

Integration of Vision and Inertial Sensors for Home-based Rehabilitation

Yaqin Tao, Huosheng Hu, Huiyu Zhou

Department of Computer Science, University of Essex
Wivenhoe Park, Colchester CO4 3SQ, United Kingdom

Email: ytao@essex.ac.uk, hhu@essex.ac.uk, zhou@essex.ac.uk

Abstract-*This paper introduces a real-time tracking system of human arm motion, specifically intent to be used for home-based rehabilitation. Both vision and inertial sensors are employed in this system to track the arm movement in 3D and in real time. Different data modalities from two sensors are fused by using arm structure relationship and geometry information. The occlusion problem, which is one of the most difficult issues in most of vision only tracking systems, is well dealt with by combining the geometry information of the subject and multiple sensors in our proposed system. The experimental results show our system can track the 3D arm motion in real time and has acceptable accuracy, compared with the commercial marker-based system, CODA.*

Keywords: Human motion tracking, multiple sensors integration, rehabilitation.

1. Introduction

Visual tracking and analysis of human motion is currently one of the most popular research topics in computer vision [1][2][3][4]. The motivation is driven by its wide application domains such as surveillance tracking, athletic/medical performance analysis, and perceptual interface. Most existing visual tracking systems can be classified into two categories: marker-based visual tracking systems and marker-free visual tracking systems. Using markers to indicate the target objects partially simplifies the human motion tracking problem and a lot of commercial marker-based systems are available on the market such as CODA [5] and Qualisys [6]. These systems are accurate and have been successfully used in many areas, like the athletic performance analysis and the motion capture for animation. However, it is expensive and only runs in a supervised environment.

It is always desirable to develop a marker-free visual tracking system instead of the intrusive marker-based ones, because they are more applicable and consistent. Additionally, they can be relatively cheaper, especially compared with commercial marker-based tracking systems. However, visual tracking of human motion is a non-trivial task because of many difficulties [7] such as depth ambiguities, occlusion and kinematics. In order to simplify the human motion tracking problem, most human motion tracking algorithms made a number of assumptions [3] or required prior knowledge about the appearance of the subject, the geometry of the subject, the kinematics and

dynamics of the subject; using multiple cameras etc. A model-based method is one of the most popular methods in human motion tracking. The shape model of a subject varies from the simplest skeleton model [8] to 2D patches models [9], and to the sophisticated 3D volumetric models [4][10][11]. The tracking is conducted by matching image features (edge, optical flow, silhouette/contour, template, colour, or blobs) with the shape model under different assumptions and prior knowledge of human body and motion. Model-based human motion tracking is high level processing and normally can be formatted into the Bayesian framework [7][10][11], where different estimators such as Kalman filter [12] and Condensation [13] are frequently employed. The problems with this kind of systems are: 1) they are computationally expensive and difficult to run in real time; 2) most of them are initialised manually, which is unrealistic in real applications; 3) the shape models used by these systems are subject specific, which is difficult to generalise; 4) the accuracy of tracking results is also a problem to be further investigated.

Non-vision motion tracking is also an active research area. Different sensors such as inertia, acoustic, and magnetic sensors [14] are attached to the human body to collect movement information. Some non-vision tracking systems, for example magnetic systems, are able to accurately track subject's movement. However, non-vision sensors have some limitations. For instance, inertial sensors have drift problems, acoustic sensors are sensitive to noise and magnetic sensors are complex and wired.

In general, each sensor has unique strengths and weakness. The recent tendency is to integrate different sensors into a motion tracking system, to compensate for their shortcomings and produce robust performance of the system. Among different sensors, a combination of vision and inertial sensors is a popular choice for human motion tracking. An inertial sensor produces accurate linear acceleration (accelerometers) and rate of turn (gyros), but the differential inertial measurements have a drift problem, which is corrected by fusing visual measurements. A lot of attempts have been made to use visual and inertial sensors to track subject's motion. Foxlin et al. [14] used two inertial sensors and three cameras to track the pilot's head motion accurately in cockpit for enhanced vision. You et al. [16] integrated visual and inertial sensors for tracking in augmented reality applications. Data fusion is regarded

as an image stabilization problem. While Lang et al. [17] used extended Kalman filters to fuse the different data modalities from visual and inertial sensors. It is also used as a mobile augmented reality (AR).

