Human Posture Sequence Estimation Using Two Un-calibrated Cameras

Ruixuan Wang and Wee Kheng Leow
School of Computing, National University of Singapore
3 Science Drive 2, Singapore 117543, Singapore
{wangruix, leowwk}@comp.nus.edu.sg

Abstract

3D Human posture sequence estimation from single or multiple image sequences is essential in many applications, such as vision-based sport coaching and physical rehabilitation. However, 3D posture sequence cannot be accurately estimated from single image sequence due to depth ambiguity and self-occlusion, and pre-calibration is often required when estimating 3D posture sequence from multiple image sequences. In this paper, we present an algorithm to accurately estimate 3D human posture sequence from two un-calibrated image sequences by combining a modified Nonparametric Belief Propagation (mNBP) method with an improved camera self-calibration method. The mNBP estimates posture even when there is partial self-occlusion and when the human model scale is different from that of body image in image sequences. The improved self-calibration guarantees to find the optimal rotation and relative scale between two fixed but un-calibrated scaled orthographic cameras, without a nonlinear optimization process. Quantitative and qualitative experiment results show that the algorithm is able to estimate 3D posture sequence from a pair of un-calibrated image sequences.

1 Introduction

3D Human posture sequence estimation from one or more image sequences is essential in many human motion analysis applications, such as vision-based sport coaching and physical rehabilitation. In this paper, we present an algorithm to accurately estimate 3D human posture sequence from two un-calibrated image sequences by combining a modified Nonparametric Belief Propagation (mNBP) with a camera self-calibration.

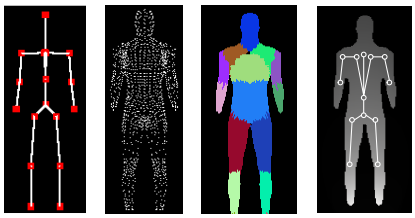
There has been much work on posture sequence estimation in articulated human body tracking [3, 1, 2, 9, 10, 16], which sequentially estimates 3D or 2D human posture sequence from monocular or multiple image sequences. From monocular image sequence [1, 9, 10, 16], 3D postures cannot be accurately estimated due to depth ambiguity and self-occlusion, and even 2D postures are not easy to be estimated due to self-occlusion and body rotation in depth. From multiple image sequences [3, 2], 3D postures may be estimated quite accurately based on the pre-calibration of multiple cameras. However in practice, the number of cameras may be limited to two or three, and the camera information is often unknown beforehand [6]. As a result, self-calibration is necessary to alleviate the issues in estimating body posture from a limited number of (i.e. two here) cameras.

Several work has been done on camera self-calibration from two human motion image sequences [6, 15]. However, they assume that 2D posture in each image is known in advance. In fact, it is not trivial and even difficult to get 2D posture from each image. In this paper, we introduce a mNBP method to automatically estimate 2D posture from the images, and at the same time, we develop an improved self-calibration method in which the optimal camera information can be directly found, without requiring a nonlinear optimization process which is the case in [6, 15].

In the following, based on a graphical model (Section 2), the mNBP algorithm is briefly introduced (Section 3) to estimate 2D (or 3D) posture from single (or multiple) image. Then one self-calibration method is improved (Section 4). By iterating the mNBP and the self-calibration two or three times, accurate 3D posture sequence can be estimated from a pair of un-calibrated image sequences, which has been shown by the experiments (Section 6).

2 Articulated Human Body Model

A human skeleton model (Figure 1(a)) is used to represent body joints and bones, and a triangular mesh model (Figure 1(b)) is used to represent the body shape. Each vertex in the mesh is attached to the related body part (Figure 1(c)). For each body part's shape, two parameters (length and width) are used to represent the size.



(a) (b) (c) (d)

Figure 1: Human body model.

Human body posture $\mathscr{X}$ is represented by a set of body parts' poses, $\mathscr{X} = \{\mathbf{x}_i | i \in \mathscr{V}\}$, where $\mathscr{V}$ is the set of body parts. Body part pose $\mathbf{x}_i = (\mathbf{p}_i, \theta_i)$ represent the i^{th} body part's 3D position $\mathbf{p}_i$ and 3D orientation θ_i. Given the shape and size of human body, any body posture $\mathscr{X}$ can be rendered and projected to generate a synthetic image observation. During posture estimation, each synthetic observation will be used to compare with a real image observation $\mathscr{Z} = \{\mathbf{z}_i | i \in \mathscr{V}\}$, where $\mathbf{z}_i$ represents the real image observation for the i^{th} body part. The relationship between $\mathbf{x}_i$ and $\mathbf{z}_i$ is represented by the observation function $\phi_i(\mathbf{x}_i, \mathbf{z}_i)$. In addition due to the articulation, every pair of adjacent body parts $\mathbf{x}_i$ and $\mathbf{x}_j$ must be connected. This constraint is enforced by the potential function $\psi_{ij}(\mathbf{x}_i, \mathbf{x}_j)$.

