Vol.2017-CG-168 No.36
Vol.2017-DCC-17 No.36
Vol.2017-CVIM-209 No.36
2017/11/9

A Motion Capture System Using a Smartphone and a Smartwatch

Chuanhua Lu[1] Maxime Deleau[2] Hideaki Uchiyama[1] Diego Thomas[1] Rin-ichiro Taniguchi[1]

Abstract: We propose a simple but effective motion capture system comprising a fixed smartphone used as a camera and a smartwatch wore by a target person for absolute scale human pose estimation. Owing to the recent advance of machine learning techniques, 3D human pose can be inferred from a single image. However, the estimated pose suffers from scale ambiguity because of the essential nature in the 3D estimation from a 2D image. Therefore, we propose a simple setup such that a target person wears a natural, non-invasive, and easy-to-wear smartwatch that can measure inertial data, and is captured by a fixed smartphone. To estimate absolute human poses, the trajectory of a wrist where the smartwatch is attached is first computed from both images and inertial data. Then, the trajectory from images is metrically scaled by computing the ratio between the two trajectories. By estimating the scale for human poses from images, the absolute human poses can be computed.

Keywords: Motion capture system, Smartphone, Smartwatch, IMU, Deep learning

1. Introduction

Vision based human motion capture systems have commonly been used to measure the behavior of people in various situations such as sports, film making, and virtual reality [10]. VICON[*1] and OptiTrak[*2] are commercial systems that can accurately measure 3D metric human motions. In their systems, multiple cameras are first fixed and carefully calibrated, and then some fiducial markers are attached on a target person so that the 3D positions of the markers can be estimated for the motion capture. Since their system configuration is complicated, simpler systems have been developed in the literature.

With the advent of RGB-D cameras (e.g. Microsoft Kinect and Intel RealSense), the 2D/3D pose of a human body can be computed from a single depth image in real-time with machine learning techniques [13]. However, the depth ranges of RGB-D cameras are normally limited, and the cameras cannot acquire depth images in outdoor environments because of the use of infrared lights. To overcome these drawbacks, motion capture systems using only a single RGB camera have also been proposed [1,4]. In recent years, the accuracy is drastically improved owing to the advance of deep learning based techniques [2,9]. One drawback of a single camera based approach is that the estimated 3D human pose is basically up to scale because the inference of 3D pose from a 2D image is essentially ill-posed.

In this paper, we propose a simple but effective motion capture system that can measure absolute scale human poses. The system simply comprises a fixed single camera to capture a target person, and an IMU attached to the person to estimate the trajectory of a body part in metric. This system can be realized by using a smartphone as a camera and a smartwatch as an IMU. A smartwatch has been introduced in recent years, and can be used as a natural, non-invasive and easy-to-wear watch attached to a wrist of the target person. Therefore, we propose an algorithm that human pose is first computed by using images, and then its absolute scale is computed by using inertial data from the IMU. In other words, the trajectory of a wrist is first computed by using a state-of-the-art method on 3D human pose estimation from images. Simultaneously, the same trajectory is accurately computed in metric by using inertial data based on a technique of pedestrian dead reckoning (PDR). Finally, the trajectory from images is metrically scaled by computing the ratio between the two trajectories. In Table 1, the advantage of our motion capture system is described compared with RGB-D camera or RGB camera based systems.

Table 1: Advantage of our proposed system

Configuration	Metric	Wide range
RGB-D camera	✓	×
RGB camera	×	✓
RGB camera + IMU (ours)	✓	✓

2. Overview

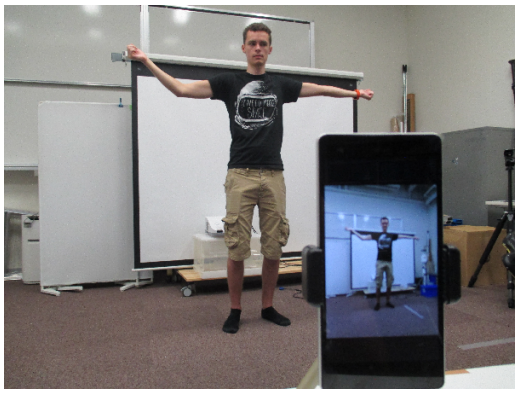
2.1 System configuration

As illustrated in Figure 1, our system comprises a fixed smartphone as a camera to capture a target person, and a smartwatch wore by the person. We used an android smartphone[*3] as both a camera device and a receiver of IMU data transferred from an android smartwatch[*4]. The frequency of acquiring images and inertial data is 10 Hz and 100 Hz, respectively, and they are synchronized by the time stamp. The resolution of the image is 640×360

