

Retrieval and Constraint-Based

Human Posture Reconstruction from a Single Image ¹

Chih-Yi Chiu ², Chun-Chih Wu ^{*}, Yao-Cyuan Wu ^{*}, Ming-Yang Wu ^{*},
Shih-Pin Chao ^{*}, and Shi-Nine Yang ^{*}

*Institute of Information Science
Academia Sinica
Nankang, Taipei, Taiwan 115*

** Department of Computer Science
National Tsing Hua University
101, Section 2 Kuang Fu Road
Hsinchu, Taiwan 300*

¹This study was supported partially by the MOE Program for Promoting Academic Excellence of Universities under the Grant No. 89-E-FA04-1-4 and the National Science Council, Taiwan under the Grant NSC92-2213-E-007-081.

²Corresponding Author: Chih-Yi Chiu.

Email: cychiu@iis.sinica.edu.tw TEL: +886-912-558735

Abstract

In this study, we present a novel model-based approach to reconstruct the 3D human posture from a single image. The approach is guided by a posture library and a set of constraints. Given a 2D human figure, i.e., a set of labeled body segments and estimated root orientation in the image, a 3D pivotal posture whose 2D projection is similar to the human figure is first retrieved from the posture library. To facilitate the retrieval process, a table-lookup technique is proposed to index postures according to their 2D projections with respect to designated view directions. Next physical and environmental constraints, including segment length ratios, joint angle limits, pivotal posture reference, and feet-floor contact, are automatically applied to reconstruct the 3D posture. Experimental results show the effectiveness of the proposed approach.

Keywords: posture retrieval and reconstruction, posture library, physical and environmental constraint.

List of Figures

- Fig. 1.** The reconstruction procedure of the proposed approach.
- Fig. 2.** The hierarchical human model.
- Fig. 3.** Eight sampling view directions $\{D_k | k = 1, 2, \dots, 8\}$ and their corresponding projections of a 3D library posture.
- Fig. 4.** Scaled orthographic projection for view direction D_k . Body segment \overrightarrow{ov} is projected on image plane I to obtain $\overrightarrow{o_I v_I}$. The angle α is measured counterclockwise from the positive U axis to $\overrightarrow{o_I v_I}$ on image plane I . We use angle α to represent body segment \overrightarrow{ov} with respect to view direction D_k .
- Fig. 5.** Indexing left upper arm of the 3D posture in the posture library.
- Fig. 6.** The ERO of the 2D human figure in the image.
- Fig. 7.** Range search in the posture table of the left upper arm.
- Fig. 8.** Posture reconstruction for the j -th body segment on reference plane R .
(a) Front view. (b) Top view.
- Fig. 9.** The feet-floor contact constraint.
(a) Reconstructed posture. (b) Set the floor and floor fulcrum. (c) Apply IK.
- Fig. 10.** Experimental results obtained by applying our reconstruction approach to some images.
- Fig. 11.** A sequence of 2D and 3D key postures of Tai Chi Chuan motion – “Grasp the Swallow’s Tail.”
- Fig. 12.** Experimental results for some test images.

List of Tables

Table 1. Segment length ratios of three subjects.

Table 2. Average RMSE of pivotal postures and reconstructed postures.

Table 3. Average RMSE of reconstructed postures based on the physical constraint
and environmental constraints.

1. Introduction

We seek to reconstruct 3D postures of a human actor from given 2D images. This issue has drawn a great attention due to its variety of applications, such as motion capture [1-2], user interface [3-4], character animation [5-6], etc. In these applications, the source image data can be a single image or single/multi-view video. In this paper, we confine ourselves to the single image case, which is also required for initialization in the video case.

Suppose that a 2D human figure, i.e., a set of labeled body segments and estimated root orientation in the image, is given by a user. To reconstruct the 3D posture of the 2D human figure, the main challenge is to determine depth information of the human figure elements. That is, since an image does not record 3D depth, each foreshortened body segment can be pointed either towards or away from the viewer with respect to the image plane. Consequently, the number of possible postures grows exponentially with the number of body segments. For example, if there are n body segments in the human figure, the number of possible 3D postures according to the given image is 2^n in general. To solve the depth ambiguity problem, several methods have been proposed. In general, there are two main approaches, namely, model-based and learning-based. A brief review of the two approaches is given in the following subsection. Moreover, posture reconstruction from a single image and motion recovery from single-view video are discussed in the following review.

1.1. Related Work

The model-based approach uses an articulated human model to generate possible

3D postures that match the 2D human figure. In order to obtain the best 3D solution, a set of physical, environmental, or dynamic constraints is then applied to cull invalid 3D postures generated initially. Lee and Chen [7] first extract the camera extrinsic parameters through geometric calibration and then generate a set of 3D postures for the given 2D human figure image. These 3D postures are verified by using joint angle limits, body segment lengths, collision detection, and heuristic motion rules to prune infeasible ones. Bregler and Malik [8] introduce the twists and product of exponential maps to model the kinematic relationship of an articulated human model. Based on this model, the 3D posture of the first video frame is acquired by minimizing the difference between the projected 3D posture and the given 2D human figure. Difranto *et al.* [9] propose a Scaled Prismatic Model (SPM) [10] to track 2D joint positions. They formulate a batch optimization function that involves a series of SPM measurements and constraints, including kinematic constraints, joint angle limits, dynamic smoothing, and 3D key frames. The optimization function is solved iteratively to recover 3D articulated motion. Taylor [11] presents an analysis to show that solutions for the 3D posture reconstruction problem can be parameterized by a scale factor under scaled orthographic projection. He further deduces the lower bound on the scale factor. Parameswaran and Chellappa [12] further extend Taylor's work by using the perspective projection model, and Loy *et al.* [13] apply Taylor's method to reconstruct long action sequences. Barron and Kakadiaris [14] estimate the anthropometry and 3D posture simultaneously for the given 2D human figure by minimizing a cost function subject to joint angle limits and segment length ratios. Park *et al.* [15] exploit 3D motion data given by users to recover motion from video. These motion data are expected to provide a good initial guess in the objective function for estimating joint orientations and the root trajectory.