A tree-structured graphical model (Figure 1(d)) is used to represent the articulated human body model. The tree consists of a set of nodes $\mathscr{V}$ and a set of edges $\mathscr{E}$. Each node $i \in \mathscr{V}$ is associated with $\mathbf{x}_i$ and $\mathbf{z}_i$ of the i^{th} body part, and each edge $(i, j) \in \mathscr{E}$ is associated with the potential function $\psi_{ij}(\mathbf{x}_i, \mathbf{x}_j)$.

3 Posture Estimation by mNBP

A modified NBP (mNBP) method is introduced that can cope with partial self-occlusion and different body image sizes in estimating 2D (or 3D) human posture from single (or multiple) image.

NBP [11, 12, 4] can be used to estimate each body part's pose. However, it assumes that observation of each part can be obtained independently [11, 12]. This limits it to cases where there is no self-occlusion. We modify NBP to handle occlusion by changing the joint probability of body posture $\mathscr{X}$ and image observation $\mathscr{Z}$ to Equation (1). Similar to NBP [11], we may calculate marginal distributions by Equations (2) and (3),

$$p(\mathscr{X},\mathscr{Z}) = \alpha_1 \prod_{(i,j)\in\mathscr{E}} \psi_{ij}(\mathbf{x}_i,\mathbf{x}_j) \prod_{i\in\mathscr{V}} \phi_i(\mathscr{X},\mathbf{z}_i) \quad (1)$$

$$m_{ij}^n(\mathbf{x}_j) \propto \alpha_2 \int_{\mathbf{x}_i} \psi_{ij}(\mathbf{x}_i,\mathbf{x}_j)\phi_i(\mathbf{x}_i,\tilde{\mathscr{X}}_{-i}^{n-1},\mathbf{z}_i) \prod_{k\in\Gamma(i)\setminus j} m_{ki}^{n-1}(\mathbf{x}_i)d\mathbf{x}_i \quad (2)$$

$$\hat{p}^n(\mathbf{x}_j|\mathscr{Z}) \propto \alpha_3\phi_j(\mathbf{x}_j,\tilde{\mathscr{X}}_{-j}^{n-1},\mathbf{z}_j) \prod_{i\in\Gamma(j)} m_{ij}^n(\mathbf{x}_j) \quad (3)$$

where $m_{ij}^n(\mathbf{x}_j)$ is the message propagated from node i to j in iteration n. $\Gamma(i)=\{k|(i,k)\in\mathscr{E}\}$ is the neighbor of node i, and $\Gamma(i)\setminus j$ is the neighbor of i except j. $\tilde{\mathscr{X}}_{-i}^{n-1}$ is the set of body parts' pose estimations except the i^{th} body part from the $(n-1)^{th}$ iteration.

When body part i is partially occluded by some others, its image observation $\mathbf{z}_i$ is generated by both this part and the others. Together with the other body parts' estimations $\tilde{\mathscr{X}}_{-i}^{n-1}$ coming from previous iteration, each estimate of $\mathbf{x}_i$ can generate corresponding observations to measure the observation functions.

As another limitation, NBP assumes that the size ratio between 3D human model and each body image is known such that each body posture can be rendered and compared with the real image in the same scale. However, the body image size may often change overtime due to the body translation in depth. The mNBP can cope with such case by updating human model scales when estimating each $\mathbf{x}_i$. For each possible model scale, $\mathbf{x}_i$ and $\tilde{\mathscr{X}}_{-i}^{n-1}$ are used to generate a corresponding observation to measure the observation functions. Based on the measurement, the scale can be updated in each iteration.

Compared to these existing NBP algorithms [12, 4], the mNBP can cope with self-occlusion and the change of body image size overtime. Furthermore, the mNBP embeds the simulated annealing idea into the algorithm by using a decreasing factor λ to modify potential functions in each NBP iteration. For more detail about how to design potential function $\psi_{ij}(\mathbf{x}_i,\mathbf{x}_j)$ and observation function $\phi_i(\mathbf{x}_i,\tilde{\mathscr{X}}_{-i}^{n-1},\mathbf{z}_i)$, and how to implement the modified NBP, please refer to [13].

4 Self-calibration of Two Cameras

Based on the estimated 2D posture in each image, a self-calibration method is improved to reconstruct 3D body postures and camera relative rotation and scale from two uncalibrated image sequences by using kinematic constraints of human body [6, 15]. In the following, we introduce our method by assuming that the two cameras are scaled orthographic and the two camera scales are fixed throughout the image sequences. Then the method is easily extended to the case of changing scales of scaled orthographic cameras.

