Human Shape Estimation in a Multi-Camera Studio

J.Starck, A.Hilton and J.Illingworth
Centre for Vision, Speech and Signal Processing
University of Surrey
Guildford, GU2 7XH, UK
j.starck@eim.surrey.ac.uk

Abstract

This paper addresses the problem of estimating the shape of an actor in a multi-camera studio for arbitrarily positioned cameras and arbitrary human pose. We adopt a seamless articulated mesh model and introduce a novel shape matching technique to automatically transform the projected shape of the model to match multiple captured image silhouettes. Our approach treats the projected mesh as a deformable model and constrains the model to follow a smooth shape transformation to match each image silhouette. Multiple 2D transformations are integrated in 3D to update the shape of the model to match the actor. We assess the technique using virtual views generated for 3D scanned human data-sets and present preliminary results in a studio.

1 Introduction

There is increasing interest in the use of photo-realistic models of people for applications such as advertising, computer games, video conferencing, user-interface agents, virtual environments and clothing retail. Our interest lies in the generation of 3D virtual humans for 3D-TV, the broadcast of 3D virtual television productions. Current virtual production makes use of actors filmed against a chroma-key background, and composites real and virtual scene elements to provide a 2D programme. The extension to production and transmission of full 3D content is now under investigation [12].

3D-TV ultimately requires "photo-realistic" 3D content to be acceptable as an alternative to conventional 2D video. The creation of highly realistic 3D animated models of humans is however a time intensive task requiring the skills of experienced designers and animators. Our goal is to develop a technique to automatically acquire animated models of actors suitable for 3D broadcast, within a production studio. In this paper we present a method to automatically deform a model to the shape of an actor using multiple calibrated cameras in a conventional virtual studio.

Stereo has been used to reconstruct the shape of the head [5] and whole-body [10] from short baseline camera images. A 51 camera set-up was used in the "Virtualised Reality" project to capture static 3D models of people in a studio [10]. Stereo matching can provide dense 3D information across the body given sufficient texture information, but can be restricted in terms of accuracy and robustness. Shape from silhouette has also

BMVC 2001 doi:10.5244/C.15.59

been used for human model reconstruction and coarse volumetric models have been generated from the volume intersection of 5 [2] and 6 [4] image silhouettes in multi-camera studios. The reconstruction of animated whole-body models from image silhouettes has been introduced by Hilton et. al. [8]. This approach is currently limited to orthogonal silhouette images and a predefined subject pose.

This paper extends current approaches to recovering animated models of people from image silhouettes [8] to both arbitrary pose and multiple, arbitrarily positioned cameras in a multi-camera studio. The actor foreground is accurately segmented from camera images using chroma-keying techniques. An animated seamless generic model is then sized and posed to match the actor using a limited set of manually defined feature points in the images. A novel 2D shape transformation process is then used to automatically match the generic model to the image silhouettes. Finally, the 2D matching is integrated in 3D to change the shape of the generic model to approximate the actor. The shape reconstruction is assessed using virtual views generated for scanned 3D human data and preliminary results are presented in a three camera studio.

2 Model Initialisation

The aim of our work is a fully automated process for generating models of actors in a studio. We start here with a manual method of initialising a generic model in order to apply the proposed automatic shape estimation technique. The problem of manually initialising a human body model in an arbitrary pose from multi-view calibrated images has been addressed in the human motion tracking literature [1]. Here we use a similar approach to match an articulated humanoid body model to images of an actor in the studio.

2.1 Articulated Generic Humanoid Model

Standardised methods of representing 3D human models are provided by the VRML Humanoid Animation Working Group (H-Anim) [15]. A draft H-Anim 1.2 format seamless model is adopted that can be visualised with any VRML-97 compliant browser and animated using JAVA. The model consists of a single seamless mesh defining the 3D body shape, attached to an underlying skeleton structure with 17 joints to synthesise the gross movements of the body and texture map(s) attached to change appearance.