Most integrated tracking systems are used to track a rigid object at 6 degrees of freedom for augmented reality applications. Our interest in this paper is to integrate visual and inertial sensors for tracking the motion of articulated objects--human body. The purpose is to develop a motion tracking system, which is cheap, accurate and run in real time, for a home-based rehabilitation project. Traditionally patients who sustain a stroke take physiotherapy with the help of physiotherapists or well-trained carers to diagnose their rehabilitation activities. We propose to develop a sensor system to support the rehabilitation program for the patients at home environments so that the burden of hospitals and the physiotherapists could be relieved.

A video camera and an inertial sensor are employed in our tracking system to capture the subject's motion. We are currently focusing on tracking the movement of human upper limbs. The geometry and structure information are used in this paper to fuse the data from two sensors.

The rest of the paper is organised as follows. Section 2 introduces inertial sensor tracking briefly. Section 3 describes visual-based colour object tracking. The integration of vision and inertial sensors for tracking the arm movement is presented in Section 4. Our experimental results on the accuracy test of the hybrid motion tracking method are shown in Section 5. Finally, conclusions and future work are presented in Section 6.

2. Inertial Tracking

The conventional usage of inertial sensors in motion tracking is to calculate the relative change of a moving target in position ΔP^t and orientation ΔO^t between two consecutive sampling times, based on measurements of acceleration a^t and angular velocity ω^t from the inertial sensor. The superscript t means at sampling time t.

In this paper, we are only using three Euler angles $\{\phi, \theta, \varphi\}$, which mean roll, pitch, yaw of the reference coordinate (W) (subsequent rotation around X, Y, Z axis), calculated from the angular velocity of a 3-axis inertial sensor MT9 [18]. The rotation rate has less drift problem than the acceleration data, and the effectiveness has been verified in [26]. The acceleration data will be exploited later in future work.

The MT9 inertial sensor has a body fixed Cartesian

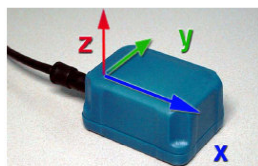


Fig. 1 MT9 with body fixed coordinate system M [18]

coordinate system (see Fig. 1) and is defined as the sensor coordinate system (M). If we define a reference world coordinate system (W), three Euler angles of the output from a MT9 sensor represent the orientation between the MT9 body fixed coordinate system (M) and a user defined reference coordinate (W) (see Fig. 2).

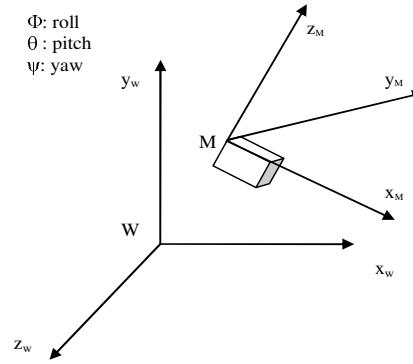


Fig. 2 Inertial sensor coordinates M relative to reference coordinate W

3. Vision based Tracking

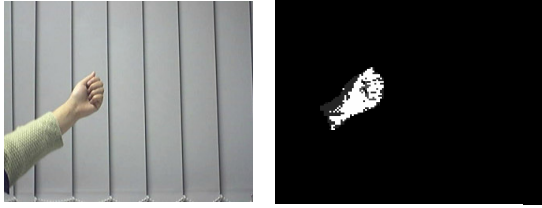
The images captured by a video camera can offer different features to assist object segmentation and tracking. Some of the most useful features are edges, colour, contour, optical flow etc. In our method, only colour feature (skin colour) of the target object (the hand) is employed. The reason is to achieve real-time performance and reduce the possibility of occlusion problems.

3.1. Colour Object Tracking

Skin colour is one of the natural image feature and very useful in human motion tracking applications. There are a lot of different methods (parametric or non parametric models) [20] to detect skin colour objects. In our method, we employ a Continuously Adaptive Mean Shift (CAMSHIFT) algorithm [19] to segment and track the colour object. Unlike the skin colour specified algorithm, the CAMSHIFT algorithm is able to track any kind of target colours by building a histogram distribution of the H channel in HSV colour space from the region of interests selected by users at the initial stage. The histogram distribution is later used to segment a target object from the background image. Fig. 3 (a) shows the original image. Fig. 3 (b) is the segmented target object by using the histogram colour model built from user selected region in (a). As we can see from Fig.3 (b), there is noise in the segmented image. The CAMSHIFT is robust to noise image and it's also possible to make the environment less clutter in the rehabilitation application [21].

