorithms, such as body, gesture, gaze, and facial expression recognition algorithms, into several crossroad research areas, including psychology, artificial intelligence, and many others. These algorithms turned 2D human body feature detection into an intensive research field and applied to areas such as the facial feature point-recognition method (Xiong and De la Torre 2013) and single- or multiple-person posture recognition (Zhou *et al.* 2013). Xiong and De la Torre (2013) applied the facial feature recognition method and supervised descent method (SDM) to an image sequence. The SDM was able to recognize the facial features in the image sequence with favorable accuracy. Zhou *et al.* (2013) presented a gesture tracking and recognition algorithm, which allowed near real-time processing. Their experiments indicated that the running frame rate reached five frames per second (fps).

Recently, the accelerated advancement of GPU technology and the evolution of multithreading-capable CPUs have led to the popularity of deep learning approaches (Ranjan *et al.* 2017), such as mask regional-based convolutional neural networks (R-CNNs) (Abdulla 2017), OpenPose (Cao *et al.* 2018), and regional multi-person pose estimation (RMPE) (Fang *et al.* 2017). Ranjan *et al.* (2017) presented an algorithm called HyperFace, which allowed simultaneous face detection and posture estimation using deep CNN. However, HyperFace required three seconds to process one image, which limited its potential for real-time human feature extraction. OpenPose and RMPE have also made it possible to evaluate and extract 2D features of human postures in real time. Fang *et al.* (2017) used the benchmark Max Planck Institute for Informatics (MPII) human pose data set (Andriluka *et al.* 2014) and Microsoft Common Objects in COntext (MSCOCO) data sets (Veit *et al.* 2016) to compare popular leading-edge human pose estimators based on the mean average precision (mAP) score. Table 1 provides an overview of these popular human pose estimators. According to Fang *et al.* (2017), deep learning-based object-detection and pose-evaluation algorithms accurately obtained the 2D key points of human posture. Among the assessed algorithms, RMPE was the most reliable and accurate multi-person pose estimator with an overall mAP of 80+ and a processing rate of 20+ (fps). The OpenPose algorithm had an mAP of almost 70+ but only achieved approximately 10+ fps running on the same platform (Cao *et al.* 2018). Due to high process efficiency and accuracy, deep learning approaches are particularly suitable for real-time 2D human posture evaluation and feature extraction.

3D Human Posture Feature Extraction

In recent years, the rapid developments of computer hardware (D'Apuzzo 2002) and affordable RGB-D cameras (Zimmermann *et al.* 2018) have expanded the study of human posture evaluation and feature extraction from 2D to 3D space. D'Apuzzo (2002) proposed a method using photogrammetry to recover 3D human body features from synchronized video sequences captured from multiple cameras at different locations and dynamically tracked their trajectories. The creation of a 3D human kinematic descriptor (Zanfir *et al.* 2013) moved the study of 3D human body gesture recognition and feature extraction from part-based posture retrieval methods (Zimmermann and Brox 2017; Sridhar *et al.* 2013) to whole-body human pose estimation (Srivastav *et al.* 2018). Even though these studies have brought human feature detection into 3D space, they were in general time-consuming and had not been implemented for real-time 3D human feature detection.

RGB-D cameras offer 3D information in a direct way and have been used for extracting 3D human posture and features in recent years. For example, Carraro *et al.* (2018) and Huang and Nguyen (2019) used the OpenPose to RGB-D camera to obtain 3D human feature points by integrating the 2D features extracted by deep learning with the depth information measured by the depth sensor. Srivastav *et al.* (2018) used an RGB-D camera to obtain 3D human body key points for indoor posture-estimation and tracking. However, the use of RGB-D cameras is limited by their short measurement ranges and narrow FOVs (Haggag *et al.* 2013).

This paper presents a real-time photogrammetric system consisting of a stereo pair of RGB cameras. The system incorporates a novel multi-threading strategy and GPU acceleration as well as an advanced deep learning method to extract and measure 3D human kinematics in real-time. The main contributions of the presented work are as follows:

1. In order to achieve real-time processing, the system adopts four threads, responsible for 3D scene reconstruction, human feature extraction, kinematic information calculation, and result visualization, respectively. Each thread works independently without waiting for other threads to complete their tasks, and thus the processing latencies of the procedures are reduced. In addition, for the thread of 3D scene reconstruction, which is a bottleneck problem and requires the most computations, a GPU-accelerated strategy is used to achieve real-time efficiency of 3D reconstruction.
2. To avoid complicated 3D computations, we develop a strategy that combines a 2D human body skeleton extraction algorithm with the projection relationships between 2D and 3D spaces. A mature deep learning method is adopted to ensure the reliability and efficiency of 2D human feature extraction, and then the 2D results are converted into 3D space based on the projection relationships, enabling 3D human body kinematic analysis.

System Development

Multi-Threading Design

The real-time photogrammetric system had four threads. Each thread performed as an individual model that handled different tasks, as shown in Figure 1. Thread 1 loaded the stereo RGB images with timestamps and known orientation parameters, and then delivered images to semi-global matching (SGM) (Hirschmuller 2007) to generate a disparity map. A 3D map was retrieved by triangulation based on the disparities and orientation parameters of the camera. A GPU-acceleration solution was also used in the first thread to speed up the 3D scene reconstruction processing rate. Thread 2 first extracted 2D

Table 1. Comparison of 2D human detection and tracking algorithms (all values are mAP scores).

	Head	Shoulder	Elbow	Wrist	Hip	Knee	Ankle	Total
Fang *et al.* (2017) (RMPE)	88.4	86.5	78.6	70.4	74.4	73.0	65.8	76.7
Iqbal and Gall (2016)	58.4	53.9	44.5	35.0	42.2	36.7	31.1	43.1
Insafutdinov *et al.* (2016) (DeeperCut)	78.4	72.5	60.2	51.0	57.2	52.0	45.4	59.5
Levinkov *et al.* (2017)	89.8	85.2	71.8	59.6	71.1	63.0	53.5	70.6
Insafutdinov *et al.* (2017) (ArtTrack)	88.8	87.0	75.9	64.9	74.2	68.8	60.5	74.3
Cao *et al.* (2018) (OpenPose)	91.2	87.6	77.7	66.8	75.4	68.9	61.7	75.6
Newell *et al.* (2017)	92.1	89.3	78.9	69.8	76.2	71.6	64.7	77.5

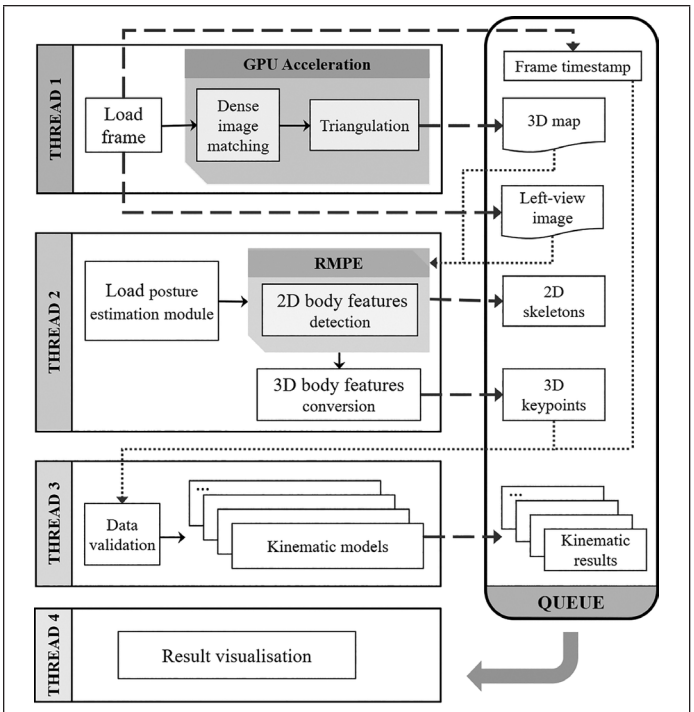


Figure 1. Workflow of the real-time photogrammetric system.

human body skeletons from the left-view images using RMPE and extended the 2D skeletons to 3D body features based on the 3D map array produced by thread 1. Thread 3 computed the following human kinematic parameters: moving velocity, step length, and joint motion angles. These parameters were based on 3D body features. The products of each thread were stored in the same queue for data exchange, and the results were loaded into thread 4 from the queue for the system visualization in real-time.