The model state is defined by the translation, limb-lengths, and pose of the skeleton. The limb-lengths are constrained to remain symmetric and 9 degrees-of-freedom (dof) are introduced for the dimensions of the skeleton. Joint rotations are defined as 1, 2 or 3dof according to the principal rotations of the anatomical joint [7] and 28dof are introduced to generate the pose of the model. We make use of the exponential map representation of rotations to avoid the singularity, or gimbal-lock, inherent in Euler angles and the non-linear unit-length constraint necessary for quaternions [6]. The combined skeleton state $\underline{\phi}$ has a total of 40dof.

2.2 Location, Pose and Dimension Estimation

A simple user interface has been developed to allow a user to select points on the generic model mesh or skeleton and define the corresponding feature points in multiple images captured in a studio. In practice we make use of the 17 skeleton joints and limb-tips where

visible in the images. We then minimise the image plane error between the projected location of the model points and the image features with respect to the model state $\underline{\phi}$. The cost function for minimisation is given by Equation 1 where $\underline{e}_{ij}$ is the image plane error for the ith feature in the jth image. An ideal pin-hole camera model is used through-out to define the image plane projection.

$$min \sum_{i,j} \underline{e}_{ij}(\underline{\phi})^T \underline{e}_{ij}(\underline{\phi}) \tag{1}$$

The function is minimised using a bound constrained, BFGS non-linear solver [16]. Bounds on pose are applied according to the expected rotation limits as presented by Grosso et. al. [7]. Bounds on dimensions are applied according to the anthropometric extremes of the population as characterised by the the 5^{th} percentile British female and the 95^{th} percentile British male [11].

3 Automatic Shape Estimation

In this section we describe a method to automatically change the shape of a generic model to match multiple image silhouettes of a person. It is assumed at this stage that the model has the approximate dimensions of the subject and is located and posed to match the subject in the images. The problem is treated first as a 2D shape transformation task to match the projected model to the silhouette in each image. The new 2D vertex locations in each image are then integrated in 3D to derive a new 3D shape for the generic model.

3.1 2D Shape Transformation

Previous approaches to matching a generic model to silhouettes of people are based on defined feature correspondences between the model silhouettes and the image silhouettes [8]. A limited set of features can be reliably extracted in the presence of varying shape and clothing for a specific pose and orthogonal camera views [8]. However, reliable feature matching cannot be achieved across arbitrary poses and camera positions. The problem we address here is to automatically define the 2D shape transformation without the need for explicit extraction of features. To achieve this we treat the projected model as a deformable mesh and constrain the mesh to follow a 2D shape morphing process between the projected model silhouette and the apparent image silhouette. Here we make the assumption that a smooth 2D deformation is likely to produce the most pleasing transformation of the model to match each image silhouette.

3.1.1 Deformable Model

Terzopoulos et. al. [13] introduced fitting of deformable models to images to recover shape and non-rigid motion and this has since gained widespread interest as a data fitting approach. A deformable model acts as an elastic body and deforms dynamically in response to applied forces and constraints derived from the data. We treat the projected generic mesh as a 2D dynamic deformable model and in Section 3.1.2 we define the applied constraints to deform the mesh to match the image silhouettes.

The dynamic behaviour of the deformable model is described by Equation 2 where the first term represents internal forces due to mass, the second term damping, third term

elasticity and the final term the external forces on the model derived from the data. The finite difference method can be applied to obtain an approximate solution, whereby the continuous model is treated as a set of discrete points at the projected mesh vertices. The dynamics can also be simplified by taking zero mass m_i and unit damping λ_i at each vertex i. The finite difference equation for the deformation of each vertex position x_i is then given by Equation 3 where f_i is the external force on each vertex and g_i is the internal elastic force on each vertex.

$$m(\underline{x})\frac{d^2\underline{x}}{dt^2} + \lambda(\underline{x})\frac{d\underline{x}}{dt} + g(\underline{x}) = f(\underline{x}) \tag{2}$$

$$\frac{d\underline{x}_i}{dt} = f_i - g_i \tag{3}$$

We treat each edge in the projected mesh as a spring that applies a restoring force to the mesh vertices to preserve the original projected edge length. The deformable model will then attempt to preserve the original projected geometry of the generic model under deformation. The spring force s_{ij} at each vertex i exerted by the connection to vertex j is given by Equation 4 where k_{ij} is the spring stiffness and l_{ij} is the original projected edge length. We assign all connections equal stiffness k and the total elastic force on each vertex i is given by Equation 5.