The correspondence problem of motion tracking is solved by assuming a small motion between two consecutive frames. The spatial mean position $X_T = \{x_T, y_T\}$ of the detected foreground object is

regarded as the 2D position of the target object at time step t . The output of the colour object tracking is the position trajectory in a 2D image plane, which is used later in Section 4 to calculate the wrist joint position.



(a) Original image (b) Segmented image
 Fig. 3 Colour object detection using CAMSHIFT method

3.2. Camera Calibration

The video camera used in our method is calibrated using a pin-hole camera model. Both intrinsic and extrinsic parameters are calculated. Fig. 4 shows camera parameters and the imaging system.

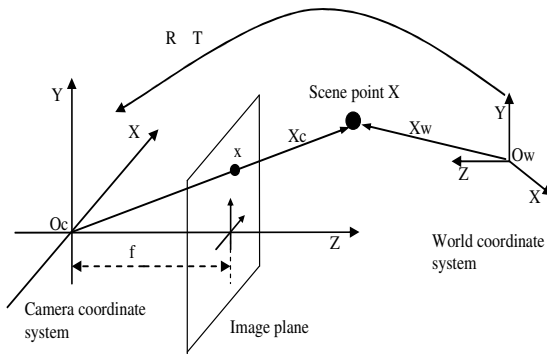


Fig. 4 Camera model and imaging process

Scene point X , represented in the world coordinate system as X_w , is projected on the image plane x under the pin-hole camera model according to (1):

$$x \sim PX_w \quad (1)$$

where P is the projective matrix, which includes both the intrinsic and extrinsic parameters.

4. Hybrid Arm Motion Tracking

The arm motion tracking is achieved by integrating the visual and inertial sensing, which thus makes the tracking system more robust and applicable. This section describes our hybrid arm motion tracking system.

4.1. Arm Model

The arm is modelled as a skeleton structure, which consists of two segments linked by a revolute joint (see Fig. 5). No shape model, (for example, to model the limb as truncated cones [22]), is required currently in our method. Three degrees of freedom (DOF) are assigned to

the upper arm and 1 DOF is assigned to the fore arm. The hand and the fore arm are assumed to be one rigid segment. Therefore, the resulted configuration space is four dimensional. The 4DOF arm model is widely used in the area of computer vision motion tracking [22][23], and is able to approximate arm motion relatively accurate.

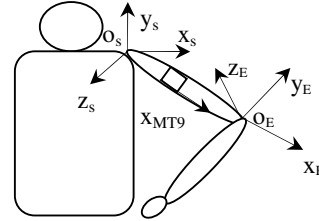


Fig. 5 Arm configuration and coordinate system

4.2. System Configuration

In our system, a MT9 sensor is attached on the upper arm of the subject and the x axis of the inertial sensor is aligned with the upper arm. The subject is required to face or be oblique to the camera view. The upper limb is represented in the shoulder coordinate system (S), and the fore limb is represented in the MT9 body attached coordinate system (M), which means that the elbow coordinate system (E) coincide with the coordinate system (M) in Fig. 5.

In order to simplify motion tracking, a few constraints are imposed in our approach. First, the shoulder is assumed to be fixed during the arm movement. Second, the geometry parameters, the length of the upper arm UL and fore arm FL are assumed to be known. Third, the world coordinate system (W) is defined at the shoulder position for both the camera tracking system and the inertial tracking system. All our measurements can thus be represented or transformed into the shoulder coordinate system.

We are interested in tracking the motion trajectories of human body joints, which contain the essential motion and structure information of human movement. It has been proved by Johansson's moving light displays (MLD) psychology experiments in [24].

4.3. Elbow Position Calculation

As we mentioned in Section 2, the output from MT9 in our method is three Euler angles of the rotated MT9 body attached coordinate system referred to a pre-defined reference coordinate system, which is the shoulder coordinate system (see Fig. 5). The x axis direction of the MT9 sensor at each time step represents the direction of the upper limb. Given the length of the upper arm, the elbow position can be uniquely determined.

