A REAL-TIME PHOTOGRAMMETRIC SYSTEM FOR MONITORING HUMAN MOVEMENT DYNAMICS

Long Chen, Bo Wu*, Yao Zhao

Department of Land Surveying and Geo-informatics, The Hong Kong Polytechnic University, Hung Hom, Hong Kong – bo.wu@polyu.edu.hk

Commission II, WG II/5

KEYWORDS: Real-time photogrammetry, 3D body feature, Motion tracking, GPU, Multithreading

ABSTRACT:

The human body posture is rich with dynamic information that can be captured by algorithms, and many applications rely on this type of data (e.g., action recognition, people re-identification, human-computer interaction, industrial robotics). The recent development of smart cameras and affordable red-green-blue-depth (RGB-D) sensors has enabled cost-efficient estimation and tracking of human body posture. However, the reliability of single sensors is often insufficient due to occlusion problems, field-of-view limitations, and the limited measurement distances of the RGB-depth sensors. Furthermore, a large-scale real-time response is often required in certain applications, such as physical rehabilitation, where human actions must be detected and monitored over time, or in industries where human motion is monitored to maintain predictable movement flow in a shared workspace. Large-scale markerless motion-capture systems have therefore received extensive research attention in recent years.

In this paper, we propose a real-time photogrammetric system that incorporates multithreading and a graphic process unit (GPU)-accelerated solution for extracting 3D human body dynamics in real-time. The system includes a stereo camera with preliminary calibration, from which left-view and right-view frames are loaded. Then, a dense image-matching algorithm is married with GPU acceleration to generate a real-time disparity map, which is further extended to a 3D map array obtained by photogrammetric processing based on the camera orientation parameters. The 3D body features are acquired from 2D body skeletons extracted from regional multi-person pose estimation (RMPE) and the corresponding 3D coordinates of each joint in the 3D map array. These 3D body features are then extracted and visualised in real-time by multithreading, from which human movement dynamics (e.g., moving speed, knee pressure angle) are derived. The results reveal that the process rate (pose frame-rate) can be 20 fps (frames per second) or above in our experiments (using two NVIDIA 2080Ti and two 12-core CPUs) depending on the GPU exploited by the detector, and the monitoring distance can reach 15 m with a geometric accuracy better than 1% of the distance.

This real-time photogrammetric system is an effective real-time solution to monitor 3D human body dynamics. It uses low-cost RGB stereo cameras controlled by consumer GPU-enabled computers, and no other specialised hardware is required. This system has great potential for applications such as motion tracking, 3D body information extraction and human dynamics monitoring.

1. INTRODUCTION

Human body dynamics and posture evaluation have been an intensive research area for decades, in areas such as facial feature point-recognition algorithms (Ranjan et al., 2017; Xiong and De la Torre, 2013) and single- or multiple-person gesture recognition (Ghidoni and Munaro, 2017; Zanfir et al., 2013). Another area of interest is human-computer interaction (Jaimes and Sebe, 2007), which has been further specific to target hand keypoint recognition (Sridhar et al., 2013; Zimmermann and Brox, 2017). The next key technology integration will be posture estimation from whole-body data (Cao et al., 2018), and with the development of computer hardware technology, such as smart cameras (Carraro et al., 2016) and affordable RGB-depth sensors (Wu et al., 2019; Tang et al., 2016), some researchers have switched from developing static human posture recognition from a single image (Shotton et al., 2013) to image sequence and dynamic human posture recognition from video (Jalal and Kim, 2014).

Accelerated advances in graphic processing unit (GPU) technology and the advent of multithreading-capable CPUs have recently led to the popularity of deep learning approaches, as exemplified by algorithms for real-time human posture evaluation, such as mask regional-based convolutional neural network (R-CNN) (Abdulla, 2017), OpenPose (Cao et al., 2018) and regional multi-person pose estimation (RMPE) (Fang et al., 2017). These deep learning-based object-detection and pose-evaluation algorithms can accurately obtain the 2D keypoints of human posture. RMPE, also called 'AlphaPose', is the most reliable and accurate multi-person pose estimator, with a mean average precision (mAP) of 80+ on the common objects in context (COCO) dataset and can achieve 20+ frames per second (fps) on the fast PyTorch version (Fang et al.,