3D Scene Reconstruction from Stereo RGB Images with GPU Acceleration

This section describes the algorithms used in thread 1 for dense image matching and the triangulation process with GPU acceleration. The GPU-accelerated procedure is shown in Figure 2. First, stereo RGB images with known interior and exterior orientation parameters were loaded from the stereo camera and stored in the host (CPU). The device (GPU) then copied the stereo RGB images from the host and split them into left-view and right-view images. The dense image matching algorithm SGM and triangulation process were performed on the GPU with an acceleration solution for reconstructing the 3D information in real-time. Simultaneously, the left-view image and disparity map obtained from SGM were stitched together as the background image prepared for the visualization in thread 4.

GPU-Accelerated SGM for Disparity Estimation

A GPU-accelerated SGM method was applied for the real-time stereo estimation of the disparity map. Figure 2 shows the generation of a consistent disparity map. Two cost items, matching cost and smoothed cost, were computed in the GPU device. The matching cost measures the probability that two pixels on the left- and right-view images correspond to the same point in the object space. Features were first extracted from the left- and right-view images, and a similarity comparison was used to generate a local-matching cost and potential disparity for each pixel. A center-symmetric census transform (CSCT) (Hernandez-Juarez *et al.* 2016) configured with a fixed-sized (e.g., 9 × 7) window was used to extract the features by moving the window on the left- and right-view images, respectively. The extracted features were presented as bit-vectors. The similarity between two corresponding pixels in the left- and right-view images was defined as the Hamming distance between their CSCT bit-vector features.

The smoothed cost was introduced to deal with nonunique or incorrect correspondences of similarity, resulting in an inaccurate estimation of disparity. In SGM, the smoothed cost was computed by considering the similarity between neighboring points and disparities along paths across the image. The global solution was approximated as one-dimensional minimization problems along these paths. For each path direction, SGM aggregated a cost that considered the cost of neighboring points and disparities. After the disparities of all pixels were estimated, a 3 × 3 median filter (Brownrigg 1984) was applied to remove outliers.

GPUs are massively parallel devices containing tens of streaming multiprocessors (SMs), and vector computation operations are highly utilized and pipelined in SMs to optimize the computational efficiency. The compute unified device architecture (CUDA) programming model (Nvidia 2019) allows for defining a massive number of threads deployed in SMs of the same program code. The SGM was coded using a two-level identifier in CUDA to specialize in each thread for disparities estimation. The code in this research was deployed following the method of Hernandez-Juarez *et al.* (2016).

Triangulation for Generation of 3D Map

Triangulation was used to generate a 3D map (point cloud) from the disparity derived from the stereo camera. Figure 3 shows the geometry of the stereo camera system. C_1 and C_2 are the perspective centers of the left and right cameras, respectively, and IP_1 and IP_2 are the respective image planes. f_1 and f_2 are the focal lengths of the left and right cameras. They are the same for the camera system used in this research.

Figure 3 illustrates the geometric relationships between the object point P and the stereo cameras C_1 and C_2, and the colinear relationship amongst the object point, image point,

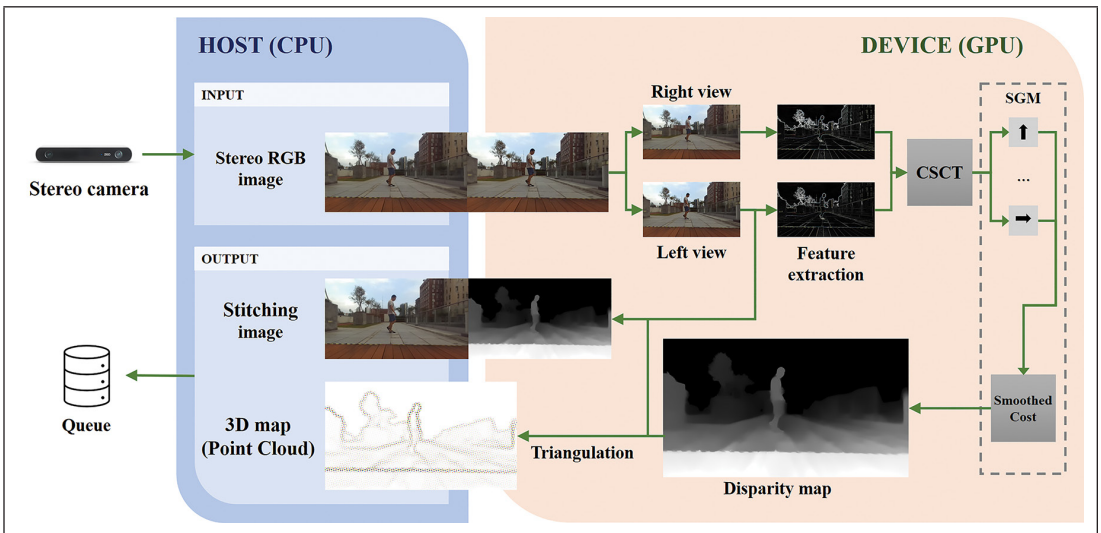


Figure 2. GPU-accelerated procedure of dense image matching and 3D map generation.

and camera perspective center (e.g., P, p_1, and C_1; or P, p_2, and C_2). Based on the geometric relationships, the 3D coordinates of the point P can be calculated using the following equation (Kaehler and Bradski 2016):

$$\begin{bmatrix} X_P \\ Y_P \\ Z_P \end{bmatrix} = \begin{bmatrix} u_1 - c_{x1} \\ v_1 - c_{y1} \\ f_1 \end{bmatrix} \cdot \frac{b}{d} \quad (1)$$

where (X_p, Y_p, Z_p) are the coordinates of the point P in object space. (u_1, v_1) are the image coordinates of P in the left image. (C_{x1}, C_{y1}) are the coordinates of the principal point in the left image. f_1 is the focal length of the left camera. b is the baseline between the left and right cameras, and d is the disparity as denoted by $u_1 - u_2$. It should be noted that Equation 1 is used here instead of the complex collinearity equations for the purpose of more efficient calculation. Equation 1 is based on the epipolar geometry, which can be derived from the collinearity equations (Fraser 1997; Gruen and Beyer 2001), assuming a fixed relationship between the left and right cameras. In the actual experiments, the camera system used has been calibrated by the manufacturer already. The focal length of the camera, the position of the principal point, and lens distortions are provided. The images have been rectified based on the lens distortion parameters. A fundamental matrix defining the relative orientation of the left and right cameras is also provided, which allows the determination of epipolar geometry between the left and right cameras.

Based on Equation 1, each pixel in the disparity map can be transferred to a 3D point in the object space. The calculated 3D points can be used to generate a 3D map, and the RGB information of the 3D points can be obtained from the corresponding 2D coordinates for visualization purposes. Figure 4 shows an experimental result of 3D map visualization.

Extraction of 3D Human Body Features

Thread 2 handled the extraction of 3D human body features. The system took advantage of the mature 2D body skeleton extraction algorithm, RMPE (AlphaPose) (Fang *et al.* 2017), and then extended the 2D body skeleton into 3D body features based on the projection relationship between the image space and object space. RMPE is an open-source CNN-based multi-person pose estimator used in conventional pictorial structure models for posture estimation. RMPE has been evaluated on two standard multi-person data sets with large occlusion cases: MPII (Andriluka *et al.* 2014) and MSCOCO 2016 Keypoints Challenge data set (Veit *et al.* 2016). MPII data set contains more than 28 000 training samples for single person pose estimation, while the MSCOCO data set consists of 105 698 training and around 80 000 testing human instances. The results of RMPE on MPII data sets indicated that it achieved an average accuracy of 72 mAP on identifying difficult joints such as wrists, elbows, ankles, and knees. The results of RMPE on MSCOCO data sets also proved that RMPE achieved state-of-the-art performance compared with other popular detectors (Fang *et al.* 2017). Since RMPE has been trained extensively on large data sets and performed well in identifying human body features, this research adopted it for real-time 2D human feature detection. The pretrained RMPE yields 17 default key joint points representing human body

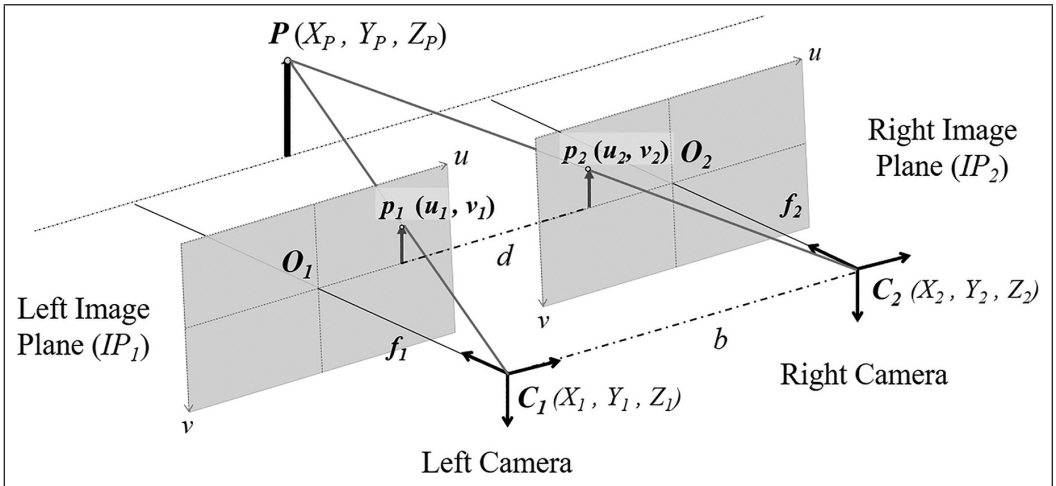


Figure 3. Illustration of geometric triangulation of the stereo camera system.

