Automatic Creation of Three-Dimensional Avatars

Maria-Cruz Villa-Uriol*, Miguel Sainz, Falko Kuester and Nader Bagherzadeh

Visualization and Interactive Systems Group
Department of Electrical and Computer Engineering
University of California, Irvine

ABSTRACT

Highly accurate avatars of humans promise a new level of realism in engineering and entertainment applications, including areas such as computer animated movies, computer game development interactive virtual environments and tele-presence. In order to provide high-quality avatars, new techniques for the automatic acquisition and creation are required. A framework for the capture and construction of arbitrary avatars from image data is presented in this paper. Avatars are automatically reconstructed from multiple static images of a human subject by utilizing image information to reshape a synthetic three-dimensional articulated reference model. A pipeline is presented that combines a set of hardware-accelerated stages into one seamless system. Primary stages in this pipeline include pose estimation, skeleton fitting, body part segmentation, geometry construction and coloring, leading to avatars that can be animated and included into interactive environments. The presented system removes traditional constraints in the initial pose of the captured subject by using silhouette-based modification techniques in combination with a reference model. Results can be obtained in near-real time with very limited user intervention.

Keywords: 3D avatars, human body reconstruction, virtual environments

1. INTRODUCTION

Realistic avatars are becoming increasingly important in many applications ranging from virtual reality and teleconferencing environments to computer animated movies and games. With consumer-level graphics boards readily reaching speeds of tens of millions of texture mapped triangles per second, ever more detailed representations become possible. Provided with this new level of rendering performance, the acquisition of accurate and visually attractive avatars is crucial.

Traditionally, the acquisition of three-dimensional avatars from real humans relied on the utilization of body suits, optical or magnetic markers, or complex and expensive three-dimensional scanners. In general, entertainment companies, especially those dedicated to computer games and computer-animated movies, are the ones that have invested into and leveraged from these techniques. Some of the main drawbacks that these capture systems introduce are equipment cost, portability and the required degree of human intervention to ensure compelling visual results. Different techniques have been investigated by the computer vision community, aimed at reconstructing human body models from video sequences or sets of static images. However, the primary focus of these techniques generally was on motion analysis and tracking whereas reconstruction of realistic avatars was only a secondary objective.

While the objective of the presented work is the creation of photorealistic avatars from still images or video sequences, our data processing pipeline shares common problems that needs to be addressed such as the estimation of the pose of the captured character. Although some authors assume that the initial pose is known[1,2], and others manually adjust a simple skeleton to the input data[3,4], in general, some sort of manual user intervention is necessary to specify key features

*mvillaur@ece.uci.edu; phone 1 949 824-2481; fax 1 949 824-3779; http://vis.eng.uci.edu/~mvillaur; Visualization and Interactive Systems Group, Electrical and Computer Engineering Department, University of California Irvine, Irvine, CA, USA 92697-2625

that control the process[1,2,3,4]. Solving the problem of pose estimation is particularly useful in applications related to the study of gesture, biomechanics, anthropometry and human gait analysis.

Different techniques have been developed for body pose recovery from image data. Frequently, these methods use geometrically defined models of humans as a reference. Moeslund and Granum[5] present a classification describing how different systems incorporate knowledge about humans into their processing. These systems can be grouped into three main classes: (1) model-free systems that do not use any a priori human model[6,7], (2) indirect model systems because the reference model constrains and guides the interpretation of the measured data, and (3) reference model based systems that are continuously updated from observations[8]. The type of input data is fundamental when determining the best technique for resolving a pose. Some authors use monocular still images[9], while others focus on multi-camera video streams[1-4,8,10]. Some limit their work to the use of calibrated views, while others choose non-calibrated images. The most common approach consists of relying on a well defined model and basing pose estimation on modifying the reference information to fit the input data, usually the extracted silhouette[11]. Knowledge about the overall structure of the human body can then be used to further constraint the problem. For example Kakadiaris and Metaxas[6] ask their test subjects to follow a protocol of movements to detect where the different limbs are located, using a spatiotemporal analysis of the deforming apparent contours.

