Virtual People:
Capturing human models to populate virtual worlds

Adrian Hilton, Daniel Beresford, Thomas Gentils, Raymond Smith and Wei Sun

Centre for Vision, Speech and Signal Processing
University of Surrey, Guildford GU25XH, UK
a.hilton@surrey.ac.uk
http://www.ee.surrey.ac.uk/Research/VSSP/3DVision/VirtualPeople

Abstract

In this paper a new technique is introduced for automatically building recognisable moving 3D models of individual people. Realistic modelling of people is essential for advanced multimedia, augmented reality and immersive virtual reality. Current systems for whole-body model capture are based on active 3D sensing to measure the shape of the body surface. Such systems are prohibitively expensive and do not enable capture of high-quality photo-realistic colour. This results in geometrically accurate but unrealistic human models. The goal of this research is to achieve automatic low-cost modelling of people suitable for personalised avatars to populate virtual worlds.

A model-based approach is presented for automatic reconstruction of recognisable avatars from a set of low-cost colour images of a person taken from four orthogonal views. A generic 3D human model represents both the human shape and kinematic joint structure. The shape of a specific person is captured by mapping 2D silhouette information from the orthogonal view colour images onto the generic 3D model. Colour texture mapping is achieved by projecting the set of images onto the deformed 3D model. This results in the capture of a recognisable 3D facsimile of an individual person suitable for articulated movement in a virtual world. The system is low-cost, requires single-shot capture, is reliable for large variations in shape and size and can cope with clothing of moderate complexity.

Keywords: Avatar, Virtual Human, Whole-body Modelling, Humanoid Animation, Virtual Reality, VRML, Vision Techniques, 3D Reconstruction

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

1. Introduction

There is increasing demand for a low-cost system to capture both human shape and appearance. 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. For instance each participant in a multi-user virtual environment could be represented to others as an 'avatar' which is a realistic facsimile of the persons shape, size and appearance. The key requirements for building models of individuals for use in virtual worlds are:

- Realistic appearance
- Animatable movements
- Low-cost (automatic) acquisition

These requirements contrast with previous objectives of whole-body measurement systems which were principally designed to obtain accurate metric information of human shape. Such systems typically capture low-resolution colour and have restrictions on surface properties which result in no measurements for areas of dark colours and hair. Current whole-body measurement systems are highly expensive and require expert knowledge to interpret the data and build animated models [13]. These systems are suitable for capturing measurements of individual people for clothing applications but are not capable of capturing recognisable models for VR or photo-realistic models for computer animation. Recent research has addressed reconstructing realistic animated face models [1, 3, 10, 14] and whole-body models of kinematic structure [4, 6] from captured images. The objective of this research is to extend this work to address the reconstruction of whole-body models of shape and appearance from captured images.

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 [1, 2, 10]. A major feature of our approach is that we can reconstruct recognisable colour models of people who are fully clothed. The aim is to capture accurate appearance together with approximate shape information and not to accurately measure the underlying body dimensions. This work generates models in the VRML-2 Humanoid Animations [15] standard which can be viewed in any VRML-2 compliant browser. It is envisaged that the commercial availability of low-cost whole-body capture will open up a mass market for personalised plug-ins to multimedia and games packages.

There is a considerable body of literature addressing the goal of realistic modelling of the head and face of individual people. Techniques have been presented [1, 9, 10, 2, 14, 16] which use captured 2D images to modify the shape of a 3D generic face model to approximate a particular individual. Photogrammetric techniques are used to estimate the 3D displacement of points on the surface of a generic model from multiple camera images. Texture mapping of the captured images is then used to achieve a recognisable 3D face model. Reconstruction of animated face models from dense 3D surface measurements has been demonstrated [3, 11, 17]. Face modelling techniques using multiple images are similar to the approach presented in this paper for whole-body modelling. A difference in our approach is the use of silhouette data to register the images with a generic model and estimate the 3D shape. Techniques for facial modelling [2, 10, 14] could be used in conjunction with whole-body reconstruction to achieve improved facial modelling. However, current image based techniques for face modelling require a full resolution image to enable automatic feature labelling. In addition, current face modelling techniques may fail to reliably reconstruct face shape automatically for large variations in shape and appearance due to hair, glasses and beards.

Recent research has addressed the image based reconstruction of whole-body shape and appearance from sets of images [4, 5, 6, 7, 12]. Reconstruction of coarse 3D shape and appearance of a moving person from multi-view video sequences has been demonstrated [7, 12]. Modelling of human shape and kinematic structure has been addressed for captured images sequences [4, 6]. Unlike previous whole-body modelling techniques the approach presented in this paper aims to reconstruct a recognisable model of a persons shape and appearance. The captured silhouette images of a person in a single pose are used to modify the shape of a generic humanoid model to obtain an estimate of the kinematic structure. Techniques for estimating kinematic structure [4, 6] could be combined with the current approach to accurately estimate joint positions using images of a person in multiple poses. This would significantly improve the accuracy of the reconstructed kinematic structure for large variations in shape, size and clothing.