Since reconstruction solely based on a single image is in general insufficient to

solve the depth ambiguity problem thoroughly, extra information is needed to obtain the desired 3D posture. Therefore, either particular motion types such as unidirectional walking are presented to reduce the reconstruction complexity [7-8, 16-18], or some extra visual cues about the 2D human figure are provided by users. For example in Difrancio's method [9], users are asked to set several keyframes of the video sequence and guess initial 3D coordinates of body joints with respect to these keyframes. In Taylor's method [11], users have to specify, for each body segment, the joint that is nearer to the viewer. In Barron's method [14], users must locate those segments being nearly parallel to the image plane for anthropometry estimation. In Park's method [15], users first prepare appropriate 3D motion data for the given video clip and then mark corresponding keyframes between the video clip and motion data for motion synchronization. All these methods require complicate human perceptions and interactions to provide extra visual cues. Some studies [19-20] propose fully automatic methods to locate body segments in an image. However, the accuracy of their methods is still far from user expectation.

For learning-based approaches, they try to derive mapping functions between features in the 2D image and that in the 3D posture through stochastic learning processes. It requires a large set of training data to learn prior knowledge of specific postures and motion. Pavlović *et al.* [21] describe a switching linear dynamic system (SLDS) to learn figure dynamics of fronto-parallel motion from video. A novel Viterbi approximation algorithm for inference in the SLDS is derived to overcome exponential complexity of motion classification, tracking, and synthesis. Brand [22] and Elgammal and Lee [23] use dynamic manifolds to model high-dimension human motion. Given a 2D silhouette in a video sequence, 3D motion and orientation are inferred through the dynamic manifolds. Howe *et al.* [24] divide motion data into short motion elements called snippets that are used to build a probability density

function. To reconstruct 3D motion, they divide the 2D tracking data into snippets and then find the best 3D snippet for each 2D observation using maximum-a-posteriori estimation. Tomasi and Kanade [34] propose a factorization technique that decomposes rigid shapes in image sequences to generate basis shapes. Then given 2D tracking data, these basis shapes can recover corresponding 3D information. Bregler *et al.* [35] further extend Tomasi's work on non-rigid shapes. Rosales and Sclaroff [25] design the Specialized Mappings Architecture (SMA) that maps 2D image features onto the 3D body posture parameters. Mapping functions in SMA are learned through the EM algorithm. Agarwal and Triggs [26] apply Relevance Vector Machine (RVM), which regresses 55D vectors of 3D body joint angles from 100D vectors of the human image silhouette, to learn 2D-3D mapping functions. Grochow *et al.* [36] present a novel model called a Scaled Gaussian Process Latent Variable Model (SGPLVM) to learn the probability density function of motion capture postures. The SGPLVM model can be learned automatically from a small training data set, and it works well in real-time animation applications. These above-mentioned methods only search databases to find the postures that are most similar to the given 2D images. No extra mechanism is provided to tune the found postures. Besides, the learning-based approach spends much time to learn 2D-3D mapping functions from large amount of training data. When the training data is modified, these mapping functions have to be recomputed again.

To conclude, 3D posture reconstruction from a single image is ill-posed due to insufficient spatial information. Using domain constraints or knowledge can moderate the underconstrained depth ambiguity problem. Both model-based and learning-based approaches do have their own merits and can provide feasible solutions under particular considerations. By taking the guiding data set in the learning-based approach and a priori knowledge of human model and constraints in the model-based

approach, we propose a novel algorithm for the reconstruction problem.

1.2. Our Approach

In this section, we present a novel approach to reconstruct the human posture from a single image. To overcome the depth ambiguity problem, we exploit a posture library and constraints to guide the reconstruction work. Suppose that a 2D human figure, i.e., a set of labeled body segments and an estimated root orientation in that image, is given. The proposed approach will first retrieve from the library an appropriate candidate whose 2D projection is similar to the human figure in the image. Since the candidate solution is from a large volume of the posture library, the effectiveness of the approach highly depends on the efficiency of the retrieval process. Therefore, we propose a table-lookup technique to index 3D human postures in the library. Each of library postures is projected onto several sampling view directions and the corresponding projection features are extracted. These features are stored in corresponding array elements for future retrieval. Next physical and environmental constraints, including segment length ratios, joint angle limits, pivotal posture reference, and feet-floor contact, are automatically applied step by step to reconstruct the 3D human posture for the given 2D human figure. Fig. 1 shows the reconstruction procedure of the proposed approach, where the word “ERO” beneath the image is the abbreviation of “Estimated Root Orientation.”

Our approach effectively integrates the techniques of model-based approach and the postures of guiding data set used in learning-based approach. Compared with the requirement of providing extra visual cues in existing model-based methods, our approach only asks users to label body segments on the image (the same requirement in existed model-based methods), and no further complicated indication required. The

posture library is exploited to deal with the depth ambiguity problem automatically. Compared with the learning-based approaches, our approach can further refine the retrieved posture automatically according to given constraints rather than outputting the retrieved posture only. Besides, a table-lookup index mechanism is proposed to speed up the retrieval process. This index mechanism does not need to spend time for data training.

Note that the posture library is assumed to contain data that are similar to the posture implied by the given image. This assumption is reasonable for most corpus-based applications. In other words, we assume that users have an appropriate posture library that records the same motion type implied by the given image. For example, if users want to reconstruct some postures of Tai Chi Chuan from images, they will use the posture library containing posture data of Tai Chi Chuan.

This paper is organized as follows. Section 2 presents preprocessing for the posture library, including posture feature representation and posture table creation. Section 3 describes the posture reconstruction process, including pivotal posture retrieval and constraint-based reconstruction. Section 4 shows our experimental results. Section 5 gives some conclusions and future work.