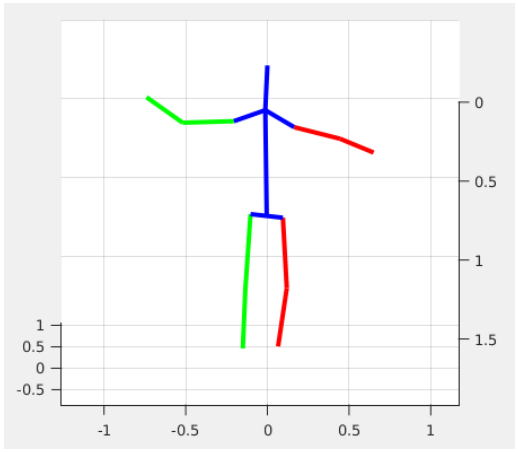
[1] Kyushu University, Japan, http://limu.ait.kyushu-u.ac.jp/e/
[2] Telecom Saint-Etienne, French
[*1] https://www.vicon.com/
[*2] http://optitrack.com/
[*3] VAIO Phone A
[*4] ASUS ZenWatch 2

© 2017 Information Processing Society of Japan

Vol.2017-CG-168 No.36
Vol.2017-DCC-17 No.36
Vol.2017-CVIM-209 No.36
2017/11/9



(a) System configuration



(b) Estimated human pose in metric

Fig. 1: Motion capture system using a smartphone and a smartwatch

pixels. In the smartwatch, the orientation is computed from both acceleration and angular velocity by using Madgwick's IMU filter [8], and then is transferred to the phone with the acceleration data. Note that the video and the IMU data are post-processed in our current implementation, but the implementation can easily be modified for online systems.

2.2 Algorithm

The basic idea is to use a state-of-the-art method for 3D human pose estimation using an image as external libraries, and use the inertial data from the IMU to estimate the metric scale of the human pose. This can be considered a loosely-coupled approach for the scale estimation, as similar to scale estimation for monocular visual SLAM using a face size [7], because the scale estimation is not tightly incorporated into human pose estimation.

As described in Fig. 2, we first compute the 3D human poses only from images, and extract the trajectory of a wrist from the estimated poses. In this process, there is the scale ambiguity for the poses. Simultaneously, the 3D metric trajectory of an IMU attached to the wrist is accurately computed from both images and inertial data based on zero velocity update used in pedestrian dead reckoning (PDR). By computing the ratio between the two wrist trajectories, the 3D human pose can be metrically scaled.

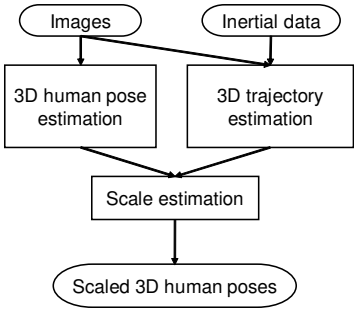


Fig. 2: Flow of Algorithm.

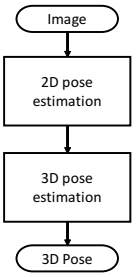


Fig. 3: Flow of 3D human pose estimation.

3. 3D human pose estimation using an image

3.1 A deep learning based approach

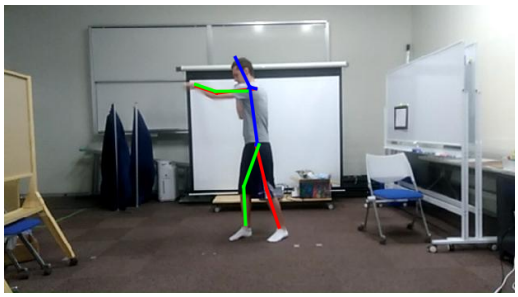
We briefly summarize a deep learning based 3D human pose estimation from an image. As illustrated in Figure 3, 2D human pose in an image is first computed from an image [5]. The accuracy of this process became higher owing to deep neural network(DNN). Then, the 3D pose of the 2D pose can be determined based on a machine learning technique with a large dataset [2]. For the dataset, a large number of pairs of a 3D pose and its 2D pose in an image are prepared, and then they are used for training the DNN so that 3D pose can be inferred from a 2D pose. This indicates that the 3D pose of a person in an image can be determined only from an image. Since it is not possible to distinguish between a small person standing near the camera and a tall person standing far from the camera, the estimated 3D pose basically has scale ambiguity.