4.1 Related Methods

Our improved self-calibration method is based on the ideas in [6, 15]. Suppose $\mathbf{p}_{ij}$ ($i=1,...,F$; $j=1,...,N$) is the unknown 3D position of the j^{th} body joint in the i^{th} frame, $\mathbf{y}_{ij1}$

and $\mathbf{y}_{ij2}$ are the two correspondingly estimated 2D body points in image sequence 1 and 2 respectively. Then each sequence of estimated 2D body joints can be viewed as the scaled orthographic projection of FN body joints in a static scene [6]. Assuming that $\mathbf{p}_{ij}$ and $\mathbf{y}_{ij}$ are centralized, Equation (4) can be obtained by SVD [14]

$$\begin{aligned}\mathbf{W} &= \begin{bmatrix} \mathbf{y}_{111} & \cdots & \mathbf{y}_{1N1} & \mathbf{y}_{211} & \cdots & \cdots & \mathbf{y}_{FN1} \\ \mathbf{y}_{112} & \cdots & \mathbf{y}_{1N2} & \mathbf{y}_{212} & \cdots & \cdots & \mathbf{y}_{FN2} \end{bmatrix} \\ &= \begin{bmatrix} \hat{\mathbf{R}}_1 \\ \hat{\mathbf{R}}_2 \end{bmatrix} \begin{bmatrix} \hat{\mathbf{p}}_{11} & \cdots & \hat{\mathbf{p}}_{1N} & \hat{\mathbf{p}}_{21} & \cdots & \cdots & \hat{\mathbf{p}}_{FN} \end{bmatrix} \end{aligned} \tag{4}$$

where $\hat{\mathbf{R}}_1$ and $\hat{\mathbf{R}}_2$ are the reconstructed two camera projection matrices, and $\hat{\mathbf{p}}_{ij}$'s are the reconstructed 3D body joints. All the reconstructions are up to an unknown affine transformation $\mathbf{A}$, i.e.,

$$\begin{bmatrix} s_1\mathbf{R}_1 \\ s_2\mathbf{R}_2 \end{bmatrix} \cdot \mathbf{p}_{ij} = \begin{bmatrix} \hat{\mathbf{R}}_1 \\ \hat{\mathbf{R}}_2 \end{bmatrix} \mathbf{A}^{-1} \cdot \mathbf{A}\hat{\mathbf{p}}_{ij} \tag{5}$$

where $\mathbf{A}$ is a full rank 3×3 matrix, s_1 and s_2 are the unknown scales of the two cameras, and $\mathbf{R}_1 = (\mathbf{R}_{11}\ \mathbf{R}_{12})^T$ and $\mathbf{R}_2 = (\mathbf{R}_{21}\ \mathbf{R}_{22})^T$ are the unknown true projection matrices whose rows (e.g., $\mathbf{R}_{11}^T$) are unit vectors. To reconstruct the structure (i.e., $\mathbf{p}_{ij}$) and camera motion (i.e., $s_1\mathbf{R}_1$ and $s_2\mathbf{R}_2$), a reasonable affine transformation $\mathbf{A}$ has to be found.

By QR decomposition, $\mathbf{A} = \mathbf{SU}$, where $\mathbf{S}$ is an orthonormal rotation matrix and $\mathbf{U}$ is an upper triangular matrix. In general, $\mathbf{S}$ can be ignored and $\mathbf{U}$ may only be obtained up to a scale [15].

In order to find $\mathbf{U}$ that has six unknown (but five independent) parameters, two camera views are not enough when the cameras are scaled orthographic [7]. Assuming that the two cameras have zero skew and unit aspect ratio, only four camera constraints can be obtained. Therefore, other constraints have to be used. In the method of Liebowitz and Carlson [6, 15], two kinds of rigid link constraints are used: (1) every two body parts have a constant length ratio, and (2) each body part has a constant length over time.