$$s_{ij} = k_{ij}\left(\|\underline{x}_i - \underline{x}_j\| - l_{ij}\right) \times \frac{\left(\underline{x}_i - \underline{x}_j\right)}{\|\underline{x}_i - \underline{x}_j\|} \tag{4}$$

$$g_i = \frac{k}{n}\sum_{j=1}^{n} s_{ij} \tag{5}$$

3.1.2 Silhouette Morphing

The deformable model paradigm makes use of external forces derived from data to evolve the model to fit the data. Kakadiaris et. al. [9] used deformable models to estimate body segment shape from image silhouettes using external forces derived from a modified nearest neighbour algorithm. Here we make use of a single deformable model for the whole body and avoid the necessity for explicit matching by constraining the model to follow a smooth transition between the projected model shape and the target image silhouette shape. We treat the problem as a 2D shape morphing task and apply the shape transformation as a hard constraint on the deformable model.

Shape morphing techniques have received a great deal of interest as a visual effects tool and here we make use of a technique introduced by Turk and O'Brien [14], based on variational implicit surfaces, to provide a smooth transition between 2D shapes. This approach transforms between 2D shapes by treating the shape contours as slices through a single 3D surface. The 2D shapes are stacked on parallel planes in 3D and a single 3D implicit function is generated using the 2D contours as constraints.

The 3D implicit function $h(\underline{x})$ is given in Equation 6 [14], where $\underline{x}$ is an image plane position augmented by an out-of-plane component to stack the model and image contours in 3D. The function consists of a summation of weighted radial basis functions $\psi(\underline{x})$ placed at constraint points $\underline{c}_i$ on the contours and internal points of the 2D shapes. The implicit function is assigned a value of zero at the contour constraints and a value of one

at the internal constraints. The zero value isosurface $h(\underline{x}) = 0$ then provides a 3D surface that blends between the 2D shape of the model and the 2D shape of the image silhouette as shown in Figure 1(b). The 3D thin-plate radial basis function $\psi(\underline{x}) = \|\underline{x}\|^3$ is used to minimise the thin-plate energy in 3D and provide a globally smooth transformation. The implicit function has a degree one polynomial $P(\underline{x})$ term to span the constant and linear portions of the shape transformation.

$$h(\underline{x}) = \sum_{i=1}^{N} d_i \psi(\underline{x} - \underline{c}_i) + P(\underline{x}) \tag{6}$$

We generate the implicit function by selecting a number of boundary points, evenly spaced around each silhouette, and for each boundary point we also select an interior constraint point directed along the inner normal of the contour [14]. We obtain satisfactory results using 256 points around each boundary and stacking the silhouettes at a distance of 10% of the subject image height in 3D. Given the value $h(\underline{c}_i)$ for each constraint $\underline{c}_i$, either one or zero, we can form a linear system of equations in the unknowns d_i and p_j as given in Equation 7. This is a symmetric, positive semi-definite linear system that we solve via LU decomposition to derive the implicit function.

$$\begin{bmatrix} \psi(\underline{c}_1 - \underline{c}_1) & \dots & \psi(\underline{c}_1 - \underline{c}_N) & 1 & \underline{c}_1^T \\ \vdots & & \vdots & 1 & \vdots \\ \psi(\underline{c}_N - \underline{c}_1) & \dots & \psi(\underline{c}_N - \underline{c}_N) & 1 & \underline{c}_N^T \\ 1 & 1 & 1 & 0 & \underline{0}^T \\ \underline{c}_1 & \dots & \underline{c}_N & 0 & \underline{0}^T \end{bmatrix} \begin{bmatrix} d_1 \\ \vdots \\ d_N \\ p_0 \\ p_1 \\ p_2 \\ p_3 \end{bmatrix} = \begin{bmatrix} h(\underline{c}_1) \\ \vdots \\ h(\underline{c}_N) \\ 0 \\ 0 \end{bmatrix} \tag{7}$$