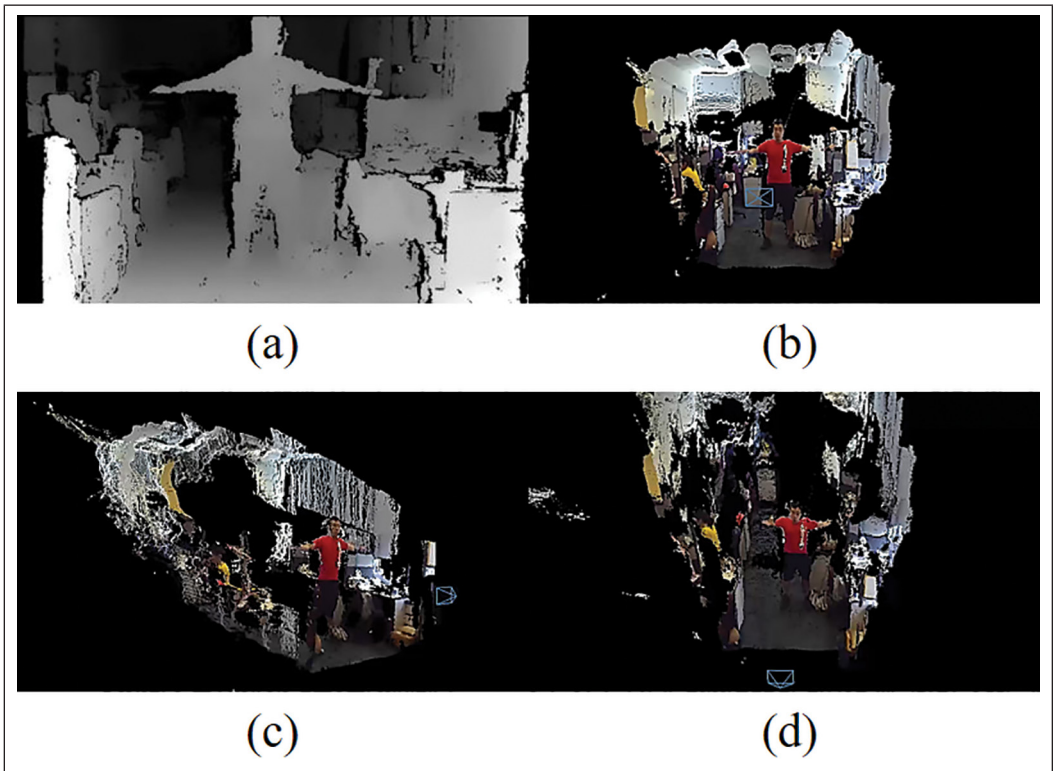


Figure 4. Disparity map converted to a 3D map. (a) Disparity map generated by SGM. (b)– (d) Colored point clouds are displayed from different perspectives.

parts (Figure 5). The key joint points and their corresponding human body parts are listed in Table 2.

The 2D body skeletons extracted from the images using RMPE are represented in Equations 2, 3, and 4:

$$\bar{E} = \{\bar{S}_1, \bar{S}_2, \cdots, \bar{S}_k\} \tag{2}$$

$$S = \{j_i \mid 0 \le i \le m\},\ 0 \le m \le 16, S \in \bar{E} \tag{3}$$

$$j_i = (x_i,\ y_i),\ 0 \le i \le m, \tag{4}$$

where $\bar{E}$ is a set of human body skeletons $\bar{S}_i (i \in \{1,2, \ldots, k\})$ of k people detected by RMPE in the image. Each skeleton S is a set of 2D joint points $j_i (i \in \{1,2, \ldots, m\})$ that contain 2D coordinates (x_i, y_i), which correspond to the left-view image. m is the total number of body parts listed in Table 2. Each pixel in a 3D map contains both 2D image coordinates and 3D coordinates.

The 2D body skeletons were converted to 3D body features by finding the 3D coordinates corresponding to the 2D joint points from the 3D map using the 2D image coordinates as the index. Thus, a set of 3D body features containing depth information was derived at this stage and saved in the queue for further analysis. These 3D body features were used to evaluate human kinematics, which will be discussed in the next section.

Analysis and Visualisation of 3D Human Kinematics

Threads 1 and 2 worked continuously with the frames loaded from the stereo camera. The 3D body features extracted from a series of stereo camera frames then facilitated the kinematic analysis of a 3D human body over time in thread 3. This study focused on typical 3D human kinematics, including the velocity of movement (moving speed and direction), step length, knee flexion angle, and arm swing angles. Table 3 lists detailed descriptions of the considered human kinematics based on the default 17 key joint points (Table 2).

The human center of mass is maintained or altered close to the midpoint of the left and right hips (Vlutters *et al.* 2016). Therefore, the midpoint of the left and right hips was used to calculate moving speed and direction. The moving speed was calculated based on the 3D coordinates of the midpoint at the initial and final positions of a person's movement during a time interval. The moving direction was treated based on trigonometry that movement can be in any direction in a 360° arc starting from the direction in which the person faces the camera. The 360° were divided into groups to represent different directions. The moving direction in this system was classified into four directions: forward, backward, left, and right (Figure 6). In Figure 6, *Pi* (i = 0) indicates the possible initial position, and *Pi* (i = 1, 2, 3, 4) illustrates the possible final positions in each direction in the next frame. The direction was determined by calculating the angle θi between the vector from an initial position to a final position and the *XY*-plane of the camera system based on the 3D coordinates of two hips. The step length was expressed as the vector length from one ankle to the other. The system calculated the step length using the 3D coordinates of both ankles while calculating the direction and speed of movement.

Human joint motion measurements include knee pressure angle and arm swing angle. The latter was quantified into two indices: upper-arm angle and elbow angle. The calculation based on the geometry is shown in Figure 7. The knee flexion angle was calculated using the angle between the knee-angle and knee-hip vectors in 3D coordinates. The upper-arm angle was represented by the angle between the shoulder-elbow and shoulder-hip vectors. The elbow angle was calculated as the angle between the elbow-shoulder and elbow-wrist vectors.

The visualization was performed by an individual thread (thread 4) that loaded all of the information saved in the queue and displayed it on the screen. Once the threading detected that the queue was full of the stitching and 3D maps generated in thread 1, the 3D body features extended from the 2D skeleton in thread 2 and the kinematic results computed from the 3D body features in thread 3, it automatically

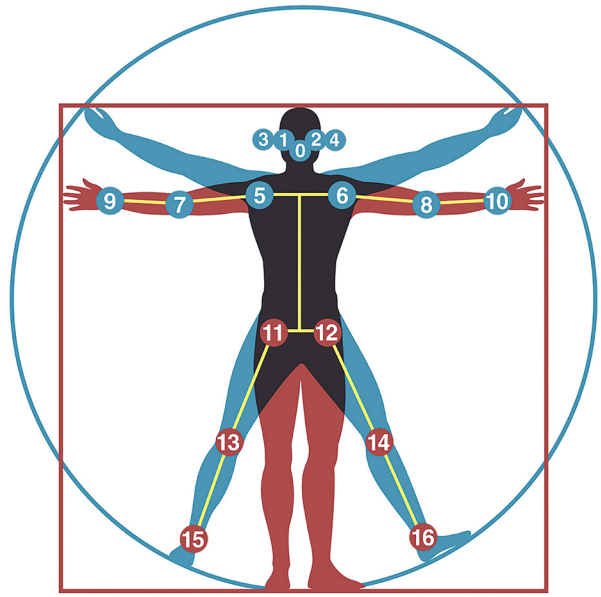


Figure 5. Default 2D skeletons of human body parts by RMPE.

