

The International Journal of Robotics Research

<http://ijr.sagepub.com>

Integration of Vision and Inertial Sensors for 3D Arm Motion Tracking in Home-based Rehabilitation

Yaqin Tao, Huosheng Hu and Huiyu Zhou

The International Journal of Robotics Research 2007; 26; 607

DOI: 10.1177/0278364907079278

The online version of this article can be found at:
<http://ijr.sagepub.com/cgi/content/abstract/26/6/607>

Published by:

 SAGE Publications

<http://www.sagepublications.com>

On behalf of:



Multimedia Archives

Additional services and information for *The International Journal of Robotics Research* can be found at:

Email Alerts: <http://ijr.sagepub.com/cgi/alerts>

Subscriptions: <http://ijr.sagepub.com/subscriptions>

Reprints: <http://www.sagepub.com/journalsReprints.nav>

Permissions: <http://www.sagepub.com/journalsPermissions.nav>

Yaqin Tao
Huosheng Hu
Huiyu Zhou

Department of Computer Science, University of Essex,
Wivenhoe Park, Colchester CO4 3SQ, U.K.

ytao@essex.ac.uk

hhu@essex.ac.uk

zhou@essex.ac.uk

Integration of Vision and Inertial Sensors for 3D Arm Motion Tracking in Home-based Rehabilitation

Abstract

The integration of visual and inertial sensors for human motion tracking has attracted significant attention recently, due to its robust performance and wide potential application. This paper introduces a real-time hybrid solution to articulated 3D arm motion tracking for home-based rehabilitation by combining visual and inertial sensors. Data fusion is a key issue in this hybrid system and two different data fusion methods are proposed. The first is a deterministic method based on arm structure and geometry information, which is suitable for simple rehabilitation motions. The second is a probabilistic method based on an Extended Kalman Filter (EKF) in which data from two sensors is fused in a predict-correct manner in order to deal with sensor noise and model inaccuracy. Experimental results are presented and compared with commercial marker-based systems, CODA and Qualysis. They show good performance for the proposed solution.

KEY WORDS—sensor fusion, extended Kalman filter, inertial sensor, human motion tracking, home-based rehabilitation

1. Introduction

Traditionally, stroke patients take physiotherapy in a hospital or care centre with the help of physiotherapists or well-trained carers to diagnose if they are performing rehabilitation correctly. However, because of staffing shortages in the National Health Service, and the need for a prolonged period of time for rehabilitation exercise, patients are not receiving enough treatment. It is desirable to develop a novel motion tracking system

to support the rehabilitation programme for patients in home environments, so that the burden on hospitals and physiotherapists can be relieved. Standard methods for clinical human motion analysis involve marker-based motion tracking. Tracking results from these systems are quite accurate because they mark objects of interest. There are many commercial marker-based human motion capture systems that can be employed for tracking a patient's motion; such as CODA (Charnwood Dynamics Ltd) and Qualisys (Qualisys). However, in addition to the difficulty of calibrating cameras and markers, these systems are too expensive for daily deployment in stroke patient's homes, and too complicated for physiotherapists to interpret patients' motion tracking results.

Recently, an attempt has been made to design a visual-based marker-free tracking system for human motion capture (Aggarwal and Cai 1999; Gavrilu 1999; Moeslund and Granum 2001; Wang et al. 2003). Marker-free tracking systems are very attractive, because instead of special cameras and intrusive markers they require only conventional cameras. However, designing a video system to track human motion is a non-trivial task because of a number of difficulties (Sminchisescu 2002), including depth ambiguities, occlusion, and kinematics singularities, etc. In order to simplify the human motion tracking problem, most human motion tracking algorithms employ a shape model of the subject to support tracking. The shape model of a subject varies from simple skeleton models (Chen and Lee 1992), to 2D patch models (Ju et al. 1996), and sophisticated 3D volumetric models (Deutscher et al. 2000; Sidenbladh et al. 2000).

However, current model based human pose estimation methods have to specify models for each subject. This makes the generalization of subject tracking in such systems difficult. Some methods are based on prior knowledge (statistical data or learned offline), e.g. the appearance of a subject, the geometry of a subject, or the kinematics and dynamics of a subject's motion (Sidenbladh, Black and Fleet 2000). These methods par-

The International Journal of Robotics Research

Vol. 26, No. 6, June 2007, pp. 607–624

DOI: 10.1177/0278364907079278

©2007 SAGE Publications

Figures 1, 2, 4, 5, 7, 8, 10, 14–19 appear in color online: <http://ijr.sagepub.com>

tially limit the performance and application domain of tracking algorithms by making them either computationally costly, or too inaccurate for real-life application. Alternatively, systems that use multiple cameras to track motion in 3D and deal with the occlusion problem (Gavrila and Davis 1995; Moeslund and Granum 2000a) have robust performance, but multiple camera systems are difficult to calibrate and are computationally expensive.

The current trend in human motion tracking is to estimate 3D poses of a human body using monocular vision, which is a very challenging task. Multiple visual features, prior knowledge, and subject models, are normally combined in such applications (Sminchisescu 2002; Sidenbladh et al. 2000). Although current tracking results are promising, their computational cost is high, and they suffer from the same problem as model-based methods. Their accuracy needs more investigation before they be employed in real-world applications. Ideally, it is desirable to have a real-time pose estimation system for human motion tracking which adopts a single camera, and has no need to employ a shape model, a uniform background, or prior learnt knowledge. However, due to the limitations of visual sensors as mentioned above, such a system does not yet exist.

Inertial tracking is also an active research area, and has been applied recently to human motion analysis. Popular inertial sensors are accelerometers and gyros. In previous work, accelerometer sensors (Tetrad et al. 2002; Bussmann 2000; Veltink et al. 1996) or gyro sensors (Bachmann 1999) were applied separately to a human body to detect and analyse human motion. However, only limited information was obtained. Recently, the trend has been to integrate accelerometers and gyros into single inertial sensor units such as the MT9 (Xsens), in which both acceleration and rate of turn can be obtained for motion analysis. Zhou and Hu (2005) used an MT9 sensor to track 3D human arm motion in real time. Although a simulated annealing algorithm was employed to correct the inertial sensor drift problem, the method only worked for a simple flexion-extension motion of upper limbs. Inertial tracking suffers from severe drift problems, due to sensor noise and bias, and cannot provide accurate position information during continuous operation.

In general, no sensor is perfect, and each sensor has its strengths and limitations. Pure vision based tracking has low jitter and stability merits, but lacks fast motion performance due to motion blur and occlusions. Inertial tracking is good for fast motion tracking, but its long term stability is affected by severe drift problems. Hybrid tracking based on visual and inertial sensors, offers not only fast motion tracking and good stability, but robust performance over occlusions. Currently, we are developing a real-time motion tracking system to track the motion of articulated objects (human upper limbs in home-based rehabilitation), which is cheap and easy to use in comparison to commercial marker-based systems. It is based on the integration of both visual and inertial sensors. The main

focus of this paper is on the development of a suitable pose estimation method.

The rest of this paper is organised as follows. We begin in Section 2 with a literature review of previous research in motion tracking using inertial and visual sensors. An overview of our proposed hybrid approach is presented in Section 3. Section 4 describes the inertial tracking part and Section 5 outlines the visual tracking part. State estimation of arm motion via the fusing of data from both vision and inertial sensors is presented in Section 6. Some experimental results on the performance of these hybrid motion tracking methods are shown in Section 7. Finally, conclusions and future work are presented in Section 8.

2. Related Work

Motion tracking methods based on an integration of visual and inertial sensors have received more and more attention recently. They have a wide range of real-world applications, such as Augmented Reality (AR), Ego-motion estimation for robot navigation, and helmet-tracking systems (HTS), etc. Foxlin et al. (2004) used two inertial sensors and three cameras to track a pilot's head motion accurately in a cockpit for enhanced vision. Three cameras were employed in an inside-outside-in mode, and data fused with inertial data using a decentralized Kalman filter (Foxlin 2002). The system tracks a head's motion accurately and stably, but is very expensive. Furthermore, complicated sensor deployment makes the method difficult to generalize to other applications.

You et al. (1999) integrated visual and inertial sensors for tracking in augmented reality applications. Data fusion was regarded as an image stabilization problem. Visual and gyro data was fused using an extended Kalman filter. This approach was subsequently employed in (You and Neumann 2001) to track a six-degree-of-freedom object pose. Visual data was obtained by detecting and tracking known artificial fiducials. Only gyro data was used in their method, which partially simplified the problem. The experimental results were quite promising in a 2D image plane, but not yet verified in 3D cases. Lang et al. (2002) also used extended Kalman filters to fuse different data modalities from visual and inertial sensors. In this work, both gyro and acceleration data from inertial sensors were employed. The visual input of the EKF estimation method was calculated by using a Perspective-n-points method (Lu et al. 2000). The experimental results based on one axis translation and one axis rotation motion were obtained separately. This is far from applicable in the proposed mobile augmented reality application.

Recently, developments in visual and inertial sensor fusion have included an estimation of object pose, and the structure of a scene from motion. Chen and Pinz (2004) estimated structure and motion by fusing visual and inertial sensor data using an EKF. It is based on You and Neumann's work (2001), but

extended to use acceleration and estimate structure in addition to motion. Chai et al. (2002) used two extended Kalman filters linked by a recursive loop: one for motion estimation and the other for structure from motion. Strelow and Singh (2003) proposed two methods for motion and structure estimation from visual and inertial measurements, namely a batch method using the Levenberg–Marquardt method and an online method employing an EKF. These methods include visual feature detection, selection, and tracking used for environments without known fiducials.

