Markerless Motion Capture using Appearance and Inertial Data

Charence Wong, Zhiqiang Zhang, Benny Lo and Guang-Zhong Yang

Abstract— **Current monitoring techniques for biomechanical analysis typically capture a snapshot of the state of the subject due to challenges associated with long-term monitoring. Continuous long-term capture of biomechanics can be used to assess performance in the workplace and rehabilitation at home. Noninvasive motion capture using small low-power wearable sensors and camera systems have been explored, however, drift and occlusions have limited their ability to reliably capture motion over long durations. In this paper, we propose to combine 3D pose estimation from inertial motion capture with 2D pose estimation from vision to obtain more robust posture tracking. To handle the changing appearance of the human body due to pose variations and illumination changes, our implementation is based upon Least Soft-Threshold Squares Tracking. Constraints on the variation of the appearance model and estimated pose from an inertial motion capture system are used to correct 2D and 3D estimates simultaneously. We evaluate the performance of our method with three state-of-the-art trackers, Incremental Visual Tracking, Multiple Instance Learning, and Least Soft-Threshold Squares Tracking. In our experiments, we track the movement of the upper limbs. While the results indicate an improvement in tracking accuracy at some joint locations, they also show that the result can be further improved. Conclusions and further work required to improve our results are discussed.**

I. INTRODUCTION

Markerless optical motion capture and inertial motion capture systems are becoming increasingly popular for recording human biomechanics as smaller devices and more robust processing techniques are developed. Biomechanical analysis considers the description of motion and the cause of motion, and has been used to better understand the state of the human body. Motion capture technology has recently enabled precise measurement of human movement. Current monitoring techniques, however, based upon snapshots of the subject in the laboratory may not be representative of normal biomechanics.

Motion capture technologies are popular in entertainment, healthcare, and sports. It has been widely used in film to capture the movement of actors and within consumer electronics to enhance human computer interaction.

In healthcare, motion capture has been used to diagnose pathologies and evaluate the rehabilitation of patients. Mechanical instruments, such as goniometers, are used to evaluate joint movement, and optical marker-based and inertial systems have been used to assess gait. For healthy individuals, gait follows a regular, balanced, and precise pattern. As deviations from a normal gait may indicate underlying health problems, numerous gait assessments have been formalised to diagnose pathologies and monitor recovery. The assessment of orthopaedic patients pre-operatively and post-operatively, for example, has enabled the progress of patient recovery to be quantified and enabled earlier identification of complications that may have arisen post-surgery where a lack of recovery is found.

In sports, the performance of athletes can be evaluated during training using motion capture. While traditional systems confine biomechanical analysis to the laboratory, developments in markerless optical systems and inertial motion capture have enabled movement to be studied across a more diverse range of environments, such as on the rowing lake [1], tennis court [2], and ice rink [3].

Long-term noninvasive biomechanical analysis is also important in the workplace. In the operating theatre and hospital wards, workflow analysis studies have considered the biomechanics of staff in order to assess performance and identify lapses in patient safety [4][5]. However, current studies have been limited to extremely coarse movement information. Wearable inertial sensors and cameras in operating theatres and wards can be utilised to study the natural movements of staff with greater accuracy.

For optical and inertial motion capture systems, occlusions and drift remain key challenges that need to be resolved to ensure accuracy. Recent works have proposed zero velocity updates (ZUPT) [6] and constant velocity updates (CUPT) [7] to reduce drift in lower limb movement studies.

Multi-sensor fusion techniques that combine vision and inertial measurements have been explored to improve resilience against occlusions and drift. While high frequency inertial sensors enable accurate tracking of fast movements without line-of-sight, vision-based tracking enables stable and drift-free estimation of pose for slower movements. Zhou et al. [8] uses this complementary information for tracking hand motion to recognise hand gestures, demonstrating the improved accuracy attained when vision and inertial measurements are fused together using an extended Kalman filter (EKF).

Consistent identification of markers in marker-based optical motion capture has been achieved through motion tracking. However, unlike markers, the appearance of body parts can change. Changes in human posture, illumination, camera pose, and differences between different subjects affect the appearance of tracked parts of the body, such as the hand.

Adaptive appearance models used in Incremental Visual Tracking (IVT) [9], [10], [11], Multiple Instance Learning (MilTrack) [12], Tracking-Learning-Detection (TLD) [13], and Least Soft-Threshold Squares Tracking (LSST) [14]

This work is supported by the UK's Engineering and Physical Sciences Research Council (ESPRIT Programme Grant, EP/H009744/1).