Table 2. Order number of human body parts.

Order No.	Body Part
0	Nose
1	Left eye
2	Right eye
3	Left ear
4	Right ear
5	Left shoulder
6	Right shoulder
7	Left elbow
8	Right elbow
9	Left wrist
10	Right wrist
11	Left hip
12	Right hip
13	Left knee
14	Right knee
15	Left ankle
16	Right ankle

Table 3. 3D human kinematic measurements considered in thread 3.

Name of Measurement	Description	Body Parts Used	Body Part No.
Moving velocity	A vector quantity that measures the position changes in a time interval, including moving speed and direction.	Left and right hip	11, 12
Step length	The distance covered when people start walking and take one step.	Left and right ankle	15, 16
Knee flexion angle	A measurement of knee joint motion when people are moving.	Left and right hip	11, 12
		Left and right knee	13, 14
		Left and right ankle	15, 16
Arm swing angle	An essential index of the human moving pattern, including the upper-arm angle and elbow angle.	Left and right shoulder	5, 6
		Left and right elbow	7, 8
		Left and right wrist	9, 10

displayed all of the results by thread 4 in a window. As shown in Figure 8, the background is the stitching image with the left-view image of the camera and colored disparity map. Red colors on the disparity map indicate objects closer to the camera, whereas darker blue colors represent objects further away from the camera. Each joint is connected by different-colored lines. The distance of each body joint was loaded from 3D information in the queue and drawn on the left side of the background beside each body joint using 2D coordinates. All kinematic results were loaded from the queue and displayed on the colored disparity map for real-time monitoring of human locomotion.

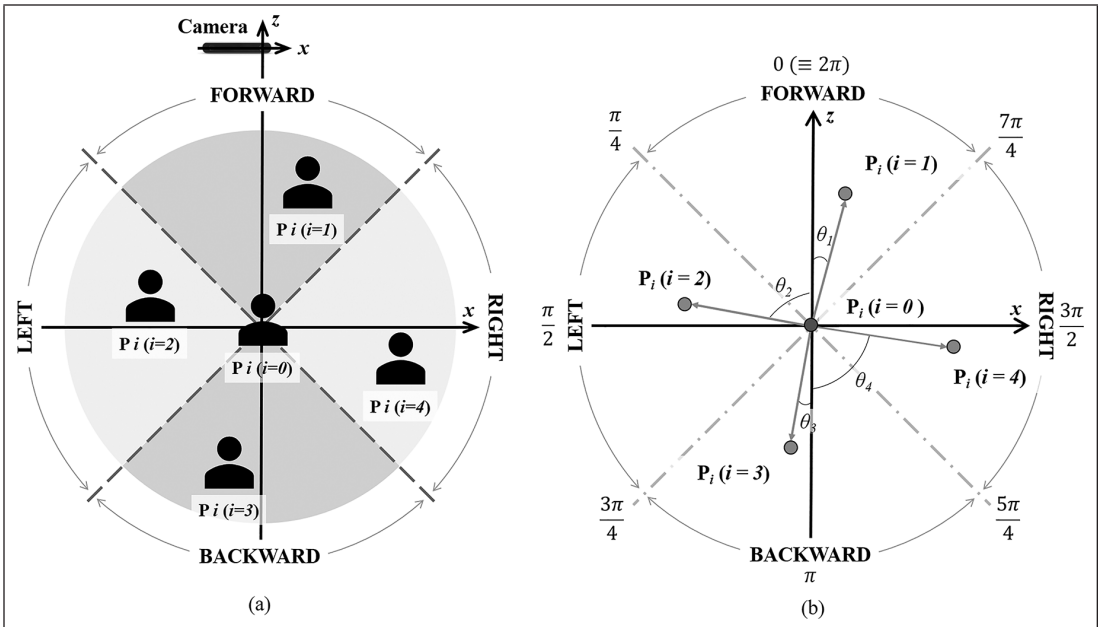


Figure 6. The geometric model of human moving direction. (a) The possible initial position and final position of a locomotory action. (b) Geometry between an initial position and each possible final position.

System Implementation and Evaluation

Hardware Configuration of the System

The camera system used in this research is a ZED camera, which includes a stereo pair of RGB cameras of the same model on the same mainboard. The baseline between the two cameras is 12 cm, and each camera has a horizontal FOV of 90° and a vertical FOV of 60°. The left and right cameras each have a focal length of 5.6 mm. The image resolution is 672 × 376 pixels for each camera, with a pixel size of 4 µm. The camera system was calibrated by the manufacturer. The camera interior orientation parameters, including the focal length, offset of the principal point, and lens distortions, and a fundamental matrix defining the relative orientation of the stereo cameras, are provided and ready for use. We used a local coordinate system in the experiment with the origin at the perspective center of the left camera, *X*-axis along the baseline, *Y*-axis pointing downwards, and *Z*-axis pointing to the range direction (see Figure 3). The camera system was run on a computer equipped with two NVIDIA RTX 2080Ti graphics cards, 64 GB of RAM, and two 12-core CPUs.

Evaluation of the System Capacity

The capabilities of the developed system were evaluated by assessing the processing rate and effective detection distance of a person moving in front of the stereo camera. During the assessment, 6000 frames were captured within 300 seconds

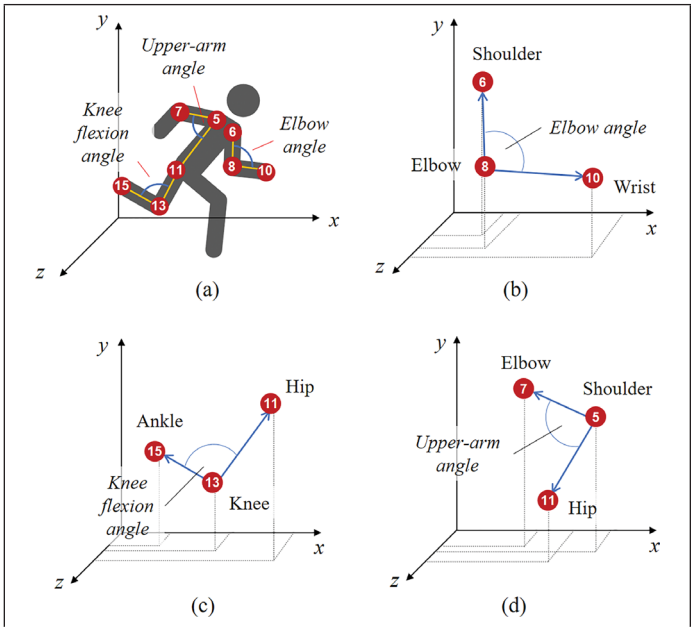


Figure 7. Geometric model of joint motion monitoring. (a) Body parts used in joint motion monitoring. For the corresponding order and name, refer to Table 1. (b)–(d) Geometric model for calculating elbow angle, knee flexion angle, and upper-arm angle.

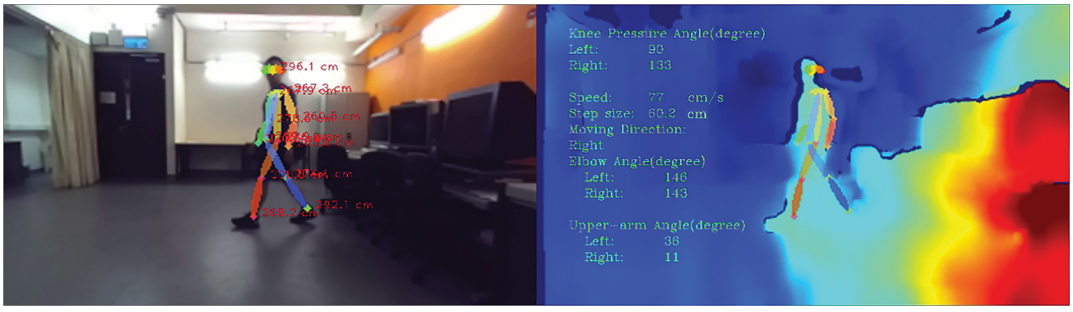


Figure 8. Visualization of the real-time photogrammetric system for human kinematics.