2. Posture Library Preprocessing

The objective of this section is to build an index structure for effectively retrieving pivotal posture from the posture library. It consists of two parts, namely, posture feature representation and posture table creation. In the posture feature representation part, we introduce the definitions and notations of 3D human postures in the posture library. In the posture table creation part, we propose a table-lookup

technique to index 3D human postures. The index structure of lookup table is easy to update when the data set is modified. However, it suffers exponential storage and computation costs in high-dimension indexing and retrieval. To overcome the “curse of dimensionality,” Li *et al.* [27] and Sundaram and Chang [28] proposed algorithms to decompose a high-dimension feature vector into several low-dimension ones. Thus the indexing and retrieval costs can be greatly reduced. In this study, we divide the whole human body into nine separate segments and create nine corresponding posture tables. Details are described in the following.

2.1. Posture Feature Representation

Let Ω be the given 3D posture library which is obtained from motion capture devices. In Ω , a set of posture parameters, e.g., 3D positions of joints of body segments and the torso facing direction, is stored with respect to the human model. The human model is a hierarchical structure, which is defined according to the MPEG-4 Body Definition Parameters (BDPs) [29] standard. In this study, we simplify the human model to nine major body segments and a root orientation, as shown in Fig. 2. These body segments are the torso, upper arms and legs, lower arms and legs, and their associated joints (e.g., pelvis, chest, etc.). The root is defined at the pelvis joint and the root orientation is defined as follows. Let P be the plane passing through the root and parallel to the XZ plane, and t be the vector starting from the root and parallel to the torso facing direction. The projection of t on P is defined to be the root orientation.

Denote $\Omega = \{\omega_i \mid i = 1, 2, \dots, N\}$, where ω_i is the i -th posture and N is the total number of postures in Ω . Each posture ω_i is defined as $\omega_i = (B_i, V_{i,1}, V_{i,2}, \dots, V_{i,9})$,

where B_i is the root orientation and $V_{i,j}$ is the orientation vector of the j -th body segment (i.e., from its parent joint to its child joint). For simplicity, all postures in Ω are aligned so that their root orientations B_i are $(0, 0, 1)$.

2.2. Posture Table Creation

Let F be a given 2D human figure and $C(F)$ be its 3D reconstruction. Our goal is to find a 3D posture $\omega^* \in \Omega$ such that ω^* is the best approximation of $C(F)$ among all $\omega \in \Omega$. The basic notion of our approach is that for every $\omega \in \Omega$, we compute its projections with respect to a set of sampling view directions $\{D_k \mid k = 1, 2, \dots, 8\}$. Here eight view directions are sampled because they are easy to be described by users in the later reconstruction process. These view directions are parallel to the XZ plane and positioned circularly around posture ω , as shown in Fig. 3. By comparing F with these sampling projections, we retrieve the most appropriate $\omega^* \in \Omega$ as the approximation of $C(F)$. The key issue is to design an efficient index structure to facilitate the retrieval process. This issue is discussed in the following.

We apply scaled orthographic projection [11, 13] to model the projection relationship between the 3D space and the 2D image. This model is simply based on orthographic projection, plus a scale factor to represent the length ratio of a 3D world to a 2D image. Scaled orthographic projection is appropriate when the range of the object depth is small with respect to the distance between the object and the camera. It is less appropriate in images that have significant perspective effects. However, the scaled orthographic projection model does not need to acquire camera parameters through calibration, which is a difficult task for a single image. Fig. 4 shows a body segment and its projection under scaled orthographic projection. As a human posture

is composed of several individual body segments, we first introduce the single segment index structure under scaled orthographic projection.

Let o be the root of the human model and \overrightarrow{ob} be the root orientation vector. For a body segment vector V , we define the vector $\overrightarrow{ov} = (x, y, z)$ to be the equivalent vector of V , i.e., $\overrightarrow{ov} // V$ and $\|\overrightarrow{ov}\| = \|V\|$. Let $\{D_k = \overrightarrow{d_k o} \mid k = 1, 2, \dots, 8\}$ be the view direction set (see Fig. 3), where

$$d_k = \left(\sin\left(\frac{k-1}{8} \cdot 2\pi\right), 0, \cos\left(\frac{k-1}{8} \cdot 2\pi\right) \right).$$

Consider the scaled orthographic projection for the k -th view direction D_k and its associated image plane I (see Fig. 4). The reference plane of \overrightarrow{ov} with respect to the projection is defined as the plane R passing through the joint o and parallel to I . Moreover, the Cartesian coordinate system $(\overrightarrow{o_I U}, \overrightarrow{o_I V})$ and $(\overrightarrow{o_{X_R}}, \overrightarrow{o_{Y_R}})$ are defined in I and R respectively such that $\overrightarrow{o_I U} // \overrightarrow{o_{X_R}}$ and both are parallel to the XZ plane. Then the coordinates of $\overrightarrow{ov} = (x, y, z)$ in the $X_R Y_R Z_R$ coordinate system, denoted by (x_R, y_R, z_R) , can be expressed as:

$$(x_R, y_R, z_R) = (x \cos \theta + z \sin \theta, y, -x \sin \theta + z \cos \theta),$$

where $\theta = \frac{k-1}{8} \cdot 2\pi$ with respect to the k -th view direction D_k . Under scaled orthographic projection, \overrightarrow{ov} 's projection on image plane I , denoted as $\overrightarrow{o_I v_I} = (u', v')$, can be written as follows (see Fig. 4):

$$\begin{pmatrix} u' \\ v' \end{pmatrix} = s \cdot \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix} \cdot \begin{pmatrix} x_R \\ y_R \\ z_R \end{pmatrix} = \begin{pmatrix} s \cdot x_R \\ s \cdot y_R \end{pmatrix},$$

where s is an unknown scale factor. Define $\alpha \in [0, 2\pi)$ as a radian angle measured

counterclockwise from the positive U axis to $\overrightarrow{o_I v_I}$ on image plane I . It can be computed by the following formula:

$$\begin{cases} \alpha = \cos^{-1}\left(\frac{x_R}{\sqrt{x_R^2 + y_R^2}}\right) & \text{if } y_R \geq 0, \\ \alpha = 2\pi - \cos^{-1}\left(\frac{x_R}{\sqrt{x_R^2 + y_R^2}}\right) & \text{else.} \end{cases} \quad (1)$$

Note that the scale factor s is canceled in Eq. (1). After the above process, segment vector \overrightarrow{ov} is projected onto the image plane with respect to the k -th view direction D_k . The projected vector can be expressed as angle α .