2. Overview

An overview of the model-based 3D human shape reconstruction algorithm is illustrated in Figure 1. A generic 3D humanoid model is used as the basis for reconstruction as shown in Figure 1(a). Four synthetic images are generated for orthogonal views (front,left,right, back) of the model by projection of the generic model as illustrated in Figure 1(b). To reconstruct a model of a person four orthogonal view images are captured with the subject in approximately the same posture as the generic model. This is illustrated in Figure 1(c). We will refer to captured images of a particular person as the 'data images' and to images of the generic 3D model as the 'model images'.

Silhouette extraction is performed on the model and data images and a small set of key feature points are extracted as illustrated in Figure 1(d) and (e). Initial alignment of the feature points between the model and data ensures that separate functional body parts of the generic model (arms, legs and head) are correctly mapped to corresponding parts of the captured image silhouettes. Correct correspondence of body parts is required to achieve correct animation of the reconstructed 3D model of a particular person. A 2D-to-2D linear affine mapping between the model and data image silhouettes is introduced to establish a dense correspondence for any point inside the silhouette. This correspondence can be used to map the colour information from the data image onto the model image as illustrated in Figure 1 (f).

The dense 2D-to-2D mapping for a single image is used to define the shape deformation of the 3D model in a plane orthogonal to the view direction. Applying this deformation to the 3D generic model achieves a 2D-to-3D linear mapping of the image silhouette shape onto the shape of the 3D model. This model-based 2D-to-3D mapping is the core of the technique for reconstruction of 3D human models. Integrating shape deformation information from two or more orthogonal views gives three orthogonal components of shape deformation. Applying this deformation to the generic model we can approximate the shape of a particular individual as illustrated in Figure 1(g). Combining the 3D shape with the 2D-to-2D mapping of the colour information we can obtain a colour texture mapped 3D model as illustrated in Figure 1(i). The resulting reconstructed 3D model provides a realistic representation of a particular individual. The articulated joint structure of the generic functional model can then be used to generate movement sequences for a particular individual in a virtual world as illustrated in Figures 1(h) and (j).

