Towards Model-Based Capture of a Persons Shape, Appearance and Motion

Adrian Hilton
Centre for Vision, Speech and Signal Processing
University of Surrey, UK
a.hilton@surrey.ac.uk
http://www.ee.surrey.ac.uk/Research/VSSP/3DVision

Abstract

This paper introduces a model-based approach to capturing a persons shape, appearance and movement. A 3D animated model of a clothed persons whole-body shape and appearance is automatically constructed from a set of orthogonal view colour images. The reconstructed model of a person is then used together with the least-squares inverse-kinematics framework of Bregler et al. [4] to capture simple 3D movements from a video image sequence.

1 Introduction

There is increasing demand for a low-cost system to capture both human shape, appearance and movement. Potential applications for such a system include population of virtual environments, communication, multi-media games and clothing. This paper presents a technique for capturing recognisable models of individual people for use in VR applications. The captured model is then used as the basis for tracking the persons movements from a video image sequence. Captured 3D movements are used to animate the individuals avatar. A more advanced version of such a system could be used to provide an interactive interface for multi-user virtual environments where each person is represented by a recognisable model 'avatar' of themselves and the avatars movements are driven by the movements of the real person. Key requirements for model building and motion capture for use in virtual worlds are:

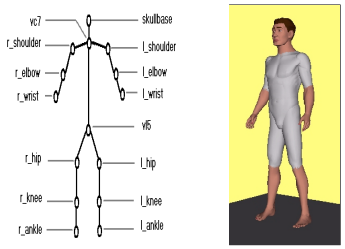
- Realistic appearance and approximate shape
- Articulated movement for animation
- Non-invasive movement capture
- Low-cost (automatic) acquisition of shape and motion

These requirements contrast with previous objectives of whole-body measurement systems which were principally designed to obtain accurate metric information of human shape. Current whole-body measurement systems are highly expensive and require expert knowledge to interpret the data and build animated models [14]. Similarly current commercial motion capture systems are focused on high accuracy motion capture for biometrics or film animation and require the placement of invasive targets on the person.

In this paper we introduce a technique for automatically building models of individual people from a set of four orthogonal view images using standard camera technology. The reconstruction from multiple orthogonal view images is analogous to previous work on facial modelling [2, 3, 11]. There is a considerable body of literature addressing the goal of realistic modelling of the head and face of individual people. A major feature of our approach is that we can automatically reconstruct recognisable whole-body articulated models of people who are fully clothed. The goal is to capture realistic appearance together with approximate shape information. Models are generated in the VRML-2 Humanoid Animation [15] standard which can be viewed in any VRML-2 compliant browser.

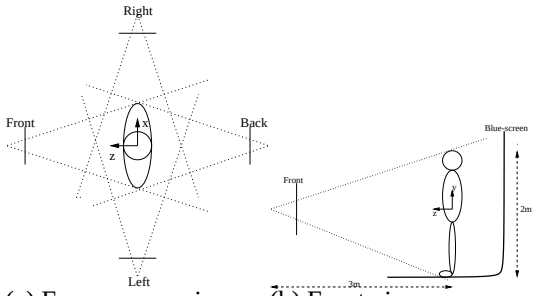
Capture of photo-realistic 3D models of a persons shape and appearance during movement is desirable for multi-media content production. Multi-camera studios have been used to capture realistic volumetric models[9, 12]. Over the past decade there has been extensive research in video based capture of human movement [4, 5, 6, 16, 17], see [1] for a comprehensive review. Many of these approaches are based on tracking movement using a model of human shape such as stick figures, overlapping spheres, cylinders, elliptical cylinders, truncated cones or super-quadrics. Recent related work [6, 8] has addressed the model-based reconstruction of whole-body shape, kinematic structure and movement from image sequences. In this paper we present an initial investigation into the use of a relatively realistic model of a persons shape and appearance as the basis for tracking movement across an image sequence. An individuals model provides non-invasive markers based on their shape and appearance. This paper presents initial results of tracking simple movements using this approach.