The rotation matrix from the shoulder coordinate system (S) to the MT9 body attached coordinate system (M) can be calculated from three Euler angles $\{\phi, \theta, \varphi\}$ and be represented by (2)

$$R_S^M = R_{\phi, \theta, \phi} = R_{z, \phi} R_{y, \theta} R_{x, \phi} \quad (2)$$

The elbow position in the MT9 local body attached coordinate system (M) is always $X_{EM} = \{UL, 0, 0\}$, where UL is the length of upper limb, and the subscript EM means the elbow position represented in the MT9 coordinate system. The elbow position in the shoulder coordinate system can be calculated according to (3)

$$X_{ES} = R_S^M X_{EM} \quad (3)$$

4.4. 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 previous sections. We actually get the candidate wrist joint positions by solving two constraint equations.

One constraint equation is from the colour object tracking and camera calibration results. In Section 3, we described the camera imaging process from 3D to 2D. What we are interested here is the inverse process: how to calculate 3D positions based on 2D image positions. Given the calibrated camera parameters, and an image point x , the corresponding line (line $O_c X_c$ in Fig. 4) in camera centered space is uniquely determined, which comprises of all world points that map to the same image point x . This is a one-to multiple mapping. The possible 3D positions of the scene point X corresponding to the image point x can be represented in the camera coordinate system by (4).

$$X_c(\lambda) = O_c + \lambda P^+ x \quad (4)$$

where P^+ is the inverse pseudo of projective matrix P . Since all our data are represented in the shoulder coordinates system, the back-projected line represented by (4) can be converted to the shoulder coordinate system using camera extrinsic parameters

$$X_{WS}(\lambda) = R^{-1} X_c(\lambda) - T \quad (5)$$

Another constraint equation is from the geometry relationship between elbow and wrist. Based on the elbow position from Section 4.2 and the length of fore arm FL , the position trajectory of the wrist joint is a sphere surface and can be represented by (6)

$$(X_{WS} - X_{ES})^2 = FL^2 \quad (6)$$

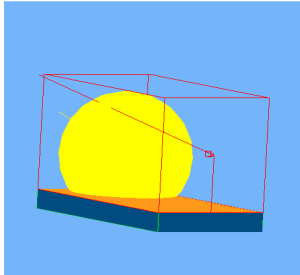


Fig. 6 the intersection of the sphere and the back projected line.

According to the analysis above, the solution space of the wrist joint can be reduced from a line and a sphere surface to at most 2 points by calculating the intersection points of the sphere and the back projected line. The solutions of (4) and (6) are the candidates of the desired 3D position. Fig. 6 shows the intersection of a sphere and a back projected line, where an arrowed line is the back projected line and the yellow ball is a sphere.

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

- (1) No intersection point. For the first situation, we introduce a variance ε for the radius r of the sphere. If the distance between the centre of the sphere and back projected line L satisfies (7), the closest point on the back projected line to the sphere centre is taken as the 3D object position. If (7) is not satisfied, the wrist position in the previous frame is used in this frame.

$$L \leq r + \varepsilon \quad (7)$$

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

- Firstly, 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. Among two intersection points, the one that contributes more to the smooth motion is selected.

- Secondly, the arm model is used to distinguish and select the correct intersection point. In Fig. 7, the coordinate system XYZ represents the MT9 body attached coordinate system at time step t . The forearm is also expressed in this coordinate system. We only assigned one degree of freedom to the forearm, which is the rotation angle α on 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 [25], the trajectory of the wrist joint is a clip circle arc (Fig. 8).

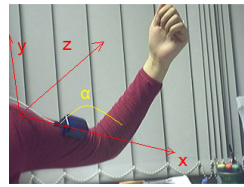


Fig. 7 Forearm coordinates system

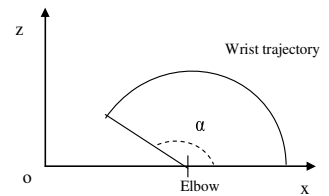


Fig. 8 Wrist trajectory

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

5. Experimental Results

Tracking performance of our proposed hybrid tracking system for arm motion is evaluated by comparing the

results with a commercial marker based motion tracking system, CODA. Results from the CODA system are accurate and can be regarded as a ground truth. If the accuracy of our tracking system can approximate the accuracy of the CODA system, even under specific situations, this would be a good start point for the initial investigation.