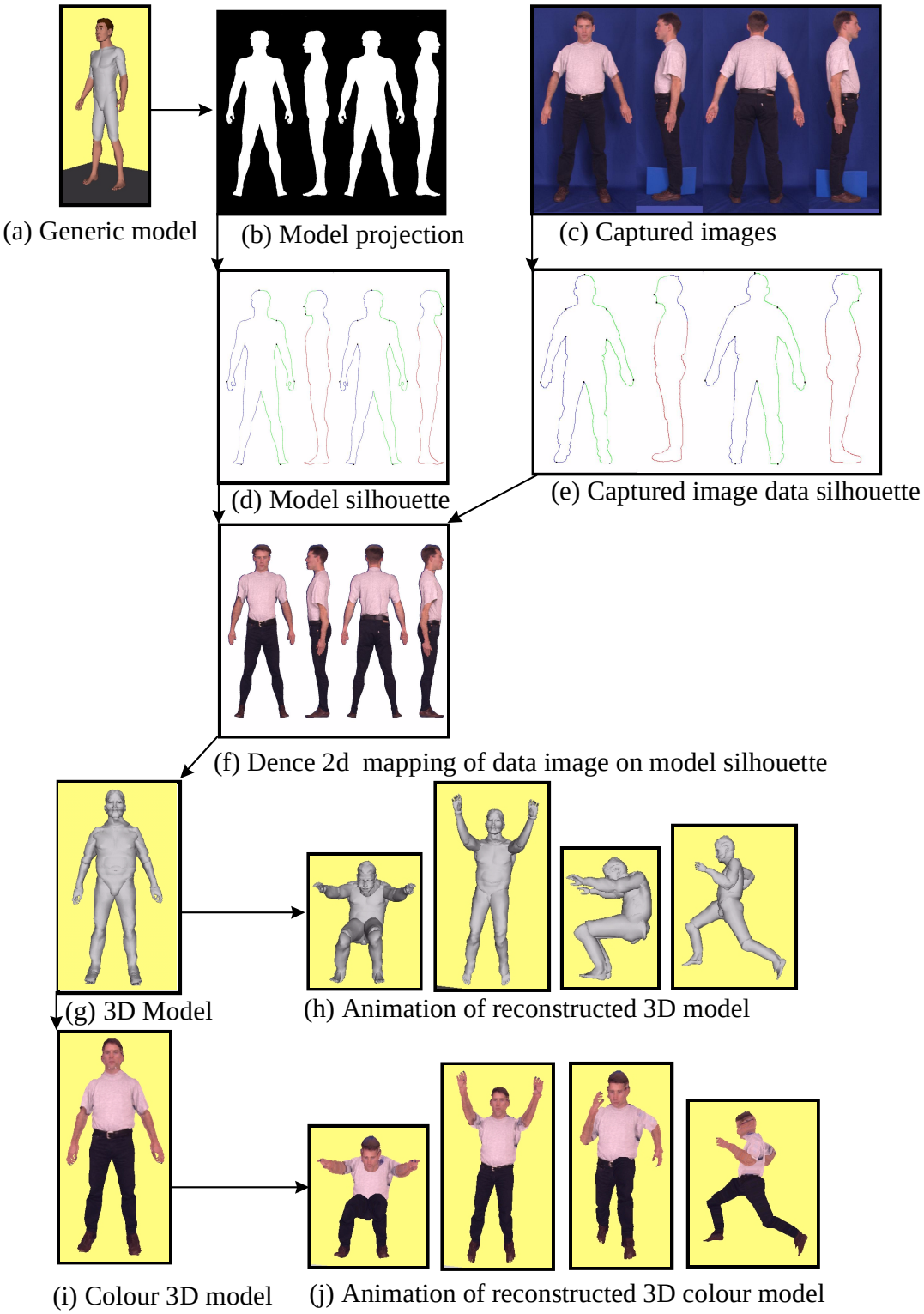


Figure 1. Overview of model reconstruction for an individual person

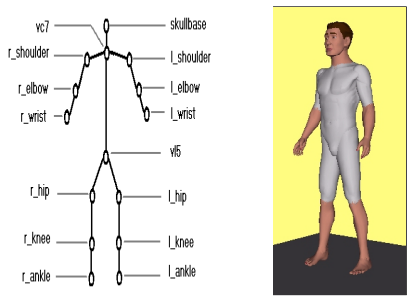
3. Model-based avatar reconstruction

3.1. Generic human model specification

Definition of a standard 3D humanoid model has recently received considerable interest for both efficient coding [8] 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. A set of 3D humanoid models based on the draft standard are publicly available from the humanoid animation web site [15]. The generic humanoid model used in this work is shown in Figure 2. The H-Anim draft standard defines a hierarchical articulated joint structure to represent the degrees-of-freedom of a humanoid. The humanoid shape is modelled by attaching either a 3D polygonal mesh segment to the joint for each body part or a single polygonal mesh surface for the whole-body. For example the articulated structure of an arm can be represented by three joints shoulder-elbow-wrist and the shape by segments attached to each joint upper-arm-forearm-hand. The shape segments can be specified with multiple levels-of-detail to achieve both efficient and realistic humanoid animation. Material properties and texture maps can be attached to each body segment for rendering the model.

The model-based reconstruction algorithm introduced in this paper can use any reasonable generic humanoid body as the initial model which is modified to approximate the shape and texture of a particular person. The reconstruction algorithm can also handle models with multiple levels-of-detail for each body part. All reconstruction results presented in this paper are based on a publicly available humanoid model which is compliant with the draft standard and gives a reasonable compromise between representation quality and animation efficiency. The joint structure for the generic humanoid model consists of fifteen joints as illustrated in Figure 2(a). The model shape consists of fifteen body segments with a total of 10K mesh vertices and 20K triangular polygons. The rendered surface model is shown in Figure 2(b). The VRML-2 specification allows movement animations based on interpolation of joint angles to be specified independent of the humanoid geometry.

The following nomenclature is used in later sections to refer to the polygonal model and associated texture map. Throughout this work the notation $\vec{x} = (x, y, z)$ refers to a 3D vector such as a mesh vertex. For each body part the polygonal mesh is specified as a list of N_v 3D vertices,$\{\vec{v}_i = (x_i, y_i, z_i)\}_{i=1}^{N_v}$, and a list of N_t polygons, $\{t_r = (\vec{v}_i, \vec{v}_j, \vec{v}_k)\}_{i=1}^{N_t}$. An image or texture map 2D coordinate is defined as $\vec{u} = (u, v)$ where u is the vertical coordinate and v is the horizontal coordinate with $u, v \in [0, 1]$



(a)Joints (b)Surface

Figure 2. Generic VRML H-Anim humanoid

and the origin at the top left-hand corner of the image. Texture mapping of an image onto a polygonal mesh is specified by a 2D texture coordinate for each mesh vertex, $\{\vec{u}_i = (u_i, v_i)\}_{i=1}^{N_v}$.

3.2. Image capture and feature extraction

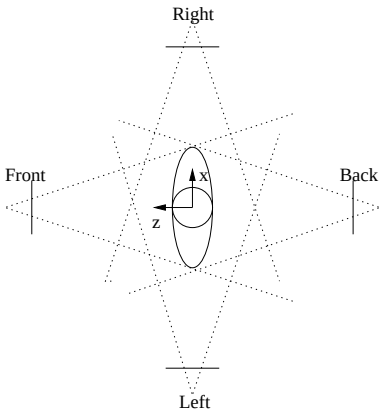
3.2.1 Image capture

An experimental system has been setup to capture whole body images of an individual from four orthogonal views (front,left,back,right). The four view camera configuration is illustrated in Figure 3(a). Colour images are captured using a Sony DXC-930P 3CCD camera with 756×582 picture elements. This gives a resolution of approximately 40×40 pixels for the subjects face. Images are taken against a photo-reflective blue screen backdrop which allows reliable foreground/background separation with arbitrary foreground lighting and most blue clothing. The subject stands in a standard pose similar to the generic model pose as shown in Figure 5(a). Currently each view is taken with a single camera with the subject rotating to present the required view to the camera. The use of a single camera may result in small changes of pose between views but has the advantage of identical intrinsic camera projection parameter for each view. The capture process results in a set of four data images, I_i^D $i = 1...4$, for orthogonal views of a specific person.