All of the above methods use multiple feature points from the visual sensor, either in a direct way (You and Neumann 2001; Strelow and Singh 2003), or indirectly based on the P-N-P method (Lang et al. 2002; Chen and Pinz 2004). Multiple visual feature points offer redundant measurements and provide good performance; they are therefore widely used. However, detecting, selecting, and tracking multiple feature points in an image sequence is a difficult task, especially when feature points are identical. Alenya et al. (2003) proposed a method for fusing visual contours with inertial data instead of feature points. However, the initialization of tracking contours can only be solved manually, and contour tracking is computationally demanding. Huster and Rock (2001) introduced a method for fusing visual and inertial data by using only one feature point. Their emphasis was on exploiting inertial information to reduce the visual information required in the tracking process because visual information is expensive to process and easily distracted by clutter or noise. However, using a single feature point is problematic because of the non-linear property of the fusion problem. A conventional non-linear EKF estimator may fail to estimate a system's state vector accurately. Huster (2003) proposed a new estimator for the one feature point and inertial data fusion problem. Experimental results on a pick up object task showed good performance.

Existing hybrid methods differ in the number of inertial and vision sensors used, and the number of visual feature points, as well as different fusion algorithms. It is always desirable to use less sensors and efficient algorithms if we can achieve the same goal. Therefore, we build our motion tracking system by using a video camera and an inertial sensor to capture a subject's motion. In order to simplify the image processing procedure and improve the robustness of the system over image noise, we adopt the concept of commercial marker-based systems that feature the object of interest by colouring the target. Only the image position of the coloured object is employed in our method, similar to Huster's method (2003). We propose two methods in this paper to exploit the data fusion of inertial and visual sensors. Detailed information is provided in the following sections.

3. System Overview

Instead of tracking whole body motion, we focus our work on upper limb motion tracking at this stage. This is because

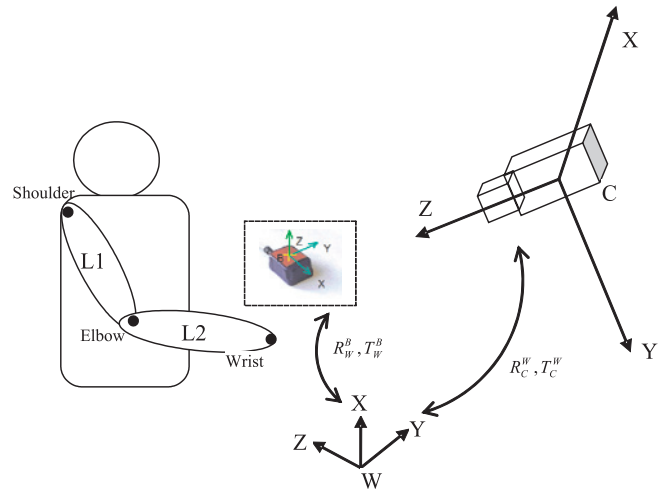


Fig. 1. Arm model and system configuration.

upper limb motion is a very important recovery process for stroke patients in home-based rehabilitation. The system will be extended to lower limb motion capture and analysis at a subsequent stage, i.e. gait analysis, which has attracted a lot of interest recently (clinical gait analysis).

3.1. Arm Model

In this work, an arm is modelled as a skeleton structure which consists of two segments linked by a revolute joint (see Figure 1). This skeleton structure is simple, but it is acceptable when recovering arm motions in rehabilitation, e.g. flexion, extension, target reaching, feeding, etc. No shape model (for example to model the limb as truncated cones (Goncalves et al. 1995)), is considered in our method. In order to simplify the tracking problem, we assume that the shoulder joint is fixed during motion and the position is known a priori; which is a realistic constraint that has been widely used in many upper-limb motion tracking systems (Goncalves et al. 1995; Moeslund and Granum 2000b). Furthermore, we assume the length of forearm L_2 , and upper arm L_1 , are known a priori.

3.2. The State Space

There are different ways to represent the pose of a human arm in state space. In this paper, we mainly use Cartesian coordinates supported by a joint angle representation. Johansson's psychology experiments in moving light displays (MLDs) (Johansson 1973) showed that a set of body joints' motion trajectories are meaningful, and can be used to analyse a subject's motion (or be used for recognition). Based on an arm model and MLD experiments, the pose of an arm can normally be

represented using six variables in Cartesian coordinates; these are the elbow position $P_{e,W} = \{x_e, y_e, z_e\}$ and wrist position $P_{w,W} = \{x_w, y_w, z_w\}$, where W represents the world frame. The state vector thus has six DOFs, represented as follows:

$$X(k) = \begin{bmatrix} P_{e,W}(k) \\ P_{w,W}(k) \end{bmatrix}. \tag{1}$$

Another representation is the joint angle based method, where the arm is modelled in a four-dimensional phase space. There are three joint angles $\{\psi, \theta, \phi\}$ in the shoulder joint and one in the elbow joint α . This representation is widely used in robot manipulator applications.

3.3. System Configuration

Figure 1 shows the system configuration and related coordinate systems of our proposed arm motion tracking system; in which three coordinate systems are adopted as follows:

- Camera frame (C): This is attached to the camera and its origin located in the camera centre. Unlike conventional hybrid systems in which camera and inertial sensors are rigidly aligned and attached to a moving object, we separate the camera and inertial sensor in different places. The video camera is fixed in the environment and used to capture a subject's arm motion; while the inertial sensor is attached to the wrist joint of the subject's arm. There are two advantages to arranging the equipment in such a way: first, the camera is too heavy to be attached to a patient's limb, and it may affect the patients' rehabilitation motion; second, captured images from the camera may provide additional information to assist articulated human motion besides fusion input, which leaves great potential for further exploitation.
- Inertial sensor frame (B): This is attached to the body of an inertial sensor, as shown in Figure 1. During measurement, the inertial sensor is attached to a human arm. The pose of the inertial sensor frame with respect to the world frame R_W^B, T_W^B changes from time to time as the arm moves.
- World coordinate system (W): This is a reference frame, where the joint positions of a human arm are tracked and represented. It is also an intermediate coordinate system that links visual and inertial measurements. It is defined by overlaying reference frame W with inertial frame B. Note that world coordinate system W and camera coordinate system C, are rigidly aligned and related by rotation matrix R_C^W and translation vector T_C^W , as shown in Figure 1. The calculation of both R_C^W and T_C^W is performed at the initialization stage.

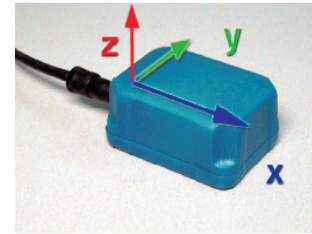


Fig. 2. MT9 with body fixed coordinate system B.

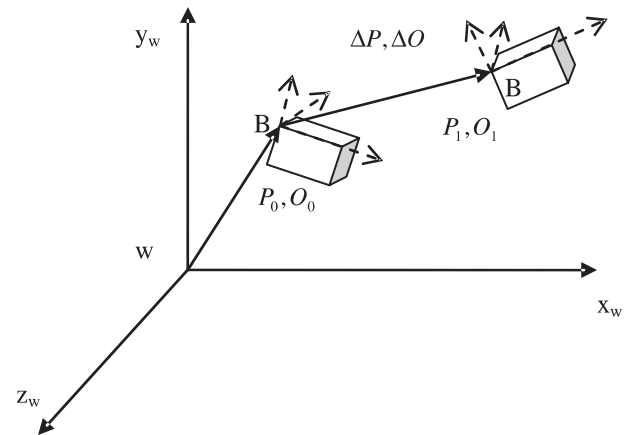


Fig. 3. Inertial tracking processes.

Before we discuss how to estimate system state using fusion methods, we will first introduce inertial and visual tracking respectively.

4. Inertial Tracking

Normally, an inertial sensor is attached to a moving object. The conventional usage of inertial sensors in motion tracking is to calculate the relative change of a moving target in position $\Delta P(k)$ and orientation $\Delta O(k)$ between two consecutive sampling times; based on measurements of acceleration $a^B(k)$ and angular velocity $\omega^B(k)$ from the inertial sensor. The k means "at sampling time k ", while superscript B means "data measured in sensor frame B". In our motion tracking system, we use an MT9 inertial sensor from Xsens, which has a body fixed Cartesian coordinate system (B) as shown in Figure 2.

If we define a reference world coordinate system (W) as shown in Figure 3, the pose of the moving upper limb (where the inertial sensor is attached) can be calculated as follows:

$$O(k+1) = O(k) + \Delta O(k) \tag{2}$$

$$P(k+1) = P(k) + \Delta P(k). \tag{3}$$

4.1. Orientation Data

Orientation information can be calculated from angular velocity $\omega^B(k)$ deduced from inertial output. Fortunately, besides angular velocity output, the MT9 inertial sensor contains a proprietary algorithm that can accurately calculate, over time, the absolute orientation of the moving sensor with respect to reference frame $O(k)$ in 3D space. The output format of orientation information $O(k)$ can be rotation matrix R_W^B , or Euler angle (roll, pitch, and yaw; represented as $\{\psi, \theta, \phi\}$), or quaternion $q = \{a, bi, cj, dk\}$. Our experiments show this orientation information is quite accurate, and can be used directly in motion tracking; which considerably simplifies the tracking problem.

4.2. Position Data

We assume a constant acceleration motion model for the position tracking problem. The system's position dynamic model can be expressed as:

$$P(k+1) = P(k) + v(k)\Delta t + \frac{1}{2}a(k)\Delta t^2 \quad (4)$$

where v represents the velocity of a moving object, and can be expressed as:

$$v(k+1) = v(k) + a(k)\Delta t \quad (5)$$

where Δt is sampling interval, and $a(k)$ is acceleration of the moving object in world frame W .

However, the acceleration output of the inertial sensor is represented in moving frame B , expressed as $a^B = (a_x^B, a_y^B, a_z^B)$. The relationship between them is:

$$a(k) = q(k) * a^B(k) * q(k)' - g \quad (6)$$

where $*$ denotes quaternion multiplication; g is gravity, and $q(k)$ is the unit quaternion that represents the orientation of frame B relative to reference frame W .

In theory, inertial tracking alone can fulfil our motion tracking tasks. However, the inertial sensor has a drift problem, and is very sensitive to noise. It is necessary to use visual data to correct its drift, which will be addressed in next section.