(Figure 9). The implementation of all threads reached ~18 fps or above with an image resolution of 672 × 376 pixels. The average processing rate of this system was 17.8 fps. Figure 9a illustrates the processing time of each frame from thread 1 to thread 4. According to Figure 9a, the processing rate sometimes exceeded 20 fps. This occurred when the person moved so fast that a ghosting effect appeared on the corresponding frames, or the illumination was so weak that the person barely disappeared from the screen. As a result, RMPE failed to extract the 2D human skeletons in the above situations. In response to such situations, thread 2 skipped the current frame and processed the next frame directly, resulting in a moderate uplift in frame rate. The entire assessment took 6000 frames (Figure 9a), and in general, the developed system achieved real-time processing during the assessment.

The system achieved an effective measurement distance of ~15 m, based on assessing a person moving back and forth from near to far along the optical axis of the left camera. During the evaluation, when the person left the camera view and returned along the same path, the system recorded all distance values from the person's waist, defined as the midpoint between the left and right hips, during this movement. As shown in Figure 9b, when the person moved ~1.5 m, the system was able to extract the 3D body features of the left and right hips and started to compute the corresponding 3D coordinates. When the person moved more than 15.7 m away from the camera, the system could not measure the distance because the person became too small on the screen to be detected by the RMPE. As the person began to move back towards the camera and moved within 14.2 m, the system was able to extract the 3D human body features again and simultaneously calculated the distance until the person moved to a distance of less than 1.1 m from the camera. Measurements were unstable starting at 14.2 m, whereas the dead zone for close-range measurements was from ~1.1 m to ~1.5 m. Thus, the effective measurement range was ~1.5 m to ~15 m, which covers a large scale of scenes.

Accuracy Evaluation of the 3D Human Body Kinematics

Distance Accuracy

To evaluate the accuracy of the distance measurements achieved by the system over a specific resolution, we had a person standing still in front of the camera at different distances (Figure 10). The distance accuracy was assessed by comparing the measured distances between the person and the camera to the ground truth. As shown in Figure 10b, the system captured 1000 frames of the person when the person was standing still in front of the camera at distances of 2.3 m,

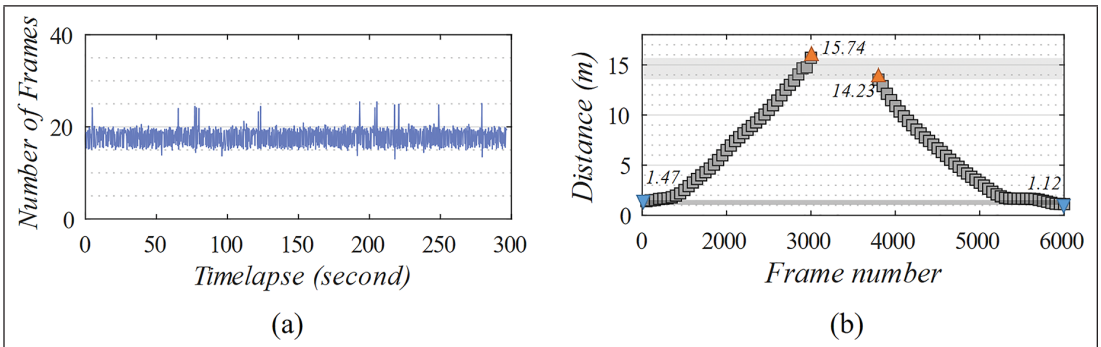


Figure 9. Efficiency assessment results of the real-time photogrammetric system. (a) Frame rate records of the system. (b) Effective measurement distance assessment results.

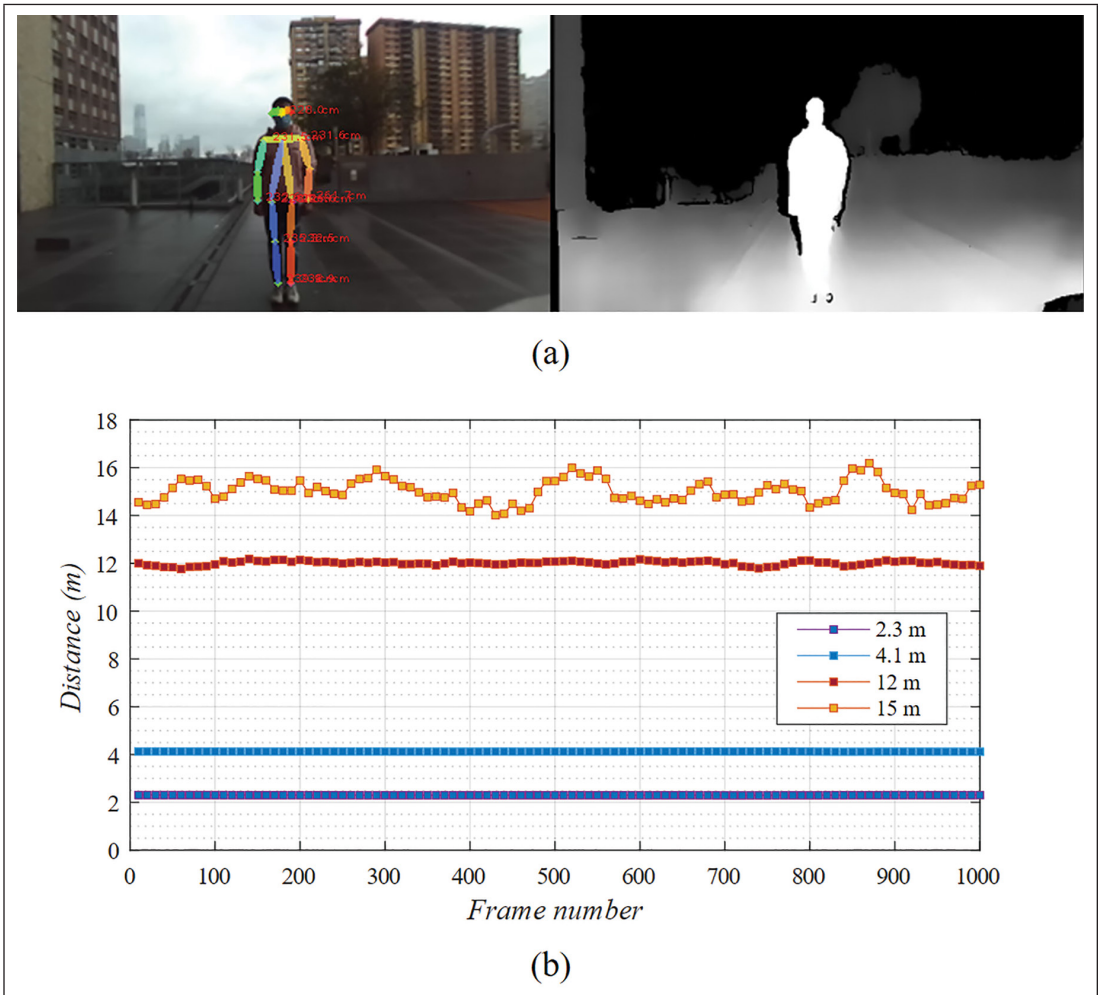


Figure 10. Evaluation of distance measurement accuracy. (a) A person standing still in front of the camera for evaluating the accuracy of the measurements. (b) Measurements of people standing in front of the camera at different distances.

4.1 m, 12 m, and 15 m. Table 4 lists the measurement averages. According to Table 4, the measurements of the system were close to the ground truth. When the person was 2.3 m and 4.1 m away from the camera, the respective root-mean-square errors (RMSEs) were 0.4 cm and 2.6 cm, respectively. The errors were 0.2% and 0.6%, respectively. As the person moved to 12 m, the measurements began to become unstable. The RMSE increased to 8.7 cm, and the error became 0.7%. When the person stood 15 m away from the camera, the measurements were even more erratic. The RMSE increased to 47.9 cm, and the error rose to 3.2%. Because the effective measurement range was ~1.5 m to 15 m (see the section "Evaluation of the System Capacity"), the measurements at 15 m were not detected on each frame. Overall, this system provided 3D human body measurements with a geometric accuracy of better than 1% of the distance within the distance range of 12 m.

Accuracy of Human Body Kinematics

The human moving direction was assessed by recording a person moving in four directions relative to the camera: left, right, forward, and backward. Figure 11 shows the method to evaluate the moving direction. Figure 11a shows the initial

Table 4. Assessment of system measurement accuracy.

Distance (m)	Mean of Measurements (m)	RMSE (m)	Error (%)
2.3	2.3	0.004	0.2
4.1	4.1	0.026	0.6
12.0	12.1	0.087	0.7
15.0	15.1	0.479	3.2

Table 5. Statistic results of moving direction identification.