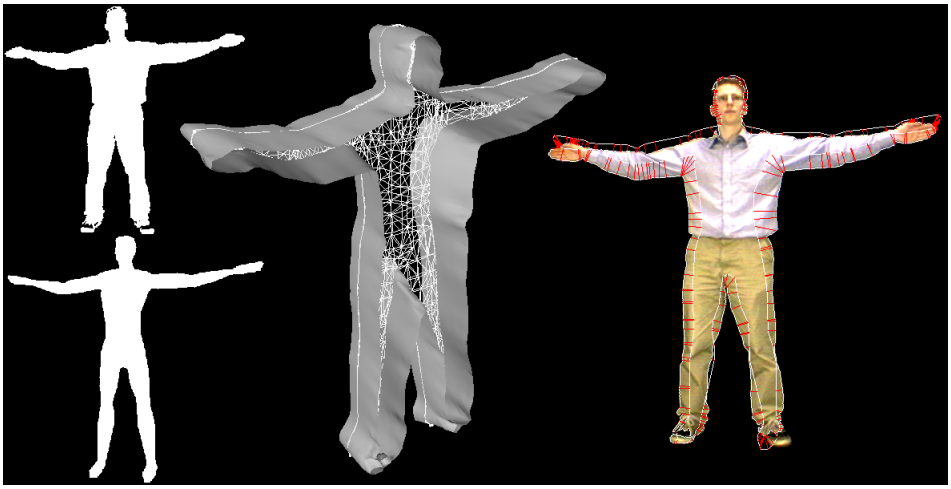
3.1.3 Morphing the Deformable Model

We assume that the vertices forming the model silhouette contour lie on the contour of the target image silhouette and we use the 3D implicit function as a hard constraint on our projected 2D deformable model. The model starts on the model shape plane in the 3D function, we then iteratively move the model between the model plane and the image silhouette plane as a 2D object by applying an equal external force at each vertex, directed normal to the planes in 3D. The contour vertices of the deformable model are constrained to move only along the isosurface of the implicit function. The resultant force $f_i - g_i$ at a contour vertex i is projected onto the tangent plane of the isosurface to give the direction of movement for the vertex $n_i(\underline{x})$ as given in Equation 8

$$n_i(\underline{x}) = (f_i - g_i) - \frac{((f_i - g_i) \cdot \nabla h(\underline{x}))\, \nabla h(\underline{x})}{\|\nabla h(\underline{x})\|^2} \tag{8}$$

The model moves at constant velocity under the applied external forces and deformation is terminated at the image silhouette plane. The isosurface provides only a smooth approximation to the image silhouette and we update contour vertices by searching along the gradient of the implicit function $\nabla h(\underline{x})$ projected onto the image silhouette plane, to find the closest actual silhouette contour point. The matched contour points are again

treated as hard constraints and the deformable model is relaxed with no applied external forces to update the position of internal vertices and release unmatched contour vertices arising from inaccuracies in the iterative morphing process. Figure 1(c) shows the match between the generic model and the target shape. The approach automatically provides contour matches based on the similarity in shape between the projected model and the image silhouette.



(a) Image and model shape (b) Implicit surface (c) Contour matches

Figure 1: Shape transformation of a 2D deformable model based on a 3D variational implicit surfaces

3.2 3D Shape Estimation

The 2D shape transformation process provides a means of updating the projected 2D vertex locations of the generic model based on the shape information contained in the image silhouettes. Where shape information is lacking across occluded regions of the body, the deformable model preserves the original relative geometry of the generic model. The transformed 2D vertex locations are integrated in 3D to minimise the reprojection error between the 3D mesh and the image plane locations of the vertices. We currently treat each vertex independently and simply minimise the image plane error between the projected vertex positions and the transformed 2D locations. Each vertex is adjusted using a Gauss-Newton non-linear solver to minimize the cost function as already given in Equation 1 where the state $\underline{\phi}$ now represents the 3D vertex locations.