Human 3D reconstruction from a set of still images or image sequences is an attractive research field since low cost digital cameras or camcorders can serve as the data acquisition system. Generally, most of the proposed techniques are based in contour analysis differing in the way they use extracted silhouettes. Hilton et al.[1,2] take four pictures of the human subject to be reconstructed and adjust a well-known 3D reference model to silhouettes extracted from the input images. Cohen and Medioni[12] use the silhouette median axes to reconstruct an articulated model based on generalized cylinders. Cheung et al.[10] pioneered the use of a 3D voxel world to acquire human motion. They use five cameras to perform the 3D reconstruction using a method known as shape from silhouettes and fit ellipsoids in real-time to the reconstructed volume. Mikic et al.[8] also present a completely voxel-based method to obtain a human body model by locating the different body parts using sequential template growing and fitting, which uses prior knowledge of average body part shapes and dimensions. Starck et al.[13] base their reconstruction process on the mesh deformation of a 3D synthetic reference model to match the reconstructed volume. Plänkers and Fua[3,14] estimate the pose by performing an optimization process over a state vector that represents the shape and position of the model in each of the frames of the input video sequence, restricting the problem to the upper body.

Several commercial systems such as Avatar-Me, 3DMeNow and Active Worlds are now available and incorporate avatars based on images of the target subject. However they still impose constraints onto the capture process and trade quality for processing speed.

2. METHODOLOGY

The presented work is an avatar construction pipeline designed to use multiple standard video cameras for the creation of realistic three-dimensional avatars that can be included into interactive virtual environments. We remove traditional constraints in the initial pose of the captured subject, combining it with a silhouette-based modification method of a reference model and generating a textured and animatable avatar with limited user intervention. An outline of the pipeline stages is given in Figure 1.

In this technique, an initial skeleton is approximately fitted to the visual hull[14] of the subject determined from the available views. This volumetric reconstruction can be described as the maximal shape that generates the same silhouette for every reference view outside of the object's convex hull. Beneficial properties of the visual hull include (1) it is guaranteed to enclose the real object, (2) depending on the number of views and the geometry of the object, it can be a tighter approximation than the convex hull, and (3) the size of the visual hull decreases monotonically with the number of different views for the object, although there is no guarantee that it will converge to physical object. Once an initial approximated volume of the captured character has been obtained, the system performs several cross-sections over the visual hull and uses fast discrete image moment computation[15] to estimate a set of parameters describing the underlying skeleton.

The complexity of this process is characterized by the observation that (1) the human body is highly articulated and (2) that information will have to be extracted from images showing arbitrarily clothed subjects. To resolve the ambiguities that come from occlusion, skeleton adjustment is performed considering the previous knowledge available about the anatomic features and structure of the human body. Once the pose is known, the geometry reconstruction process begins by aligning a synthetic reference model to the subject pose. The silhouette information is then extracted from the input images obtained from the video cameras. Next, the camera parameters, that were used to acquire the input images, are used to properly align the synthetic model and to generate the synthetic silhouettes. Then, using the knowledge about the captured body, the silhouettes are divided into regions and a 2D matching function between regions is defined. Information obtained while merging the different 2D views of the input silhouettes, provides a description of how the synthetic reference model has to be modified.