Supported by EPSRC Advanced Fellowship AF/97/2531 and EPSRC Grant GR/89518 'Functional Models: Building Realistic Models for Virtual Reality and Animation'



(a)Joints (b)Surface

Figure 1: Generic VRML H-Anim humanoid model



(a) Four camera views (b) Front view

Figure 2: Image capture setup

2 Model-based avatar reconstruction

In this section we introduce a model-based algorithm for reconstructing an animated model of a particular individual from a set of colour images. A generic 3D humanoid model is used as the basis for the reconstruction process. The shape of the model is modified based by an affine transform in multiple views which deforms the model to fit the silhouette shape of a particular person. The resulting model is then texture mapped using the captured images. This results in a recognisable model of an individual person which can be animated using the underlying kinematic structure of the generic model. For further details of this approach see [7].

2.1 Generic human model specification

Definition of a standard 3D humanoid model has recently received considerable interest for both efficient coding [10] and animation in virtual worlds [15]. In this work we have adopted the draft specification of the VRML Humanoid Animation Working Group (H-Anim) which defines a humanoid model structure which can be viewed using any VRML-2 compliant browser. The generic humanoid model used in this work is shown in Figure 1. The model consists of an articulated joint structure together with polygonal mesh segments attached to each joint to model the body-parts.

2.2 Image capture and feature extraction

An experimental system has been setup to capture whole body images of an individual from four orthogonal views (front,left,back,right), $I_j^D \ j = 1...4$, as illustrated in Figure 2(a) and (b). Video images are taken against a photo-reflective blue screen backdrop. The subject stands in a standard pose similar to the generic model pose.

The camera 3D to 2D projection can be modelled as pinhole camera expressed in homogeneous coordinates as:

$$\vec{u}'^j = P_j \vec{x}' = M E_j \vec{x}' \quad (1)$$

$$M = \begin{bmatrix} f_u & 0 & o_u & 0 \\ 0 & f_v & o_v & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix} \qquad E_j = \begin{bmatrix} R_j & \vec{t}_j \\ \vec{0}^T & 1 \end{bmatrix}$$

Where $\vec{u}' = (u, v, w)$ is a 2D point $\vec{u} = (u/w, v/w)$ in the camera image plane and $\vec{x}' = (x, y, z, 1)$ is a 3D point $\vec{x} = (x, y, z)$. P_j is the 4x3 camera projection matrix which can be decomposed into a 3x3 camera calibration matrix M and a Euclidean rigid body transform in 3D space E_j composed of R_j a 3x3 rotation matrix and $\vec{t}_j$ a 3x1 translation vector. The camera calibration parameters (f_u, f_v) are the focal lengths and (o_u, o_v) the image origin. The calibrated camera model is used to generate a set of four synthetic images of the generic humanoid model, $I_j^M \ j = 1...4$.

2.3 2D-to-2D Mapping

Chroma-key techniques are used to extract the silhouette from the captured images. The 8-connected chain of pixels on the border of the foreground image of a person is extracted for both the model and captured silhouette images. Automatic silhouette feature extraction is performed to extract three key feature points which can be accurately localised on the model (armpits and crotch). Reliable extraction and localisation of feature points has been achieved by first extracting the extremities of the silhouette and then analysing the shape of parts of the silhouette. The set of extracted features are sufficient to accurately align the seven major body parts for a captured image silhouette (head, shoulders, torso, arms, legs) with those of the generic model image. Alignment of body parts is essential to achieve correct animation of the reconstructed model.

A unique one-to-one correspondence between points inside the model and data sets for a particular body-part is established by a 2D linear (affine) mapping for each body part:

$$\vec{u}'^D = S \vec{u}'^M = \vec{u}'^M + \Delta \vec{u}' \quad (2)$$

$$S = \begin{bmatrix} s_u & s_{uv} & t_u \\ s_{vu} & s_v & t_v \\ 0 & 0 & 1 \end{bmatrix} \quad (3)$$

The vertical scale factor, s_u, and translation, t_u, for a particular body part can be computed from the vertical limits, $[u_{min}, u_{max}]$. Similarly the horizontal scale factor, s_v and translation t_v are given by the horizontal limits, $[v_{min}, v_{max}]$, for a horizontal slice ($u = const$) through the silhouette contour.