3.2 Problems of existing approach

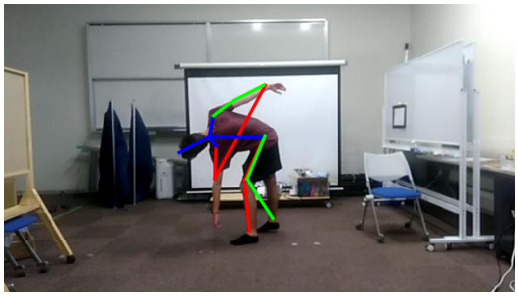
We found several problems of the 3D human pose estimation from a single image, as illustrated in Figure 4. First, the 2D pose estimation is not perfect. In fact, in some cases of occlusions or background textures, the detected joint positions in an image can be totally wrong. To lower the impact of this problem, we detect obvious wrong pose estimation by simply thresholding the wrist displacement. If the displacement of the wrist between two frames is more than a threshold, such case is classified as a wrong pose. Then, the wrong poses are replaced with poses obtained by linear regression of the neighbor poses.

The 3D pose estimation from the 2D pose also has some limitations. Basically, estimating the 3D pose only from the 2D pose is an ill-posed problem. This sometimes results in bad perspective reconstruction. Another limitation is that the depth estimation of the pose is not stable. Even though a person remains stationary,

© 2017 Information Processing Society of Japan



(a) Two arms are detected as punching whereas the right arm is not



(b) A hand region cannot be distinguished from the background

Fig. 4: Failure cases of 2D skeleton estimation from a single image.

the estimated Z value will can often change. This causes a big problem to estimate the scale since we need to compute the 3D displacement of the wrist from the estimated 3D human poses. To solve this problem, we propose to normalize the human pose at a fixed distance.

3.3 Our solution

First, we normalize the estimated 3D pose so that the length of the legs will be constant in every frame. This can compensate the effect of perspective projection which makes the 3D model bigger when the person is closer to the camera. Next, we lock the Z coordinate of the neck joint to 0 to compensate the instability of the depth estimation. With this compensation, we may lose the accuracy of the depth estimation. However, this is still reasonable if we consider that neck position does not largely move between two frames. The normalization process can be described in the following equation:

$$Pose = \frac{\begin{bmatrix} X_1 & Y_1 & Z_1 \\ X_2 & Y_2 & Z_2 \\ \dots & \dots & \dots \\ X_{14} & Y_{14} & Z_{14} \end{bmatrix} - \begin{bmatrix} 0 & 0 & Z_2 \\ 0 & 0 & Z_2 \\ \dots & \dots & \dots \\ 0 & 0 & Z_2 \end{bmatrix}}{Legs\ length}$$

where $(X_n \quad Y_n \quad Z_n)$ is the position of the n-th joint in the 3D pose, and Z_2 is the Z coordinate of the neck joint (2nd joint) in the 3D pose estimation.

4. 3D trajectory estimation using an IMU

4.1 Metric trajectory estimation

Since we take a loosely-coupled approach for absolute scale estimation, the trajectory of an IMU is independently estimated from the human pose estimation. The flow is illustrated in Figure 5.

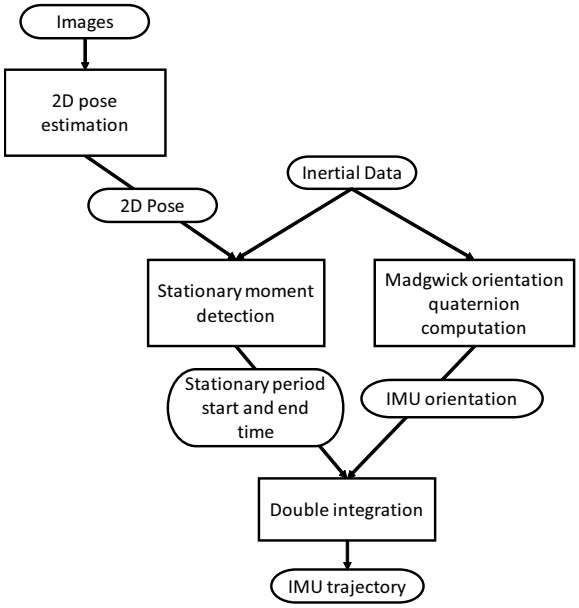


Fig. 5: Flow of the IMU trajectory estimation.