There are two reasons that we extract the angle α of a segment vector with respect to a view direction. First, the computation of α , as shown in Eq. (1), does not involve with the scale factor s , which is not easy to know with such limited information provided. Second, suppose that a postured character in an image is given for reconstruction. When a body segment is labeled on the image, α can be obtained directly according to the above formula. In the following, we will show how to use α to create posture tables for indexing.

Suppose that the range of angle $\alpha \in [0, 2\pi)$ is equally divided into M bins. A posture table is created for each of nine body segments. Denote T_j , a two-dimensional array, as the posture table of the j -th body segment. Element $T_j(a, k)$ stores a list of posture numbers, where $a = 1, 2, \dots, M$ and $k = 1, 2, \dots, 8$ represent the indices about angle α and view direction D_k respectively. Let $\Omega_j = \{V_{1,j}, V_{2,j}, \dots, V_{N,j}\}$ be the vector sequence of the j -th body segment in the posture library. For each $V_{i,j} \in \Omega_j$, compute its angle $\alpha_{i,j,k}$ with respect to view direction D_k using Eq. (1) and quantify $\alpha_{i,j,k}$ as follows:

$$a_{i,j,k} = \left\lceil \frac{M \cdot \alpha_{i,j,k}}{2\pi} \right\rceil,$$

where $\lceil \cdot \rceil$ denotes the ceiling function. Then number i is recorded in the following elements:

$$\{T_j(a_{i,j,k}, k) \mid k = 1, 2, \dots, 8\}.$$

Fig. 5 shows an example that the left upper arm of a 3D posture is indexing in the corresponding posture table.

To conclude, the proposed posture table structure has two major advantages for indexing the posture library. First, when some data are added to or deleted from the posture library, we only have to compute their indices using Eq. (1) to find corresponding elements instead of re-learning the whole data set. Second, we divide the whole human body into nine separate segments and create nine corresponding posture tables to avoid the curse of dimensionality. In the next section we will describe the retrieval algorithm by using these posture tables.

3. Human Posture Reconstruction

Suppose that an image with a postured character is given for 3D human posture reconstruction. In our approach, users are first asked to provide a 2D human figure by labeling body segments and estimating the root orientation of the postured character in the image. Then the reconstruction work is accomplished through the following two processes: pivotal posture retrieval and constraint-based reconstruction. In the pivotal posture retrieval process, a 3D pivotal posture whose 2D projection is the best approximation of the given human figure is retrieved from the posture library. Based on the proposed index structure (i.e., posture tables), we develop an effective mechanism to retrieve the desire posture in the large volume of the data repository.

Next, in the constraint-based reconstruction process, a set of constraints, including segment length ratios, joint angle limits, pivotal posture reference, and feet-floor contact is automatically applied to reconstruct the 3D human posture with respect to the 2D human figure. In the following, we detail the 2D human figure and its subsequent reconstruction processes.

3.1. Pivotal Posture Retrieval

A 2D human figure consists of nine body segments and a root orientation with respect to the postured character in the image. Denote the 2D human figure as $F = (k_F, \alpha_{F,1}, \alpha_{F,2}, \dots, \alpha_{F,9}, l_{F,1}, l_{F,2}, \dots, l_{F,9})$, where $k_F \in \{1, 2, \dots, 8\}$ indicates the estimated root orientation of F (abbreviated as ERO), $\alpha_{F,j} \in [0, 2\pi)$ is the angle feature for the j -th body segment of F , and $l_{F,j}$ is the 2D length for the j -th body segment of F , as shown in Fig. 6. We note that the angle $\alpha_{F,j}$ is measured counterclockwise from the positive U axis to the j -th body segment.

Now we describe how to find the pivotal posture. Consider the j -th body segment of F . Angle $\alpha_{F,j}$ is quantified as:

$$a_{F,j} = \left\lceil \frac{M \cdot \alpha_{F,j}}{2\pi} \right\rceil,$$

where M is the number of rows in posture table T_j . Then $\alpha_{F,j}$ is indexed into element $T_j(a_{F,j}, k_F)$. However the best solution may not be stored in this element. For example, similar 3D segment postures may be stored in neighbors of the element. Besides, both position biases in body segment labeling and root orientation estimation are inevitable. Therefore the search area must be enlarged to tolerate these cases. In

other words, element $T_j(a_{F,j}, k_F)$ together with its neighbors should be searched to find the best solution for $\alpha_{F,j}$. We utilize a Gaussian mask to perform range search in the posture table, as shown in Fig. 7. Let G be a $(2m+1) \times (2m+1)$ Gaussian mask centered at $(a_{F,j}, k_F)$. Thus the range of G on T_j is:

$$\{(X, Y) \mid X = a_{F,j} - m, a_{F,j} - m + 1, \dots, a_{F,j} + m, \\ \text{and } Y = k_F - m, k_F - m + 1, \dots, k_F + m\}.$$

Let $(x, y) \in (X, Y)$, define:

$$G(x, y) = \exp\left(-\frac{(x - a_{F,j})^2}{2\sigma_x^2} - \frac{(y - k_F)^2}{2\sigma_y^2}\right),$$

where σ_x and σ_y are the standard deviations of dimension X and Y respectively.