$$\begin{aligned} s_u &= \frac{u^D_{max} - u^D_{min}}{u^M_{max} - u^M_{min}} \\ t_u &= -s_u u^M_{min} + u^D_{min} \\ s_v(u) &= \frac{v^D_{max}(u) - v^D_{min}(u)}{v^M_{max}(u) - v^M_{min}(u)} \\ t_v(u) &= -s_v(u) v^M_{min}(u) + v^D_{min}(u) \end{aligned}$$

2.4 2D-to-3D Mapping from Orthogonal Views

The objective of the 2D-to-3D mapping is to combine the dense 2D-to-2D mapping information, $\Delta\vec{u}_i$, from multiple views, $i = 1, ..., 4$ to estimate the 3D displacement, $\Delta\vec{x}$ of a point $\vec{x}$ on the surface of the 3D model.

This is achieved by estimating the inverse projection of the displacement of the 2D point $\Delta\vec{u}_i$ in the camera image based on our knowledge of the distance to the corresponding 3D point, $\vec{x}^M$, on the generic model:

$$\begin{aligned} \vec{x}'^D &= \vec{x}'^M + \Delta\vec{x}'_i \\ &\approx \lambda_i(\vec{x}^M) E_i^{-1} M^{-1} (\vec{u'}_i^M + \Delta\vec{u'}_i) \end{aligned} \quad (4)$$

$$M^{-1} = \begin{bmatrix} \frac{1}{f_u} & 0 & -\frac{o_u}{f_u} \\ 0 & \frac{1}{f_v} & -\frac{o_v}{f_v} \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{bmatrix} \qquad E_i^{-1} = \begin{bmatrix} R_i^{-1} & -R_i^{-1}\vec{t}_i \\ \vec{0}^T & 1 \end{bmatrix}$$

Where λ_i is a scale factor equal to the orthogonal distance of the 3D point from the camera. The estimated 3D displacement component Δx_i orthogonal to the camera view direction $\vec{n}_i$ is:

$$\Delta\vec{x}_i \approx \lambda_i(\vec{x}^M) R_i^{-1} M^{-1} (\vec{u}^M + \Delta\vec{u}) - R^{-1}\vec{t} - \vec{x}^M \quad (5)$$

The distance to the true 3D point on a specific person is approximated by the distance to the corresponding point on the generic model. For a 3D point $\vec{x}^M$ in world coordinates the distance to the i^{th} camera centre with camera transform E_i gives the scale factor $\lambda_i(\vec{x}^M) = ||R_i\vec{x}^M + \vec{t}_i||$. Combining the estimated displacement components from two or more orthogonal views of a point $\vec{x}$ we obtain an estimate of the 3D displacement $\Delta\vec{x}$:

$$\Delta\vec{x} \approx \begin{bmatrix} \frac{1}{N_x}\sum_{i=0}^{N_x} \Delta x_i \\ \frac{1}{N_y}\sum_{i=0}^{N_y} \Delta y_i \\ \frac{1}{N_z}\sum_{i=0}^{N_z} \Delta z_i \end{bmatrix} \quad (6)$$

Where $N_x, N_y and N_z$ are the number of displacement estimates in a particular direction. This gives an estimate of the 3D displacement $\Delta\vec{x}$ of a point on the generic model $\vec{x}^M$. The estimated displacement is used to approximate the 3D shape of a specific person: $\vec{x}^D = \vec{x}^M + \Delta\vec{x}$.

The generic model shape is modified by estimating the 3D displacement $\Delta\vec{x}(\vec{v}_j)$ for each vertex $\vec{v}_j$. The resulting modified model is an affine transform of the shape of the original generic model to approximate the silhouette shape of the captured images. This model can be texture mapped using the known correspondence between model vertices and the captured 2D images to obtain a recognisable 3D model of an individual person.