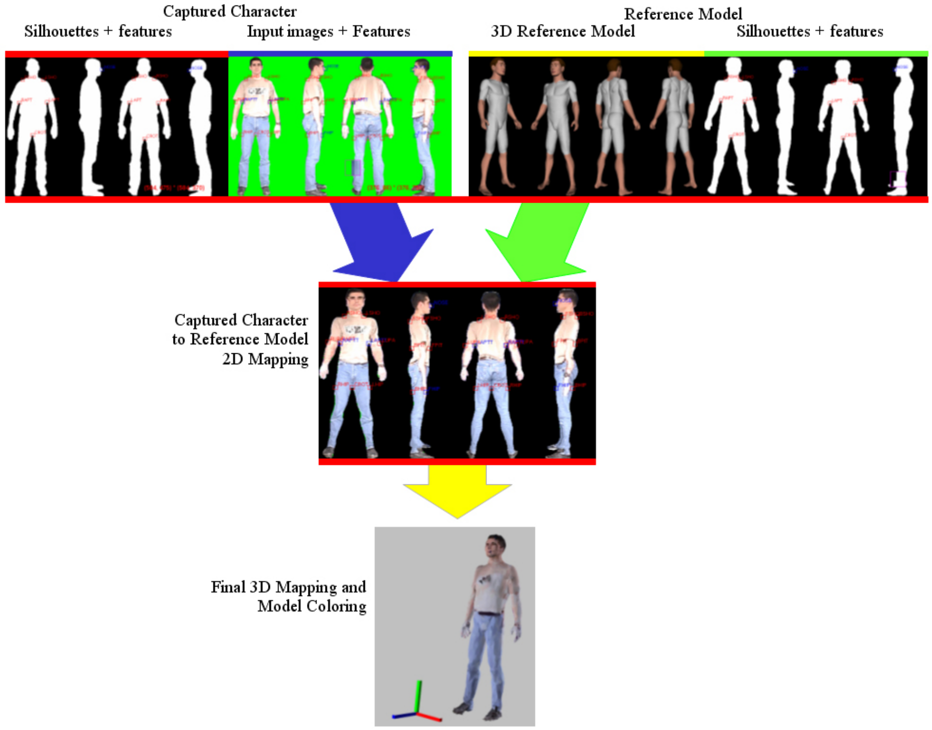

Figure 1. Geometry reconstruction overview.

3. SYSTEM OVERVIEW

The main stages of the presented avatar construction pipeline include: (1) images acquisition, (2) body pose estimation, (3) body geometry estimation and (4) texture mapping. The most important stages in the presented pipeline are body pose estimation and body geometry estimation and are covered in detail in sections 3.1 and 3.2. The avatar creation pipeline is shown in Figure 2.

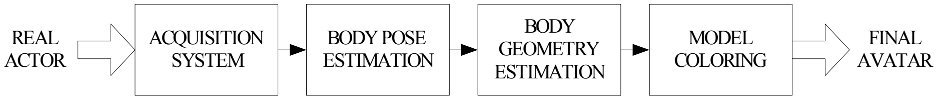

Figure 2. Avatar creation pipeline.

To emulate a multi-camera acquisition system, we use a single consumer-level digital video camcorder to take images of the subject from different angles. A turntable is used to position the actor for the appropriate views, while preserving the initial pose. A rough camera calibration is performed by measuring the distance from the camera to the center of the rotating platform and the height of the human actor. The output of this stage consists of a set of images, and camera calibration parameters.

Body pose estimation then begins with the calculation of the visual hull from the available input images and camera parameters from the acquisition stage. Next, the process of skeleton fitting to the space occupied by the input subject is performed. For this purpose, the volume is divided into slices parallel to the XZ, YZ and XY planes. Using a fast algorithm[15] for 2D image moment calculation, a set of parameters such as the area, centroid, orientation and eccentricity are computed and used to segment the reconstructed volume into different body parts. A set of lines is then adjusted by least squares fit to the cloud of centroids that belong to the underlying skeleton. A general synthetic human model used as a reference in the body geometry estimation stage is then aligned to match the pose of the captured subject. In addition, a set of key features that allows dividing the input body images in body parts are extracted from the boundaries between the homogeneously connected 3D regions obtained from the visual hull.

The body geometry estimation stage obtains a set of images from the aligned synthetic 3D reference model analogous to the one available from the captured subject. Next, a silhouette-finding algorithm is applied to all input images. In combination with the body parts division information extracted in the body pose estimation stage, a 2D mapping function is established between similar body parts of each <synthetic model, captured subject> pair of input silhouettes. This function allows establishing for each view a one-to-one correspondence between the known 3D points in the aligned synthetic reference model and the 3D points of the subject under reconstruction. A set of 2D displacements are then generated for each view and transformed into 3D displacements while merging the available views.

Finally, the recovered avatar is colored using the color information available from the captured input images.
A detailed description of body pose estimation is provided in Sect. 2.1 and body reconstruction in 2.2.

3.1 Body Pose Estimation

In order to ensure an accurate 3D reconstruction of an arbitrarily posing actor two primary components are required, (a) a synthetic 3D human reference model similarly aligned as the captured subject and (b) a set of reliable features that enable body part identification from the collection of input images. The body pose estimation stage is responsible for providing this information to the body geometry estimation stage in the avatar construction pipeline.