C. Wong, Z. Zhang, B. Lo and G.-Z. Yang are with the The Hamlyn Centre, Institute of Global Health Innovation, Imperial College London, London, UK {charence, z.zhang, benny.lo, g.z.yang} at imperial.ac.uk

978-1-4244-7929-0/14/$26.00 ©2014 IEEE

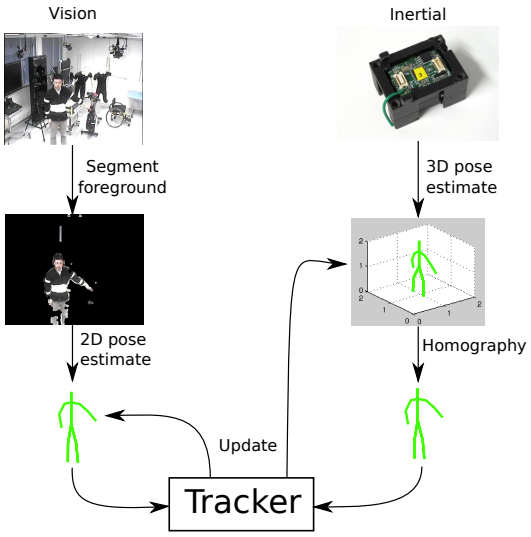

Fig. 1. System structure.

allow the tracker to learn the changing appearance of the tracked object. IVT, for example, incrementally learns a low-dimensional subspace representation of the tracked object that can adapt to appearance changes.

Adaptive models enable tracking through subtle appearance changes. Occlusions, however, also cause appearance models to adapt and drift away from the original appearance of the tracked object.

We present a pose estimation algorithm that uses an adaptive appearance model for tracking visual features on the upper body and inertial measurements from a wearable motion capture system for robust tracking of upper body motion. An unscented Kalman filter (UKF) [15] is used to merge the estimated poses and update the appearance model and inertial measurements.

This paper is structured as follows. In Section II, we introduce the motion tracking algorithm. The estimation results of our upper body motion tracker are discussed in Section III. Conclusions and future works are discussed in Section IV.

II. METHODOLOGY

The proposed system uses vision-based markerless motion tracking and a five node inertial motion capture system for estimating upper limb movement. Figure 1 presents an overview of the system structure.

A. System Setup

Upper body motion is estimated by fusing measurements from vision and a wearable inertial system. The rigid skeleton model of the upper body, tracked joints, and placement of inertial sensors is shown in Figure 2.

Standard definition (576i) ceiling-mounted Panasonic closed-circuit television cameras are used to capture the scene. A lightweight inertial human motion capture system is used to estimate the relative movement of the subject's upper body. Each node of the inertial system contains a triaxial

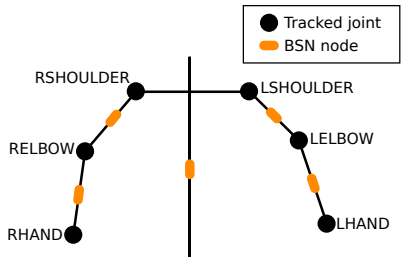

Fig. 2. Upper body model.

accelerometer, triaxial gyroscope, and triaxial magnetometer. Figure 2 shows the placement of the sensor nodes on the upper body.

The upper body human model is defined as an articulated structure of seven links. Only the tracked motion of the shoulders, elbows, and hands are evaluated in this paper. Measurements of each subject's shoulder width, upper arm length, and forearm length are taken to ensure an accurate motion estimation. 3D and 2D pose estimates of the upper body are obtained from inertial data and vision, respectively.

B. 3D Pose Estimation & Homography

The overall 3D posture of the human subject is acquired by determining the orientation of each link in the human model. In our system, sensor nodes are attached to the rigid body segments whose attitudes are to be determined. To estimate the attitude of each link, a process model and sensor measurement model for the rigid body human model are constructed. A Kalman filter is used to fuse all the sensor data. To deal with the temporary linear acceleration interference, magnetic disturbance, and provide a more reliable orientation estimate, a vector selection scheme is designed. The scheme minimises the affect of undesirable conditions, such as sudden intensive movements and magnetic disturbances, enabling the estimation algorithm to more accurately estimate orientation. To further improve orientation estimation accuracy, geometrical information is also fused within the Kalman filter.