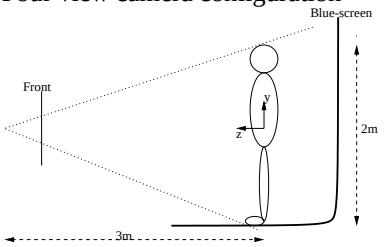
To model the image capture process we assume a pinhole camera without lens distortion. The camera 3D to 2D projection can be expressed in homogeneous coordinates as:

$$\vec{u'}_i = P_i \vec{x'} = M P_{orth} E_i \vec{x'} \quad (1)$$

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 $\vec{x} = (x, y, z)$ in world coordinates expressed in homogeneous form. P_i is the 4x3 camera projection matrix which can be decomposed into a 3x3 camera calibration matrix M,



(a) Four view camera configuration



(b) Front view configuration

Figure 3. Image capture setup

a 3x4 orthographic projection matrix from 3D to 2D P_{orth} and a Euclidean rigid body transform in 3D space E_i representing the view transform for the i^{th} camera in world coordinates. The individual matrices have the following form:

$$M = \begin{bmatrix} f_u & 0 & o_u \\ 0 & f_v & o_v \\ 0 & 0 & 1 \end{bmatrix}$$

$$P_{orth} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}$$

$$E_i = \begin{bmatrix} R_i & \vec{t}_i \\ \vec{0}^T & 1 \end{bmatrix}$$

Where R_i is a 3x3 rotation matrix and $\vec{t}_i$ a 3x1 translation vector. The camera calibration parameters (f_u, f_v) are the focal lengths and (o_u, o_v) the image origin and $\vec{0}$ a 3x1 vector of zero's. Camera calibration is based on direct estimation of the camera parameters from the size and position of the captured view plane relative to the distance from the camera (for instance $f_u = \frac{viewheight}{2 \times distance}$ giving $f_u = f_v \approx \frac{1}{3}$). Figure 3(b) illustrates the camera projection process for a single view. The calibrated camera model is used to generate a set of four synthetic images of the generic humanoid model, I_i^M $i = 1...4$. This is achieved by projection of each vertex v_i on the generic model to its corresponding image coordinates u_i using equation 1.

3.2.2 Silhouette extraction

Silhouette extraction aims to construct the chain of image pixels which lie on the boundary between the image of the person and the background. A standard chroma-key technique is used to identify background pixels based on the percentage of blue in each pixel. Given an image pixel with red, green and blue colour components (r, g, b) the percentage blue is $p_b = 100 \times (b/(r + g + b))$. A constant chroma-key threshold (50%) is used together with an intensity threshold ($|b| > 150$) to reliably generate a binary image with each pixel labelled as either foreground or background. An example of an extracted binary silhouette image is shown in Figure 5(b).

The silhouette curve, C_i^D, for each captured image, I_i^D, is extracted by following an 8-connected chain of pixels on the border of the foreground image of a person. We denote the silhouette by the chain of N_p 8-connected adjacent pixel coordinates $C_i^D = \{\vec{u}_j\}_{j=1}^{N_P}$ which is specified in counter-clockwise order with-respect-to the image view direction. An example of an extracted silhouette chain is shown in Figure 5(c). A similar process is performed on the synthetic images I_i^M of the generic model using a binary threshold to obtain a set of model silhouette curves, C_i^M.

3.2.3 Feature extraction

The objective of feature extraction is to establish the correct correspondence between the captured data and synthetic model images for body parts such as the arms, legs, head and torso. Correct correspondence is essential for realistic animation of the reconstructed model of a person based on the articulated joint structure of the generic model. We therefore require robust extraction of a set of feature points for a wide range of changes in body shape, size and clothing. To achieve this we constrain the person to stand approximately in a pre-specified pose and wear clothing which allows both the arm pits and crotch to be visible such as a shirt and trousers. Given these assumptions an algorithm has been developed for reliable extraction and localisation of a set of features based on our knowledge of the silhouette contour structure.

The algorithm for extracting feature points from the front or back silhouette contour, $C_i, i = 1, 3$, is presented in Figure 4. Initially the algorithm traverses the silhouette contour C_i to locate five extremum points, $\vec{u}_{e1-e5}$, on the contour. These correspond to the head, hands and feet as illustrated in Figure5(c). The extrema points can be reliably extracted for all silhouettes but their location varies significantly due to variation in shape and pose. Therefore, the extrema are used to identify five key feature points, $\vec{u}_{f1-f5}$, which can be accurately located even with large changes in shape and pose. The feature points correspond to the crotch, arm-pits and shoulders as shown in Figure 5(d).

This procedure gives reliable extraction of a set of key feature points for a wide range of people shape, size, clothing and hair-styles. It has been found that other potential features points such as the neck cannot be reliably localised as small changes in shape can result in a large variation in position. Resulting in a poor quality correspondence between the captured and generic model images. The set of extracted features are sufficient to accurately align the major body parts for a captured image silhouette (head, torso, arms, legs) with those of the generic model image. A similar procedure is applied for the side views to identify the tip of the nose as the left or right extremum on the head. Other body parts such as the fingers cannot be reliably identified with the image resolution used as each finger is less than three pixels across. Higher resolution images may permit correspondences to be established between additional body parts.