Let A_1 and A_2 be the two end joints of body part $\mathscr{L}_A$, B_1 and B_2 be the ends of body part $\mathscr{L}_B$. Let $\mathbf{X}_{iA} = \hat{\mathbf{p}}_{iA_1} - \hat{\mathbf{p}}_{iA_2}$ and $\mathbf{X}_{iB} = \hat{\mathbf{p}}_{iB_1} - \hat{\mathbf{p}}_{iB_2}$ denote the two estimated body parts at i^{th} frame. Given the length ratio r_{AB} of body part $\mathscr{L}_A$ to $\mathscr{L}_B$, from the first kind of rigid link constraint, there is [6] $\mathbf{X}_{iA}^T\Omega\mathbf{X}_{iA} = r_{AB}^2\mathbf{X}_{iB}^T\Omega\mathbf{X}_{iB}$, where $\Omega = \mathbf{U}^T\mathbf{U}$. Similarly, from the second kind of rigid link constraint, there is $\mathbf{X}_{i,A}^T\Omega\mathbf{X}_{i,A} = \mathbf{X}_{i+1,A}^T\Omega\mathbf{X}_{i+1,A}$. The two constraint equations are actually linear in term of the six (but five independent) unknown parameters in the symmetrical matrix Ω. By combining the equations coming from all the frames, a linear solution of Ω can be obtained by solving over-constraint linear equations [6]. Unfortunately, such solution of Ω is seldom positive definite due to the noise in estimated 2D joints [15]. Since $\mathbf{U}$ can be recovered (up to reflection transformation) by Cholesky factorization of Ω if and only if Ω is positive definite, the method of Liebowitz and Carlson often gets into trouble in practice.

Furthermore, since the scaled orthographic camera constraints are linearly related not to Ω but to Ω^{-1}, the camera constraints cannot easily be combined with the rigid link constraints. As a result, Ω has to be solved numerically by a nonlinear optimization process in [6]. By contrast, the method of Tresadern and Reid [15] can first eliminate four degrees of freedom in Ω^{-1} by considering the camera constraints, and then numerically find the other two unknown parameters of Ω^{-1} by a nonlinear optimization process.

4.2 The Improved Method

Compared with the above two methods, we observe that one of the two camera scales can be absorbed into $\mathbf{U}^{-1}$. As a result, $\mathbf{U}^{-1}$ and corresponding Ω^{-1} will have six independently unknown parameters, and at the same time camera constraints can eliminate five of the six DOFs of Ω^{-1}. We show that the remained single unknown parameter can easily be embedded into the rigid link constraints which are related to Ω, and the single unknown parameter can be directly obtained by solving an equation of one-variable six-order polynomial. Our method can guarantee to find the optimal solution $\mathbf{U}^{-1}$ (or $\mathbf{U}$) in the sense of mean squared error, without any nonlinear optimization process like others [6, 15].

Let $s = s_2/s_1$ be the camera relative scale, Equation (5) can be transformed to (6)

$$\begin{bmatrix} \mathbf{R}_1 \\ s\mathbf{R}_2 \end{bmatrix} \cdot s_1\mathbf{p}_{ij} = \begin{bmatrix} \hat{\mathbf{R}}_1 \\ \hat{\mathbf{R}}_2 \end{bmatrix} \mathbf{U}^{-1} \cdot \mathbf{U}\hat{\mathbf{p}}_{ij} \tag{6}$$

where $\mathbf{U}$ has six DOFs. From the camera constraints, there is

$$\begin{aligned} \mathbf{R}_{11}^T\Omega^{-1}\mathbf{R}_{11} &= 1 && (7)\\ \mathbf{R}_{12}^T\Omega^{-1}\mathbf{R}_{12} &= 1 && (8)\\ \mathbf{R}_{11}^T\Omega^{-1}\mathbf{R}_{12} &= 0 && (9)\\ \mathbf{R}_{21}^T\Omega^{-1}\mathbf{R}_{21} - \mathbf{R}_{22}^T\Omega^{-1}\mathbf{R}_{22} &= 0 && (10)\\ \mathbf{R}_{21}^T\Omega^{-1}\mathbf{R}_{22} &= 0 && (11) \end{aligned}$$

The five equations are actually linear equations in terms of the six independent parameters of Ω^{-1}. Denote the particular and the homogeneous solution of the under-constrained linear equations by Ω_0^{-1} and Ω_1^{-1} respectively, all the possible solutions of Ω^{-1} that satisfy the camera constraints can then be represented by $\Omega^{-1}(\beta) = \Omega_0^{-1} + \beta\Omega_1^{-1}$. Since a real matrix $\Omega^{-1}(\beta)$ is positive definite if and only if the determinants of all its top left corner submatrix are positive, and the determinants are simple functions of β, the valid range $(\beta_{min}, \beta_{max})$ of β that satisfy the positive definite property of $\Omega^{-1}(\beta)$ can be obtained.