Once the 3D posture of the human subject is known, we can find the corresponding 2D projection of the 3D pose using homography. Direct linear transformation (DLT) [16] is used to obtain the camera projection matrix, P, from 3D to 2D where the 3D points and corresponding 2D points are known. Subsequent 2D points can be found by applying a perspective project of projection matrix, P, onto new 3D scene points.

C. Implementation

First, the moving foreground in the scene is extracted using background segmentation. A Mixture of Gaussians [17] with a low learning rate is used to model the background to remove background noise and capture slow movements. For the first frame, the position and size of each tracked body part is given.

The initial 2D positions and 3D pose estimate from the wearable inertial system are used to find the camera

projection matrix, P, through DLT. Projection matrix, P, is applied onto 3D points from the pose estimate, X^{3D}, to project the 3D estimate onto the image plane. Perspective projection is applied to obtain the 2D estimate, X^{2D}, as shown in Eq. 1, where $X^{3D} = \left[x^{3D}, y^{3D}, z^{3D}, 1\right]'$.

$$X^{2D} = perspectiveProj(PX^{3D}) \tag{1}$$

$$\begin{bmatrix} x^{2D} \\ y^{2D} \\ 1 \end{bmatrix} = \frac{PX^{3D}}{z^{3D}}$$

To calculate the difference in appearence of the new tracked template with the existing dictionary template, the Chi-squared distance of the histograms of the templates are used to evaluate similarity, Eq. 2. Chi-squared distance is zero when both templates are the same and increases inversely to similarity.

$$compareHistogram(H_1, H_2) = \sum_I \frac{(H_1(I) - H_2(I))^2}{H_1(I)} \tag{2}$$

where H_1 and H_2 are the histograms of the new and existing template of the tracked position, respectively.

The Euclidean distance between 2D points is used for $positionChange$, $3dPositionChange$, and $distancesChange$. $positionChange$ is the distance between previous position estimate for the body part and the new position estimate, $3dPositionChange$ is the distance between the previous projected position estimate of the body part and the new projected position estimate from inertial data, and $distancesChange$ is a matrix of the difference between the previous and new body part separation distances between each pair of connected body parts defined in the model, Figure 2.

For each frame, the updated appearance model from the tracker is only retained when changes in appearance and position are within predefined limits. To learn the updated appearance model, the following conditions must be true:

1) The change in appearance of the updated model template must be less than the $appearanceThreshold$. The change in appearance is calculated using Eq. 2.

$$appearanceChange < appearanceThreshold$$

2) The absolute difference between the change in position of the tracked body part and estimation from inertial data should be similar. The difference must be less than the defined $positionThreshold$.

$$abs(positionChange - 3dPositionChange) < positionThreshold$$

3) The mean change in distance between each connected body part, from the previous to the current frame, must be less than the $distanceThreshold$.

$$mean(nonzeros(abs(distanceChange))) < distanceThreshold$$

To update the inertial measurements, an unscented Kalman filter (UKF) is used. In general, where the process and measurement model are nonlinear, UKF has been shown to improve upon the performance of the extended Kalman Filter (EKF) for estimation [15]. The UKF is defined as follows:

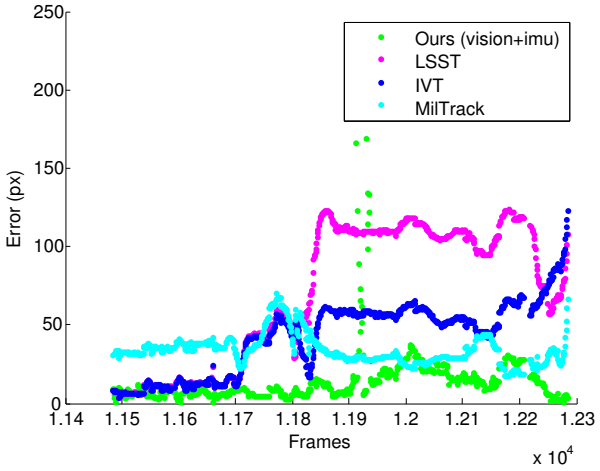

Fig. 3. Comparison of error (px) observed in state-of-the-art motion tracking algorithms for pose estimation and our implementation. Error in tracked right shoulder motion is shown.

$$q_t = f(q_{t-1}) + w_{k-1} \tag{3}$$

$$X_t^{2D} = h(q_t) + v_k \tag{4}$$

where the process model, $f(q_{t-1})$, and observation model, $h(q_t)$, are:

$$f(q_t) = q_t \times \Delta q \tag{5}$$

$$h(q_t) = perspectiveProj(P(q_t \times V \times q_t^{-1})) \tag{6}$$

Given the 2D position of the tracked body parts, X^{2D}, the quaternion rotations, q_t, from inertial motion capture can be updated to compensate for drift. Δq is obtained from inertial measurements by rearranging $q_t = q_{t-1} \times \Delta q$ to $\Delta q = q_{t-1}^{-1} \times q_t$. w_{k-1}, v_k, and V are process noise, measurement noise, and body segment vector, respectively.