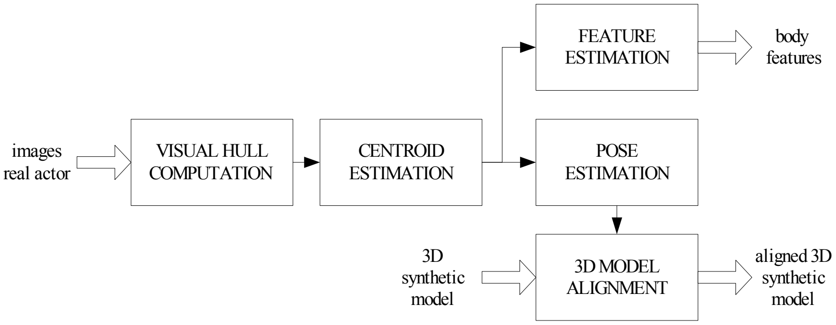

Figure 3. Body estimation pose stage.

The initially available information consists of the reference views of the actor and the camera calibration parameters. To obtain a volumetric reconstruction it is necessary to segment the views into foreground and background. This step can be performed easily when the background is constant for each of the cameras, by performing a thresholded image subtraction of a reference view of the background. Next, an initial 3D volume containing the actor is calculated by intersecting the camera frustums, and subdivided into voxels with a resolution defined by the user. The visual hull calculation consists of determining which voxels of the volume project on the foreground in every reference view. Using transparency, projective texture mapping and stenciling[16], an algorithm can be implemented to obtain the visual hull from a set of segmented images in real time. The presented algorithm sweeps a carving plane along one of the axis of the bounding box to determine which voxels are contained in the visual hull and which are outside (Figure 4). The algorithm exploits the OpenGL API and uses 3D hardware acceleration provided on today's video cards to perform the most computationally expensive operations of the visual hull approximation.

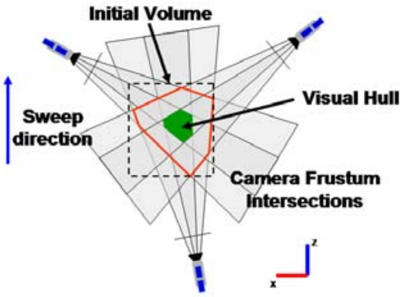

Figure 4. Visual hull with three reference views.

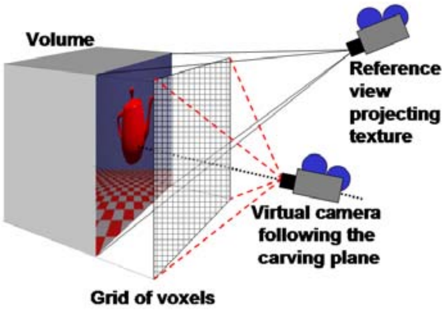
Volume
Reference view projecting texture
Virtual camera following the carving plane
Grid of voxels

Figure 5. Location of virtual camera.

Similar to the hardware accelerated voxel carving technique for 3D model reconstruction, presented by Sainz et al.[17], a virtual camera is located at a constant perpendicular distance to the carving plane (Figure 5). During each of the sweeps, the reference images are projected onto the carving plane using projective texture mapping and analyzed from the virtual camera. During the initialization stage, OpenGL buffers are cleared, stencil and alpha tests enabled to paint only the foreground of the images and background pixels are marked with an alpha value of 0 (fully transparent). Next, all views are rendered while the stencil buffer accumulates the number of times a given pixel has been painted. If all *n* views project foreground pixels to a voxel on the plane, the corresponding stencil values will be equal to the number of cameras and the voxel belongs to the visual hull. Once all *n* views have been projected, the stencil test function is changed to pass only those pixels that have the maximum value in the stencil buffer and render the carving plane.

In order to identify which voxels of the carving plane belong to the visual hull, the rendering of the plane is performed by drawing all the voxels and assigning each a unique color ID. Subsequently, every time a primitive is rendered using the calculated stencil buffer, the visual hull is queried to determine the part of the primitive that is within the real object. Next, the framebuffer is captured and scanned to determine which voxels are present. The result is stored in a 3D bitmask array to which a generic 18-limb skeleton will be fitted (Figure 7). Since holes may appear in the volume due to chosen resolution, a 3D mask voxel-based hole-filling algorithm is applied to the volume as a pre-process to eliminate them.