Noticing that $\Omega^{-1}(\beta)$ is a 3×3 matrix with parameter β, we can get $\Omega(\beta)$,

$$\Omega(\beta) = \frac{1}{f^{[3]}(\beta)} \begin{bmatrix} f_{11}^{[2]}(\beta) & f_{12}^{[2]}(\beta) & f_{13}^{[2]}(\beta) \\ f_{21}^{[2]}(\beta) & f_{22}^{[2]}(\beta) & f_{23}^{[2]}(\beta) \\ f_{31}^{[2]}(\beta) & f_{32}^{[2]}(\beta) & f_{33}^{[2]}(\beta) \end{bmatrix}$$

where $f^{[i]}(\beta)$ is an i^{th} order polynomial in terms of β.

Then from the rigid link constraints, over-constraint equations can be obtained:

$$\mathbf{E}(\beta) = \frac{1}{f^{[3]}(\beta)}\mathbf{C}\mathscr{B} = \mathbf{0} \tag{12}$$

where $\mathbf{C}$ is the constraint matrix which combine all the rigid link constraints in all frames, and $\mathscr{B} = (\beta^2 \quad \beta \quad 1)^T$. In the presence of noise in estimated 2D joints, $\mathbf{E}(\beta)$ will not be $\mathbf{0}$ and we can use $F(\beta)$ to evaluate the goodness of β,

$$F(\beta) = \mathbf{E}^T(\beta)\mathbf{E}(\beta) = f^{[4]}(\beta)/\{f^{[3]}(\beta)\}^2 \tag{13}$$

where the best β corresponds to one minimum of $F(\beta)$. All the possible minima can be found by solving the first derivative equation of $F(\beta)$,

$$F'(\beta) = f^{[6]}(\beta)/\{f^{[3]}(\beta)\}^3 = 0 \tag{14}$$

i.e., $f^{[6]}(\beta) = 0$. The roots of the sixth-order polynomial can be easily obtained, and the best β solution is the one on which $F(\beta)$ is global minimum in the valid range $(\beta_{min}, \beta_{max})$. Note that such solution must exist because the non-negative $F(\beta)$ increase to infinity at β_{min} and β_{max}.

Given the best β and then the $\Omega(\beta)$, the affine matrix $\mathbf{U}$ can be got from the Cholesky factorization of $\Omega(\beta)$ up to a reflection transformation, and then the camera relative motion and the metric 3D structure can be obtained. Note that the reflection ambiguity can be eliminated by re-measuring the observation function using the reconstructed camera motion and the estimated postures by mNBP.

4.3 The Extension of the Improved Method

In practice, each camera scale can change over time due to large motion in depth. In this case, one affine transformation $\mathbf{U}$ has to be estimated for each frame pair in the two image sequences. Since each $\mathbf{U}$ is obtained in a single frame pair, the second rigid link constraints on consecutive frames cannot be used. Even so, the constraint on cameras and the first kind of rigid constraint are enough to reconstruct each $\mathbf{U}$.

When the two cameras are fixed, the reconstructed $\mathbf{U}$ for different frame pairs should be different only in a scale factor. Although independent reconstruction of $\mathbf{U}$'s cannot assure such constraint, this issue can be easily solved in a final bundle adjustment [15].

5 Iteration Process

By combining the above two sections, an iteration process is often required to estimate more accurate 3D human postures. In the first iteration, 2D joints in each image are estimated independently of the other image sequence, due to the unknown camera relative motion. Therefore, the estimated 2D joints may not be accurate enough especially when severe self-occlusion between body parts happens. The noise in the 2D joints will then make the reconstruction of camera relative motion and 3D body joints not be accurate. As a result, an iteration process is required to improve the estimation accuracy. In the next iteration, the reconstructed (approximate) camera relative motion can be used to help estimate both 2D and 3D positions of body joints, by combining the image information coming from the two camera views. The more accurate 2D joints can be used to reconstruct 3D camera relative motion and 3D joints more accurately, which can be shown by our test results.

6 Experimental Results

Quantitative and qualitative evaluation of our algorithm are performed with three tests. The first test evaluates the modified NBP's capability for estimating human posture when the scales of 3D human model and the body image in the input image are different. For

estimating posture in the case of self-occlusion and large initial posture, please refer to our previous work [13]. The second test evaluates the accuracy and robustness of the improved self-calibration algorithm. The third one evaluates the capability of 3D posture sequence estimation by combining the mNBP and the self-calibration algorithm.

To quantitatively evaluate our algorithm, we capture human motion using Gypsy motion capture system and extract a 3D posture sequence from the motion. Every posture is mapped to a 54-DOF human skeleton model with mesh model for skin, and rendered using OpenGL from two viewpoints to get the two input image sequences. To obtain initial posture for each input image, we add some uniform random noise to joint angles of the true posture. Note that the estimated posture of previous frame can also be used as the initial posture for the current frame image.