* Corresponding author. Email: bo.wu@polyu.edu.hk

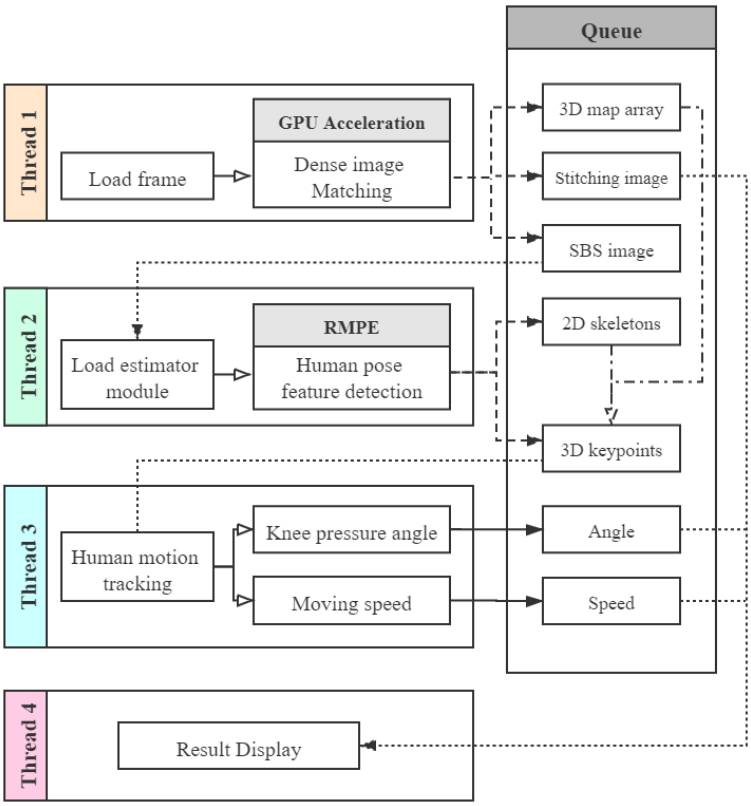


Figure 1. Workflow of the real-time photogrammetric system with multithreading and graphics-processing unit (GPU) acceleration

2017), whereas the first multi-person pose estimation algorithm OpenPose has a mAP of almost 70+ but only achieves approximately 10+ fps running on the same platform (Cao et al., 2018). Of these, only OpenPose achieves single-person real-time 3D human body keypoint-detection and posture estimation when applied to a specific stereo camera. It has a low frame rate with seconds of delay, which is equivalent to its 2D human keypoint posture-evaluation applied to a monocular camera. Additionally, the above algorithms have been used in some studies in RGB-depth sensors to obtain 3D human body keypoints for tracking (Schwarz et al., 2015) and indoor posture-estimation (Srivastav et al., 2018). However, the RGB-depth sensor is limited with respect to measurement distance and field-of-view, and sufficiently reliable only enough in for close-range real-time applications (Haggag et al., 2013).

To improve the running frame rate and efficiency of real-time 3D human body keypoint-detection and posture estimation in a large-scale real-time response, here we describe a novel real-time photogrammetric system that incorporates multithreading and GPU acceleration. This system comprises a low-cost RGB stereo-pair sensor deployed on a consumer GPU-enabled computer with two NVIDIA RTX 2080Ti graphic cards and two 12-core CPUs. The 2D human body features are extracted by RMPE on the images collected by the stereo cameras and extended to 3D human keypoints with distance computed according to the disparity generated by dense image matching from the left-view and right-view images of the stereo camera. At the same time, we use multithreading and GPU acceleration to accelerate and optimise the algorithm to achieve real-time 3D human body feature acquisition with nearly 20 fps in a specific video resolution. The effective monitoring distance can reach 15 m in the same resolution with a geometric accuracy of better than 1% of the actual distance. As the 3D body skeleton information can be applied to human-movement monitoring and tracking, this system can simultaneously obtain the distance, direction and speed information of human body movement for various applications.