Note that both dimensions are wrapped around. That is, if the range of G exceeds a boundary of the posture table, these exceeded elements are located at the opposite side of the boundary. For each of posture numbers stored in element $T_j(x, y)$, we set score:

$$S_{i,j} = G(x, y)$$

for posture number i . In other words, the projection of segment vector $V_{i,j} \in \Omega_j$,

which is indexed in $T_j(x, y)$, is similar to the corresponding segment of F and its

similarity degree is $S_{i,j}$. Recall that $V_{i,j}$ is indexed in the following elements

$\{T_j(a_{i,j,k}, k) \mid k = 1, 2, \dots, 8\}$. The range of G may cover multiple elements where $V_{i,j}$

is indexed. In this case, $S_{i,j}$ is set to its maximum of $G(x, y)$ with respect to $V_{i,j}$.

Then we calculate total score S_i of posture ω_i by summing up scores for segment

vectors of ω_i :

$$S_i = \sum_{j=1}^9 W_j \cdot S_{i,j},$$

where W_j is the weight of the j -th body segment. We assign the highest weight to the torso and the lowest weight to the lower arms and legs. This is because the movement of the torso will also affect the positions of the lower and upper limbs. On the contrary, the movement of the lower limbs will not affect the positions of the torso and upper limbs. Finally the pivotal posture, denoted as ω^* , is set to be the ι -th posture in Ω whose total score S_i is the highest.

3.2. Constraint-Based Reconstruction

After obtaining the pivotal posture, we proceed the next reconstruction for the given 2D human figure. In the reconstruction process, four constraints are used step by step, namely, segment length ratios, joint angle limits, pivotal posture reference, and feet-floor contact. For expository purpose, the constraint-based reconstruction is divided into two parts, namely, the physical constraint and the environmental constraint. The physical constraint, including segment length ratios, joint angle limits, and pivotal posture reference, is first used to reconstruct the 3D human posture for the given 2D human figure. Next the environmental constraint, i.e., feet-floor contact, is applied to further fine-tune the posture through inverse kinematics (IK). Details are described in the following two subsections.

3.2.1. Physical Constraint

Suppose that the 2D human figure F is given and length ratios of body segments are known a priori. For convenience, we denote the 3D length of the torso as l_1 and

set $l_1 = 1$. Accordingly l_j , $j = 2, 3, \dots, 9$, are the j -th segment lengths related to l_1 .

For F , we retrieve its pivotal posture ω^* from the posture library, as described in Section 3.1. To simplify the following reconstruction task, ω^* is further rotated about the Y axis so that its root orientation B^* is aligned with the estimated root orientation of F .

The reconstruction order of body segments is from the torso, upper limbs, to lower limbs. Consider the j -th body segment of F to be reconstructed. Let $\overrightarrow{ov_R}$ be its corresponding vector on reference plane R paralleled to image plane I under scaled orthographic projection. For the definition of the reference plane and the image plane, please refer to Section 2.2. The objective is to reconstruct the actual segment vector \overrightarrow{ov} from $\overrightarrow{ov_R}$. Because of the depth ambiguity problem, there are two candidates for \overrightarrow{ov} , as shown in Fig. 8. The Cartesian coordinates of \overrightarrow{ov} in the $X_R Y_R Z_R$ coordinate system, denoted by (x_R, y_R, z_R) , can be written as:

$$\begin{aligned} x_R &= l_j \cdot \cos \beta_j \cdot \cos \alpha_{F,j}, \\ y_R &= l_j \cdot \cos \beta_j \cdot \sin \alpha_{F,j}, \\ z_R &= \pm l_j \cdot \sin \beta_j, \end{aligned} \tag{2}$$

where $\beta_j \in [0, \frac{\pi}{2}]$ is the included angle for the j -th body segment between \overrightarrow{ov} and $\overrightarrow{ov_R}$, and $\alpha_{F,j}$ is the angle feature for the j -th body segment of F . In Eq. (2), l_j and $\alpha_{F,j}$ are given a priori, the only unknown parameter is β_j . To derive β_j , we first acquire β_1 for the torso by exploiting pivotal posture ω^* . Let V_1^* be the torso vector of ω^* . We project V_1^* onto reference plane R and measure the angle β_1^* between V_1^* and its projected vector. Assume that β_1^* is correct for the actual torso

and let $\beta_1 = \beta_1^*$. Based on this assumption, β_j for the j -th body segment can be derived in the following equation:

$$l_1 \cdot \cos \beta_1 : l_j \cdot \cos \beta_j = l_{F,1} : l_{F,j}, \quad (3)$$

where $l_{F,1}$ and $l_{F,j}$ are the 2D length for the torso and the j -th body segment of F , respectively. This equation formulates the relationship of segment length ratios between 2D and 3D space. Rewrite Eq. (3) as follows:

$$\beta_j = \cos^{-1} \left(\frac{l_1}{l_j} \cdot \frac{l_{F,j}}{l_{F,1}} \cdot \cos \beta_1 \right). \quad (4)$$

Recall that we set $l_1 = 1$. By substituting Eq. (4) into Eq. (2), the Cartesian coordinates of \overrightarrow{ov} can be rewritten as:

$$\begin{aligned} x_R &= \frac{l_{F,j}}{l_{F,1}} \cdot \cos \beta_1 \cdot \cos \alpha_{F,j}, \\ y_R &= \frac{l_{F,j}}{l_{F,1}} \cdot \cos \beta_1 \cdot \sin \alpha_{F,j}, \\ z_R &= \pm l_j \cdot \sqrt{1 - \left(\frac{1}{l_j} \cdot \frac{l_{F,j}}{l_{F,1}} \cdot \cos \beta_1 \right)^2}. \end{aligned} \quad (5)$$

Note that there is a special case for $\beta_j = \frac{\pi}{2}$. When $\beta_j = \frac{\pi}{2}$, $\overrightarrow{ov_R}$ is a vertical line or a point on reference plane R and the above equation is failed to deal with this case. Therefore when $\overrightarrow{ov_R}$ is a vertical line or a point, we perturb v_R by horizontally shifting it on R in a small distance so that $\beta_j \neq \frac{\pi}{2}$.