5. Vision Tracking

Images captured by a video camera can offer different features to assist upper limb motion tracking. Some useful features are edges, colour, contours, and optical flow, etc. In this work, we only employed colour features of the target object in order to achieve real-time performance and to reduce the possibility of occlusion problems.

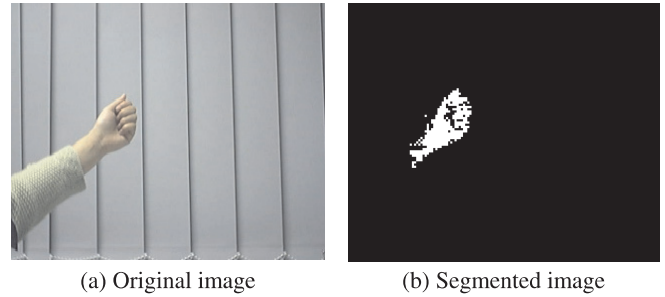


Fig. 4. Colour object detection using the CAMSHIFT algorithm.

5.1. Colour Object Tracking

Skin colour is a natural image feature and very useful in human motion tracking applications, which is used in our deterministic method. There are many different parametric or non-parametric models for detecting skin colour (Vezhnevets et al. 2003). In our approach, we employ a Continuously Adaptive Mean Shift (CAMSHIFT) algorithm (Bradski 1998) to segment and track a coloured object. This algorithm's performance is fast and robust (Tao and Hu 2004). Unlike other skin colour specified algorithms, CAMSHIFT is able to track any kind of target colour by building a histogram distribution colour model in Hue Saturation Value (HSV) colour space. It will make our tracking system easy to generalize and able to track different target colours. The histogram distribution is later used to segment a target object from a background image. Figure 4(a) shows the original image, while Figure 4(b) shows the segmented target object.

CAMSHIFT uses a fixed colour model. Once a colour model is built, it is assumed suitable for a whole image sequence. In other words, it assumes lighting conditions are static during a tracking period. This assumption does not hold in real life, lighting changes over time, and a coloured object may reflect ambient lighting, which causes it to take on another colour. Using a fixed colour model may fail or not track objects stably. Figure 5(a) shows an image at an initial stage. The colour model works well initially, however, it fails to detect the object in (b), when lighting conditions change.

Some researchers (e.g. Wren et al. 1995) use adaptive colour models to deal with this problem, which means that after building the colour model represented as M_0 in the initialization stage, the colour model is updated during each frame, by the measurement Y_k obtained, using the equation: $M_k = \alpha Y_k + (1-\alpha)M_{k-1}$. The problem with this method is that it is difficult to properly define coefficient α . In addition, the colour model may drift if measurement Y_k or the coefficient is inaccurate.

We use a simple but effective sampling method. The colour model is built based on some sampled sub-images at an initial stage. First we move a target object randomly in the camera

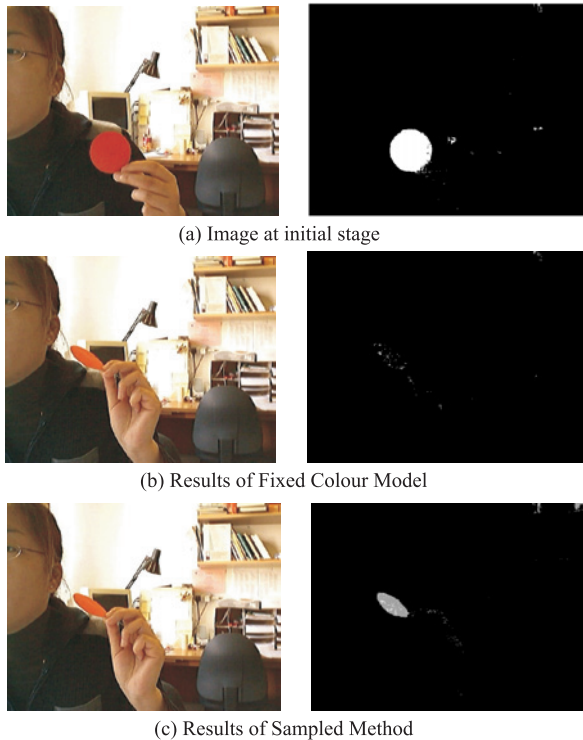


Fig. 5. Results from different colour models.

view, and select several points in the region of interest from a live video. Second, a 17×17 sub-image is extracted for each point, and these sub-images used to calculate a colour model. As can be seen in Figure 5(c), this method detects an object when a fixed colour model fails.

In the worst scenario, direct hand tracking is very noisy and unreliable. A colour belt can easily be introduced into the system by attaching it to the wrist joint to highlight its position (see Figure 14). The proposed tracking method can be used directly without any change or extra computational cost. Using a colour belt instead of a marker has several benefits. First, it doesn't increase the possibility of occlusion, as it is observable from different angles. Second, it is not affected significantly by muscle or skin movement as much as markers do.

The spatial mean position $z = (I_x, I_y)$ of the detected foreground object is regarded as the 2D position of the target object. The correspondence problem of motion tracking is solved by assuming small smooth motion between two consecutive frames. The output of colour object tracking is the motion trajectory in a 2D image plane, which will be used to fuse inertial tracking results.

5.2. 2D–3D Relationship

The video camera used in our approach is calibrated using a pin-hole camera model, including both intrinsic and extrinsic

parameters. Fig. 6 shows camera parameters and the imaging system.

The 2D visual measurements of colour tracking results are related with scene point P , by a measurement equation expressed as follows:

$$z(k) = \begin{bmatrix} I_x(k) \\ I_y(k) \\ 1 \end{bmatrix} \sim KI * \begin{bmatrix} \frac{P_{x,C}(k)}{P_{z,C}(k)} \\ \frac{P_{y,C}(k)}{P_{z,C}(k)} \\ 1 \end{bmatrix} \sim KI * \left(R_C^W * \begin{bmatrix} P_{x,W}(k) \\ P_{y,W}(k) \\ P_{z,W}(k) \end{bmatrix} + T_C^W \right) \quad (7)$$

where KI are camera intrinsic parameters calibrated offline; $P_C = (p_{x,C}, p_{y,C}, p_{z,C})$ is the scene point represented in camera frame C ; $P_W = (p_{x,W}, p_{y,W}, p_{z,W})$ is the scene point in word frame W , and ' \sim ' means "the equation up to a scale". Homogeneous coordinates are used in (7) to express an affine transformation.

6. Hybrid Arm Motion Tracking

According to the above system configuration and arm model, arm tracking is a process of essentially locating the positions of wrist and elbow joints respectively in each frame. It is achieved by integrating visual and inertial sensing, which therefore makes the tracking system more robust and applicable. We propose two methods in this section, a deterministic method and a probabilistic method, to fuse inertial and visual data for arm motion tracking.

6.1. Deterministic fusion method

6.1.1. Elbow Position Calculation

As shown in Figure 8, an inertial sensor is attached to, and its x axis aligned with, a subject's upper limb. Given the length of the upper limb, the elbow position can be uniquely determined in sensor frame B as $P_{e,B} = \{L_1, 0, 0\}$, where L_1 is the length of the upper limb. The elbow position in the reference system can be calculated according to the following equation. $P_{s,W}$ is the shoulder joint in the world frame, and assumed known a priori:

$$P_{e,W} = R_W^B P_{e,B} + P_{s,W}. \quad (8)$$

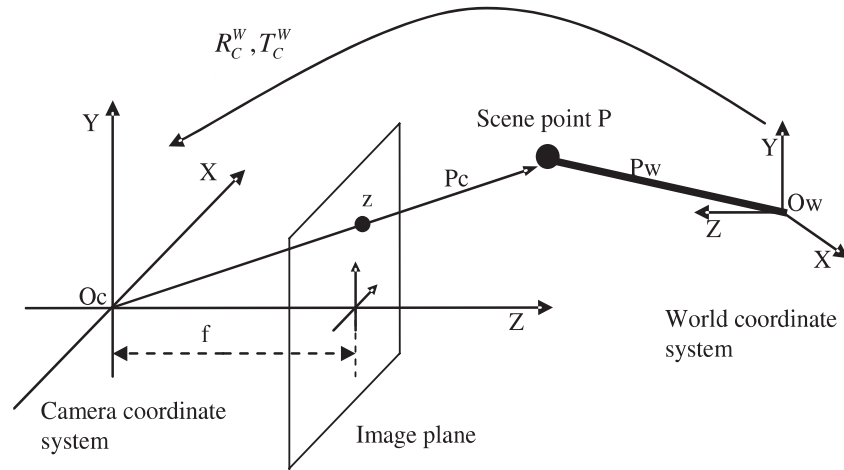


Fig. 6. Camera model and imaging process.

6.1.2. Wrist Position Calculation

The wrist position is calculated based on the results of 2D colour object tracking, camera calibration, and the elbow calculation described in the previous sections. Two constraint equations are involved in the calculation. One constraint equation is from the 2D–3D relationship in Section 5. Given the calibrated camera parameters and an image point z , the corresponding line (line $OcPc$ in Figure 7) in camera centred space is uniquely determined, which comprises all world points that map to the same image point z . This is a one-to-multiple mapping. The possible 3D positions of scene point P corresponding to image point z can be represented in the camera coordinate system as follows:

$$P_C(\lambda) = O_C + \lambda P^+ z \tag{9}$$

where P^+ is the inverse pseudo of the camera projective matrix, and λ is a parameter. Since all our data is represented in world frame W , the back-projected line can be converted to such a coordinate system using camera extrinsic parameters:

$$P_{w,W}(\lambda) = R_C^{W-1} P_C(\lambda) - T_C^W. \tag{10}$$

Another constraint equation develops from the geometry relationship between elbow and wrist joints, as the forearm length is known and fixed. The possible position of the wrist joint consists of a sphere surface, and is defined by:

$$(P_{w,W} - P_{e,W})^2 = L_2^2. \tag{11}$$

According to the above analysis, the solution space of the wrist joint can be reduced from a line and a sphere surface to at most two points by calculating the intersection points of the sphere and back projected line, as shown in Figure 7 (where an arrowed line is the back projected line).