4 Results

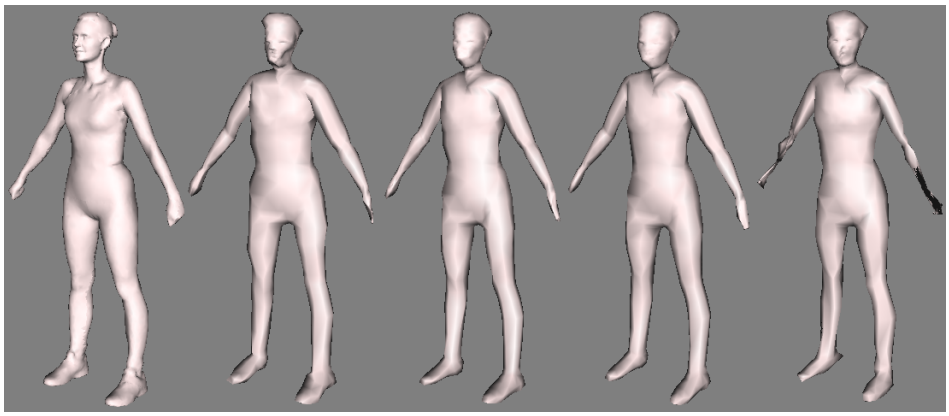
We test the proposed shape estimation technique using virtual views generated for 3D human data-sets provided by Cyberware. The models are positioned at the centre of a simulated studio with cameras located at a 3m radius around the centre on a horizontal

plane at 1m height from the floor. Ideal pin-hole camera models are used with NTSC resolution (720x486 pixels) and typical intrinsic parameters (focal length fx=680, fy=625 pixels), where the models form less than 10% of each total image area. The models are rendered for two orthogonal virtual cameras in the studio and either 6 or 8 cameras equally spaced around the centre, with the model orientated to face one camera. These images are then used to test the shape estimation of the original 3D data-sets.

We assess the error between the estimated shape of the 3D data-sets from the images and the orginal data using the RMS error between the two surfaces, calculated using the Metro tool [3]. Table 1 gives the RMS and maximum errors between the 3D data-sets and the posed generic model and the estimated shape for 2, 6 and 8 camera views. The error measurements show a reduction in the error for the shape estimation technique compared to the posed and scaled generic model, a 5mm error here approximately corresponds to a 1pixel reprojection error in the images.

RMS error /mm (max. error)	Generic Model	2 Cameras	6 Cameras	8 Cameras
Alison	19.9 (64.5)	12.2 (58.5)	12.4 (53.7)	14.5 (57.3)
Brian	32.0 (91.6)	21.2 (81.8)	16.3 (56.3)	18.9 (70.8)
Eric	19.0 (74.0)	14.6 (66.6)	12.0 (69.1)	13.2 (54.4)
John	27.9 (93.7)	20.0 (70.7)	14.8 (55.3)	18.1 (72.4)
Tammy	19.8 (62.7)	13.7 (50.8)	11.0 (46.5)	11.4 (43.8)

Table 1: Shape estimation error between the generic model and Cyberware 3D data-sets



(a) Scanned data (b) 2 views (c) 6 views (d) 8 views (e) Incorrect pose

Figure 2: Reconstructed model for Cyberware “Tammy”

It is interesting to see that a good shape estimate can be obtained using only 2 orthogonal camera views, as demonstrated by Hilton et. al. [8]. There is some additional shape information from other views and a small improvement in the visual quality of the models

as shown in Figure 2. Figure 2(e) also demonstrates the dependence of the technique on the pose of the model, a 30 degree error is introduced in the vertical rotation of the body leading to incorrect matches of the arms to the image silhouettes.

Although we have presented the errors across the estimated shape of the Cyberware models it is not our goal to obtain an accurate 3D reconstruction. We seek an approximate estimate of the shape of a subject to which we can apply texture from the images to give "photo-realism" to our models. Our technique provides a smooth transformations of our generic model that minimises the reprojection error for the multi-view image silhouettes. Figure 3 illustrates the approximate shape generated for the Cyberware subject "Brian" and the animation of the model with texture from the images. Figure 4 shows preliminary shape estimation results for a subject in a studio with an arbitrary pose, captured from 3 front-facing cameras at 3m from a subject with approximately 1m spacing. It is our intention to extend this technique to more camera views surrounding a subject in a studio.