For the two candidates of \overrightarrow{ov} , denoted as $\overrightarrow{ov_1}$ and $\overrightarrow{ov_2}$, we exam which one is invalid by applying the joint angle limitations of MPEG-4 [29] and Lee's culling method [7]. In our experimental test, the probability to filter out the invalid candidate by applying the joint angle limitations is 1/3. If both candidates are valid, pivotal posture ω^* is referred to select an appropriate candidate as follows. Let V_j^* be the

j -th segment vector of ω^* . If $\|V_j^* - \overrightarrow{ov_1}\| \leq \|V_j^* - \overrightarrow{ov_2}\|$, then set $\overrightarrow{ov} = \overrightarrow{ov_1}$; else set $\overrightarrow{ov} = \overrightarrow{ov_2}$, where $\|x\|$ is the Euclidean distance of a vector x . That is, the candidate that is close to the corresponding segment vector of ω^* is chosen as our reconstructed segment vector \overrightarrow{ov} . Then \overrightarrow{ov} is joined to its parent segment vector. The above procedures are repeated until nine body segment vectors of the human posture are reconstructed.

3.2.2. Environmental Constraint

In some cases, we observe that the interaction between the reconstructed human posture and the environment is unreasonable. For example, if both feet in the image contact the floor, then the heights of the reconstructed feet should be the same. So if the reconstruction result violates this constraint, the positions of the reconstructed lower and upper legs can be further fine-tuned so that the heights of the feet are the same. For such case, we provide a fine-tuning option for users to apply the feet-floor contact constraint, as illustrated in Fig. 9. Fig. 9a shows a reconstructed human posture. Suppose that the heights of its feet should be the same, i.e., its feet contact the floor. The following fine-tuning steps are executed:

Step 1. Set locations of floor and floor fulcrum: Since the floor location is unknown yet, the first step is to acquire the floor-related information. We project two feet and the root onto the XZ plane to determine which projected foot is nearer to the projected root. Assume the gravity center of a human body is centralized in the root and the human body is kept balance in general. The floor is set to be the plane that is parallel to the XZ plane and passing through the position of the

nearer foot, and the nearer foot is set to be the floor fulcrum, as shown in Fig. 9b.

Step 2. Apply the inverse kinematics technique: For another foot that is not the floor fulcrum, its may penetrate the floor or be suspended in the air. We apply a real-time inverse kinematics technique [30] to move the lower and upper legs of the foot so that the foot contacts the floor while the hip holds fixed, as shown in Fig. 9c.

Using the feet-floor contact constraint to fine-tune the reconstructed result is an optional choice. If users consider that the feet contact the floor in actual scenes, they can apply this constraint to obtain more reasonable 3D human postures.

4. Experimental Results

We use motion capture data of Cheng's Tai Chi Chuan [31], a traditional Chinese martial art, as our posture library. The library contains more than 20000 3D human postures captured from a professional martial art master. The proposed approach is implemented using Matlab on an Intel Pentium 4 2.4GHz computer with 512 MB memory. The posture table we used is a 12×8 array in this study. In other words, the range of angle $\alpha \in [0, 2\pi)$ is equally divided into 12 bins. The search range on the posture table, i.e., the size of the Gaussian mask, is set to 3×3 .

4.1. Performance

In our experimental tests, the average time spent for users to generate the 2D human figure in an image is about 10 seconds. The average time spent for the

computer to reconstruct the 3D human posture (including pivotal posture retrieval and constraint-based reconstruction) is less than 1 second. Fig. 10 shows a number of images scanned from the Tai Chi Chuan book [31], and corresponding 3D human postures reconstructed by our approach. The first column shows the human figures in the original images. The second and third columns show the reconstructed human postures viewed from novel vantage points. A set of video clips that demonstrate the reconstruction procedures and results is available on

<http://www.cs.nthu.edu.tw/~dr888314/Reconstruction.html>. Since the segment length ratios of these postured characters are unknown, we use the master's ratios recorded in our posture library for reconstruction. Table 1 lists the segment length ratios of the master. Fig. 11 shows a sequence of key postures of Tai Chi Chuan motion – “Grasp the Swallow’s Tail.” The first and second rows are the sequences of 2D and 3D key postures.

To verify the accuracy of the proposed approach, three subjects, including the master and two disciples, are invited to perform Cheng’s Tai Chi Chuan for test data collection. Their 3D postures are captured through motion capture devices. At the same time, these postures are photographed from different view directions. The distance between the subject and the photographer is 5 meters. Besides, the segment length ratios of these three subjects are used to reconstruct more accurate postures. Table 1 lists their heights and segment length ratios. For a test image, its pivotal posture is retrieved and the corresponding human posture is reconstructed. Then we compute the discrepancies of these two postures compared to the captured posture by finding the translation and rotation that minimize the Root Mean Square Error (RMSE) between their nine segment vectors. Table 2 summarizes average RMSE of pivotal postures and reconstructed postures. The average RMSE is normalized, i.e., it divided by the subject’s height. In Table 2, the first row lists the test image number of each

subject. The second row lists the average RMSE of the retrieved pivotal postures for these test images. There exists about 13% error rate between retrieved pivotal postures and captured 3D data in this case. Besides, if the height difference is getting larger, the error is more prominent. The third row lists the average RMSE of reconstructed postures using the physical constraint. We observe that the proposed reconstruction method can improve about 6% error rate from retrieved pivotal postures. Fig. 12 shows some images and their reconstruction results. The first column shows human figures in test images. The second column shows the retrieved pivotal postures of these human figures from the same view directions. The third and fourth columns show their reconstructed postures from the same and other view directions. Red circles indicate the main differences between the pivotal postures and the reconstructed postures. It is obvious that the reconstructed postures look more similar to the test images than the pivotal postures.