6.1 Posture Estimation under Different Human Model Scales

In the first test, the mNBP algorithm is used to estimate posture from a single image, therefore the depth information cannot be accurately estimated. As a result, 2D joint position error $E_{2D} = \frac{1}{nh}\sum_{i=1}^{n} \|\hat{\mathbf{y}}_{2i} - \mathbf{y}_{2i}\|$ is computed to assess the algorithm performance, where $\hat{\mathbf{y}}_{2i}$ and $\mathbf{y}_{2i}$ are the estimated and the true 2D image position of the i^{th} body joint respectively. h is the articulated body height and it is about 195 pixels in the test.

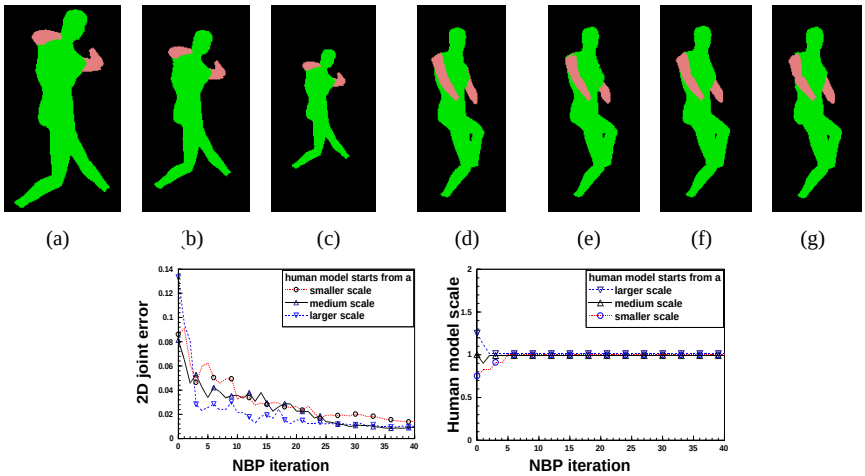


Figure 2: Test 1. (a), (b) and (c) are the images of an initial posture corresponding to three different human model scales. (d) is the input image. (e), (f) and (g) are images of the estimated postures respectively. (h) is the error E_{2D}, starting from different human model scales. (i) is the scale changing of human model.

Figures 2 (a)-(g) illustrate one example to estimate body posture from different human model scales. We can see that the estimated (projected 2D) posture is very close to the true posture, whatever the model scale is. For this example, Figure 2 (h) shows that, after 15 iterations or so, the error E_{2D} has decreased to a relatively small value. Figure 2 (i) tells us that the human body model can be modified to the approximate body size in the input image in several (i.e. 6 or so) iterations.

6.2 Reconstruction of Camera Motion and Body Posture

In the second test, a 3D posture sequence of 30 frames are used. We use the orthographic projection of the 3D posture sequence from two fixed viewpoints as the pair of input 2D posture sequences. For each 2D posture pair, uniform random noise of each 2D joint position is increased from 0 to 9% of the body height in both image row and column directions. The two fixed camera relative rotation angles are $(-10^o, -60^o, 20^o)$.

Three kinds of error measurements are used to assess the self-calibration performance: (1) 3D joint position error $E_{3D} = \frac{1}{nh}\sum_{t=1}^{l}\sum_{i=1}^{n}\|\check{\mathbf{p}}_{ti} - \mathbf{p}_{ti}\|$, where $\check{\mathbf{p}}_{ti}$ and $\mathbf{p}_{ti}$ are respectively the reconstructed and the true 3D positions of the i^{th} joint at frame t, h is the body height and it is 195pixels here; (2) camera relative scale error $E_s = \|\check{s} - s\|$, where $\check{s}$ and s are the reconstructed and the true camera relative scales, and s is 1 in this test; (3) camera relative rotation error $E_{rj} = \|\check{\theta}_j - \theta_j\|$, $j = 1,2,3$ where $\check{\theta}_j$ and θ_j are the reconstructed and the true rotation angles around the j^{th} axis direction.

In this test, two cases are tested: (1) self-calibration from a pair of 2D input sequences and (2) from a single pair of 2D input images. We use the same error measurements in the second case as in the first one, except that the errors are computed from each individual pair and then averaged over the 30 frames.