2.5 Results of model-based reconstruction

The model-based reconstruction algorithm has been applied to automatically build models of approximately twenty individuals wearing a variety of clothing. Figure 3 shows examples of reconstructed models for male and female subjects wearing a variety of clothing. A realistic 3D model is captured for all subjects. In the case of a long-skirt a realistic model is captured but the animation structure is not valid due to the misplacement of hip and knee joints. Currently the principal limitation of the whole-body reconstruction from silhouettes is the quality of the face models generated due to the absence of specific facial feature point identification [3, 11].

Figure 4 shows a simple animated scene with several virtual people walking. Common animation parameters from motion capture data are applied to each model. The result is realistic animation of simple movements such as walking, running and jumping suitable for VR application.

3 Model-based motion capture

In this section we introduce a model-based approach to reconstruction of a persons movements from video image sequences. The captured model of a particular individual provides both shape and appearance information which in used as non-invasive markers for tracking the person between image frames. Least-squares inverse kinematics [4] is employed to estimate the 3D movement from the image motion. Initial results for tracking simple movements are presented.

Figure 3: Examples of reconstructed virtual people

Figure 4: Virtual people in a virtual catwalk scene animation

3.1 Twist representation for a kinematic chain

Bregler et al. [4] introduced a linear least-squares technique for solving the inverse kinematics of an articulated chain from image measurements. Their approach introduced to the computer vision community the *twist* representation [13] for a general rigid-body transform. The twist representation has the important property that it allows the instantaneous spatial velocity of a point on a kinematic chain to be expressed as a linear function of the angular velocity of the parent joints. This property enables a linear approximation to be developed between the instantaneous velocity of a point in the image and the velocity of a point on the kinematic chain assuming an orthographic projection model. In this section we briefly review the twist representation [13] and present a framework [4] for estimating the 3D motion of a person.

The twist representation of a rigid-body transform in 3-space is a 6-vector $\xi = [\vec{v}\vec{w}]^T$ where $\vec{w} = (w_x, w_y, w_z)$ is a unit 3-vector representing the direction of the rotation axis and $\vec{v} = (v_1, v_2, v_3)$ represents the location and translation along the rotation axis. Rotation by angle θ about the axis $\vec{w}$ is represented by the product: $\xi\theta$. An arbitrary rigid-body transform g can be represented by the exponential map of the twist $\xi\theta$ as follows:

$$g = e^{\hat{\xi}\theta} = I + \hat{\xi} + \frac{\hat{\xi}^2}{2!} + \frac{\hat{\xi}^3}{3!} + ... \quad (7)$$

$$= \begin{bmatrix} e^{\hat{w}\theta} & (I - e^{\hat{w}\theta})(\vec{w} \times \vec{v}) + \vec{w}\vec{w}^T v\theta \\ 0 & 1 \end{bmatrix} \quad (8)$$

where $R = e^{\hat{w}\theta} = I + \hat{w}sin\theta + \hat{w}^2(1 - cos\theta)$ is the 3x3 rotation matrix and the twist matrix:

$$\hat{\xi} = \begin{bmatrix} \hat{w} & \vec{v} \\ \vec{0}^T & 0 \end{bmatrix} \qquad \hat{w} = \begin{bmatrix} 0 & -w_z & w_y \\ w_z & 0 & -w_x \\ -w_y & w_x & 0 \end{bmatrix}$$

The instantaneous spatial velocity of a point on a body rotating with angular velocity $\dot{\theta}$ and constant twist ξ is:

$$\dot{\vec{x}}^w = \hat{\xi}\dot{\theta}e^{\hat{\xi}\theta}\vec{x}^o = \hat{\xi}\dot{\theta}\vec{x}^w \quad (9)$$

For the k^{th} segment of an articulated kinematic chain the transform between a point in object coordinates $\vec{x}^o$ and world coordinates $\vec{x}^w$ is given by the product of the transform between individual segments:

$$\vec{x}^w = g_0 g_1 .... g_k \vec{x}^o = e^{\hat{\xi}_0\theta_0} e^{\hat{\xi}_1\theta_1} .... e^{\hat{\xi}_k\theta_k} \vec{x}^o \quad (10)$$