2. REAL-TIME PHOTOGRAMMETRIC SYSTEM

The real-time photogrammetric system has four threads, with each thread handling different tasks as an individual model, as shown in Figure 1. Thread 1 loads a side-by-side (SBS) RGB image from the camera and uses semi-global matching (SGM, Hirschmuller, 2007) as the dense-image matching method to generate a disparity map. According to the disparity of each pixel and the camera orientation parameters, a 3D map array with three-dimensional coordinates of each pixel is generated by triangulation in thread 1. An additional stitching image with a left-view of the SBS image and disparity map is simultaneously generated in the first thread and stored in a queue with the 3D map array and SBS image, for other thread use. Thread 2 reads the SBS image and 3D map array from the queue, extracts 2D human body features by RMPE from the left-view of SBS image and then extends these to 3D body-feature coordinates associated with the 3D map array. The 3D body features are store in the same queue as a list for further use. Thread 3 reads the list of

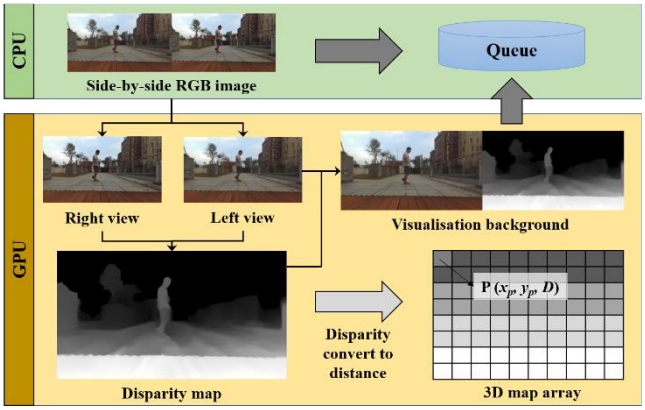


Figure 2. Dense image matching and triangulation processing

3D body features from the queue and simultaneously outputs 3D body dynamics of moving speed and knee pressure angle. All the results are stored in the same queue as in the previous thread. Thread 4 loads all the results from the queue for visualisation, namely the stitching image, 3D body features, moving speed and knee angle. These four threads work separately and individually, which means that each thread does not need to wait for the other threads to run after its task is completed, and thus can directly process the next frame and other information stored in the queue.

2.1 Disparity estimation and triangulation

The disparity estimation and triangulation are processed in thread 1, and a GPU-accelerated semi-global matching (SGM) method (Hernandez-Juarez et al., 2016) is applied to the real-time stereo estimation to obtain a disparity map. As shown in Figure 2, each frame of the stereo images in the rectified pipeline is captured by the preliminary calibrated cameras as an SBS image and saved in the host memory. The GPU device copies this image from the host memory space and splits it into left-view and right-view images, in preparation for dense-image matching by SGM. The features are extracted from the two images and used for a similarity comparison to generate a local-matching cost for each pixel and potential disparity. SGM is then used to aggregate a smoothing cost that considers the similarity of the neighbouring points and disparities along different paths, to reduce errors. In this system, the number of paths is set as four to lower computational consumption and to ensure both the quality and the performance of the real-time result. The disparity of each pixel is computed and a 3 × 3 median filter is applied to remove outliers. The resulting disparity image is copied back to the local host memory and stitched with the left-view image to form a new image array, which is then saved in the queue for visualisation.

Triangulation is used to gauge a 3D map array of each pixel from the disparity. Figure 3 shows the variables used in triangulation, which are as follows: i) the optical centre of the left camera C_1 and right camera C_2; ii) the focal length f_1 of the left camera and f_2 of the right camera; iii) the left-image plane IP1 and the right-image plane IP2; and iv) the pixel points p_1 and p_2. For any point P of the object in the real world, p_1 and p_2 are pixel-point representations of P in the IP1 and IP2 images taken from the stereo cameras at C_1 and C_2 respectively. The baseline is the offset distance between the optical centre of cameras C1 and C2. The following formula (Eq. (1)) describes the geometric relation of the above triangulation procedure:

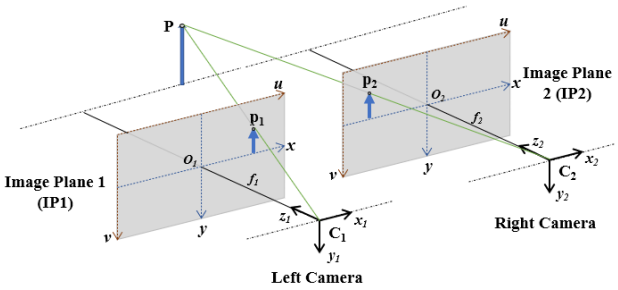


Figure 3. Triangulation of the parallel stereo camera

$$p = \begin{bmatrix} x_p \\ y_p \\ D \end{bmatrix} = \begin{bmatrix} u - c_x \\ v - c_y \\ f \end{bmatrix} \cdot \frac{b}{d} \tag{1}$$

where D represents the distance of object P in the real world,
b is baseline of the camera pair,
f is the camera focal length,
d is the disparity value of the corresponding pixel point,
(u, v) are the image plane coordinates of any pixel point,
(c_x, c_y) is the optical centre of the corresponding sensor.

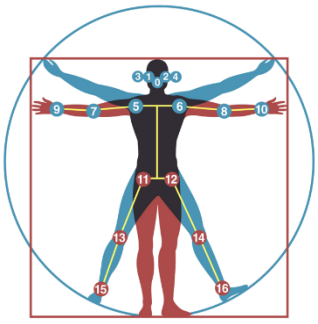
The 3D coordinates of the object in the real world in the camera coordinate system are represented by (x_p, y_p, D), with the original at the centre of the camera sensor. The focal length of a single sensor is fixed, hence the distance of different points varies solely based on its disparity component, where the disparity of each point from the left-view to the right-view images is calculated in the previous process. This disparity allows calculation of the actual distance of each pixel in the real world from the SBS images. The result of each pixel should therefore have 2D coordinates in the image plane coordinate system and 3D coordinates in the camera coordinate system. All these results are saved as a 3D map array in the queue for future use.

2.2 Extraction of 3D human body features

The 3D body feature extraction is processed in thread 2. In this system, we use the RMPE (AlphaPose) library as the pose estimator to extract and track the 2D body features of each person. This is an open-source CNN-based single person pose estimator (SPPE) method used in conventional pictorial structure models for pose estimation, and is particularly well-suited for real-time detection of RGB images. This yields a well-trained posture estimation model of the COCO dataset with 17 default keypoint outputs of human body joints, which are listed in Figure 4 with the corresponding order number.

We adopt this well-trained RMPE model in the system to extract the body features of each person in the left-view image. Each body feature is a 2D skeleton of each person in the image, which contains a set of 2D joints following the human model depicted in Figure 4. Their relationship can be represent by the following equations (Eqs. (2)):

$$\begin{aligned} &\bar{E} = \{\bar{S}_1, \bar{S}_2, \cdots, \bar{S}_k\} \\ &S = \{j_i | 0 \le i \le m\}, \quad 0 \le m \le 16, S \in \bar{E} \\ &j_i = (x_i, y_i), \quad 0 \le i \le m \end{aligned} \tag{2}$$



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

Figure 4. Default keypoints outputs of human body part in RMPE

where $\bar{E}$ is a set of human body skeletons $\bar{S}_i(i \in \{1, 2, \dots, k\})$ of k people detected by RMPE in the image. Each skeleton S is the set of 2D joints $j_i(i \in \{1, 2, \dots, m\})$ of each person that contain 2D coordinates (x_i, y_i) detected on the left-view image, and m is the total number of body parts listed in Figure 4 We use the 2D coordinates of each joint as an index to search corresponding 3D coordinates from the 3D map array. Thus, a set of 3D body features of each person containing distance information is derived in this step and saved in the queue for use in the following step.

2.3 Measurement of human movement dynamics

A dynamic model of a 3D human body captures how body parts change in 3D over a small interval of time. We therefore formulate this problem as the geometric relationship changes of human body keypoints in 3D space on thread 3. We apply 3D coordinate information of body joints derived from the previous section on knee kinematics for a straightforward illustration of human dynamics in this system.