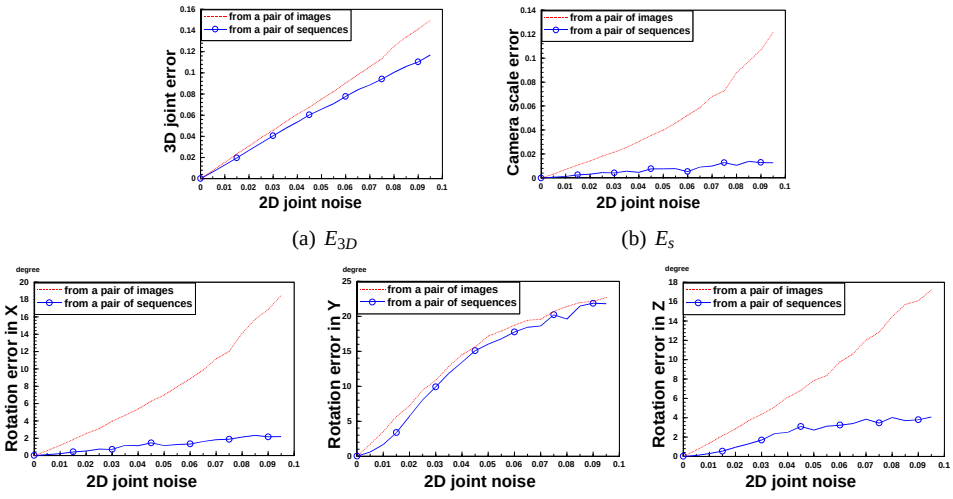
Figure 3(a) illustrates that the 3D joint error E_{3D} increases linearly when 2D joint position noise increases. Because each reconstructed 3D position is determined by the 2D joint positions, large noisy 2D joints obviously will result in large error in 3D joint reconstruction. Figure 3(b) tells us that the error E_s increases little when the camera scale is reconstructed from two sequences. However, when the scale comes from a single pair, the error E_s increase with respect to 2D joint noise. The result is reasonable because reconstruction from two sequences can capture more statistical information on camera scale compared to the reconstruction from a single pair. The same reason may explain the error E_{rj} around x-axis (Figure 3(c)) and z-axis (Figure 3(e)).

However, the error E_{rj} around y-axis (Figure 3(d)) from two sequences is not reduced much compared with that from a single pair. It is reasonable that the 2D joint noise will cause more reconstruction uncertainty in the y-axis direction around which there is a large rotation angle (i.e., -60^o here), which has been verified by our more experiments.

6.3 Combination of the mNBP and the Camera Self-calibration

In the third test, we show that the 3D posture sequence estimation can be accurately estimated from two un-calibrated cameras by iteratively using the modified NBP and the self-calibration algorithms. Almost all the time is spent in the mNBP where each mNBP iteration costs around 20 seconds. Here a pair of sequences of three images are used as the input.

Figure 4(a) illustrates the 3D joint reconstruction error with respect to the iteration. We can see that the error is still relatively large after the first iteration because each 2D posture sequence is estimated from a single image sequence. After the first iteration, the camera relative rotation and scale (Figure 4(b)(c)) are estimated by the self-calibration algorithm. The 3D joint error has been reduced largely in the second iteration because our modified NBP has been able to use two camera viewpoints' image information. Figure 4(b) and (c) tell us the camera relative rotation and scale can be estimated accurately enough to help improve the 2D posture estimation in the next iteration. Figure 4(d) and (e) respectively illustrate the true and the reconstructed 3D posture of one image pair in

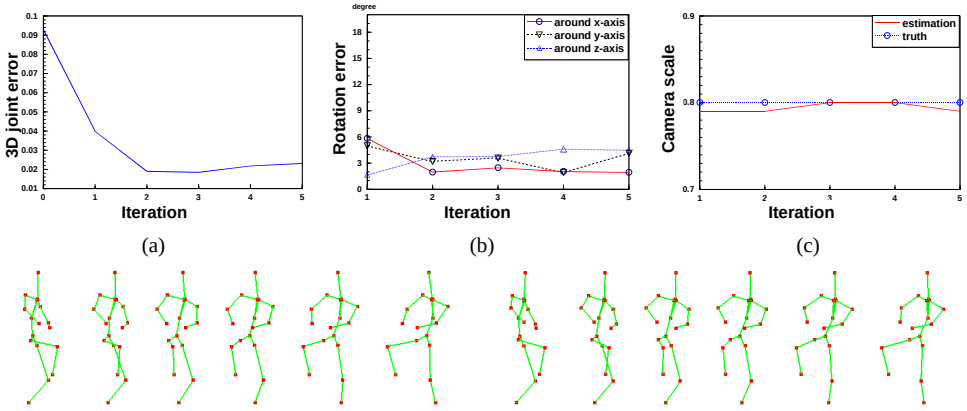


(a) E_{3D} (b) E_s

(c) E_{r1} (d) E_{r2} (e) E_{r3}

Figure 3: Test 2. The reconstruction errors.

the second iteration, from which we can see that a very similar 3D posture to the truth is obtained in the second iteration.