The instantaneous spatial velocity $\dot{\vec{x}}^w$ of a point on the k^{th} of a kinematic chain is [13]:

$$\dot{\vec{x}}^w = \sum_{i=1}^{k} \left(\hat{\xi}_i'\dot{\theta}\right) x^w \quad (11)$$

with adjoint twist $\xi_i' = Ad_{(e^{\hat{\xi}_0\theta_0} e^{\hat{\xi}_1\theta_1} .... e^{\hat{\xi}_{i-1}\theta_{i-1}})}\xi_i$:

$$g = \begin{bmatrix} R & \vec{t} \\ \vec{0} & 1 \end{bmatrix} \qquad Ad_g = \begin{bmatrix} R & \hat{\vec{t}}R \\ \vec{0} & R \end{bmatrix} \quad (12)$$

The adjoint twist ξ_i' corresponds to the i^{th} joint twist transformed by the rigid body transformation which takes the i^{th} joint frame from its reference configuration to the current configuration. Thus equation 11 represents the instantaneous spatial velocity of a point on the k^{th} segment as a linear sum of the angular velocities for each parent joint multiplied by the adjoint twist for the current configuration.

If the root of the kinematic chain is subject to an arbitrary rigid-body transform representing the rotation and translation of the model in world coordinates. This can be represented by a six parameter twist transform $g_0 = e^{\hat{\xi}_0}$ where $\vec{w}_0$ is not constrained to be unit length and the magnitude, $|\vec{w}_0|$, is the angle of rotation about the axis. The instantaneous spatial velocity $\dot{\vec{x}}^w$ of a point on the rigid body is: $\dot{\vec{x}}^w = \dot{g}_0\vec{x}^0 = \lim_{\delta t \to 0}(g_0(t + \delta t) - g_0(t))\vec{x}^0$. Let the rigid-body transform $g_0(t) = e^{\hat{\xi}}$ and $g_0(t + \delta t) = e^{\hat{\xi}}e^{\delta\hat{\xi}}$. Then from equation 7 taking the first order approximation of the exponential series the spatial velocity is:

$$\dot{\vec{x}}^w = \lim_{\delta t \to 0}(e^{\delta\hat{\xi}} - I)e^{\hat{\xi}}\vec{x}^0 \approx \delta\hat{\xi}\vec{x}^w \quad (13)$$

where the twist representing the change in rigid-body transform from time t to $(t + \delta t)$ is $\delta\xi = [\delta v_1, \delta v_2, \delta v_3, \delta w_x, \delta w_y, \delta w_z]^T$. Equation 13 gives a linear approximation of the spatial velocity of a point in 3-space subject to an arbitrary rigid body transform. This approximation is valid provided the change in rigid-body transform is small.

Combining the approximate spatial velocity for a rigid-body transform of the root equation 13 with the spatial velocity for a point on the k^{th} segment of the kinematic chain 11 gives:

$$\dot{\vec{x}}^w \approx \left(\delta\hat{\xi} + \sum_{i=1}^{k}\left(\hat{\xi}_i'\dot{\theta}\right)\right)\vec{x}^w \quad (14)$$

This is a linear approximation of the instantaneous spatial velocity of a point on the k^{th} segment of the model for a given kinematic configuration $[\xi_0, \theta_1, ..., \theta_n]$. The approximate spatial velocity is linearly dependent on the parameters for the change in the root transform and the angular velocity of parent joints $\phi = \left[\delta v_1, \delta v_2, \delta v_3, \delta w_x, \delta w_y, \delta w_z, \dot{\theta}_0, \dot{\theta}_1, ..., \dot{\theta}_k\right]^T$. Therefore, to estimate the 3D motion of an articulated body such as a person we must estimate the relative change in the position, orientation and angular configuration over time.