First, the orientation of the IMU is computed to express the acceleration of the sensor in the earth frame. In our system, we used the Madgwick's algorithm [8] to compute the orientation quaternion from the accelerometer and gyroscope data. Note that we did not use the magnetometer data because of magnetic distortion.

The next step is the double integration of the acceleration to estimate the trajectory of the IMU. However, this needs an accurate initial velocity. If the initial velocity is wrong, the wrong velocity is integrated such that the more time the movement lasts and the less reliable the computed trajectory will become. This often occurs when using a cheap IMU such as the ones integrated in other devices like smartwatches and smartphones. To estimate an accurate velocity, a technique of zero velocity update has been proposed in PDR [3, 6, 11, 12]. The idea is to set the velocity to 0 when the IMU is static. This can reset the velocity and suppress the error accumulation. To use this technique, we need to detect the stationary periods of the IMU movement, as described in the next section.

4.2 Stationary moment detection

Stationary moment detection basically relies on some signal processing on the acceleration magnitude[*5]. After removing the offset of the acceleration magnitude using a high-pass filter, it is smoothed, and then thresholded to find the stationary moments. In PDR, such method works fine if the IMU is attached at a foot. However, it is not totally adapted for our case because we track the trajectory of a wrist, which is different from the trajectory of a foot in PDR. While walking, the foot is subject to strong acceleration because of the nature of the movement. However, to track a wrist under free movement, the acceleration is not always high and do not always stand out from the noise. Specifically, in case of a linear movement, we cannot use a simple thresholding on the smoothed acceleration magnitude.

For this reason, we use the wrist displacement in pixels com-

[*5] http://x-io.co.uk/gait-tracking-with-x-imu/

© 2017 Information Processing Society of Japan

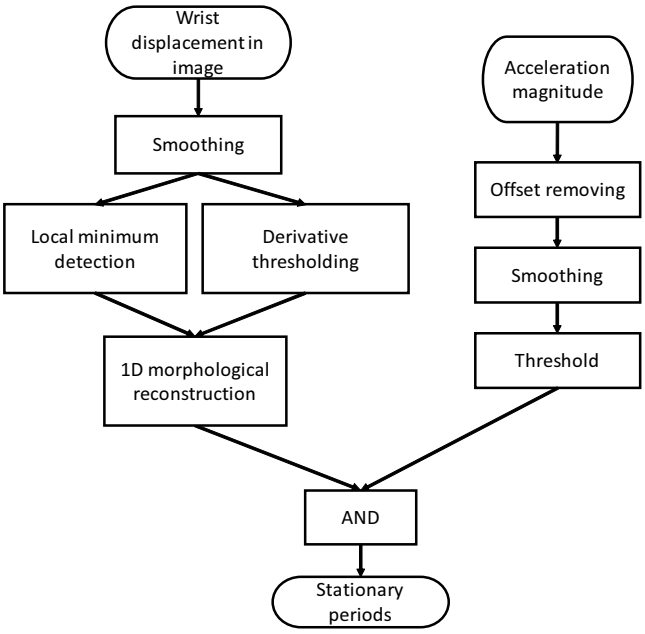


Fig. 6: Stationary moment detection.

puted from the 2D pose estimation, as illustrated in Figure 5. By finding the local minimums of the smoothed wrist displacement, we can find the instants when the wrist is not moving. Also, by thresholding the absolute value of the wrist displacement derivative, we can find the duration of each stationary period. If we only keep the periods which contain a local minimum, we will get all the stationary periods.

The next step is to fuse the two methods, as illustrated in Figure 6. A simple logical AND between the outputs of the two methods allows to get a good detection of the stationary periods. The combination of the two methods ensures that the stationary periods are well detected by checking the video data. Note that the synchronization of the data is crucial for this method to give good results.

5. Scale estimation from trajectory of wrist

The estimation of the scale is done by comparing the displacement of the wrist computed from the IMU trajectory and in the estimated 3D pose from the images. In that sense, the scale can be obtained according to the following formulation:

$$Scale = \frac{IMU\ displacement}{Image\ based\ displacement}.$$

The initial idea was to compute the scale using the displacement between the beginning and the end of each movement(between the end of a stationary period and the beginning of the next stationary period). However, because of the uncertainty of the IMU measurements and the 3D pose, the results were not stable. Therefore, our new idea was to compute the scale using the displacement of the wrist between random moments in the movements. The more time we compute the scale, the more the scale distribution takes a Gaussian shape. This was confirmed through the experiments with 100,000 samples. As illustrated in Figure 7, the scale distribution takes a Gaussian shape. The expectation of the distribution is about 1.69 m, this result is quite near to the exact height of our target size (1.77 m). By taking the mean value of that distribution, we can have a good estimate of the scale.