In Table 2, the human posture is reconstructed based on the physical constraint only. We further evaluate the improvement for the reconstruction based on the physical and environmental constraints, as summarized in Table 3. Some testing images that should satisfy the feet-floor contact constraint are selected for this experiment. In Table 3, the second row lists four entries of leg segments that will be fine-tuned in the proposed environmental constraint. These leg segments are Left Lower Leg (LLL), Left Upper Leg (LUL), Right Lower Leg (RLL), and Right Upper Leg (RUL). The third row lists the average RMSE of the reconstructed postures using the physical constraint, whereas the fourth row lists the average RMSE using the physical and environmental constraints. It is clear that applying the environmental constraint indeed reduces the overall RMSE for posture reconstruction. However, the environmental constraint may result in ill effects on the upper leg (see the fifth row in Table 3). This is because that to satisfy this constraint, the posture shape of the leg is

modified through the IK technique. Sometimes the modified posture shape does not look similar to the given image. Therefore, the IK technique should take the posture shape into consideration. Another interesting phenomenon in Table 2 and Table 3 is that the master has the greatest accuracy and improvement. It is due to that the posture library is created by using the master's motion capture data. This relationship results in better performance when reconstructing the master's test images.

4.2. Discussion

We remark the error comes from the following factors:

1. Scaled orthographic projection: Taylor [11] designed a simulation experiment to investigate the effect of the scaled orthographic projection compared to the perspective projection. According to Taylor's simulation result, there will be at least 5.88% RMSE due to scaled orthographic projection contributed in our experimental case. Compare with our experimental results in Table 2, we speculate that there is about 1%~2% error caused by other minor factors, as discussed in the following.
2. Labeling body segment: Body segments in the image may not be perfectly labeled by users. It can be regarded as input signal noise. However, the noise level for labeling is about several pixels. Its scale is relatively minimal to the segment length scale.
3. Root estimation error: Recall that the estimated root orientation is classified into one of eight directions. Therefore there is a difference between the actual root orientation and the estimated one. Since the difference may up to $\frac{\pi}{4}$, we suggest

that the search range radius in the posture table is $\frac{\pi}{4}$ at least.

4. Posture data retargetting: The posture library is created by using the master's motion capture data. For a character of different segment length ratios, it may produce inappropriate results. This is similar to the retargetting problem in computer animation, which occurs when applying existing motion to different characters [32-33]. To overcome the retargetting problem, Barron's anthropometry estimation [14] provides a nice alternative choice.
5. Assumption in reconstruction: Recall that the reconstruction is based on the assumption that torso parameter β_1 is equal to that of the retrieved pivotal posture (for details please refer to Section 3.2.1). This assumption somewhat causes error during the reconstruction phase. However, if we use an adequate posture library to reconstruct the given image, e.g., reconstruct a Tai Chi Chuan image by using a Tai Chi Chuan library, the β_1 error can be reduced.

5. Conclusions and Future Work

In this study, we present a novel model-based approach to reconstruct the 3D human posture from a single image. The approach is guided by posture library retrieval and constraint-based reconstruction. A table-lookup index structure is devised to facilitate the retrieval. Besides, the physical and environmental constraints are automatically applied to reconstruct the 3D human posture. The major contribution is that we use the posture library to avoid providing extra visual cues manually. Moreover, a complete constraint-based procedure is provided for human posture reconstruction. Our experiments report acceptable error rates and show promising results on different human actors.

For future work, we will consider perspective projection instead of scaled orthographic projection. This work can be accomplished through the camera calibration process. Besides, we want to add some conditions in IK so that it can fine-tune the leg position without affecting the posture shape of the leg as possible. Another interesting research direction is to extend our approach to 3D human motion reconstruction from video, which contains rich spatial and temporal information. Our approach can provide a good initial estimation of spatial information in motion reconstruction.

References

- [1] C. Theobalt, M. Magnor, P. Schüler, H. Seidel, Combining 2D Feature Tracking and Volume Reconstruction for Online Video-Based Human Motion Capture, *Pacific Conference on Computer Graphics and Applications*, Beijing, China, Oct. 9-11, 2002.
- [2] G. K. M. Cheung, S. Baker, T. Kanade, Shape-from-silhouette of articulated objects and its use for human body kinematics estimation and motion capture, *IEEE Computer Vision and Pattern Recognition*, Madison, Wisconsin, USA, Jun. 16-22, 2003.
- [3] K. Morimura, T. Sonoda, Y. Muraoka, A whole-body-gesture input interface with a single-view camera – a user interface for 3D games with a subjective viewpoint, *International Conference on Web Delivering of Music*, Darmstadt, Germany, Dec. 9-11, 2002.
- [4] L. Ren, G. Shakhnarovich, J. Hodgins, H. Pfister, P. Viola, Learning Silhouette Features for Control of Human Motion, *ACM SIGGRAPH Conference on Sketches & Applications*, Los Angeles, CA, USA, Aug. 8-12, 2004.

- [5] J. Lee, J. Chai, J. K. Hodgins, P. S. A. Reitsma, N. S. Pollard, Interactive control of avatars animated with human motion data, *ACM Transactions on Graphics*, 21(3), 2002, pp. 491-500.
- [6] J. Davis, M. Agrawala, E. Chuang, Z. Popović, D. Salesin, A sketching interface for articulated figure animation, *ACM SIGGRAPH/Eurographics Symposium on Computer Animation*, San Diego, California, USA, Jul. 26-27, 2003.
- [7] H. J. Lee, Z. Chen, Determination of 3D human body postures from a single view, *Computer Vision, Graphics, and Image Processing*, 30(1985), pp. 148-168.
- [8] C. Bregler, J. Malik, Tracking people with twists and exponential maps, *IEEE Conference on Computer Vision and Pattern Recognition*, Santa Barbara, California, USA, Jun. 23-25, 1998, pp. 8-15.
- [9] D. E. Difranco, T. J. Cham, J. M. Rehg, Recovery of 3D articulated motion from 2D correspondences, *Compaq Cambridge Research Laboratory Technical Report Series*, CRL 99/7, Dec. 1999.
- [10] D. D. Morris, J. M. Rehg, Singularity analysis for articulated object tracking, *IEEE Conference on Computer Vision and Pattern Recognition*, Santa Barbara, CA, USA, Jun. 23-25, 1998, pp. 289-296.
- [11] C. J. Taylor, Reconstruction of articulated objects from point correspondences in a single uncalibrated image, *Computer Vision and Image Understanding*, 80(3), 2000, pp. 349-363.
- [12] V. Parameswaran, R. Chellappa, View independent human body pose estimation from a single perspective image, *IEEE Conference on Computer Vision and Pattern Recognition*, Washington, D.C., USA, Jun. 27-Jul. 2, 2004.
- [13] G. Loy, M. Eriksson, B.J. Sullivan, S. Carlsson, Monocular 3D reconstruction of human motion in long action sequences, *European Conference on Computer Vision*, Prague, Czech Republic, May 11-14, 2004, pp.442-455.