3.2 Inverse kinematics from image motion

For a camera j the projection of a point $\vec{x}'^w$ in homogeneous world coordinates to image coordinates $\vec{u}^j$ is given by equation 1. The instantaneous image velocity of the projection of a point $\vec{x}^w$ on the k^{th} segment of a kinematic chain in the j^{th} camera can be expressed in a linear form using the twist approximation of spatial velocity Equation 14 as:

$$\dot{\vec{u}}'^j = M_j E_j \dot{\vec{x}}'^w \approx M_j E_j \left(\hat{\delta\xi} + \sum_{i=1}^{k} \left(\hat{\xi}_i \dot{\theta} \right) \right) \vec{x}^w \quad (15)$$

This gives a linear relationship between the angular velocity of the segments of the kinematic chain and the image velocity in homogeneous coordinates. Assuming a scaled orthographic projection model equation 15 gives a linear relationship between the instantaneous spatial velocity of points $\vec{x}_i$ on the model and their corresponding image velocity for the j^{th} camera view $\dot{\vec{u}}_i^j$.

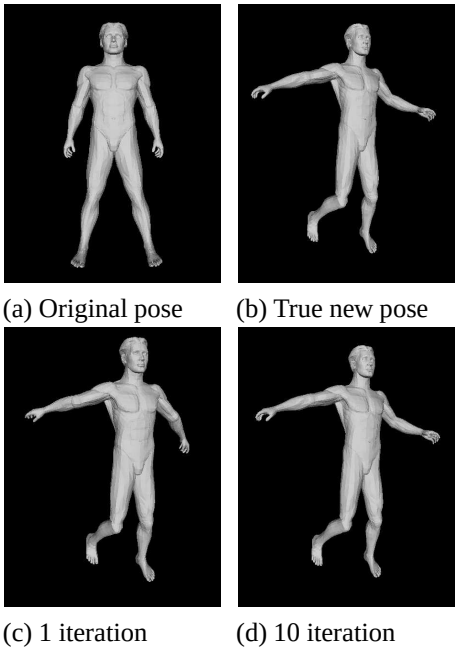
Given an image sequence we can estimate the image velocity $\dot{\vec{u}}^j$ between adjacent image fields $I(t)$ and $I(t+\delta t)$ by matching corresponding points between images. Assuming a constant angular velocity of the kinematic chain over the time interval δt then the change in angle for the r^{th} joint is $\delta\theta_r = \delta t \dot{\theta}_r$. From equation 15 for the a point $\vec{x}_i^w$ on an n-joint kinematic model we can write a linear relationship for the expected image velocity in terms of the parameter vector $\vec{\phi}' = [\delta\xi, \delta\theta_0, ..., \delta\theta_n]^T$ as:

$$\vec{h}_i^j \vec{\phi}' = \dot{u}_i^j \quad (16)$$

Therefore, given a set of estimated image velocities $\tilde{\dot{u}}_i^j$ for $i = 0, ..., n_p$ corresponding to n_p points on the model $\vec{x}_i^w$ we obtain a set of simultaneous linear measurement equations which can be expressed in matrix form as:

$$H^j \vec{\phi}' = \vec{z}^j \quad (17)$$

Where $\vec{z}^j$ is the measurement vector and H^j is the design matrix with rows $\vec{h}_i^j$ relating the point positions to the expected measurements by the parameters $\vec{\phi}'$ for the j^{th} camera view. Given a set of m synchronised cameras for each camera we can obtain a set of independent measurements for $\dot{\vec{u}}_i^j$ $i = 0...n_{p_j}$ for each camera view. For each measurement we can estimate the corresponding 3D point on the model $\vec{x}_{ij}^w$ which is related to the image coordinates by the j^{th} camera model M_j and E_j. Assuming that the camera model is a known scaled orthographic projection we can write a combined linear matrix equation of the form 17 which relates the measurements in all camera views to a single set of changes in articulation parameters $\vec{\phi}'$ for the human model: $H\vec{\phi}' = \vec{z}$ This set of equations can be solved for the parameters ϕ by a standard least-squares technique. In this work we use singular-value decomposition to obtain a solution which is insensitive to degeneracies inherent in the measurement set and kinematic chain configuration. For complex multi-dimensional problems such as the human kinematic structure singularities often occur due to insufficient measurements or configurations with multiple solutions.