3.2.4 Pose estimation

Pose estimation identifies the angle of the arms and legs for a set of captured images of a specific person. This information is used to adjust the pose of the generic model to that of a particular individual. The pose of the arms, legs and head are estimated by computing the principal axis for the contour points corresponding to each of these components. If the set of contour points for a particular body part is $C_i = \{\vec{u}_j\}_{j=r}^{s}$ then the principal axis is given by:

$$\vec{u}_{axis} = \frac{1}{(s-r+1)} \sum_{j=r}^{s} \vec{u}_j^2 - (\frac{1}{(s-r+1)} \sum_{j=r}^{s} \vec{u}_i)^2 \quad (2)$$

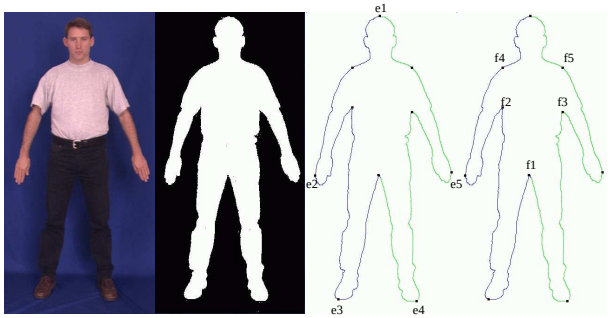
The angle of the principal axis with the vertical gives the approximate pose of the body parallel to the image plane. The body part pose is used in the mapping to correct for small variations between the generic model and the captured image set for a particular individual.

3.3. 2D-to-2D Silhouette Mapping

The objective of mapping between the generic humanoid model and the captured images is to establish a dense correspondence for mapping each model part. Dense correspondence establishes a unique one-to-one mapping between any point, $\vec{u}^M$, inside the generic model silhouette and a point on the same body part, $\vec{u}^D$, inside the captured image silhouette. This correspondence is used to modify the shape of the generic humanoid model to approximate the shape of a particular individual. For example to achieve realistic arm movement for the reconstructed model of an individual it is necessary to map the projection of the arm on the generic model image to the corresponding arm on the captured image.

1. Find the extremum points u_{e1-e5}:
   (a) Find the extremum point on the top of the head, u_{e1}, as the contour point with minimum vertical coordinate, u:
   $u_{e1} = min(\{u_j\}_{j=1}^{N_P})$ and $v_{e1} = v_j$
   (b) Find the extremum point on the left, $\vec{u}_{e2}$, and right, $\vec{u}_{e5}$, hands as the contour points with minimum and maximum horizontal coordinate, v:
   $v_{e2} = min(\{v_j\}_{j=1}^{N_P})$ and $u_{e2} = u_j$
   $u_{e5} = max(\{v_j\}_{j=1}^{N_P})$ and $u_{e5} = u_j$
   (c) Evaluate the centroid of the silhouette contour:
   $\vec{u}_C = \frac{1}{N_P} \sum_{j=1}^{N_P} \vec{u}_(i)$.
   (d) Find the extremum points on the left, $\vec{u}_{e3}$, and right, $\vec{u}_{e4}$, feet as the contour points with maximum vertical coordinate, u, either side of the centroids horizontal coordinate, v_C:
   $u_{e3} = max(\{u_j\}_{j=1}^{N_P})$ and $v_{e3} = v_j \leq v_C$
   $u_{e4} = max(\{u_j\}_{j=1}^{N_P})$ and $v_{e4} = v_j \geq v_C$
2. Find the feature points u_{f1-f5}:
   (a) Locate the key feature points corresponding to the crotch, $\vec{u}_{f1}$, and the left, $\vec{u}_{f2}$, and right, $\vec{u}_{f3}$, arm-pits as the contour points with minimum vertical coordinate, u, which are between the corresponding hand and feet extremum points:
   crotch $u_{f1} = min(\{u_j\}_{j=e3}^{e4})$ and $v_{f1} = v_j$
   left-armpit $u_{f2} = min(\{u_j\}_{j=e2}^{e3})$ and $v_{f2} = v_j$
   right-armpit $u_{f3} = min(\{u_j\}_{j=e4}^{e5})$ and $v_{f2} = v_j$
   (b) Locate feature points on the left, $\vec{u}_{f4}$, and right, $\vec{u}_{f5}$, shoulders with the same horizontal coordinate, v, as the arm-pit features u_{f2} and u_{f2}:
   left-shoulder $u_{f4} = min(\{u_j\}_{j=e1}^{e2})$ and $v_{f4} = v_{f2}$
   right-shoulder $u_{f5} = min(\{u_j\}_{j=e5}^{e1})$ and $v_{f5} = v_{f3}$

Figure 4. Algorithm for feature extraction



(a) Image (b) Silhouette (c) Extrema (d) Features

Figure 5. Silhouette and feature extraction

Body-part correspondence is established using the feature points, u_{f1-f5}, on the silhouette contours of the generic model and the captured data images. These features can be used to establish a correct correspondence for each part of the human body. Based on the five key points the human model is separated into seven functional parts: head; shoulders; left-arm; right-arm; torso; left-leg; right-leg. Separating the silhouette images into body-part allows a dense mapping to be established independently for points inside each body-part silhouette.