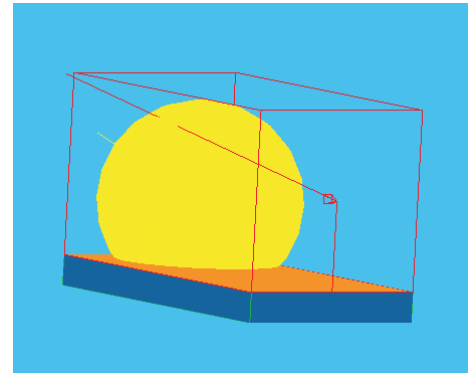


Fig. 7. Intersection of a sphere and back projected line.

There are three possible situations for the intersection of a sphere and a line:

- (1) No intersection point – For this situation, we introduce a variance ϵ for radius L_2 of the sphere. If the distance between the centre of the sphere and back projected line D satisfies (12), the closest point on the back projected line to the sphere centre is taken as the 3D object position. If (12) is not satisfied, the wrist position in the previous frame is used in this frame.

$$D \leq L_2 + \epsilon \tag{12}$$

- (2) One intersection point – This is the ideal situation, where the intersection point is regarded as the 3D position.
- (3) Two intersection points – Selecting the correct intersection point from two solutions is achieved by using constraints.

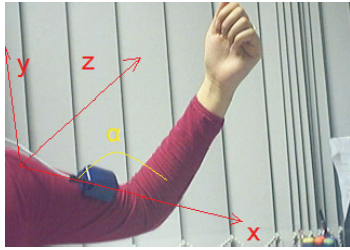


Fig. 8. Forearm coordinates system.

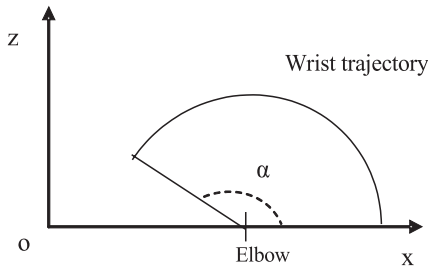


Fig. 9. Wrist trajectory.

- First, a motion smoothness constraint is employed which requires the velocity ΔP and orientation ΔO change of the target object between two consecutive frames to be small. Of the two intersection points, the one that contributes more to the smoothing motion is selected.
- Second, the arm model is used to distinguish and select the correct intersection point. In Figure 8, the coordinate system XYZ represents the sensor frame at time step k . The forearm is also expressed in this coordinate system. We only assign one degree of freedom to the forearm, which is the rotation angle α in the XOZ plane. This means the wrist joint position is only located in a XZ plane, and the trajectory is a circle instead of a sphere. Taking the geometry constraint of the range of the elbow angle (which is about $[0^\circ, 149.75^\circ]$ in Tolani et al. 2000), the trajectory of the wrist joint is a clip circle arc (Figure 9).

Using these two constraints, we can successfully select the right intersection point from two candidates.

6.1.3. Error Propagation Study

There are many noise sources in a deterministic hybrid tracking system. The first is from visual tracking. Visual tracking is only performed in a 2D image plane in our method. Inferring

3D scene properties from 2D image measurements is an under-constrained task due to a lack of depth information. A small 2D image measurement error may dramatically affect the 3D reconstructed optical line, and therefore affect tracking accuracy. The second is from inertial tracking. The movements of an upper arm muscle may affect MT9 output, and therefore affect elbow position calculation. This error will accumulate in wrist joint position calculations. Other error sources are, for example, image segmentation, camera calibration, etc. Here we perform a simple version error propagation study by analysing two main error sources from visual and inertial tracking respectively. To simplify the problem, we assume that visual tracking error is independent of inertial tracking error.

Visual tracking error propagation

Visual output is the 2D image position $z = (I_x, I_y)$. Output error is represented as $(\partial I_x, \partial I_y)$. The error propagation study is used to find out how visual tracking error affects a target object's position accuracy $\partial P_{w,w} = (\partial x_w, \partial y_w, \partial z_w)$. A target object's position $P_{w,w} = (x_w, y_w, z_w)$ is related to visual input according to equation (7), (9), and (10). We expand these equations as follows:

$$\begin{aligned}
 P_{w,w} &= \begin{bmatrix} x_w \\ y_w \\ z_w \end{bmatrix} = \begin{bmatrix} f_x(I_x, I_y) \\ f_y(I_x, I_y) \\ f_z(I_x, I_y) \end{bmatrix} \\
 &= R_C^{W-1} \left(O_C + \lambda P^+ \begin{bmatrix} I_x \\ I_y \\ 1 \end{bmatrix} \right) - T_C^W \\
 &= \underbrace{R_C^{W-1} O_C - T_C^W}_{KS_{3 \times 1}} + \underbrace{R_C^{W-1} \lambda P^+}_{KC_{3 \times 3}} \begin{bmatrix} I_x \\ I_y \\ 1 \end{bmatrix} \quad (13)
 \end{aligned}$$

where KS is a 3×1 vector and KC a 3×3 matrix. Both have constant values. In this study, the extrinsic R_C^W, T_C^W , intrinsic P^+ and λ parameters are assumed known and accurate. In reality, the calibration procedure is not 100% accurate. These errors are called systematic errors, and the careful design of an experiment will allow us to eliminate or correct systematic error (uncertainties and error propagation).

We illustrate the error propagation calculation in the x coordinate using eq. (14). It is the same calculation procedure in the y and z coordinates.

$$\begin{aligned}
 x_w &= f_x(I_x, I_y) = KS_{11} + KC_{11}I_x + KC_{12}I_y + KC_{13} \\
 \partial x_w &= \sqrt{\left(\frac{\partial f_x}{\partial I_x} \partial I_x\right)^2 + \left(\frac{\partial f_x}{\partial I_y} \partial I_y\right)^2} \quad (14)
 \end{aligned}$$

$$\frac{\partial f_x}{\partial I_x} = KC_{11} \quad \frac{\partial f_x}{\partial I_y} = KC_{12}.$$

The maximum visual output error in our experiment is $\max(\partial I_x) = 10 \text{ pixels}$, $\max(\partial I_y) = 10 \text{ pixels}$. We can calculate the worst target position deviation in our experiment as $\partial x_w = 1.495 \text{ cm}$, $\partial y_w = 0.914 \text{ cm}$, $\partial z_w = 1.17 \text{ cm}$.

It should be noted that in (14), target tracking accuracy is unbounded if visual tracking is very noisy.

Inertial tracking error propagation

We use orientation output, roll, pitch, and yaw $\{\psi, \theta, \phi\}$ from an inertial sensor. A target object position is governed by the equation:

$$\begin{aligned} P_{w,w} &= \begin{bmatrix} x_w \\ y_w \\ z_w \end{bmatrix} = \begin{bmatrix} f_x(\psi, \theta, \phi) \\ f_y(\psi, \theta, \phi) \\ f_z(\psi, \theta, \phi) \end{bmatrix} \\ &= P_{e,w} \pm L_2 = R_W^B P_{e,B} + P_{s,w} \pm L_2 \\ &= \begin{bmatrix} C\phi C\theta & C\phi S\theta S\psi - S\phi C\psi & C\phi S\theta C\psi + S\phi S\psi \\ S\phi C\theta & S\phi S\theta S\psi + C\phi C\psi & S\phi S\theta C\psi - C\phi S\psi \\ -S\theta & C\theta S\psi & C\theta C\psi \end{bmatrix} \\ &\quad \times \begin{bmatrix} L_1 \\ 0 \\ 0 \end{bmatrix} + P_{s,w} \pm L_2 \quad (15) \\ \partial x_w &= \sqrt{\left(\frac{\partial f_x}{\partial \psi} \partial \psi\right)^2 + \left(\frac{\partial f_x}{\partial \theta} \partial \theta\right)^2 + \left(\frac{\partial f_x}{\partial \phi} \partial \phi\right)^2} \\ \frac{\partial f_x}{\partial \psi} &= 0 \quad \frac{\partial f_x}{\partial \theta} = -S\theta C\phi L_1 \quad \frac{\partial f_x}{\partial \phi} = -C\theta S\phi L_1 \\ \partial x_w &= \sqrt{(S\theta C\phi L_1 \partial \theta)^2 + (C\theta S\phi L_1 \partial \phi)^2} \\ &< L_1 \sqrt{0.5 \partial \theta^2 + 0.5 \partial \phi^2}. \end{aligned}$$

According to Xsens, $\partial \psi$, $\partial \theta$, $\partial \phi$ dynamic accuracy is 3 degrees. Therefore we can calculate $\partial x_w < 6 \text{ cm}$, $\partial y_w < 6 \text{ cm}$, $\partial z_w < 6 \text{ cm}$.

The error propagation results coincide with the experimental results shown in Section 7. The deterministic method is analytical; it searches completely the state space, runs fast, and converges correctly. However, this method is sensitive to noise, as errors are not modelled and compensated for in tracking. Error from different sources may accumulate as time passes, and degenerate system performance.



Fig. 10. Inertial position.

6.2. Probabilistic Fusion Method

We propose using a probabilistic fusion method, an extended Kalman filter (EKF), for arm motion tracking. Process noise, measurement noise, and model noise, are normally modelled and accounted for in probabilistic methods. These kinds of system are expected to achieve more robust performance over noise.

System configuration is slightly changed from the deterministic method. The inertial sensor is now attached to the position of the wrist joint by a colour belt, and one of the axes of sensor frame B is aligned with the forearm as shown in Figure 10. The reason for moving the inertial sensor from the upper to forearm is to reduce the affection of muscle movement on the output of the inertial sensor. The colour belt is tracked with a visual camera using CAMSHIFT. The advantage of using a colour belt is that it will reduce the possibility of occlusion.