(a) Original pose (b) True new pose

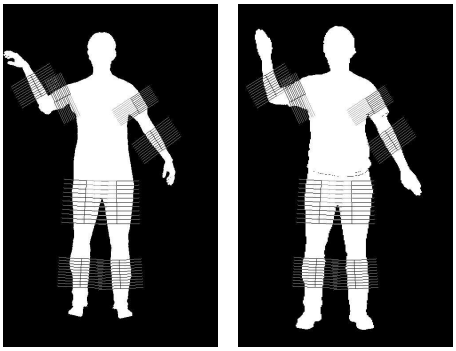
(c) 1 iteration (d) 10 iteration

Figure 5: Least-squares inverse kinematics from measurements on hands and feet only

Equation 17 enables a solution for the change in joint position between image fields which satisfies the kinematic constraints and best fits the set of image velocity measurements. For instance if a set of image measurements is taken on the extremities of the body (hands and feet) the solution to equation 17 gives a valid solution for the inverse kinematics of the parent joints (elbow and knee) even if there are multiple possible configurations. Figure 5 illustrates the reconstruction of the change in kinematic structure for the whole body given measurements of the change in position for the hands and feet only.

3.3 Single-view human motion estimation

In this section we present some initial results of model-based motion capture using the least-squares inverse kinematics framework [4] outlined in the previous section. The captured model of a particular person reconstructed using the technique presented in section 2 is used as the basis for 3D motion estimation. The reconstruction al-



(a) Previous estimate (b) New image

Figure 6: Image grid projection for estimated model $I^{\tilde{M}}(t)$ and new $I^D(t+\delta t)$ silhouette images

gorithm results in a model whose projection $I^M(0)$ is aligned with the persons projection in the first image $I^D(0)$ of the video image sequence.

To track the change in position and articulated joint angles between captured image frames $I^D(t)$ and $I^D(t+\delta t)$ we estimate the change in position of a set of points on the silhouette of the articulated figure. Silhouette boundary points in isolation are not ideal as they do not constrain all degrees of freedom. They have been used for the initial investigation as they are easily extract from the input image sequence. For the current estimated model $\tilde{M}(t)$ we obtain a projection image $I^M(t)$. Based on the current estimated joint positions we project a grid onto the model image for each body-part as illustrated in Figure 6(a). For each grid line we locate the image points on the boundary of the model silhouette which gives a set of 2D points $\{\vec{u}_i^M(t), i = 0, ..., n_p^M\}$. The corresponding 3D points on the current model $\{\vec{x}_i^M(t)\}$ are identified by back projecting the image ray and locating the intersection with the current model $\tilde{M}(t)$.

Estimates of the corresponding points in the new image $I^D(t+\delta t)$ are obtained by projecting the joint grid for each body part based on the model at $M^D(t)$ and locating the points on the boundary of the persons silhouette $\{\tilde{\vec{u}}_i^D(t+\delta t), i = 0, ..., n_p^M\}$. The least-squares inverse kinematics Equation 17 is then applied to obtain the least-squares estimate for the change in model parameters to obtain a new model estimate $M^D(t+\delta t)$.

Figure 7 shows the estimated motion from a single camera sequence of 100 images of a simple arm movement. The 3D model contains 21 degrees of freedom: 6 for the humanoid root, 3 for the spine and 3 for each arm and leg. The simple silhouette grid tracking algorithm introduced above provides estimates of the silhouette point correspondence between images which are sufficient to reconstruct the motion. The arms rotations about an axis in the view direction are correctly reconstructed. However, the rotation of the forearm which is orthogonal to the view direction is not correctly reconstructed resulting in the hand not rotating correctly. This error in the reconstructed 3D motion is due to the simple silhouette tracking algorithm. Correlation of points on the arm surface between frames is required to reconstruct full rotations and translations between views. The second limitation of the current silhouette based approach is that it cannot cope with self-occlusions which are of considerable importance in human motion capture. A more sophisticated correlation based matching approach is required to estimate the motion in the presence of self-occlusion.

These results demonstrate the principle of model-based motion capture. Current limitations are due to the simple silhouette based 2D tracking approach. The silhouette based approach exploits 3D shape information from the captured model of a particular person to perform the tracking. Future work will also exploit the colour information to reconstruct more complex movements with self-occlusion.