- [14] C. Barron, I. A. Kakadiaris, Estimating anthropometry and posture from a single image, *IEEE Conference on Computer Vision and Pattern Recognition*, South Carolina, USA, Jun. 13-15, 2000.
- [15] M. J. Park, M. G. Choi, S. Y. Shin, Human motion reconstruction from inter-frame feature correspondences of a single video stream using a motion library, *ACM SIGGRAPH Symposium on Computer Animation*, San Antonio, Texas, Jul. 21-22, 2002.
- [16] K. Rohr, Towards model-based recognition of human movements in image sequences, *CVGIP: Image Understanding*, 59(1), 1994, pp. 94-115.
- [17] X. Liu, Y. Zhuang, Y. Pan, Video based human animation technique, *ACM International Conference on Multimedia*, Orlando, Florida, USA, Oct. 30-Nov. 5, 1999.
- [18] H. Ning, T. Tan, L. Wang, W. Hu, Kinematics-based tracking of human walking in monocular video sequences, *Image and Vision Computing*, 22(5), 2004, pp. 429-441.
- [19] M. W. Lee, I. Cohen, Proposal maps driven MCMC for estimating human body pose in static images, *IEEE Conference on Computer Vision and Pattern Recognition*, Washington, D.C., USA, Jun. 27-Jul. 2, 2004.
- [20] G. Mori, X. Ren, A. A. Efros, J. Malik, Recovering Human Body Configurations: Combining Segmentation and Recognition, *IEEE Conference on Computer Vision and Pattern Recognition*, Washington, D.C., USA, Jun. 27-Jul. 2, 2004.
- [21] V. Pavlović, J. M. Rehg, T. J. Cham, K. P. Murphy, A dynamic Bayesian network approach to figure tracking using learned dynamic models, *IEEE International Conference on Computer Vision*, Kerkyra, Corfu, Greece, Sep. 20-25, 1999, pp. 94-101.
- [22] M. Brand, Shadow puppetry, *IEEE International Conference on Computer Vision*,

Kerkyra, Corfu, Greece, Sep. 20-25, 1999, pp. 1237-1244.

- [23] A. Elgammal, C. S. Lee, Inferring 3D body pose from silhouettes using activity manifold learning, *IEEE Conference on Computer Vision and Pattern Recognition*, Washington, D.C., USA, Jun. 27-Jul. 2, 2004.
- [24] N. R. Howe, M. E. Leventon, W. T. Freeman, Bayesian reconstruction of 3D human motion from single-camera video, *Neural Information Processing Systems*, Denver, Colorado, USA, Nov. 29-Dec. 4, 1999.
- [25] R. Rosales, S. Sclaroff, Learning body pose via specialized maps, *Neural Information Processing Systems*, Vancouver, British Columbia, Canada, Dec. 3-8, 2001.
- [26] A. Agarwal, B. Triggs, 3D human pose from silhouettes by relevance vector regression, *IEEE Conference on Computer Vision and Pattern Recognition*, Washington, D.C., USA, Jun. 27-Jul. 2, 2004.
- [27] C. S. Li, J. R. Smith, L. D. Bergman, V. Castelli, Sequential processing for content-based retrieval of composite objects, *SPIE Storage and Retrieval of Image and Video Databases*, San Jose, CA, Jan. 28-30, 1998, pp. 2-13.
- [28] H. Sundaram, S. F. Chang, Efficient video sequence retrieval in large repositories, *SPIE Storage and Retrieval of Image and Video Databases*, San Jose, CA, Jan. 26-29, 1999.
- [29] MPEG-4 Overview, ISO/IEC JTC1/SC29/WG11 N4668, March 2002,
<http://mpeg.telecomitalia.com/standards/mpeg-4/mpeg-4.htm>
- [30] D. Tolani, A. Goswami, N. Badler, Real-time inverse kinematics techniques for anthropomorphic limbs, *Graphical Models*, 62(5), 2000, pp. 353-388.
- [31] S. McFarlane, *The Complete Book of T'ai Chi*, Dorling Kindersley Limited, London, 1999.
- [32] M. Gleicher, Retargetting motion to new characters, *ACM SIGGRAPH*, Orlando,

Florida, USA, Jul. 19-24, 1998, pp. 33-42.

- [33] K. J. Choi, H. S. Ko, Online motion retargetting, *The Journal of Visualization and Computer Animation*, 11(5), 2000, pp. 223-235.
- [34] C. Tomasi, T. Kanade, Shape and motion from image streams under orthography: a factorization method, *International Journal of Computer Vision*, 9(2), 1992, pp. 137-154.
- [35] C. Bregler, A. Hertzmann, H. Biermann, Recovering non-rigid 3D shape from image streams, *IEEE Conference on Computer Vision and Pattern Recognition*, Jun. 13-15, 2000.
- [36] K. Grochow, S. L. Martin, A. Hertzmann, Z. Popović, Style-based Inverse Kinematics, *ACM Transactions on Graphics*, 25(3), 2004, pp. 522-531.