6.2.1. State Space Pruning

It is always desirable that the size of a state vector should be as small as possible while capturing all the essential properties of a system. This not only saves computational time, but also ensures a search method can converge to a global minimum/maximum. Moreover, the EKF (Bar-Shalom and Fortmann 1988) is a linearised filter, and does not work well with large sizes of state vectors in highly nonlinear situations. Finally, since we are using one visual feature point in measurements, it is important to estimate with as small variability as possible, to make the system observable. Moeslund and Granum (2000) proposed an alternative phase space representation for arm motion capture, in which only two variables are required to represent arm pose, i.e. the z coordinate of a wrist joint z_e and swivel angle α . Various constraints, such as anthropometric, kinematic, and collision, are employed to prune

impossible z_e and α . However, the problem with this representation is that $z_e-\alpha$ space isn't directly related to the visual measurement silhouette. Each time, phase space variables have to be converted from $z_e-\alpha$ state space to Cartesian coordinates in order to match silhouette data. This is computationally expensive and can introduce unavoidable errors.

In our arm tracking system, the size of a state vector can be reduced from six to three in Cartesian coordinates. The main idea is to simplify arm motion tracking by only tracking the pose of the wrist joint so that the elbow joint is then inferred from wrist joint tracking and arm geometry information. This state space is linked to the visual and inertial measurements directly, and does not need any space conversion. The detailed procedures are similar to those of calculating elbow positions in the deterministic method. The elbow joint in local frame B is constant. Given the pose of local frame B with respect to reference frame W, elbow joint position can be represented in reference frame W using:

$$P_{e,W} = R_W^B P_{e,B} + P_{w,W}(k). \quad (16)$$

The system's state vector is now down to 3DOF by tracking the wrist joint position:

$$X(k) = P_{w,W}(k) = (x_w(k), y_w(k), z_w(k))^T \quad (17)$$

6.2.2. Fusion

The goal of fusion filtering is to estimate the pose of a wrist joint from the measurement of inertial and visual sensors. Sensor fusion is implemented with an EKF that operates using a predictor-corrector model, as shown in Figure 11. Inertial tracking represented by (4), serves as the dynamic motion model of the system, and is used to predict space states. Predicted states are corrected using the visual measurement (7). The system's state estimate $\hat{X}(k)$ and associated state covariance matrix $PP(k)$ involve (18), and equal the prediction function in Figure 11.

$$X(\hat{k} + 1)^- = f(\hat{X}(k)) \quad (18)$$

$$PP(k + 1)^- = F(k)PP(k)F(k)^T + Q(k).$$

Measurement correction equations are given by:

$$\hat{Z}(k) = f(\hat{X}(k)^-) \quad (19)$$

$$K(k) = PP(k)^- H(k)^T (H(k)PP(k)^- H(k)^T + R(k))^{-1} + R(k))^{-1}$$

$$X(\hat{k} + 1) = X(k + 1)^- + K(k)(Z(k) - \hat{Z}(k))$$

$$PP(k + 1) = (I - K(k)H(k))PP(k)^-$$

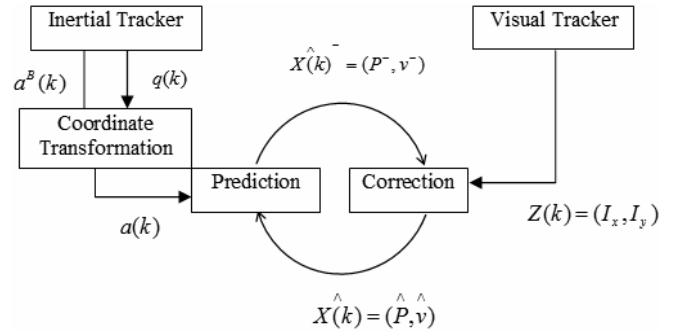


Fig. 11. Sensor fusion framework.

where $Z(k)$ is visual measurement at time k ; $\hat{Z}(k)$ is predicted visual measurement given the current predicted state according to (7); $F(k)$ and $H(k)$ are the gradient of dynamics (4) and measurement (7) respectively, at the current state estimate. $Q(k)$ is a process noise covariance matrix. It is assumed in our experiments that accelerometer measurements are Gaussian noise, and isotropic with $\sigma = 0.4 \text{ m/s}^2$. $R(k)$ is a measurement covariance matrix. Visual measurement noise is also assumed to be Gaussian and isotropic with $\sigma = 4$ pixels.

The visual camera and inertial sensors have different sampling frequencies. In our method, we simply scale the sampling frequency of the inertial sensor 100 Hz/s to equal that of the visual camera, which generates 25 frames per second. Our system can easily be extended to fuse data from asynchronous sensors by employing a modified EKF (Welch 1996).

In our arm motion tracking method, the inertial sensor is used to obtain the orientation and position of a wrist joint with respect to reference frame W. The visual sensor is used to track the 2D image projection of the wrist joint; and the tracking results then input into a fusion estimator, an EKF, to correct drift error in the inertial sensor. Elbow position calculation is a straightforward task, based on wrist joint calculation and the geometry information of a human arm.

6.2.3. Occlusions

Due to the optical character of a camera, it requires a target object in sight of view. When a target object is out of view or occluded by another object, visual information is no longer available. How to deal with the occlusion problem is one of the criteria used to evaluate the robustness of a tracking algorithm. In our proposed tracking system occlusion is very rare, and occurs only for a short time. First, from an application point of view a subject's arm is normally required to be observable to a physiatrist when performing arm rehabilitation. Second, we use a colour belt and wrap it around a subject's wrist, which makes the target observable from different angles. Third, an inertial sensor is employed to track the target

object with a visual camera. When a target object is not observable by the camera, the sensor fusion EKF relies more on the inertial tracking; which doesn't suffer from occlusion. We use a simple strategy to deal with short time occlusion in our system.

When occlusion happens at time k , visual information is not available. We synthesize the input of correction stage z_k in the EKF by choosing from the following two values:

- (1) The visual tracking results at the previous time instant:

$$z_k^1 = z_{k-1}. \quad (20)$$

- (2) The interpolated visual results from the previous two frames:

$$z_k^2 = \begin{bmatrix} I_x(k) \\ I_y(k) \end{bmatrix} = \begin{bmatrix} I_x(k-1) \\ I_y(k-1) \end{bmatrix} + \begin{bmatrix} I_x(k-1) \\ I_y(k-1) \end{bmatrix} - \begin{bmatrix} I_x(k-2) \\ I_y(k-2) \end{bmatrix}. \quad (21)$$

The choosing criterion is calculated according to the following equations:

$$z_k = \min(f(z_k^1), f(z_k^2)) \quad (22)$$

$$f(z_k) = [z(k) - \hat{z}(k|k-1)]S^{-1}(k)[z(k) - \hat{z}(k|k-1)] \leq \gamma \quad (23)$$

where S is the covariance matrix in the EKF, and (23) represents the validation region.

This strategy works well for short time occlusions and smooth motions. However, when sudden motion occurs, or the synthesized visual information is incorrect, tracking performance may degenerate. We propose a further accuracy improvement model in the next section to improve tracking performance.

6.2.4. Accuracy Improvement Model

Experimental results in Section 7 show that the fusion method can successfully estimate a system's state based on inertial and visual data. However, the filter's performance may degenerate when both inertial and visual measurements are poor (or noise parameter settings not properly established). An additional optimization model can be included in the system to improve accuracy in these situations. It is called an accuracy improvement model, and is based on two constraints.

The first constraint relies on the geometry of the human arm, in that the length of an upper arm L_1 is fixed:

$$(P_{s,w} - P_{e,w}(X(k)))^2 - L_1^2 = 0. \quad (24)$$

The second constraint uses the perspective projection property of the pin-hole camera model. The re-projection position of the wrist joint in a 2D image plane should be the same as that for the colour tracking results.

$$I_x - \text{proj}(X(k))_x = 0, \quad I_y - \text{proj}(X(k))_y = 0. \quad (25)$$

We formulate the two constraints into a cost function:

$$f(X(k)) = \frac{1}{2}((P_{s,w} - P_{e,w}(X(k)))^2 - L_1^2)^2 + (I_x - \text{proj}(X(k))_x)^2 + (I_y - \text{proj}(X(k))_y)^2. \quad (26)$$

The first term in (26), means that the elbow joint lies on a sphere surface in which the centre is located in the shoulder joint and the radius is the length of the upper arm. The second and third term is to restrict the wrist joint, lying on a back-projected line through the camera projection centre and the 2D image position of colour tracking. Now, the task is to find $X(k)$ which minimizes (27):

$$X(k) = \text{argmin}_{X(k)} f(X(k)). \quad (27)$$

This is a nonlinear least square problem. The famous optimization algorithm Levenberg–Marquardt (LM) method is employed here to find the system's state vector $X(k)$ at time k that minimizes the cost function. Using an LM algorithm to solve the minimization problem requires the initial value of the variable. It is an iterating method; so it is important to provide it with a good initial value. This may reduce the number of times the algorithm iterates, and ensure convergence to a correct minimum. We combine the sensor fusion method described in the last section and an LM algorithm to achieve improved tracking accuracy. The procedures are as follows:

- (1) Fuse data from the inertial and visual sensor using an EKF algorithm.
- (2) The results from the EKF method are used as the initial value in (26) to find the minimum that meets (27).

6.2.5. Initialization

Initialization work is required to define the world frame, and calculate the orientation and position of reference frame W relative to camera frame C , represented as R_C^W, T_C^W in Figure 12. This is also necessary in order to relate the inertial and visual measurements in a consistent coordinate frame in preparation for fusion. As we have mentioned, reference frame W is defined as the original frame of inertial local frame B . We employ a visual pose estimation algorithm (known as the Perspective n point (P-N-P) method) to initialize our tracking system.

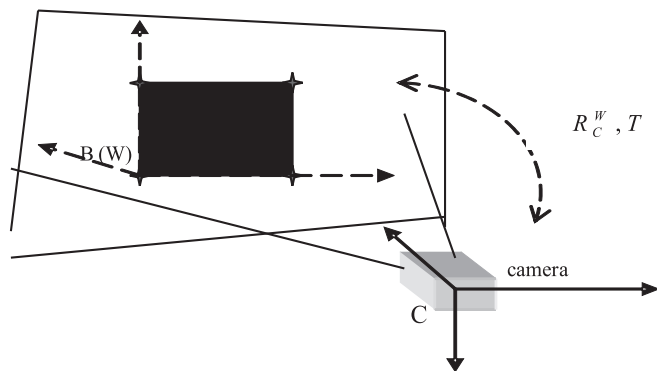


Fig.12. Illustration of visual pose estimation.