The knee is the most affected site during walking and running injuries, some of which are believed to be caused by abnormal knee motion (Lysholm and Wiklander, 1987). Thus, 3D joint information is useful for performing an effective assessment of knee kinematics during human movement, as it reveals potential injuries in which knee angles play an important role. Specifically this involves estimation of muscle activation and investigation of the possible influence of different knee angles on muscle inhibition (Suter and Herzog, 1997). There are six essential joints in 3D body features that are used to formulate the knee angle computation, as shown in Figure 5. The calculation of knee angle is simplified to a geometric problem as follows (Eq (3)):

$$\cos \alpha_k = \overrightarrow{KH} \cdot \left\|\overrightarrow{KH}\right\|^{-1} \cdot \overrightarrow{KA} \cdot \left\|\overrightarrow{KA}\right\|^{-1} \qquad (3)$$

where α_k represents the shortest angle between two vectors of knee-hip and knee-ankle as the knee angle,
$\overrightarrow{KH}$ is vector from knee to hip,
$\overrightarrow{KA}$ is the vector from knee to ankle.

The magnitude of the two vectors changes with the hip, knee and ankle during movement, hence the knee angle α_k varies on the basis of the 3D coordinates of the above joints. As before, all the results are saved in the queue for visualisation in the next section.

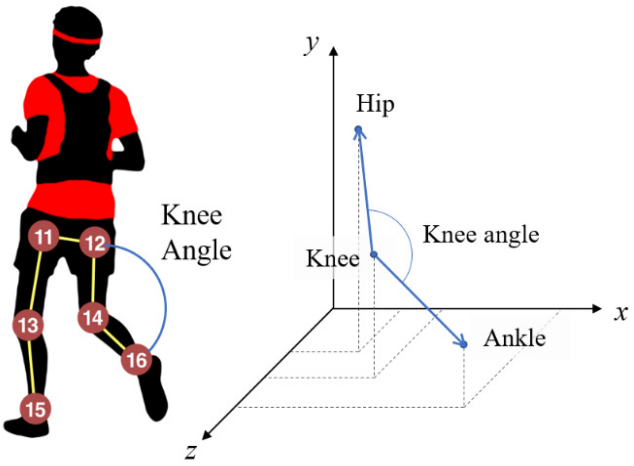


Figure 5. Geometry between knee, hip and ankle used in moving dynamic computation

3. EXPERIMENTAL EVAULATION

The visualisation is an individual thread that loads all the information saved in the queue and displays it on the screen. Once the threading detects that the queue is full of the stitching map from thread 1, the 3D body features extended from the 2D skeleton in thread 2, and the knee angles computed from the 3D body features in thread 4, it automatically displays all of the results in a window. As shown in Figure 6, all the results are loaded and visualised in thread 4 from the queue. The stitching map is loaded as the background of the visualisation window. All the 3D body features are loaded from the queue and drawn on the left side of the background, according to the image-plane coordinates of each joint. Each joint is connected by different-coloured lines, and the distance information is directly displayed beside each joint. The knee angles of the left and right legs are also loaded from the queue and directly shown on the right side of the background enable for real-time human dynamic monitoring. The system is run on a computer equipped with two NVIDIA RTX 2080Ti graphics cards, 64 GB of RAM and two 12-core CPUs. The achievable average framerate is 20 fps at a resolution of 1377 × 376 pixels (with each view being 677 × 376 pixels).

The maximum effective measurement distance of this system reaches 15 m, as assessed by a person moving back and forth from near to far along the direction of the optical axis of the left camera. The assessment result is shown in Figure 7. The system captures 6,000 frames and records the distance value of the waist, which is the middle point of the left and right hips when a person moves away from the camera and returns along the same route. When the person moves to ~100 cm, the system can extract the 3D body features of the left and right hips and start to record their 3D coordinates. The distance cannot be measured when the person moves more than 1500 cm away, because the person becomes so small on the screen that the system cannot extract 2D body features. Therefore, the maximum measurement distance is 1570 cm, and the minimum distance is 111 cm.