Figure 11 panel	Expected Behavior	Test Times	Average θ (°)	Average Speed (cm/s)	Accuracy of Correct Identification (%)
(a)	Standing still	30	0.2	0	93
(b)	Moving left	30	88.1	52	87
(c)	Moving right	30	-89.7	55	90
(d)	Moving forward	30	-10.2	43	83
(e)	Moving backward	30	-2.4	61	93

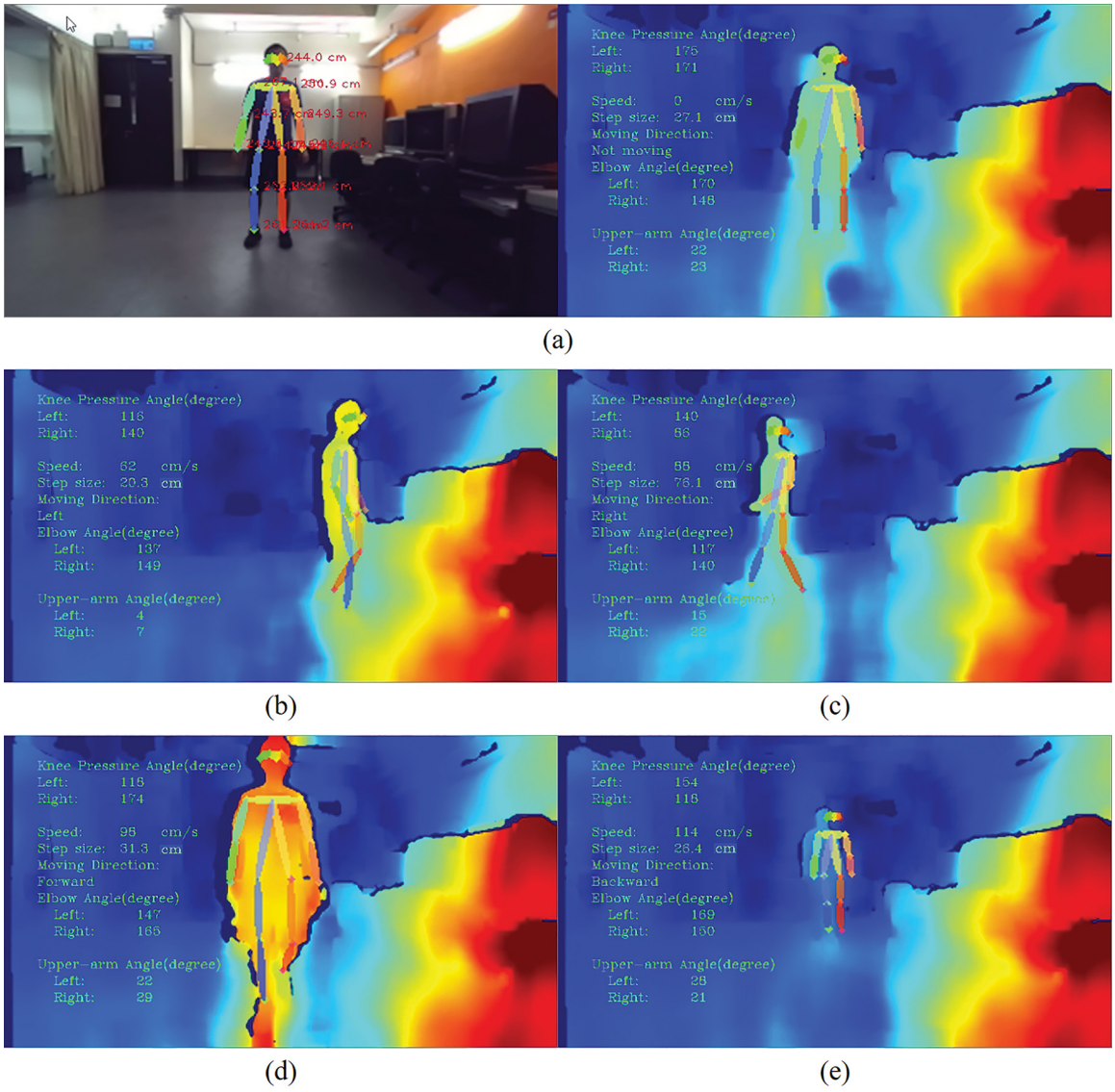


Figure 11. Results of monitoring human movement direction. The direction of movement is determined relative to the position of the camera. (a) The initial position of the human. (b) Identified result of moving left. (c) Identified result of moving right. (d) Identified result of moving forward. (e) Identified result of moving backward.

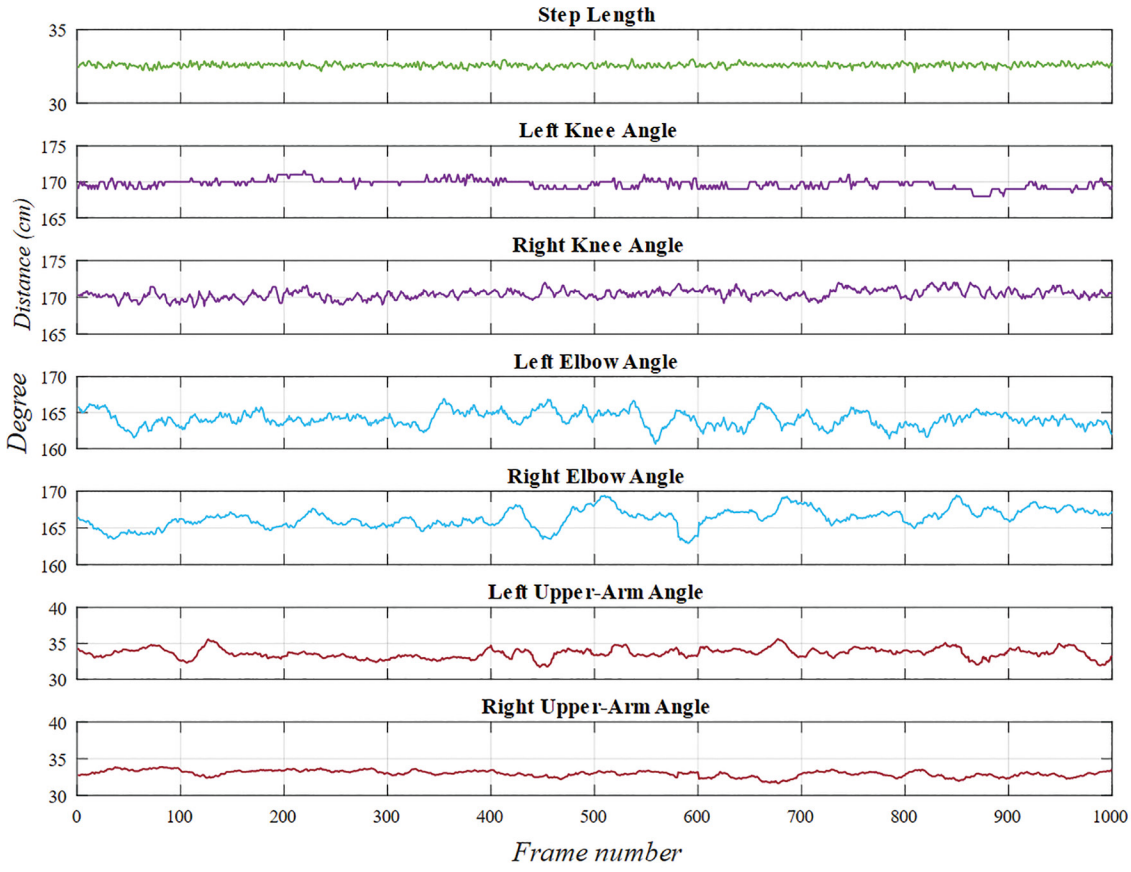
Table 6. Analysis results of kinematic applications.