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 mapping based on the relative dimensions of the silhouette. This is equivalent to an 2D affine transform in the image plane (rotation,scale,shear and translation). The mapping between corresponding points inside the silhouette for a particular body part is given as follows in homogeneous coordinates:

$$\vec{u'}^{D} = S\vec{u'}^{M} \tag{3}$$

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

The components s represent the rotation, shear and scale between the parts and t translation between body parts. If the Principal axis of the body parts are aligned $s_{uv} = s_{vu} = 0$ and s_u, s_v represent the horizontal and vertical scale factors respectively. 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. The vertical and horizontal scale factors and translations are given by:

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

This mapping enables us to evaluate a unique one-to-one correspondence of points inside the data silhouette $\vec{u}_D$ for any point inside the model silhouette $\vec{u}_M$. This allows 2D information such as the colour from the captured model to be mapped to the silhouette of the generic model as illustrated in Figure 1 (f). The mapping achieves an exact correspondence at the feature points and a continuous mapping elsewhere including across boundaries between different body parts. The change in position $\Delta\vec{u}$ between model and data silhouette is given by:

$$\vec{u}^D = \vec{u}^M + \Delta\vec{u} \tag{4}$$

This 2D change in position in the image plane for a particular view can be used to estimate the change in position of a 3D point orthogonal to the view direction.

3.4. 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.

3.4.1 Single view 3D displacement

The 2D-to-2D mapping for the i^{th} view gives an estimate of the displacement of a 3D point $\vec{x}$ between the projection of the generic model $\vec{u}^M_i$ and the projection of the surface of a real person $\vec{u}^D_i$. The 2D image plane displacement, $\Delta\vec{u}_i$, defined by equation 4, can be used to estimate the 3D displacement component, $\Delta\vec{x}_i = (\Delta x_i, \Delta y_i, \Delta z_i)$, of the projected 3D point, $\vec{x}$, on the generic model orthogonal to the i^{th} image view direction. This is achieved by estimating the inverse projection of the displacement of the 2D point $\Delta\vec{u}_i$ in the camera image. The inverse projection can be estimated uniquely from our knowledge of the distance to the corresponding 3D point, $\vec{x}^M$, on the generic model. This approximates the unknown distance to the corresponding 3D point, $\vec{x}^D$, on the captured person which we want to estimate. The distance to the generic model is a reasonable approximation as the distance between the camera and person, $(\approx 3m)$, is large relative to the difference in 3D surface position, $(\approx 0.1m)$, between the model and person.

Estimation of the 3D displacement component orthogonal to the view direction is illustrated in figure 6(a). A single view image I^D_i gives an approximation of the component of 3D displacement $\Delta\vec{x}_i$ of a known 3D point $\vec{x}^M$ on the generic model orthogonal to the i^{th} view direction $\vec{n}_i = (n_x, n_y, n_z)$ such that $n_i \bullet \Delta\vec{x}_i = 0$. For example for the front image the view direction normal $\vec{n}_0 = (0, 0, 1)$ consequently the displacement of an image point $\Delta\vec{u}_0 = (\Delta u, \Delta v)$ gives an approximation of the 3D displacement of the corresponding point $\Delta\vec{x}_0 = (\Delta x_0, \Delta y_0, 0)$. Similarly a left side view with view normal $\vec{n}_1 = (1, 0, 0)$ gives an estimate of the corresponding 3D displacement $\Delta\vec{x}_1 = (0, \Delta y_1, \Delta z_1)$.

A point in the 2D camera image plane corresponds to an infinite ray in 3D space. Therefore, inverting the camera projection equation 1 gives the equation of a line in 3D space. From equation 1 we obtain the inverse projection in homogeneous coordinates for view direction i:

$$\begin{aligned} \vec{x'}^D &= \vec{x'}^M + \Delta x_i \\ &\approx \lambda_i(\vec{x}^M) E_i^{-1} P_{orth}^{-1} M^{-1} (\vec{u'}_i^M + \Delta \vec{u'}_i) \end{aligned} \quad (5)$$

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 is orthogonal to the camera view direction $\vec{n}_i$. The inverse camera calibration and transform matrices are given by:

$$M^{-1} = \begin{bmatrix} \frac{1}{f_u} & 0 & -\frac{o_u}{f_u} \\ 0 & \frac{1}{f_v} & -\frac{o_v}{f_v} \\ 0 & 0 & 1 \end{bmatrix}$$

$$E_i^{-1} = \begin{bmatrix} R_i^{-1} & -R_i^{-1}\vec{t}_i \\ \vec{0}^T & 1 \end{bmatrix}$$

Thus the 3D point on the model in real coordinates $\vec{x}^D$ is on the 3D line represented by:

$$\Delta \vec{x}_i \approx \lambda_i(\vec{x}^M) R_i^{-1} M^{-1} (\vec{u}^M + \Delta \vec{u}) - R^{-1}\vec{t} - \vec{x}^M \quad (6)$$

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 $\lambda_i(\vec{x}^M) = ||R_i\vec{x}^M + \vec{t}_i||$. Equation 6 gives an approximation of the 3D displacement component $\Delta\vec{x}_i$ orthogonal to the view direction $n_i \bullet \Delta\vec{x}_i = 0$. This 3D displacement component can be evaluated for each 3D vertex $\vec{v}_j$ on the generic model.