To assess the accuracy of the distance measurements achieved in this system over a specific resolution, we had one person stand still

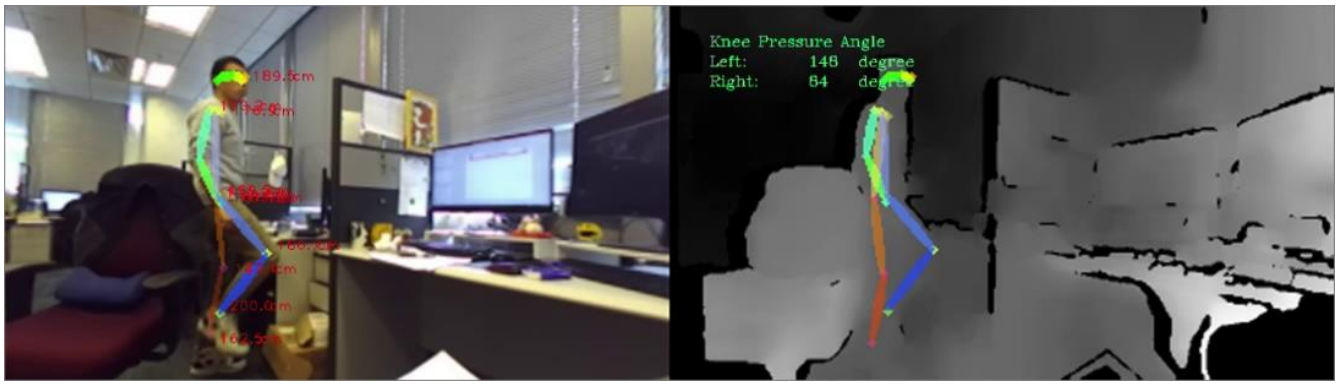


Figure 6. Experimental result and visualisation of all the information stored in the queue

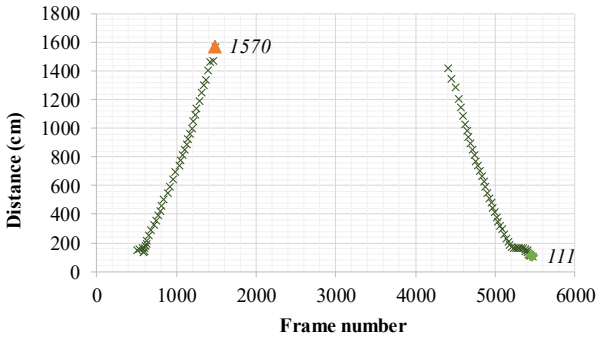


Figure 7. Maximum and minimum effective measurement distances

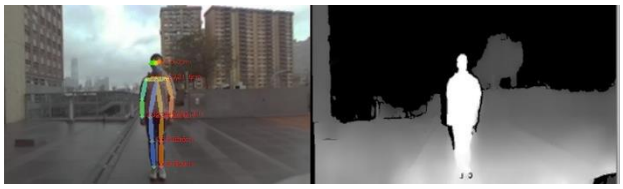


Figure 8. A person stands still while in front of the camera to enable the distance value of the waist to be recorded

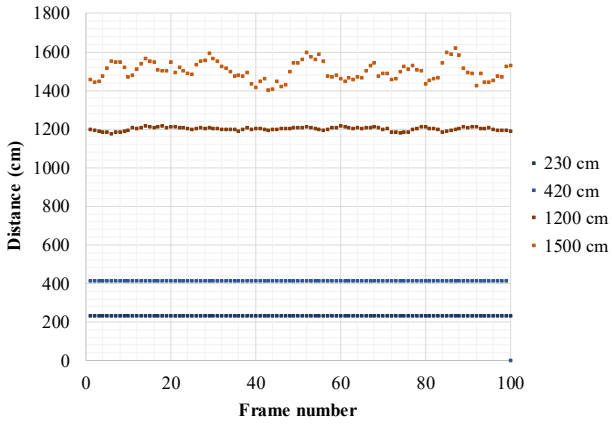


Figure 9. Accuracy assessment of the results from distance measurements of a person standing still at different distances in front of the camera

in front of the camera at different distances (Figure 8) and compare the distance measurements with the ground truth. As shown in Figure 9, the system captures 100 frames of the person standing still in front of the camera at distances of 230, 410, 1200 and 1500 cm. The measurements are close to the ground truth when the person is 230 and 410 cm away with respective root mean square errors (RMSEs) of 0.37 and 2.56 and accuracies of 0.16% and 0.63%. When the person moves to 1200 cm, the measurements begin to destabilise, as shown by an RMSE of 8.74 and an accuracy of 0.73%. When the person stands 1500 cm away from the camera, the measurements are extremely unstable, as shown by an RMSE of 47.88 and an accuracy of 3.2%. Therefore, this system provides an average geometry accuracy of better than 1% of the distance within an effective measurement distance of 15 m.