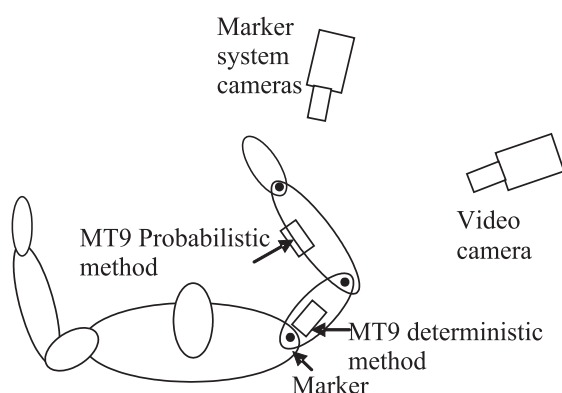


Fig. 13. Top view of the experimental set up.

The principle of the P-N-P method is that given n 3D points and their corresponding 2D projections in an image plane, the pose between the cameras coordinate C and object coordinate system B can be calculated (see Figure 12) using the algorithm proposed by Lu et al. (2000).

The minimum number of feature points needed to uniquely determine the pose between the two frames is four. In our method, we use a square patch and the four corners are used as feature points for pose estimation. The feature point's detection is achieved by first using the CAMSHIFT colour tracking algorithm to find the region of interest; feature points are then detected and selected using LKFtracker (Shi and Tomasi 1994). These feature point detection results are input into the P-N-P algorithm to obtain pose information.

7. Experimental Results

The tracking performance of our proposed hybrid tracking system for arm motion is evaluated by comparing its results with a commercial marker-based motion tracking system. We used the CODA marker tracking system when performing

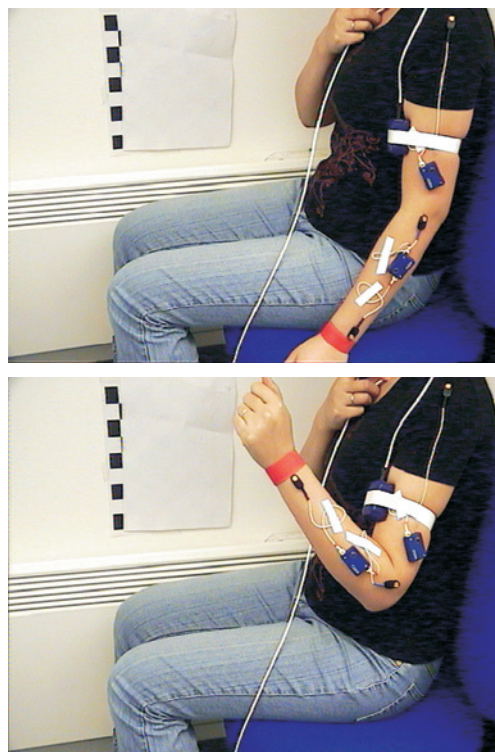


Fig. 14. Subject wearing an MT9 and CODA markers.

deterministic tracking method validation, and the Qualysis marker tracking system when performing probabilistic tracking method validation. This is due to the availability of the marker tracking systems. Results from the marker-based tracking system are accurate, and can be regarded as a ground truth. If the accuracy of our tracking system can approximate the accuracy of the marker-based tracking systems (even under specific situations), this would be a good starting point for our initial investigation.

7.1. Experimental Set-up

In order to evaluate the accuracy of our tracking system, we capture a subject's motion by using a marker-based tracking system and our hybrid tracking system simultaneously. Figure 13 shows the top view of our experimental set up. A subject wears both an MT9 sensor and markers. The markers are attached to the three arm joints of shoulder, elbow, and wrist; while an inertial sensor is attached to the upper arm for the deterministic method, and on the wrist joint for the probabilistic method. The video camera and marker system camera capture a subject's motion at the same time. Data from the two systems is compared in order to evaluate the performance of our proposed tracking method.

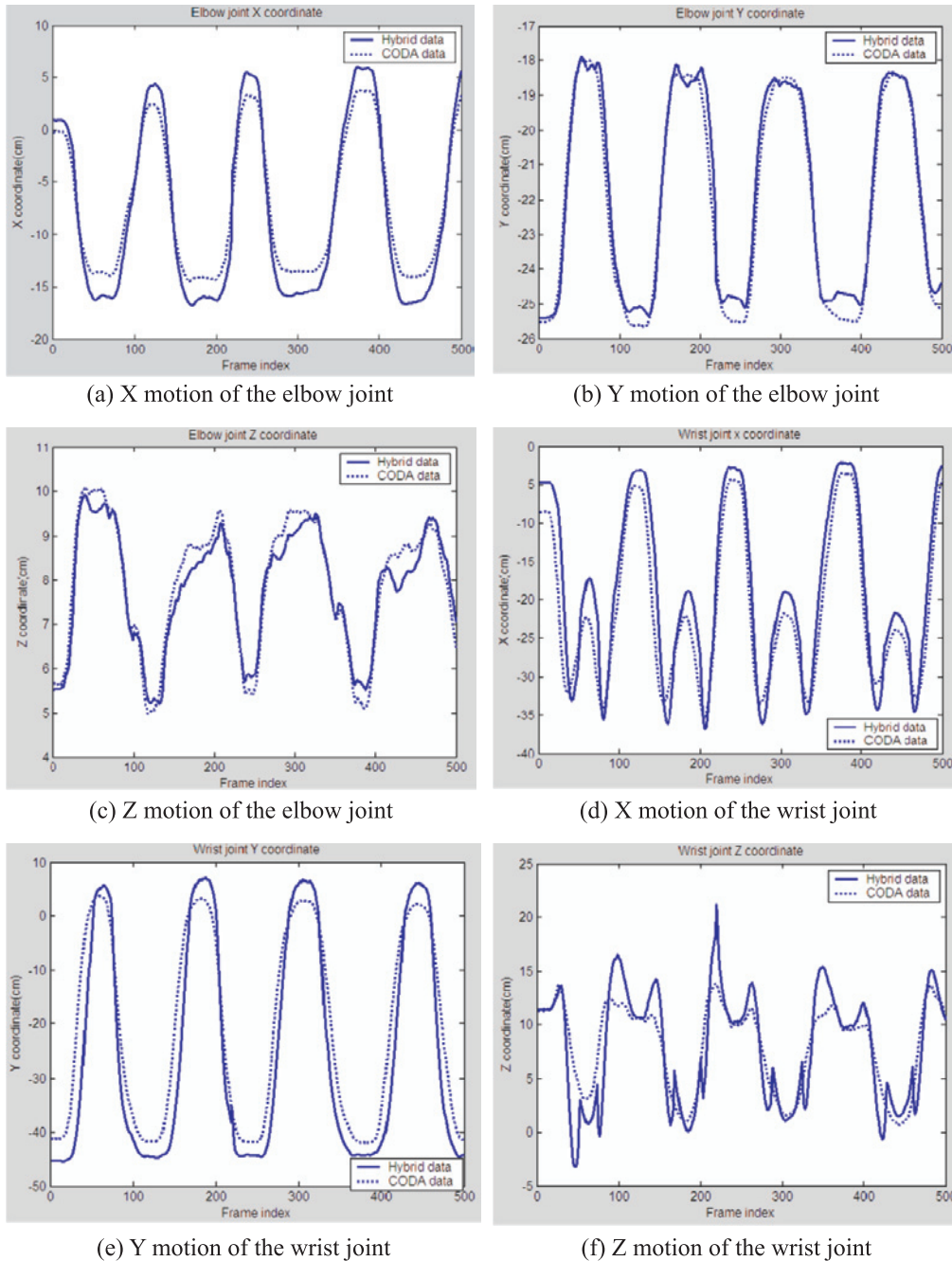


Fig. 15. Comparison of the performance of a CODA system and our hybrid system for arm motion tracking.

7.2. Experiment Results – Deterministic Method

A number of repeated motion patterns were tested in order to establish comprehensive comparisons. Figure 14 shows a subject wearing an MT9 and Markers, performing flexural and extensible arm motions. Figure 15 illustrates a full arm motion sequence for about 20 seconds. Shoulder position is as-

sumed to be fixed, therefore only elbow and wrist positions represented in the world coordinate system are illustrated. The three coordinates x , y and z of the position trajectories of the elbow and wrist joints from each tracking system are plotted respectively in Figure 15. Dotted lines represent data from the CODA system, and solid lines from our hybrid system. The three figures (a), (b), and (c) show the coordinates of the el-

bow joint, while (d), (e), and (f) show the coordinates of the wrist joint. As can be seen from our experimental results, our proposed hybrid tracking method is reasonably accurate and efficient.

As can be seen in Figure 15, the data from our hybrid tracking system is very close to the CODA results and match well, especially for the elbow joint. The difference between the two systems for the elbow joint is around 2 cm. The data for the wrist joint is noisier than the elbow joint; but the difference between the two systems is still within 4–6 cm, which is quite promising and has great potential for use in home-based rehabilitation.

The deterministic method is a straightforward closed-form method and easy to implement. It runs in real time, and provides accurate tracking results as shown above. However, the characteristics of the deterministic method means that the tracking system has limited tolerance to noise, therefore the performance of the deterministic method is sensitive to noise. It works well when motion is simple, however, when tracking more complicated motions, or when sensor measurement is noisy, a more robust method is required. The proposed probabilistic method accounts for this noise by modelling it in a probability distribution. It therefore produces more stable and robust results in cluttered situations. Experiment results are shown in the next section.