Applying the displacement component to each vertex on the 3D generic model results in an affine transform of the 3D model orthogonal to the view direction. Reprojecting the modified model results in a silhouette which approximates the captured silhouette shape. The surface shape for 3D points whose projection is inside the silhouette is a 2D affine transform of the shape of the generic model.

3.4.2 Multi-view 3D displacement

Combining the estimated displacement components from two or more orthogonal views of a point $\vec{x}$ we obtain an estimate of the 3D displacement $\Delta\vec{x}$. Estimation of the 3D displacement by combining components from multiple views is illustrated in Figure 6(b). Displacement components from multiple views can be combined by averaging to estimate the 3D displacement:

$$\Delta\vec{x} = \begin{bmatrix} \frac{1}{N_x}\sum_{i=0}^{N_x} \Delta x_i \\ \frac{1}{N_y}\sum_{i=0}^{N_y} \Delta y_i \\ \frac{1}{N_z}\sum_{i=0}^{N_z} \Delta z_i \end{bmatrix} \quad (7)$$

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:

$$\vec{x}^D = \vec{x}^M + \Delta\vec{x} \quad (8)$$

The generic model shape is modified by estimating the 3D displacement $\Delta\vec{x}(\vec{v}_j)$ for each vertex $\vec{v}_j$. The model vertex position, $\vec{v}_j$, is projected to the 2D image plane using the camera model equation 1 to obtain the 2D coordinates $\vec{u}_i^M(\vec{v}_j)$. This point is then mapped to a corresponding point on the captured data image $\vec{u}_i^D(\vec{v}_j)$ using equation 3. The 2D displacement $\Delta\vec{u}_i(\vec{v}_j)$ is then used to estimate the corresponding 3D displacement component orthogonal to the view direction $\Delta\vec{x}_i(\vec{v}_j)$ from equation 5. The 3D displacement components from each view are then combined using equation 7 to estimate the 3D displacement $\Delta\vec{x}(\vec{v}_j)$ for a vertex on the model.

Applying the estimated vertex displacement to all vertices on the generic model results in a modified model which approximates the shape of a specific person. The modified 3D model for a particular individual will produce silhouette images with approximately the same shape as the captured image silhouettes for each view direction. Points that do not project to the silhouette boundary are subject to a linear transform based on the position of the corresponding point on the 3D model. For example con-cavities on the generic humanoid model are scaled according to the 2D affine mapping to produce con-cavities in the final model. The resulting modified 3D model shape is a first order linear approximation of the shape of a particular person based on the local surface shape information represented in the 3D generic humanoid model.

3.5. Colour texture mapping

The 2D-to-2D mapping for a single view enables the colour texture of the captured data image to be mapped onto the projected model image as illustrated in Figure 1(f). For all points, $\vec{u}^M$, inside the projected generic model we know the corresponding point on the captured data image $\vec{u}^D$. To texture map the 3D model we project each vertex $\vec{x}_i^M$ to obtain the image coordinates $\vec{u}_i^M$ and from the 2D-to-2D mapping obtain the corresponding data image coordinates $\vec{u}_i^D$. The 2D data coordinate value is then the texture coordinate for the modified 3D model vertex, $\vec{x}_i^D$.

For each body-part we obtain a single cylindrical texture map by back projecting and integrating the four overlapping images as in previous work on face modelling [1, 10, 16]. Integration of the texture map is based on the approximate 3D shape information for the reconstructed model. As the reconstructed model is only an approximation of the 3D

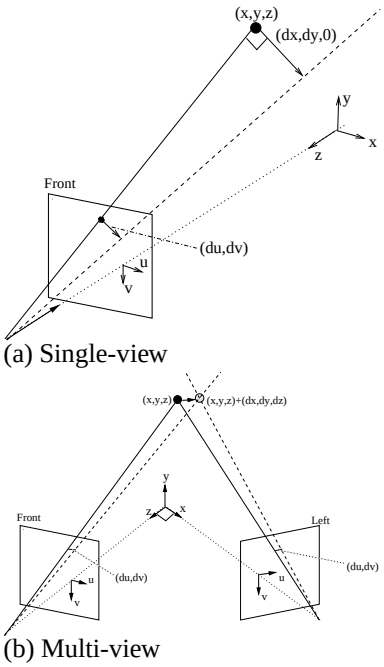


Figure 6. 3D Displacement estimation

shape of a particular person based on the orthogonal view silhouette outlines the overlapping images do not exactly correspond. However, the approximate shape information has been found to be sufficient to obtain a reasonable integrated texture map by blending overlapping regions between the front, sides and back. In addition, for some parts of the body no image information is available from the side views due to occlusion such as the torso resulting in a jump in the texture map. Figure 7 shows examples of integrated texture maps obtained for several body parts. All four views are integrated for the head and shoulder texture maps. For the torso only the front and back images are used as the sides are not visible resulting in a discontinuity in the image texture.

4. Results

The model-based reconstruction algorithm has been used to capture models of approximately twenty individuals wearing a variety of clothing. Subjects were constrained to wear clothing such as trousers and shirt which allows the location of the arm pits and crotch to be visible in the front view. Having captured a set of images the reconstruction algorithm was applied to automatically build a model of a particular individual without any manual intervention. The model-based reconstruction from silhouettes is applied to all body parts except the hands and feet. Hands and feet are modelled by scaling the corresponding part of the generic model as there is insufficient information on the silhouette

Figure 7. Integrated texture maps

images to identify feature points.