4. CONCLUSIONS AND DISCUSSION

The novel real-time photogrammetric system described herein provides a solution for 3D feature extraction of different human body parts and potential applications. However, 3D body features cannot be extracted if the person stands more than 15 m from the camera, because at this distance the resolution is too low for computing the disparity value. Furthermore, the body features cannot be detected, as the person becomes smaller on the screen with increasing distance.

These problems will be solved by optimising algorithms and programming code, which we will undertake in future work. We expect that this system will be operable at high resolution to increase the measurement distance and geometric accuracy. It is suited for use on a portable integrated processing unit for application of real-time photogrammetry to a wider range of scientific fields and industries.

ACKNOWLEDGEMENTS

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*. URL: https://github.com/matterport/Mask_RCNN

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

Carraro, M., Munaro, M., Menegatti, E., 2016. A powerful and cost-efficient human perception system for camera networks and mobile robotics. *International Conference on Intelligent Autonomous Systems*. Springer, pp. 485-497.

Fang, H.-S., Xie, S., Tai, Y.-W., Lu, C., 2017. Rmpe: Regional multi-person pose estimation. *Proceedings of the IEEE International Conference on Computer Vision*, pp. 2334-2343.

Ghidoni, S., Munaro, M., 2017. A multi-viewpoint feature-based re-identification system driven by skeleton keypoints. *Robotics and Autonomous Systems* 90, 45-54.

Haggag, H., Hossny, M., Filippidis, D., Creighton, D., Nahavandi, S., Puri, V., 2013. Measuring depth accuracy in RGBD cameras. *2013, 7th International Conference on Signal Processing and Communication Systems (ICSPCS). IEEE*, pp. 1-7.

Hernandez-Juarez, D., Chacón, A., Espinosa, A., Vázquez, D., Moure, J.C., López, A.M., 2016. Embedded real-time stereo estimation via semi-global matching on the GPU. *Procedia Computer Science* 80, 143-153.

Hirschmuller, H., 2007. Stereo processing by semiglobal matching and mutual information. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 30, 328-341.

Jaimes, A., Sebe, N., 2007. Multimodal human–computer interaction: A survey. *Computer Vision and Image Understanding* 108, 116-134.

Jalal, A., Kim, Y., 2014. Dense depth maps-based human pose tracking and recognition in dynamic scenes using ridge data. *2014 11th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS). IEEE*, pp. 119-124.

Lysholm, J., Wiklander, J., 1987. Injuries in runners. *American Journal of Sports Medicine* 15, 168-171.

Ranjan, R., Patel, V.M., Chellappa, R., 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.

Schwarz, M., Schulz, H., Behnke, S., 2015. RGB-D object recognition and pose estimation based on pre-trained convolutional neural network features. *2015 IEEE International Conference on Robotics and Automation (ICRA). IEEE*, pp. 1329-1335.

Shotton, J., Sharp, T., Kipman, A., Fitzgibbon, A., Finocchio, M., Blake, A., Cook, M., Moore, R., 2013. Real-time human pose recognition in parts from single depth images. *Communications of the ACM* 56, 116-124.

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

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

Suter, E., Herzog, W., 1997. Extent of muscle inhibition as a function of knee angle. *Journal of Electromyography and Kinesiology* 7, 123-130.

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

Wu, B., Ge, X., Xie, L., Chen, W., 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., De la Torre, F., 2013. Supervised descent method and its applications to face alignment. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 532-539.

Zanfir, M., Leordeanu, M., Sminchisescu, C., 2013. The moving pose: An efficient 3d kinematics descriptor for low-latency action recognition and detection. *Proceedings of the IEEE International Conference on Computer Vision*, pp. 2752-2759.

Zimmermann, C., Brox, T., 2017. Learning to estimate 3d hand pose from single rgb images, *Proceedings of the IEEE International Conference on Computer Vision*, pp. 4903-4911.