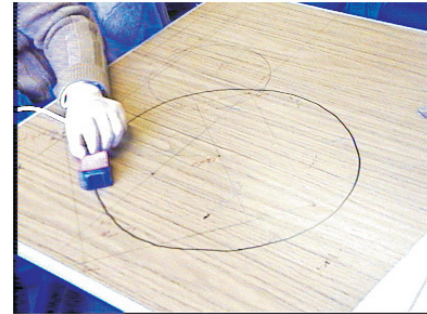
7.3. Experiment Results – Probabilistic Method

In this experiment, we performed two types of arm motion to test the performance of the probabilistic fusion method. The motion patterns of the subject are a rectangle and a circle respectively; as shown in Figure 16(a) and Figure 16(b). These motion patterns are used frequently in the stroke rehabilitation area (GENTLE/S).

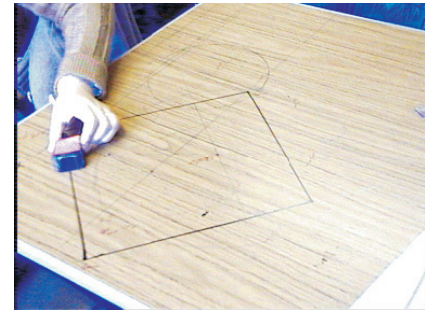
The inertial sensor is attached to a wrist joint, and the coloured patch on its top tracked by a video camera. Three markers from Qualysis are attached on the arm joint of the shoulder, elbow, and wrist respectively. Figure 17 shows the subject wearing both the inertial sensor and Qualysis markers during probabilistic experiments.

The results from our proposed fusion method and the marker-based tracking system are illustrated in Figure 18 and Figure 19 for the circular and rectangle motion respectively. We describe the tracking performance in two ways: the first is the three-coordinate trajectory of the wrist joint; the second is the 3D reconstructed trajectories of the wrist joint. The bold lines in the two figures represent the tracking results from the marker-based system, and are regarded as the motion ground truth. Solid lines are the results of the EKF fusion method, and dashed lines are the results of the EKF+Accuracy improvement model.

It is clear from the results that the use of the EKF alone produces noisy results. When the accuracy improvement model



(a) Circle motion



(b) Rectangle motion

Fig. 16. Desired motion patterns.

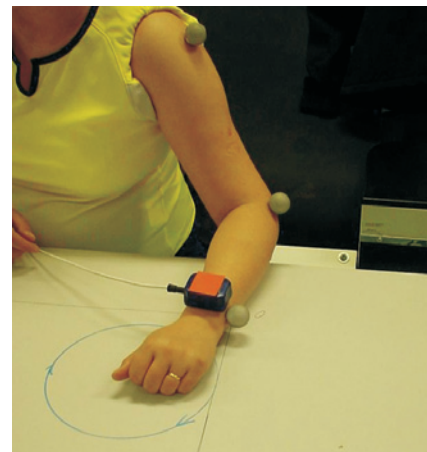


Fig. 17. Subject wearing both inertial sensor and markers.

is applied, tracking performance improves significantly and is very close to the ground truth. The reason that the EKF performs poorly in some cases is that it linearizes non-linear arm motion, and assumes the probability distributions of noise are Gaussian. When this assumption is not satisfied, the performance of the EKF degrades. The accuracy improvement model works well in these situations, and provides accurate results.

To further evaluate the proposed fusion method, we calculated the statistical performance of the algorithms. The statis-

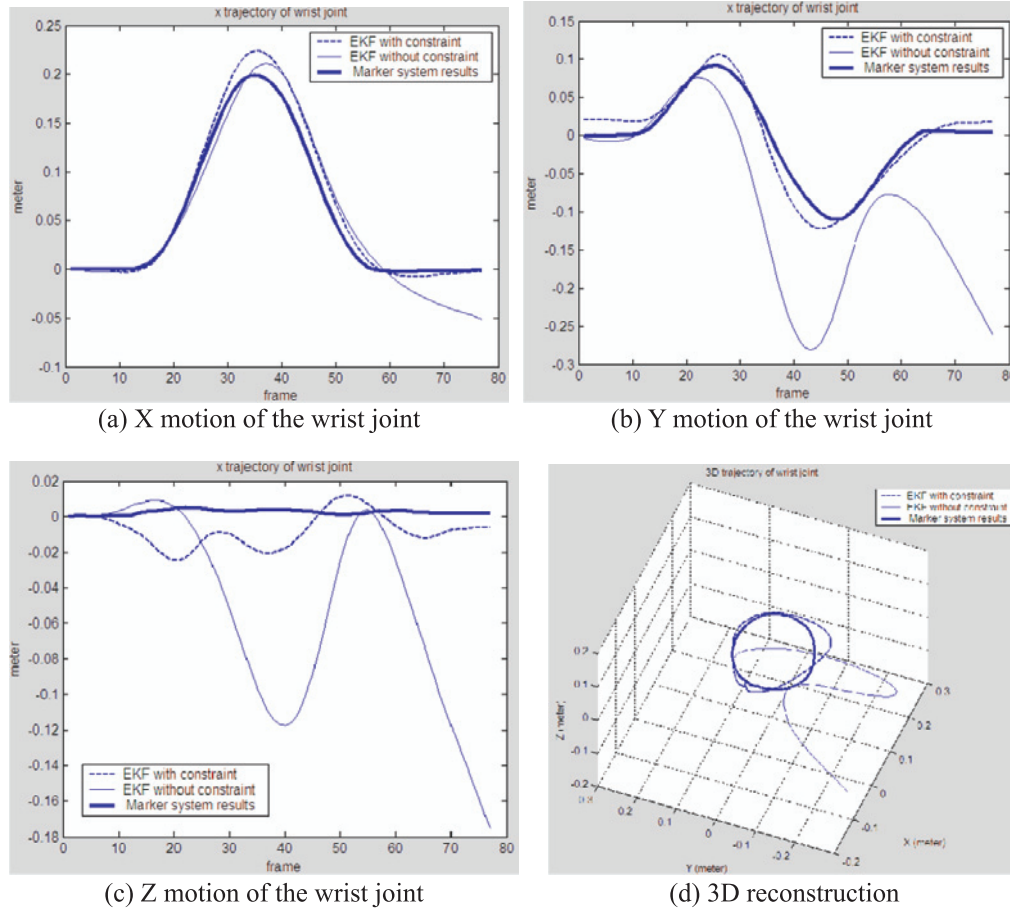


Fig. 18. Comparison of the performance of a marker system with our hybrid system – circular motion.

Table 1. Mean Error and Standard Deviation in Each Axis of a Circle Motion for Different Algorithms

Circle motion	X coordinate error (cm)	Y coordinate error (cm)	Z coordinate error (cm)
Inertial only	58.80/70.58	71.04/110.25	-64.66/82.96
EKF	-5.01/8.46	-29.92/40.22	-19.52/27.07
EKF with constraint	0.73/1.07	0.26/1.81	0.65/1.24

Table 2. Mean Error and Standard Deviation in Each Axis of a Rectangle Motion for Different Algorithms

Rectangle motion	X coordinate error (cm)	Y coordinate error (cm)	Z coordinate error (cm)
Inertial only	19.63/32.43	2.6/9.8	31.85/33.49
EKF	0.19/1.33	6.0/8.5	-2.6/4.25
EKF with constraint	0.27/1.09	-0.41/2.19	-1.6/1.250.19

tical properties, average mean, and standard deviation for each motion pattern were calculated in our experiments. A number of the same motion patterns were captured and compared with the marker-based tracking results. We calculated the mean error and standard deviation of each motion axis for the circle and rectangle motion respectively. Tables 1 and 2 show the statistical properties of the different algorithms. It is clear that the inertial only tracking method suffers severe drift error. The

EKF greatly improved tracking results by fusing inertial data with visual data, but its accuracy is not sufficient. The EKF with an accuracy improvement model provides the best performance of the three algorithms. It has great potential for application in home-based rehabilitation.

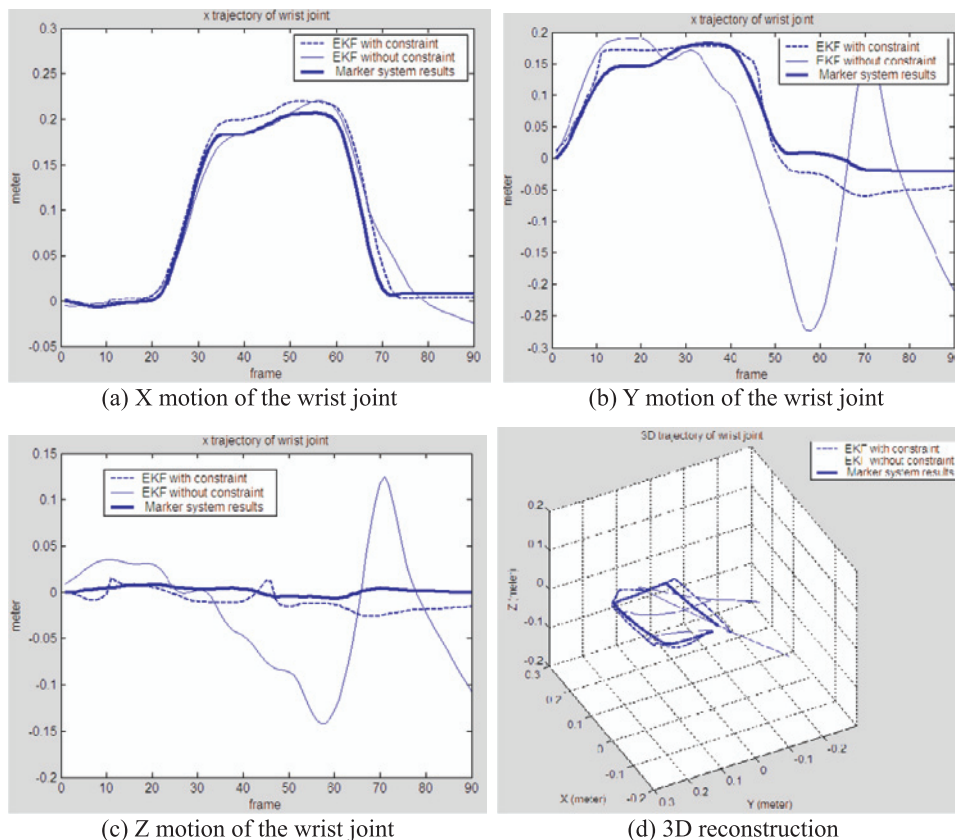


Fig. 19. Comparison of the performance of a marker system with our hybrid system – rectangle motion.