4 Conclusions

Model-based technique have been demonstrated for:

1. Reconstructing recognisable articulated models of individual people from a set of captured images.

2. Capturing simple movements of a person from a video image sequence based on shape.

The technique for reconstruction of a persons shape and appearance enables automatic acquisition of a model of a person suitable for populating shared virtual environments. This approach provides low-cost capture of recognisable models of people with a wide variety of shape, size and clothing. Further work will address improved reconstruction of facial models based on feature points [11].

Model-based reconstruction of a persons movements enables automatic capture of simple 3D movements from a video image sequence. The initial investigation demonstrates the feasibility of using a realistic model of shape for tracking. Further work is required to evaluate if a realistic model will provide sufficient information for non-invasive 3D capture of a persons movements.

References

[1] J. Aggarwal and Q. Cai. Human motion analysis: A review. In *IEEE Workshop on Non-Rigid and Articulated Motion*, pages 90—102, 1997.

[2] T. Akimoto, Y. Suenaga, and R. Wallace. Automatic creation of 3d facial models. *IEEE Computer Graphics and Applications*, 13(5):16—22, 1993.

[3] Andy Mortlock, Dave Machin, Stephen McConnell and Phil Sheppard. Virtual conferencing. Technical report, BT Technology Journal, 1997.

(a) Frames 1,20,50,70,99 (b) Estimated 3D model

Figure 7: Model-based capture of simple arm movement

[4] C. Bregler and J. Malik. Tracking people with twists and exponential maps. In *Int. Conf. on Computer Vision and Pattern Recognition*, 1998.

[5] Q. Cai and J. Aggarwal. Automatic tracking of human motion in indoor scenes across multiple synchronized video streams. In *IEEE International Conference on Computer Vision*, pages 356—363, 1998.

[6] P. Fua, A. Gruen, R. Plankers, N. Apuzzo, and D. Thalmann. Human body modeling and motion analysis from video sequences. In *Proc.Int.Symp. on Real-Time Imaging and Dynamic Analysis*, 1998.

[7] A. Hilton, D. Beresford, T. Gentils, R. Smith, and W. Sun. Virtual people: Capturing human models to populate virtual worlds. In *IEEE International Conference on Computer Animation*, pages 174—185, May 1999.

[8] I. Kakadiaris and D. Metaxas. Three-dimensional human body model acquisition from multiple views. *International Journal of Computer Vision*, 30(3):191—218, 1998.

[9] T. Kanade and P. Rander. Virtualized reality: Constructing virtual worlds from real scenes. *IEEE Multi-media*, 4(2):34—47, 1997.

[10] R. e. Koenen. *Coding of Moving Pictures and Audio*. http://drogo.cselt.stet.it/mpeg/standards/mpeg-4.htm, 1996.

[11] W.-S. Lee and N. Magnenat-Thalmann. Head Modeling from Pictures and Morphing in 3D with Image Metamorphosis Based on Triangulation. In *Modelling and Motion Capture Techniques for Virtual Environments - Magnenat-Thalmann,N. and Thalmann,D. (Eds.)*, pages 254—268, 1998.

[12] S. Moezzi, L.-C. Tai, and P. Gerard. Virtual view generation for 3d digital video. *IEEE Multi-media*, 4(2):18—26, 1997.

[13] R. Murray, Z. Li, and S. Sastry. *A Mathematical Introduction to Robotic Manipulation*. CRC Press, 1994.

[14] S. Paquette. 3d scanning in apparel design and human engineering. *IEEE Computer Graphics and Applications*, 16(9):11—15, 1996.

[15] B. Roehl. *Draft Specification for a Standard VRML Humanoid, Version 1.0*. http://ece.uwaterloo.ca/ h-anim/, 1997.

[16] S. Wachter and H.-H. Nagel. Tracking of persons in monocular image sequences. In *IEEE Workshop on Non-Rigid and Articulated Motion*, pages 2—9, 1997.

[17] C. Wren, A. Azarbayejani, T. Darrell, and A. Pentland. Pfi nder: Real-time tracking of the human body. In *IEEE Workshop on Face and Gesture Recognition*, pages 51—56, 1996.