Reconstructed 3D models for three individuals are shown in Figure 8. The left-hand column Figure 8(a) shows the original 756×582 colour photo captured from the front view. Columns Figure 8(b) and (c) show the reconstructed 3D model rendered and colour texture mapped. These results demonstrate that the automatic reconstruction generates a recognisable 3D facsimile of the real person. Some artifacts can be seen in the shape near the feet due to poor segmentation of the feet from the legs in the front view. The 3D shape approximation is of sufficient accuracy to give a recognisable model when texture mapped with the image colour information. Further examples of reconstructed models for male and female subjects are presented in Figure 9. Currently the principal limitation of the reconstruction from silhouettes is the quality of the face models generated. The absence of feature point labelling results in misalignment of the face image with the generic model. Previous feature based approaches to face modelling [2, 10] could be used to improve face modelling if full resolution face images were captured. However, current techniques for face modelling may also fail to automatically reconstruct face shape in the presence of hair and glasses. Figure 10 shows reconstructions for the same person wearing different clothing. This example demonstrates that the approach can be used to generate set of models for a particular individual suitable for multiple virtual world applications (business,sports,leisure).

Animation of the reconstructed 3D models for particular individuals performing standard movements is illustrated in Figure 8(d). Figure 11 shows a simple virtual catwalk scene with several animated virtual people. The articulated structure of the generic humanoid model is modified for a particular individual by mapping the 3D joint positions using the 2D-to-3D mapping algorithm presented in the previous section. The animation parameters based on joint angle interpolation are the same as for the generic model. Animation

of movements such as walking, running and jumping using a common set of parameters results in reasonable movements of a particular individual for VR applications. Currently the VRML animation is based on a rigid 3D model resulting in visible artifacts. A more sophisticated seamless humanoid model which incorporates skin and clothing deformations is required to realistically animate movements of people with a wide variety of clothing.

5. Conclusions

A model-based approach has been introduced for automatic reconstruction of an articulated 3D colour model of a particular person from a set of colour images. Results demonstrated that this approach achieves recognisable models of individuals with a wide range of shape, size and clothing. The reconstructed model represents the 3D shape, colour texture and articulation structure required for animating movements. This approach enables low-cost capture of models of people in a VRML H-Anim avatar format suitable for populating virtual worlds.

The results presented demonstrate the feasibility of model-based reconstruction of realistic representations of individual people from sets of images. However, several issues should be addressed in future work to improve on the current system:

- Facial feature point labelling for accurate modelling and animation [2, 10].
- Capture of a person in multiple poses for accurate reconstruction of kinematic structure [4, 6].
- High-resolution image acquisition for improved photorealism.
- Synchronous image acquisition from multiple views to avoid movement.
- Increased number of views to reduce occlusion.
- Multiple levels-of-detail for efficient shape represention.
- Seamless VRML model for improved animation quality.

Further development of this system and integration with previous work on face and body modelling will give incremental improvements in the quality of the reconstructed models. The results presented in this paper demonstrate the potential of a low-cost whole-body system for capturing recognisable 3D models of individual people from sets of colour images.

References

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

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

[3] M. Escher and N. Magnenat-Thalmann. Automatic 3D Cloning and Real-Time Animation of a Human Face. Technical report, MIRALab, University of Geneva, Switzerland, 1997.

[4] 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.

[5] J. Gu, T. Chang, I. Mak, S. Gopalsamy, H. Shen, and M. Yuen. A 3D Reconstruction System for Human Body Modeling. In *Modelling and Motion Capture Techniques for Virtual Environments - Magnenat-Thalmann,N. and Thalmann,D. (Eds.)*, pages 229—240, 1998.

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

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

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

[9] T. Kurihara and A. Kiyoshi. A transformation method for modeling and animation of the human face from photographs. *State of the Art In Computer Animation, Springer*, pages 45–57, 1990.

[10] 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.

[11] Y. Lee, D. Terzoupoulos, and K. Walters. Realistic modeling for facial animiation. In *SIGGRAPH*, pages 55—62, 1995.

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

[14] P.Fua and C.Miccio. From Regular Images to Animated Heads: A Least Squares Approach. Technical report, Computer Graphics Lab (LIG), EPFL, Lausanne, Switzerland, 1998.

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

[16] G. Sannier and N. Magnenat Thalmann. A user-friendly texture-fi tting methodolgy for virtual humans. In *Computer Graphics International*, pages 167—176, 1997.

[17] D. Terzopoulos. From physics-based representation to functional modeling of highly complex objects. In *NSF-ARPA Workshop on Object Representation in Computer Vision*, pages 347—359, 1994.

(a) Photo (b) 3D Model (c) 3D Colour Model (d) Running

Figure 8. Reconstructed 3D texture mapped models for individual people with animated movements

Figure 9. Examples of reconstructed virtual people

Figure 10. Models reconstructed for the same person with different clothing

Figure 11. Virtual people in a virtual catwalk scene animation