(a) (b) (c)

(d) true posture (e) reconstructed posture in the second iteration

Figure 4: Test 3. (a) 3D joint error. (b) Camera relative rotation error. (c) The estimated camera relative scale. (d) True posture, and (e) Reconstructed posture, viewed from six viewpoints respectively.

7 Conclusion and Future Work

This paper introduces a modified NBP algorithm and presents an efficient camera self-calibration algorithm. By combining the two algorithms, 3D posture sequence can be

estimated from a pair of image sequences captured by two un-calibrated but fixed cameras. Quantitative and qualitative evaluation on the algorithms show that (1) the modified NBP can estimate posture even if the human model scale is different from the body image size, (2) the self-calibration algorithm can efficiently find the rotation and relative scale between two scaled orthographic cameras by solving an equation of single-variable six-order polynomial, without requiring a nonlinear optimization process, and (3) accurate 3D posture sequence can be estimated by iterating the two algorithms quite a few times. In the future work, tests on real image sequences will be performed. Also, the computation cost of mNBP should be reduced in order to apply our algorithm to the long image sequences.

Acknowledgement

Thank Saurabh Garg, Sheng Zhang and Hanna Kurniawati for their valuable suggestions.

References

[1] T.J. Cham and J.M. Rehg, "A multiple hypothesis approach to figure tracking," *Proc. IEEE Int. Conf. Computer Vision and Pattern Recognition*, vol. 2, pp. 239-245, 1999.

[2] J. Deutscher, A. Blake and I. Reid, "Articulated body motion capture by annealed particle filtering," *Proc. IEEE Int. Conf. Computer Vision and Pattern Recognition*, vol. 2, pp. 126-133, 2000.

[3] D.M. Gavrila and L.S. Davis, "3-D model-based tracking of humans in action: a multi-view approach," *Proc. IEEE Int. Conf. Computer Vision and Pattern Recognition*, pp. 73-80, 1996.

[4] G. Hua and Y. Wu, "Multi-scale visual tracking by sequential belief propagation," *Proc. IEEE Int. Conf. Computer Vision and Pattern Recognition*, vol. 1, pp. 826-833, 2004.

[5] M. Isard, "PAMPAS: Real-valued graphical models for computer vision," *Proc. IEEE Int. Conf. Computer Vision and Pattern Recognition*, vol. 1, pp. 613-620, 2003.

[6] D. Liebowitz and S. Carlsson, "Uncalibrated motion capture exploiting articulated structure constraints," *Proc. IEEE Int. Conf. Computer Vision*, vol. 2, pp. 230-237, 2001.

[7] L. Quan, "Self-calibration of an affine camera from multiple views," *International Journal of Computer Vision*, 19(1):93-110, 1996.

[8] L. Sigal, S. Bhatia, S. Roth, M.J. Black and M. Isard, "Tracking loose-limbed people," *Proc. IEEE Int. Conf. Computer Vision and Pattern Recognition*, vol. 1, pp. 421-428, 2004.

[9] C. Sminchisescu and B. Triggs, "Covariance scaled sampling for monocular 3D body tracking," *Proc. IEEE Int. Conf. Computer Vision and Pattern Recognition*, vol. 1, pp. 447-454, 2001.

[10] C. Sminchisescu and B. Triggs, "Kinematic jump processes for monocular 3D human tracking," *Proc. IEEE Int. Conf. Computer Vision and Pattern Recognition*, vol. 1, pp. 69-76, 2003.

[11] E.B. Sudderth, A.T. Ihler, W.T. Freeman and A.S. Willsky, "Nonparametric belief propagation," *Proc. IEEE Int. Conf. Computer Vision and Pattern Recognition*, vol. 1, pp. 605-612, 2003.

[12] E.B. Sudderth, M.I. Mandel, W.T. Freeman and A.S. Willsky, "Visual hand tracking using nonparametric belief propagation," *Proc. IEEE CVPR Workshop on GMV*, 2004.

[13] R. Wang and W.K. Leow, "Human body posture refinement by nonparametric belief propagation," *Int. Conf. Image Processing*, 2005.

[14] C. Tomasi and T. Kanade, "Shape and motion from image streams under orthography: A factorization approach," *International Journal of Computer Vision*, 9(2):137-154, 1992.

[15] P. Tresadern and I. Reid, "Uncalibrated and Unsynchronized Human Motion Capture: A Stereo Factorization Approach," *Proc. IEEE Int. Conf. Computer Vision and Pattern Recognition*, vol. 1, pp. 128-134, 2004.

[16] Y. Wu, G. Hua and T. Yu, "Tracking articulated body by dynamic Markov network," *Proc. IEEE Int. Conf. Computer Vision*, vol. 2, pp. 1094-1101, 2003.