5.1. Experimental Set-up

In order to evaluate the accuracy of our tracking system, we capture the subject's motion by using CODA and our hybrid tracking system simultaneously. Fig. 9 shows the top view of the experimental set up. The subject wears both a MT9 sensor and CODA markers (see Fig. 10 for illustration). The video camera and CODA cameras capture the subject's motion at the same time.

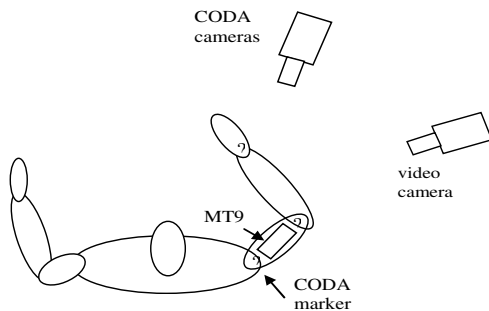


Fig. 9 Top view of the experimental set up

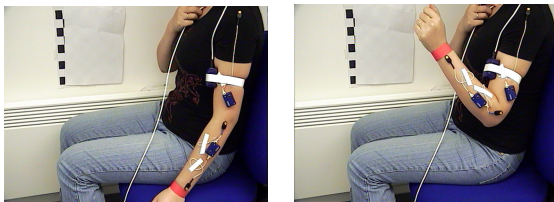


Fig. 10 Subject wearing MT9 and CODA markers

In this experiment, the video camera is used to track the colour belt attached at the wrist joint (see Fig. 10) instead of tracking the skin colour, in order to exclude CODA marker disturbing. For daily usage, it is only required to attach a MT9 on the upper arm and move the hand in front of camera, shown in Fig. 9.

5.2. Data Comparison

A number of repeated motion patterns were tested in order to do comprehensive comparisons. Fig. 11 shows a full arm motion sequence in about 20 seconds. Since the shoulder position is assumed to be fixed, only elbow and wrist positions represented in the shoulder coordinate system are illustrated. Three coordinates x , y , z of the position trajectories of elbow and wrist joints from each tracking system are plotted respectively in Fig.11. Red lines represent the data from the CODA system and blue lines from our hybrid system. The three figures in first row

shows the coordinates of the elbow joint, and the ones in the second row shows the coordinates of the wrist joint.

As we can see in Fig.11, 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 of the wrist joint is much noisier than the elbow joint, but the difference between two systems is still within 4~6cm, which is quite promising and has great potential for home-based rehabilitation.

There are mainly three reasons affecting the tracking accuracy of our hybrid tracking system.

- The first reason is the movement of the upper arm muscle, which may affect the MT9 output, thus affect the elbow position calculation.
- The second reason is that the wrist joint is calculated based on the elbow joint. So the error of the elbow joint is accumulated to the calculation of the wrist joint.
- The third reason is that our visual tracking is only performed in 2D image plane. Inferring 3D scene properties from 2D image measurements is an under-constrained task due to the lack of depth information. Image segmentation and camera calibration also introduce errors.

As can be seen from our experimental results, our proposed hybrid tracking method has a reasonable accuracy and efficiency since physiotherapists may not be able to accurately tell the difference of the subject motion using their eyes in a range of 5cm.

6. Conclusions and Future Work

This paper presents a new arm motion tracking system based on integration of both vision and inertial sensors. The system is able to track the non-rigid human arm motion in 3D and runs in real time. The human geometry structure and information is used to fuse different data modalities from two sensors. Unlike conventional visual tracking systems that suffer huge computational cost and poor accuracy, our proposed method is able to track the arm movement in real time and accurately.

Our future work will be focused on two aspects. One is to release the fixed shoulder position constraint by exploiting the acceleration output from the MT9 sensor. The other is to extend the method from tracking an arm movement to the whole upper body movement. The two main challenge issues are:

- how to solve the drift problem of MT9 acceleration output using image features, and
- how to build a proper kinematic upper body model for whole upper body tracking.