In image analysis, Hu[18] introduced the use of moment-based techniques for object-characterization. In general, a meaningful subset of moment values (m_{pq}) or central moments (μ_{pq}) can be defined such that they contain sufficient information to uniquely express an entity,

$$m_{pq} \equiv \sum_{y=0}^{M-1}\sum_{x=0}^{N-1} x^p y^q g(x,y) \quad , \quad \mu_{pq} \equiv \sum_{y=0}^{M-1}\sum_{x=0}^{N-1} (x-\bar{x})^p (y-\bar{y})^q g(x,y) \tag{1}$$

where *g(x,y)* represents the intensity of a discrete image of dimensions *M*x*N*, and ($\bar{x}$, $\bar{y}$) the center of mass of the shape.

More specifically, the low-order moment values represent well-known, fundamental geometric properties of an arbitrary object. In our case, we are interested in the zeroth, first and second order moments because they describe respectively the area of the analyzed entity,

$$m_{00} = \sum_{y=0}^{M-1}\sum_{x=0}^{N-1} g(x,y) \tag{2}$$

and the center of mass (centroid),

$$\bar{x} = \frac{m_{10}}{m_{00}} \quad , \quad \bar{y} = \frac{m_{01}}{m_{00}} \tag{3}$$

and the parameters of the ellipse that best approximates the studied 2D shape.

$$\phi = \frac{1}{2}\tan^{-1}\left(\frac{2\mu_{11}}{\mu_{20}-\mu_{02}}\right) \quad , \quad \alpha = \left(\frac{2\left(\mu_{20}+\mu_{02}+\sqrt{(\mu_{20}-\mu_{02})^2+4\mu_{11}{}^2}\right)}{\mu_{00}}\right)^{\frac{1}{2}} \quad , \quad \beta = \left(\frac{2\left(\mu_{20}+\mu_{02}-\sqrt{(\mu_{20}-\mu_{02})^2+4\mu_{11}{}^2}\right)}{\mu_{00}}\right)^{\frac{1}{2}} \qquad (4)$$

The general centroids estimation process consists of (1) sequentially slicing the visual hull; initially along the Y direction, (2) incrementally examining all the blobs found in each of the XZ planes. A fast moments computation algorithm[15] is applied to each of the blobs contained in every slice of the volume, calculating their area, centroid and approximated ellipse parameters. The collection of centroids provides an initial estimation towards determining where the skeleton of the captured subject lies in the visual hull. Studying their distribution in space as shown in Figure 6 enables determining their spatial connectivity and provides the initial topological information.

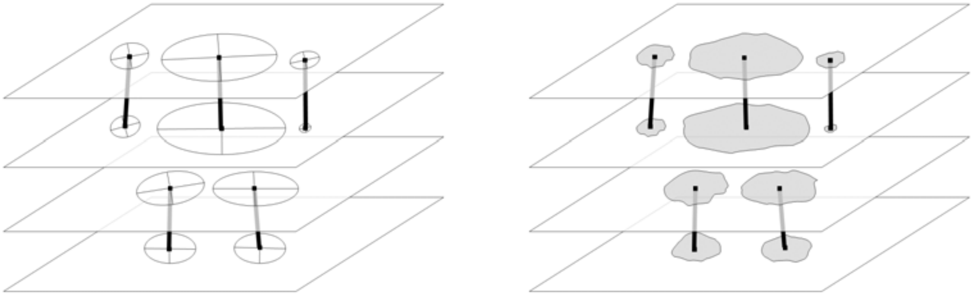

Figure 6. Stacks of slices with centroids and fitting ellipses

In order to estimate body part division, the presented technique exploits the fact that spatial continuity is an intrinsic characteristic of the human body. As the current slice is scanned, the information about its blobs is compared with data available from the previous slice. Two criteria are used to decide if two blobs in two consecutive slices belong to the same body part, (a) the distance between both centroids is below a predetermined threshold and (b) there is not a significant change in the blob area and/or shape with respect to the corresponding blob in the previous slice. Wherever a spatial discontinuity is found in the volume, the corresponding blobs are marked and stored as potential candidates to be considered as boundary body parts. A primary purpose of these blobs is to help in determining the body part identification in the set of input images. This step is fundamental for the avatar geometry estimation stage described in Sec. 3.2.