(a) Cyberware "Brian" (b) 8 view shape (c) Textured model

Figure 3: Generating an animated model from 8 synthetic camera views

5 Conclusions

This paper describes a technique for estimating the shape of an actor in a multi-camera calibrated studio. We currently make use of a manual interface to initialise a generic humanoid model to match the actor pose and anthropometric dimensions. We then introduce a new 2D shape matching process to automatically transform the projected shape of the generic model to match each captured image silhouette. We treat the projected model as a deformable model that is constrained according to a smooth shape transformation between the model shape and each image silhouette. The updated 2D vertex locations of the generic model are used to change the 3D positions of the model vertices to minimise the reprojection error in the images. The result is an animated humanoid model that approximates the shape of the actor in the studio. This technique allows us to estimate the shape of an actor from multiple image silhouettes without the requirement for specific feature

Figure 4: Estimating shape in a 3 camera studio

matches, predefined pose or specific camera views. We have found that we are able to obtain reasonable shape estimates in our artificial test cases and preliminary tests in a 3 camera studio. Further work is required to improve the integration of the transformed vertex locations in 3D based on surface smoothness of the 3D model and to incorporate uncertainty in the 2D shape transformation for different viewpoints and poses. We must also address specific feature matching in order to apply model texture for highly detailed areas of the body such as the face and hands.

6 Acknowledgements

This work is supported by EPSRC Grant GR/M88075 and sponsored by the BBC and BT. The authors would like to thank Cyberware for the provision of human 3D scanned data-sets.

References

[1] C. Barron and I.A. Kakadiaris. Estimating anthropometry and pose from a single image. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 669–676, 2000.

[2] G.K.M. Cheung, T. Kanade, J.Y. Bouguet, and M. Holler. A real time system for robust 3D voxel reconstruction of human motions. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 714–720, 2000.

[3] P. Cignoni, C. Rocchini, and R. Scopigno. Metro: measuring error on simplified surfaces. *Computer Graphics Forum*, 17(2):167–174, 1998.

[4] L. Davis, E. Borovikov, R. Cutler, D. Harwood, and T. Horprasert. Multi-perspective analysis of human action. In *Proceedings of the 3rd International Workshop on Cooperative Distributed Vision*, Kyoto, Japan, 1999.

[5] P. Fua. Using model-driven bundle-adjustment to model heads from raw video sequences. In *IEEE International Conference on Computer Vision*, 1999.

[6] F.S. Grassia. Practical parameterization of rotations using the exponential map. *The Journal of Graphics Tools*, 3(3), 1998.

[7] M.R. Grosso, R. Quach, E. Otani, J. Zhao, S. Wei, P.H. Ho, J. Lu, and N.I. Badler. *Anthropometry for computer graphics human figures. Technical Report MS-CIS-89-71*. University of Pennsylvania, Dept. of Computer and Information Science, Philadelphia, PA, 1989.

[8] 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*, May 1999.

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

[10] T. Kanade, P. Rander, and P.J. Narayanan. Virtualized reality: Constructing virtual worlds from real scenes. *IEEE Multimedia*, 4(1):34–47, 1997.

[11] S.T. Pheasant. Anthropometry estimates for british civilian adults. *Ergonomics*, 25(11):993–1001, 1982.

[12] M. Price and G.A. Thomas. 3D virtual production and delivery using MPEG-4. In *International Broadcasting Convention, IBC 2000*, Amsterdam, September 2000.

[13] D. Terzopoulos, A. Witkin, and M. Kass. Constraints on Deformable Models: Recovering Shape and Nonrigid Motion. *Artificial Intelligence*, 36(1):91–123, 1988.

[14] G. Turk and J.F O'Brien. Shape transformation using variational implicit functions. In *SIGGRAPH Conference Proceedings*, pages 335–342. ACM SIGGRAPH, August 1999.

[15] VRML Humanoid Animation Working Group, http://www.h-anim.org/. *H-ANIM 1.1 Specification*, 1999.

[16] C. Zhu, R. H. Byrd, and J. Nocedal. L-BFGS-B: Algorithm 778: L-BFGS-B, FORTRAN routines for large scale bound constrained optimization. *ACM Transactions on Mathematical Software*, 23(4):550–560, 1997.