Kinematic Application	Mean of Measurements	Ground Truth	RMSE	Error (%)
Step length (cm)	32.6	33.1	0.3	0.8
Knee angle (°)				
Left	169.7	176.0	6.3	3.6
Right	170.4	176.0	5.7	3.2
Elbow angle (°)				
Left	164.1	161.0	5.4	3.4
Right	166.4	160.0	7.1	4.4
Upper-arm angle (°)				
Left	33.6	35.0	2.6	7.3
Right	32.9	31.0	2.3	7.5

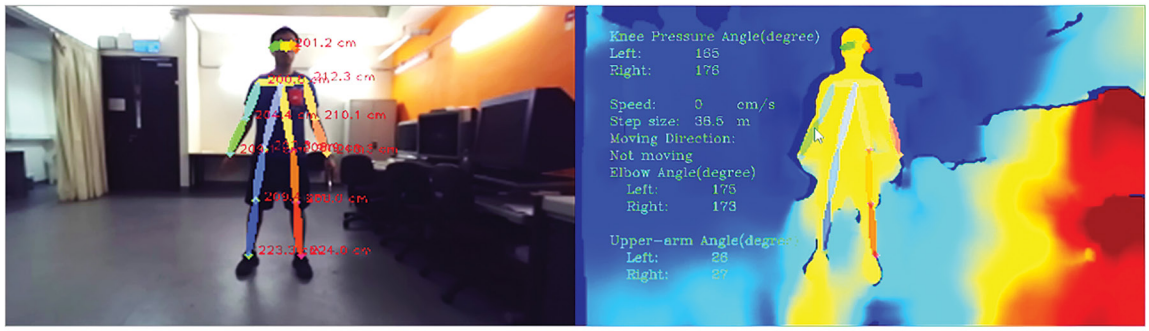
position of the person, and Figures 11b–11e display the monitoring results of the person moving in four different directions, with moving speed computed in real time.

The identification of moving direction was performed following the geometry shown in Figure 6. We evaluated each direction and repeated the measurements 30 times for each direction. The results are listed in Table 5. According to the results, the identified moving direction is generally consistent with the expected behavior in each direction, with an accuracy of over 83%. Figure 11 shows examples of the results.

Table 6 and Figure 12 present the kinematic analysis results, including step length, knee angles, elbow angles, and upper-arm swing angles, measured and recorded from 1000 frames by letting a person stand in front of the camera a while. Ground truth data were manually measured by a ruler for the step length and a protractor for the angles. An RMSE of 0.3 cm



(a)



(b)

Figure 12. Kinematic analysis of system measurements. (a) 1000-frame measurements of step length, knee flexion angles, and arm swing angles. (b) The system-measured kinematic results of a person standing still in front of the camera for a moment.

is calculated for measured step length with an error of 0.8%. The left and right knee angles fluctuate slightly around 170°, with RMSEs of 6.3° and 5.7°, respectively. The error remained at about 3%. The elbow and upper-arm angle measurements are unstable due to the high illumination at their positions on the image. This uncertainty results in an inaccuracy of the RMPE in extracting the 2D features and 3D feature conversion. The mean of the elbow angle measurements on both sides hovers at 164.1° and 166.4°, respectively. The RMSE was 5.4° and 7.1°, respectively, with an error of less than 5% for both. These values are relatively stable overall. The RMSE and errors of the left and right upper-arm angles are 2.6° (error 7.3%) and 2.3° (error 7.5%). This result indicates that the measured angles were slightly discrepant relative to the ground truth value.

Conclusions and Discussion

This paper proposed a novel real-time photogrammetric system for 3D human body feature extraction with potential applications for human kinematics. The run-time frame rate of all frameworks, including 3D map generation, 2D and 3D human body feature extraction, and human kinematic analysis, was improved by multi-threading on the CPU and CUDA programming on the GPU. The 3D map was derived from disparities using the GPU-accelerated SGM method and photogrammetry method. Human body features in 2D and 3D were extracted using the deep-learning-based RMPE method and were run in an individual thread. Several geometric models were introduced as an example of human kinematic analysis.

The experimental results presented in this paper quantitatively evaluate the efficiency and accuracy of each measurement for human kinematic analysis. The process rate (pose framerate) reached ~18 fps. The effective detection distance reached 15 m, with a geometric accuracy of better than 1% of the distance within a range of 12 m. The accuracy for real-time measurement of human body kinematics ranged from 0.8% to 7.5%. Our system achieved large-scale 3D human body feature detection. The integration of deep learning methods let the system accurately recognize 3D human body features for human kinematic analysis. With the help of multi-threading and GPU-acceleration technology, this system improved the running framerate and achieved real-time 3D human monitoring at a large scale.

There are some limitations in the experiments. The RMPE failed to detect the 2D human features when the person was moving very fast (a ghosting effect appeared on the screen) or when the illumination was dark (the person almost disappeared from the screen). Similarly, the light intensity in the environment was not constant, and the SGM did not accurately obtain the disparity value in a very high-lighting environment, such as an area near a lamp, or low-lighting environments, such as shadows. The 3D information was not extracted in these cases. Moreover, 3D body features were not extracted if a person was standing more than 15 m from the camera. At this distance, the person was so small in the image that the 2D human detection algorithm was unable to extract human skeletons. These problems can be improved by optimizing the algorithms to support higher image resolutions. The clearer outline of a person in a higher-resolution image allows the deep learning method to recognize the body features at farther distances. It should be noted that the current system is only able to process image sequences of a resolution of 672 × 376 pixels in real-time. With proper optimization of the software in our future works, real-time processing of image sequences of higher resolutions can be expected. It should also be noted that the moving directions that can be identified in the current system only allow four main directions. The algorithms will be further improved in our future work to allow the identification of more sophisticated moving directions.

In this study, although several applications of 3D human kinematics, including joint angles, movement directions, etc., were selected for demonstration and evaluation of their accuracy, the system was not limited to these applications. We expect that this study provides an insight into the potential applications of real-time 3D photogrammetry. We also hope that this system would be integrated into a portable device that could extend real-time photogrammetry to a wider range of scientific fields and industries in the future.

Acknowledgments

This work was supported by grants from the Hong Kong Polytechnic University (Project No. 1-ZVN6) and the National Natural Science Foundation of China (Project No. 41671426).

References

Abdulla, W. 2017. Mask R-CNN for object detection and instance segmentation on Keras and TensorFlow. *GitHub Repository*. <https://github.com/matterport/Ma-sk_RCNN> Accessed 18 August 2020.

Agarwal, A. and B. Triggs. 2006. Recovering 3D human pose from monocular images. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 28 (1):44–58. doi: 10.1109/TPAMI.2006.21.

Andriluka, M., L. Pishchulin, P. Gehler and B. Schiele. 2014. 2d human pose estimation: New benchmark and state of the art analysis. Pages 3686–3693 in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (CVPR), held in Columbus, Ohio.

Brownrigg, D. R. 1984. The weighted median filter. *Communications of the ACM* 27 (8):807–818. doi: 10.1145/358198.35822.

Cao, Z., G. Hidalgo, T. Simon, S.-E. Wei and Y. Sheikh. 2018. OpenPose: Real-time multi-person 2D pose estimation using Part Affinity Fields. *arXiv preprint*. arXiv:1812.08008.

Carraro, M., M. Munaro, J. Burke and E. Menegatti. 2018. Real-time marker-less multi-person 3D pose estimation in RGB-Depth camera networks. Pages 534–545 in *Proceedings International Conference on Intelligent Autonomous Systems*. Cham, Switzerland: Springer. doi: 10.1007/978-3-030-01370-7_42.

Carraro, M., M. Munaro and E. Menegatti. 2016. A powerful and cost-efficient human perception system for camera networks and mobile robotics. Pages 485–497 in *Proceedings International Conference on Intelligent Autonomous Systems*. Cham, Switzerland: Springer. doi: 10.1007/978-3-319-48036-7_35.

D'Apuzzo, N. 2002. Surface measurement and tracking of human body parts from multi-image video sequences. *ISPRS Journal of Photogrammetry and Remote Sensing* 56 (5–6):360–375. doi: 10.1016/S0924-2716(02)00069-2.

Fang, H.-S., S. Xie, Y.-W. Tai and C. Lu. 2017. RMPE: Regional multi-person pose estimation. Pages 2334–2343 in *Proceedings of the IEEE International Conference on Computer Vision*, arXiv preprint, arXiv:1612.00137.

Fraser, C. S. 1997. Digital camera self-calibration. *ISPRS Journal of Photogrammetry and Remote Sensing* 52 (4):149–159. doi: 10.1016/S0924-2716(97)00005-1.

Gholami, M., A. Rezaei, T. J. Cuthbert, C. Napier and C. Menon. 2019. Lower body kinematics monitoring in running using fabric-based wearable sensors and deep convolutional neural networks. *Sensors* 19 (23):5325. doi: 10.3390/s19235325.

Gruen, A. and H. A. Beyer. 2001. System calibration through self-calibration. In *Calibration and Orientation of Cameras in Computer Vision*, 163–193. Berlin, Heidelberg, Germany: Springer-Verlag. doi: 10.1007/978-3-662-04567-1_7.

Haala, N. 2013. The landscape of dense image matching algorithms. In *Proceedings Photogrammetric Week 2013*, held in Stuttgart, Germany. doi: 10.1.1.396.1856.