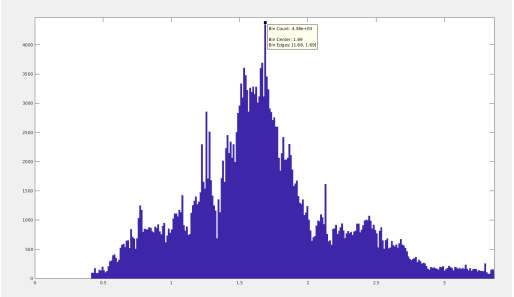


Fig. 7: Distribution of estimated scales.

6. Conclusion

We presented a simple motion capture system using a smartphone and a smartwatch. Since 3D pose estimation from an image has the scale ambiguity, the ambiguity was solved using a smartwatch wore by a target person. The evaluation of scale estimation needs to be more investigated.

References

[1] Agarwal, A. and Triggs, B.: Recovering 3D human pose from monocular images, *IEEE transactions on pattern analysis and machine intelligence*, Vol. 28, No. 1, pp. 44–58 (2006).

[2] Chen, C.-H. and Ramanan, D.: 3D Human Pose Estimation = 2D Pose Estimation + Matching, *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2017).

[3] El-Gohary, M. and McNames, J.: Human joint angle estimation with inertial sensors and validation with a robot arm, *IEEE Transactions on Biomedical Engineering*, Vol. 62, No. 7, pp. 1759–1767 (2015).

[4] Guan, P., Weiss, A., Balan, A. O. and Black, M. J.: Estimating human shape and pose from a single image, *Computer Vision, 2009 IEEE 12th International Conference on*, IEEE, pp. 1381–1388 (2009).

[5] Insafutdinov, E., Pishchulin, L., Andres, B., Andriluka, M. and Schiele, B.: Deepercut: A deeper, stronger, and faster multi-person pose estimation model, *European Conference on Computer Vision*, Springer, pp. 34–50 (2016).

[6] Jiménez, A. R., Seco, F., Prieto, J. C. and Guevara, J.: Indoor pedestrian navigation using an INS/EKF framework for yaw drift reduction and a foot-mounted IMU, *Positioning Navigation and Communication (WPNC), 2010 7th Workshop on*, IEEE, pp. 135–143 (2010).

[7] Knorr, S. B. and Kurz, D.: Leveraging the User's Face for Absolute Scale Estimation in Handheld Monocular SLAM, *Mixed and Augmented Reality (ISMAR), 2016 IEEE International Symposium on*, IEEE, pp. 11–17 (2016).

[8] Madgwick, S. O., Harrison, A. J. and Vaidyanathan, R.: Estimation of IMU and MARG orientation using a gradient descent algorithm, *Rehabilitation Robotics (ICORR), 2011 IEEE International Conference on*, IEEE, pp. 1–7 (2011).

[9] Mehta, D., Sridhar, S., Sotnychenko, O., Rhodin, H., Shafiei, M., Seidel, H.-P., Xu, W., Casas, D. and Theobalt, C.: VNect: Real-time 3D Human Pose Estimation with a Single RGB Camera, Vol. 36, No. 4, (online), available from ⟨http://gvv.mpi-inf.mpg.de/projects/VNect/⟩ (2017).

[10] Poppe, R.: Vision-based human motion analysis: An overview, *Computer vision and image understanding*, Vol. 108, No. 1, pp. 4–18 (2007).

[11] Sabatini, A. M.: Quaternion-based extended Kalman filter for determining orientation by inertial and magnetic sensing, *IEEE Transactions on Biomedical Engineering*, Vol. 53, No. 7, pp. 1346–1356 (2006).

[12] Sagawa, K., Abo, S., Tsukamoto, T. and Kondo, I.: Forearm trajectory measurement during pitching motion using an elbow-mounted sensor, *Journal of Advanced Mechanical Design, Systems, and Manufacturing*, Vol. 3, No. 4, pp. 299–311 (2009).

[13] Shotton, J., Sharp, T., Kipman, A., Fitzgibbon, A., Finocchio, M., Blake, A., Cook, M. and Moore, R.: Real-time human pose recognition in parts from single depth images, *Communications of the ACM*, Vol. 56, No. 1, pp. 116–124 (2013).

© 2017 Information Processing Society of Japan