Acknowledgments

We would like to thank Charnwood Dynamics Ltd. to allow us to use their CODA motion tracking system. Our thanks also go to Dr Nigel Harris at Bath University and the other members of EPSRC EQUAL Smart Rehabilitation Consortium for useful discussion.

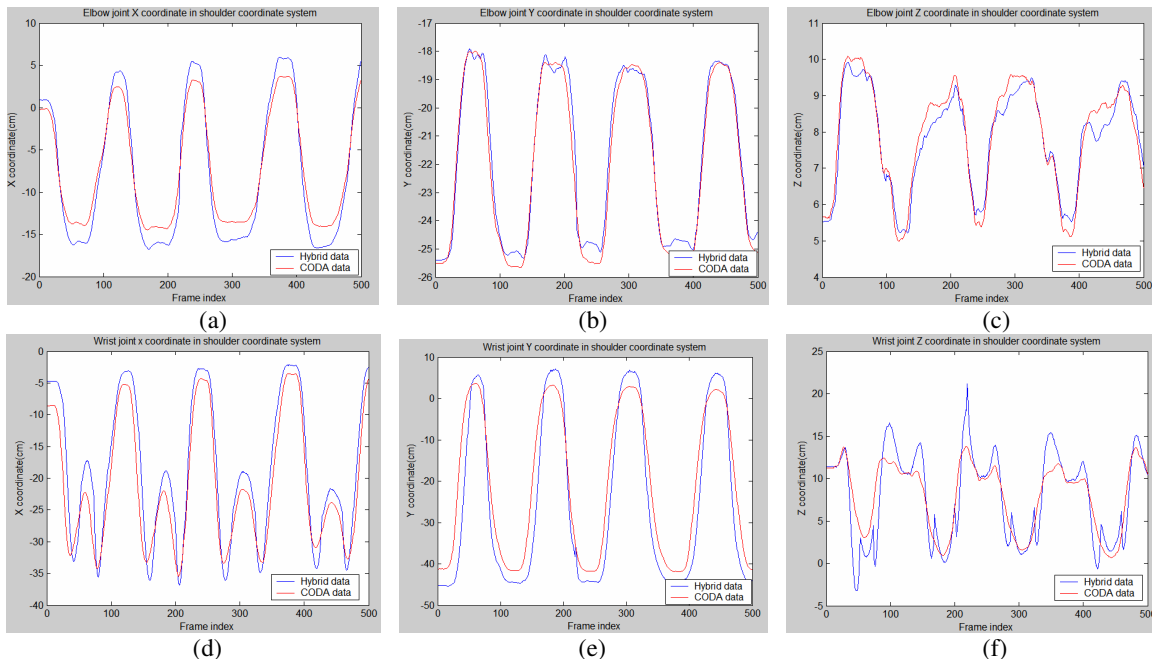


Fig. 11 Comparisons of the performance of a CODA system and our hybrid system for arm motion tracking. The first row, (a), (b), (c), shows the motion of the elbow joint and the second row, (d), (e), (f), shows the motion of the wrist joints.