Haggag, H., M. Hossny, D. Filippidis, D. Creighton, S. Nahavandi and V. Puri. 2013. Measuring depth accuracy in RGBD cameras. Pages 1–7 in *IEEE 7th International Conference on Signal Processing and Communication Systems (ICSPCS)*, held in Gold Coast, Australia. doi: 10.1109/ICSPCS.2013.6723971.

Hernandez-Juarez, D., A. Chacón, A. Espinosa, D. Vázquez, J. C. Moure and A. M. López. 2016. Embedded real-time stereo estimation via semi-global matching on the GPU. *Procedia Computer Science* 80:143–153. doi: 10.1016/j.procs.2016.05.305.

Hirschmuller, H. 2007. Stereo processing by semiglobal matching and mutual information. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 30:328–341. doi: 10.1109/TPAMI.2007.1166.

Huang, C. C. and M. H. Nguyen. 2019. Robust 3D skeleton tracking based on OpenPose and a probabilistic tracking framework. Pages 4107–4112 in *2019 IEEE International Conference on Systems, Man and Cybernetics (SMC)*, held in Bari, Italy. doi: 10.1109/SMC.2019.8913977.

Insafutdinov, E., M. Andriluka, L. Pishchulin, S. Tang, E. Levinkov, B. Andres and B. Schiele. 2017. Arttrack: Articulated multi-person tracking in the wild. Pages 6457–6465 in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, arXiv preprint, arXiv:1612.01465.

Insafutdinov, E., L. Pishchulin, B. Andres, M. Andriluka and B. Schiele. 2016. Deepercut: A deeper, stronger, and faster multi-person pose estimation model. Pages 34–50 in *Proceedings of the European Conference on Computer Vision*. Berlin: Springer. doi: 10.1007/978-3-319-46466-4_3.

Iqbal, U. and J. Gall. 2016. Multi-person pose estimation with local joint-to-person associations. Pages 627–642 in *Proceedings of the European Conference on Computer Vision*. Berlin, Germany: Springer. doi: 10.1007/978-3-319-48881-3_44.

Jaimes, A. and N. Sebe. 2007. Multimodal human–computer interaction: A survey. *Computer Vision and Image Understanding* 108:116–134. doi: 10.1016/j.cviu.2006.10.019.

Jalal, A. and Y. Kim. 2014. Dense depth maps-based human pose tracking and recognition in dynamic scenes using ridge data. Pages 119–124 in *IEEE 11th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, held in Seoul, Korea. doi: 10.1109/AVSS.2014.6918654.

Kaehler, A. and G. Bradski. 2016. *Learning OpenCV 3: Computer Vision in C++ with the OpenCV Library*, 703–760. Sebastopol, Calif.: O'Reilly Media, Inc.

Karunarathne, M. S., S. A. Jones, S. W. Ekanayake and P. N. Pathirana. 2014. Remote monitoring system enabling cloud technology upon smart phones and inertial sensors for human kinematics. Pages 137–142 in *IEEE Proceedings of the 2014 IEEE Fourth International Conference on Big Data and Cloud Computing*. doi: 10.1109/BDCloud.2014.62.

Levinkov, E., J. Uhrig, S. Tang, M. Omran, E. Insafutdinov, A. Kirillov, C. Rother, T. Brox, B. Schiele and B. Andres. 2017. Joint graph decomposition & node labeling: Problem, algorithms, applications. Pages 6012–6020 in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, arXiv preprint, arXiv:1611.04399.

Newell, A., Z. Huang and J. Deng. 2017. Associative embedding: End-to-end learning for joint detection and grouping. Pages 2277–2287 in *Proceedings of the Advances in Neural Information Processing Systems*, held in Long Beach, Calif.

Nvidia, CUDA. 2019. *CUDA C programming guide, version 10.1.* Santa Clara, Calif.: NVIDIA Corp.

Ranjan, R., V. M. Patel and R. Chellappa. 2017. Hyperface: A deep multi-task learning framework for face detection, landmark localization, pose estimation, and gender recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 41:121–135. doi: 10.1109/TPAMI.2017.2781233.

Seitz, L., C. Haas, M. Noack and S. Wieprecht. 2018. From picture to porosity of river bed material using structure-from-motion with multi-view-stereo. *Geomorphology* 306:80–89. doi: 10.1016/j.geomorph.2018.01.014.

Seo, J., S. Han, S. Lee and H. Kim. 2015. Computer vision techniques for construction safety and health monitoring. *Advanced Engineering Informatics* 29 (2):239–251. doi: 10.1016/j.aei.2015.02.001.

Shotton, J., A. Fitzgibbon, M. Cook, T. Sharp, M. Finocchio, R. Moore, A. Kipman and A. Blake. 2011. Real-time human pose recognition in parts from single depth images. Pages 1297–1304 in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, held in Colorado Springs, Colo. doi: 10.1109/CVPR.2011.5995316.

Sridhar, S., A. Oulasvirta and C. Theobalt. 2013. Interactive markerless articulated hand motion tracking using RGB and depth data. Pages 2456–2463 in *Proceedings of the IEEE International Conference on Computer Vision*, held in Sydney, Australia.

Srivastav, V., T. Issenhuth, A. Kadkhodamohammadi, M. de Mathelin, A. Gangi and N. Padoy. 2018. MVOR: A multi-view RGB-D operating room dataset for 2D and 3D human pose estimation. *arXiv preprint*, arXiv:1808.08180.

Tang, S., Q. Zhu, W. Chen, W. Darwish, B. Wu, H. Hu and M. Chen. 2016. Enhanced RGB-D mapping method for detailed 3d indoor and outdoor modeling. *Sensors* 16 (10):1589. doi:10.3390/s16101589.

Tang, S., Q. Zhu, Y. Li, W. Chen, B. Wu, R. Guo, X. Li, C. Wang and W. Wang. 2020. Trajectory drift—Compensated solution of a stereo RGB-D mapping system. *Photogrammetric Engineering & Remote Sensing* 86 (6):359–372. doi: 10.14358/PERS.86.6.359.

Veit, A., T. Matera, L. Neumann, J. Matas and S. Belongie. 2016. Coco-text: Dataset and benchmark for text detection and recognition in natural images. *arXiv preprint*, arXiv:1601.07140.

Vlutters, M., E. H. Van Asseldonk and H. Van der Kooij. 2016. Center of mass velocity-based predictions in balance recovery following pelvis perturbations during human walking. *Journal of Experimental Biology* 219 (10):1514–1523.

Wang, C., Y. Wang, Z. Lin and A. L. Yuille. 2019. Robust 3D human pose estimation from single images or video sequences. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 41 (5):1227–1241. doi: 10.1109/TPAMI.2018.2828427.

Wu, B., X. Ge, L. Xie and W. Chen. 2019. Enhanced 3D mapping with an RGB-D sensor via integration of depth measurements and image sequences. *Photogrammetric Engineering & Remote Sensing* 85 (9):633–642. doi: 10.14358/PERS.85.9.633.

Xiong, X. and F. De la Torre. 2013. Supervised descent method and its applications to face alignment. Pages 532–539 in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (CVPR), held in Portland, Ore.

Zanfir, M., M. Leordeanu and C. Sminchisescu. 2013. The moving pose: An efficient 3D kinematics descriptor for low-latency action recognition and detection. Pages 2752–2759 in *Proceedings of the IEEE International Conference on Computer Vision* (ICCV), held in Sydney, Australia.

Zhou, S., F. Fei, G. Zhang, J. D. Mai, Y. Liu, J. Y. Liou and W. J. Li. 2013. 2D human gesture tracking and recognition by the fusion of MEMS inertial and vision sensors. *IEEE Sensors Journal* 14 (4):1160–1170. doi: 10.1109/JSEN.2013.2288094.

Zhou, X., M. Zhu, S. Leonardos, K. G. Derpanis and K. Daniilidis. 2016. Sparseness meets deepness: 3D human pose estimation from monocular video. Pages 4966–4975 in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (CVPR), held in Las Vegas, Nev.

Zimmermann, C. and T. Brox. 2017. Learning to estimate 3D hand pose from single RGB images. Pages 4903–4911 in *Proceedings of the IEEE International Conference on Computer Vision*, arXiv preprint, arXiv:1705.01389.

Zimmermann, C., T. Welschehold, C. Dornhege, W. Burgard and T. Brox. 2018. 3D human pose estimation in RGBD images for robotic task learning. Pages 1986–1992 in *Proceedings 2018 IEEE International Conference on Robotics and Automation (ICRA)*, held in Brisbane, Australia. doi: 10.1109/ICRA.2018.8462833.