III. RESULTS

In our experiments, movement was unconstrained. Subjects were allowed to move freely witin the laboratory to reflect a natural environment.

For the preliminary evaluation of pose estimation accuracy of our proposed method, two subjects were recruited. The ground truth of each tracked body part has been manually extracted from camera frames. Parameters for the tracker have been set based upon empirical analysis of tracking performance on a random sample from over 12,000 frames.

A comparison of the tracking accuracy of IVT, MilTrack, LSST, and our method is shown in Figure 3. The graph shows that our proposed method has the least error for most of the frames tracking the right shoulder. Initially, error for IVT and LSST is low, however, an occlusion around 1.18×10^4 causes both trackers to lose track of the shoulder. We observe that both IVT and LSST do not recover the shoulder's position in later frames. In our implementation, the appearance model

TABLE I

MEAN ERRORS (PX) OF TRACKED BODY PARTS

	Right Hand	Right Elbow	Right Shoulder	Left Hand	Left Elbow	Left Shoulder
Ours	102.7 ± 53.25	49.5 ± 25.88	18.6 ± 46.97	87.8 ± 43.60	50.3 ± 45.05	13.2 ± 36.84
LSST	90.3 ± 50.43	70.1 ± 28.24	74.5 ± 63.25	74.7 ± 29.82	56.5 ± 46.19	49.3 ± 54.5
IVT	104.6 ± 26.52	58.5 ± 24.83	47.2 ± 46.33	85.5 ± 35.31	60.2 ± 39.12	44.2 ± 36.5
MilTrack	93.1 ± 37.98	76.7 ± 28.58	38.7 ± 41.97	92.5 ± 41.37	121.6 ± 53.87	66.2 ± 48.13

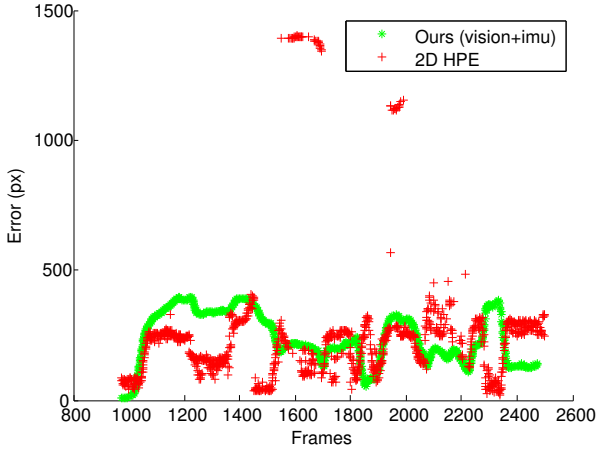

Fig. 4. Comparison of error (px) observed between our method and 2D articulated human pose estimation (2D HPE) [18] for the left hand.

template is not updated when large appearance changes are detected.

The mean error for each tracked body part is presented in Table I. Error is defined as the Euclidean distance between the ground truth and tracked 2D position. As expected, the table shows that there is least error locating the shoulders and errors for hand position estimates are the greatest, since there is typically more hand movement. For elbow and shoulder positions, our proposed method perform better than IVT, MilTrack, and LSST, however, LSST achieves the lowest average error for tracking hand position.

However, we can see in Figure 4 that human pose estimation [18] methods that evaluate pose frame by frame can achieve better estimates as they are not affected by drift in the learnt appearance and position, despite a lower median error of 216px in our method compared to 244px.

IV. CONCLUSIONS

In this paper, we have presented a comparison of state-of-the-art human pose detection and tracking algorithms. We have shown how adaptive appearance based tracking algorithms can be used to track the movement of body parts for markerless upper body motion capture and how movement constraints on appearance changes, movement, distance between other body parts, and inertial pose estimation can be used to improve the resilience of pose tracking under occlusion. Future research will explore full body pose estimation, tracking of multiple subjects for capturing multi-person interaction, and further work to better detect and track hands.