8. Conclusions and Future Work

Currently, stroke patients take physiotherapy with the help of physiotherapists or well-trained carers to diagnose their rehabilitation activities. We proposed the development of a sensor system to support rehabilitation programmes for patients in their home environment so that a burden on hospitals and physiotherapists could be relieved. This paper presents a new arm motion tracking system based on an integration of vision and inertial sensors, including both deterministic and probabilistic approaches. The deterministic method uses human arm geometry information to fuse different data modalities, while the probabilistic method fuses data using the popular extended Kalman filter. The performance of an EKF is further improved by an accuracy model based on arm geometry constraints. Experiment results show that our proposed method is able to track arm movement accurately and in real time, in comparison to commercial marker-based motion tracking systems.

Our future work will focus on two aspects. One is to release the fixed shoulder position constraint. The other is to extend the method from tracking arm movement to upper body movement. The two main challenging issues to be addressed are:

- how to exploit more useful image features such as contours and edges; and
- how to build a proper kinematics model for upper body tracking.

Acknowledgements

We would like to thank Charnwood Dynamics for their CODA motion tracking system, and Dr Martin H. Sellens of the Biological Science Department at the University of Essex for allowing us to use their Qualysis motion tracking system. Our thanks also go to the other members of the EPSRC EQUAL Smart Rehabilitation Consortium for useful discussion.

References

- Aggarwal, J. K. and Cai, Q. (1999). Human motion analysis: a review. *Journal of Computer Vision and Image Understanding*, **73**(3): 428–440.

- Alenya, G., Martinez, E., and Torras, C. (2003). Fusing visual contour tracking with inertial sensing to recover robot egomotion. *Workshop on Integration of Vision and Inertial Sensors Proceedings of International Conference on Advanced Robotics*, Coimbra, Portugal.
- Bachmann, E. R. (1999). Orientation tracking for humans and robots using inertial sensors. *Proceedings of International Symposium on Computational Intelligence in Robotics and Automation*, Monterey, CA, USA.
- Bar-Shalom, Y. and Fortmann, T. E. (1988). *Tracking and Data Association*. Academic Press.
- Bradski, G. R. (1998). Computer vision face tracking for use in a perceptual user interface. *Intel Technology Journal*, **Q2**:15.
- Bussmann, H. (2000). Ambulatory monitoring of mobility-related activities in rehabilitation medicine. PhD Thesis, Erasmus University Rotterdam. Delft, Eburon.
- Chai, L., Hoff, W. A., and Vincent, T. (2002). 3D motion and structure estimation using inertial sensors and computer vision for augmented reality. *Tele-operators and Virtual Environments*, **11**(5): 474–492, MIT Press.
- Charnwood Dynamics Ltd. URL: <http://www.charndyn.com/>
- Chen, J. and Pinz, A. (2004). Structure and motion by fusion of inertial and vision-based tracking. *Proceedings of the 28th OAGM/AAPR Conference*, Vol. 179 of Schriftenreihe, 55–62. OCG.
- Chen, Z. and Lee, H. J. (1992). Knowledge-guided visual perception of 3D human gait from a single image sequence. *IEEE Transactions On Systems, Man, and Cybernetics*, **22**(2): 336–342.
- Clinical Gait Analysis. URL: <http://guardian.curtin.edu.au:16080/cga/>
- Deutscher, J., Blake, A., and Reid, I. (2000). Articulated body motion capture by annealed particle filtering. *IEEE Conference on Computer Vision and Pattern Recognition*, Vol. 2, 126–133.
- Foxlin, E. (2002). Generalized architecture for simultaneous localization, auto-calibration, and map-building. *IEEE/RSJ International Conference on Intelligent Robots & Systems*, Lausanne, Switzerland.
- Foxlin, E. et al. (2004). FlightTracker: a novel optical/inertial tracker for cockpit enhanced vision. *IEEE/ACM International Symposium on Mixed and Augmented Reality (ISMAR 2004)*, Washington, DC, USA.
- Gavrila, D. M. (1999). The visual analysis of human movement: a survey. *Journal of Computer Vision and Image Understanding*, **73**(1): 82–98.
- Gavrila, D. M. and Davis, L. S. (1995). 3-d model-based tracking of human upper body movement: A multi-view approach. In *Int. Symposium on Computer Vision*, 253–258.
- GENTLE/S. URL: <http://www.gentle.rdg.ac.uk/>
- Goncalves, L. et al. (1995). Monocular tracking of the human arm in 3D. *Proceedings of International Conference on Computer Vision*, Cambridge, MA, USA, June, 764–770.
- Huster, A. (2003). *Relative Position Sensing by Fusing Monocular Vision and Inertial Rate Sensors*. PhD thesis. Department of Electrical Engineering, Stanford University, USA.
- Huster, A. and Rock, S. M. (2001). Relative position estimation for manipulation tasks by fusing vision and inertial measurements. In *Oceans 2001 Conference*, Honolulu, USA, November, Vol. 2, 1025–1031.
- Johansson, G. (1973). Visual perception of biological motion and a model for its analysis. *Perception and Psychophysics*, **14**(2): 210–211.
- Ju, S. X., Black, M., and Yacoob, Y. (1996). Cardboard people: a parameterised model of articulated motion. *2nd International Conference on Automatic Face and Gesture Recognition*, Killington, Vermont, USA, 38–44.
- Lang, P., Ribo, M., and Pinz, A. (2002). A new combination of vision-based and inertial tracking for fully mobile, wearable and real-time operation. *Proceedings of 26th Workshop of the Austrian Association for Pattern Recognition (GM/AAPR)*, Graz, Austria, Vol.160, 141–148.
- Lu, C.P., Hager, G. D., and Mjolsness, E. (2000). Fast and globally convergent pose estimation from video images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **22**(6): 610–622.
- Moeslund, T. and Granum, E. (2000a). Multiple cues used in model-based human motion capture. *The fourth International Conference on Automatic Face and Gesture Recognition*, Grenoble, France.
- Moeslund, T. and Granum, E. (2001). A survey of computer vision-based human motion capture. *Journal of Computer Vision and Image Understanding*, **81**(3): 231–268.
- Moeslund, T. B. and Granum, E. (2000b). 3D human pose estimation using 2D-data and an alternative phase space representation. *Workshop on Human Modelling, Analysis and Synthesis at CVPR*, Hilton Head Island, South Carolina, USA.
- Qualisys. URL: <http://www.qualisys.com/>
- Shi, J. and Tomasi, C. (1994). Good features to track. *IEEE Conference on Computer Vision and Pattern Recognition*, Seattle, USA, 593–600.
- Sidenbladh, H., Black, M., and Fleet, D. (2000). Stochastic tracking of 3D human figures using 2D image motion. *European Conference on Computer Vision*, Dublin, Ireland, 702–718.
- Sminchisescu, C. (2002). *Estimation Algorithms for Ambiguous Visual Models 3D Human Modeling & Motion Reconstruction in Monocular Video Sequences*. PhD Thesis, Institute National Polytechnique de Grenoble INRIA, France.
- Strelow, D. and Singh, S. (2003). Online motion estimation from image and inertial measurements. Online motion estimation from image and inertial measurements. *Workshop on Integration of Vision and Inertial Sensors (INERVIS 2003)*, Coimbra, Portugal.

- Tao, Y. and Hu, H. (2004). Colour-based human motion tracking for home-based rehabilitation. *Proceedings of IEEE International Conference on Systems, Man and Cybernetics*, The Hague, The Netherlands, 773–781.
- Tetrud, J.W., Sabelman, E. E., and Yap, R. (2002). Accelerometer identification of freezing-of-gait in Parkinson's syndrome. *Proceedings of the 7th International Congress on Parkinson's Disease and Movement Disorders*, Miami Beach, FL, November.
- Tolani, D., Goswami, A., and Badler, N. I. (2000). Real-time inverse kinematics techniques for anthropomorphic limbs. *Graphical Models*, **62**(5): 353–388.
- Uncertainties and Error Propagation (2007). <http://www.rit.edu/~upphysics/uncertainties/Uncertaintiespart1.html>, Accessed on 18 January 2007.
- Veltink, P. H. et al. (1996). Detection of static and dynamic activities using uniaxial accelerometers. *IEEE Trans Rehabil Eng*, **4**(4): 375–385.
- Vezhnevets, V., Sazonov, V., and Andreeva, A. (2003). A survey on pixel-based skin colour detection techniques. *Graphicon-2003*, Moscow, Russia.
- Wang, L., Hu, W., and Tan, T. (2003). Recent developments in human motion analysis. *Pattern Recognition*, **36**(3): 585–601.
- Welch, G.F. (1996). SCAAT: Incremental Tracking with Incomplete Information. PhD thesis, University of North Carolina at Chapel Hill, USA.
- Wren, C., Azarbayejani, A., Darrel, T., and Pentland, A. (1995). Pfinder: Real-time tracking of the human body. *Proceedings of SPIE*, Bellingham, WA, USA.
- Xsens Motion Technologies. (2007). URL: <http://www.xsens.com/>. Accessed on 12 January 2007.
- You, S. and Neumann, U. (2001). Fusion of vision and gyro tracking for robust augmented reality registration. *IEEE Virtual Reality*, Japan
- You, S., Neumann, U. and Azuma, R. (1999). Hybrid inertial and vision tracking for augmented reality registration. *Proceedings of IEEE VR '99*, Houston, TX, 260–267.
- Zhou, H. and Hu, H. (2005). Inertial motion tracking of human arm movements in home-based rehabilitation. *Proceedings of IEEE International Conference on Mechatronics and Automation*, Sheraton Fallsview Hotel, Niagara Falls, Ontario, Canada.