References

- [1] D. M. Gavrila, $\text{\textcircled{r}}$ The Visual Analysis of Human Movement: A Survey; *Journal of Computer Vision and Image Understanding*, Vol.73, No1, pages 82-98, 1999.
- [2] J. K. Aggarwal, and Q. Cai, $\text{\textcircled{r}}$ Human Motion Analysis: A Review; *Journal of Computer Vision and Image Understanding*, 1999.
- [3] T. Moeslund, and E. Granum, "A Survey of Computer Vision-Based Human Motion Capture", *Computer Vision and Image Understanding* (81), No 3, pages 231-268, 2001.
- [4] L. Wang, W. Hu and T. Tan, $\text{\textcircled{r}}$ Recent Developments in Human Motion Analysis; *PR(36)*, No. 3, March 2003, pp. 585-601
<http://www.chamdyn.com/>
- [5] <http://www.qualisys.com/>
- [6] C. Sminchisescu, $\text{\textcircled{r}}$ Estimation Algorithms for Ambiguous Visual Models 3D Human Modeling & Motion Reconstruction in Monocular Video Sequences. PhD Thesis, Institute National Polytechnique de Grenoble (INRIA), July 2002.
- [7] Z. Chen, and H. J. Lee, $\text{\textcircled{r}}$ Knowledge-guided Visual Perception of 3D Human Gait from a Single Image Sequence; *IEEE Transactions on Systems, Man, and Cybernetics*, 22(2):336-342, 1992.
- [8] S. X. Ju, M. Black, and Y. Yacoob, $\text{\textcircled{r}}$ Cardboard people: A Parameterised Model of Articulated Motion; *2nd Int. Conf. on Automatic Face- and Gesture-Recognition*, Killington, Vermont, pages 38-44, Oct 1996.
- [9] J. Deutscher, A. Blake, and I. Reid, $\text{\textcircled{r}}$ Articulated Body Motion Capture by Annealed Particle Filtering; *IEEE International Conference on Computer Vision and Pattern Recognition*, Vol. 2, pages 126-133, 1996.
- [10] H. Sidenbladh, M. Black, and D. Fleet, $\text{\textcircled{r}}$ Stochastic Tracking of 3D Human Figures Using 2D Image Motion; *European Conference on Computer Vision*, 2000.
- [11] Y. Bar-Shalom and T.E. Fortmann, *Tracking and Data Association*. Academic Press, 1988.
- [12] M. Isard and A. Blake, *Contour Tracking by Stochastic Propagation of Conditional Density*; *Proceedings of European Conference on Computer Vision*, Vol. 1, pp. 343-356, Cambridge UK, (1996).
- [13] H. Zhou and H. Hu, $\text{\textcircled{r}}$ A Survey-Human Movement Tracking and Stroke Rehabilitation; Technical Report, CSM-420, ISSN1744-8050, Department of Computer Science, University of Essex, 2004.
- [14] E. Foxlin, Y. Altshuler, L. Naimark and M. Harrington, *FlightTracker: A Novel Optical/Inertial Tracker for Cockpit Enhanced Vision*; *IEEE ACM Int Symposium on Mixed and Augmented Reality*, Washington, D.C., 2-5 November 2004.
- [15] S. You, U. Neumann, and R. Azuma, *Hybrid Inertial and Vision Tracking for Augmented Reality Registration*. *Proceedings of IEEE VR '99* (Houston, TX, 13-17 March 1999), 260-267.
- [16] P. Lang, M. Ribo, and A. Pinz, *A New Combination of Vision-based and Inertial Tracking for Fully Mobile, Wearable and Real-time Operation*. In Proc. of 26th Workshop of the Austrian Association for Pattern Recognition (GM/AAPR), volume 160, pages 141-148, Graz, Austria, September 2002.
- [17] <http://www.xsens.com/>.
- [18] G. R. Bradski, $\text{\textcircled{r}}$ Computer Vision Face Tracking For Use in a Perceptual User Interface; *Intelligent Technology Journal*, 998(2).
- [19] V. Vezhnevets, V. Sazonov, and A. Andreeva, "A Survey on Pixel-Based Skin Colour Detection Techniques". *Graphicon-2003*, Moscow, Russia, September 2003.
- [20] Y. Tao and H. Hu, *Colour-based Human Motion Tracking for Home-based Rehabilitation*, *Proceedings of IEEE International Conference on Systems, Man and Cybernetics*, The Hague, The Netherlands, October 10-13 2004, pages 773-781.
- [21] L. Goncalves, E. D. Bernardo, E. Ursella, and P. Perona, $\text{\textcircled{r}}$ Monocular Tracking of the Human Arm in 3D; *ICCV*, 1995.
- [22] T. Moeslund and E. Granum, "Multiple Cues used in Model-Based Human Motion Capture", *The 4th Int. Conference on Automatic Face and Gesture Recognition*, Grenoble, France, March 2000
- [23] G. Johansson, *Visual Perception of Biological Motion and a Model for its Analysis*, *Perception and Psychophysics*, 14(2):210-211, 1973.
- [24] D. Tolani, A. Goswami, N. I. Badler, *Real-time Inverse Kinematics Techniques for Anthropomorphic Limbs*. *Graphical models*, 62(5):353-388, 2000.
- [25] H. Zhou, Y. Tao and H. Hu, "Upper limb motion estimation from inertial measurements in stroke rehabilitation", unpublished.