REFERENCES

[1] M. Tesconi, A. Tognetti, E. Pasquale Scilingo, G. Zupone, N. Carbonaro, D. De-Rossi, E. Castellini, and M. Marella, "Wearable sensorized system for analyzing the lower limb movement during rowing activity," in *Industrial Electronics, 2007. ISIE 2007. IEEE International Symposium on*, 2007, pp. 2793–2796.

[2] G. Abrams, A. Sheets, S. Corazza, T. Andriacchi, and M. Safran, "Injury potential evaluation of the upper extremity and torso of three tennis serve types using a novel markerless motion system," *British Journal of Sports Medicine*, vol. 45, no. 4, p. 333, Apr. 2011.

[3] J. E. Boyd, A. Godbout, and C. Thornton, "In Situ Motion Capture of Speed Skating: Escaping the Treadmill," in *Computer and Robot Vision (CRV), 2012 Ninth Conference on*, May 2012, pp. 460–467.

[4] M. Vankipuram, K. Kahol, T. Cohen, and V. L. Patel, *Toward automated workflow analysis and visualization in clinical environments*. Elsevier, Jun. 2011, vol. 44, no. 3.

[5] A. Nara, K. Izumi, H. Iseki, T. Suzuki, K. Nambu, and Y. Sakurai, "Surgical Workflow Monitoring Based on Trajectory Data Mining," in *New Frontiers in Artificial Intelligence*, ser. Lecture Notes in Computer Science, T. Onada, D. Bekki, and E. McCready, Eds. Springer Berlin Heidelberg, 2011, vol. 6797, pp. 283–291.

[6] Y. Li and J. J. Wang, "A robust pedestrian navigation algorithm with low cost IMU," in *Indoor Positioning and Indoor Navigation (IPIN), 2012 International Conference on*, Nov. 2012, pp. 1–7.

[7] Y. Li, J. J. Wang, S. Xiao, and X. Luo, "Dead reckoning navigation with Constant Velocity Update (CUPT)," in *Control Automation Robotics Vision (ICARCV), 2012 12th International Conference on*, 2012, pp. 160–165.

[8] *2D Human Gesture Tracking and Recognition by the Fusion ofMEMS Inertial and Vision Sensors*, vol. PP, no. 99, 2013.

[9] D. Ross, J. Lim, and M.-H. Yang, "Adaptive Probabilistic Visual Tracking with Incremental Subspace Update," in *Computer Vision - ECCV 2004*, ser. Lecture Notes in Computer Science, T. Pajdla and J. Matas, Eds. Springer Berlin Heidelberg, 2004, vol. 3022, pp. 470–482.

[10] J. Lim, D. Ross, R.-s. Lin, and M.-h. Yang, "Incremental learning for visual tracking," in *Advances in Neural Information Processing Systems*, 2005, pp. 793–800.

[11] D. Ross, J. Lim, R.-S. Lin, and M.-H. Yang, "Incremental Learning for Robust Visual Tracking," *International Journal of Computer Vision*, vol. 77, no. 1-3, pp. 125–141, May 2008.

[12] B. Babenko, M.-H. Yang, and S. Belongie, "Visual tracking with online multiple instance learning," in *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, June 2009, pp. 983–990.

[13] Z. Kalal, K. Mikolajczyk, and J. Matas, "Tracking-Learning-Detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 7, pp. 1409–1422, 2012.

[14] D. Wang, H. Lu, and M.-H. Yang, "Least Soft-Threshold Squares Tracking," in *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*, Jun. 2013, pp. 2371–2378.

[15] E. A. Wan and R. Van der Merwe, "The unscented Kalman filter for nonlinear estimation," in *Adaptive Systems for Signal Processing, Communications, and Control Symposium 2000. AS-SPCC. The IEEE 2000*, 2000, pp. 153–158.

[16] E. Trucco and A. Verri, *Introductory Techniques for 3-D Computer Vision*. Upper Saddle River, NJ, USA: Prentice Hall PTR, 1998.

[17] Z. Zivkovic and F. van der Heijden, "Efficient adaptive density estimation per image pixel for the task of background subtraction," *Pattern Recogn. Lett.*, vol. 27, no. 7, pp. 773–780, May 2006.

[18] M. Eichner, M. Marin-Jimenez, A. Zisserman, and V. Ferrari, "2D Articulated Human Pose Estimation and Retrieval in (Almost) Unconstrained Still Images," *Int. J. Comput. Vision*, vol. 99, no. 2, pp. 190–214, Sep. 2012.