Slicing and analyzing the volume in the direction of the Y axis is sufficient for an initial guess of the body pose but it is not sufficient to detect it precisely. For this reason, all regions where a spatial discontinuity was detected are analyzed in order to evaluate if further analysis is necessary along the X and Z axes directions. If required, the same moment estimation method is applied to the other slicing directions.

Once the complete volume has been examined and the body parts distribution has been determined, an initial cloud of points is obtained and fitted to a set of lines using least squares minimization of the orthogonal errors[19]. When these lines are combined with the knowledge available from the reference skeleton (Figure 7), a body pose can be estimated.

This process targets the identification of key limbs such as head, feet, torso, arms and legs using information about spatial location and connectivity. For example, the head is very likely to be found in the upper part of the volume and will be connected to the torso and arms. Feet will be in the lower part of the volume, connected to the legs, providing an initial estimation of the chest orientation. The usual proportions between body parts also helps in identifying which lines belong to each body part, aligning the reference skeleton (Figure 7) limbs to match the estimated lines.

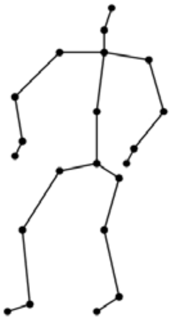

Figure 7. Reference skeleton.

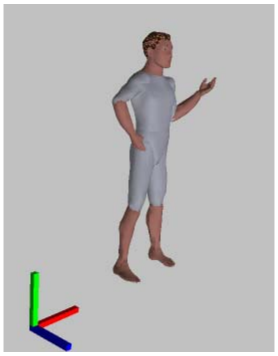

Figure 8. Aligned synthetic skeleton.

Once the pose of the captured subject is estimated and the generic skeleton (Figure 7) obtained, a 3D synthetic reference model (Figure 8) is aligned accordingly. Additionally, the collection of blobs describing remarkable discontinuities in the visual hull prone to be considered as body part boundaries, is provided to the avatar geometry estimation stage. These blobs usually reveal key body features such as the armpits, crotch, neck, elbows, knees, and wrists.

3.2 Body Reconstruction

The 3D reconstruction method is silhouette-based and uses the geometry of the aligned 3D synthetic model to match the available silhouette information of the real captured actor. Initially, the available information for the avatar geometry reconstruction is comprised of a set of arbitrary images of the real actor and the corresponding camera calibration parameters. The body reconstruction phase requires aligning a generic 3D synthetic model to match the pose of the captured subject. In addition, it needs from a collection of body features that enables body part identification and segmentation of the captured subject's input images (Figure 9). As discussed in Sect. 2.1, the acquisition method for this information is based on the analysis of the visual hull defined by the input images of the captured subject.

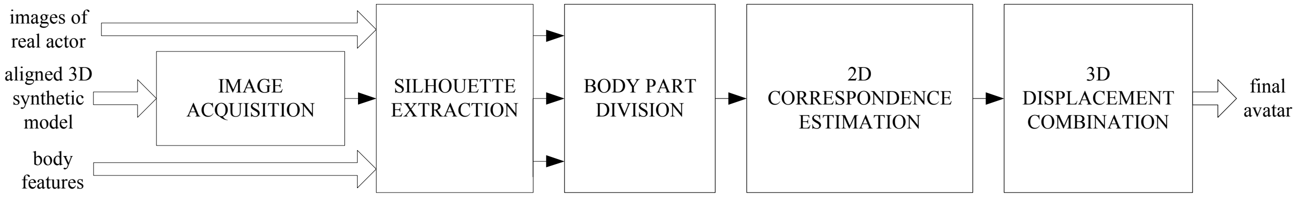

Figure 9. Body reconstruction stage.

Using the available set of camera calibration parameters, the first step consists of obtaining a set of images of the 3D aligned synthetic model, analogous to the one available from the real actor. The images are organized per views, in <captured subject, synthetic model> image pairs. Between each pair of images, a body part correspondence is defined after examining the collection of blobs obtained from the visual hull of the captured subject and the aligned synthetic model. These are the same blobs that were considered in the body pose estimation stage as candidates to be boundaries between spatially homogeneous 3D regions. In order to define 2D correspondence, each pair of images needs to be fragmented into the same number of body parts (Figure 10). Body features that subdivide the input images are obtained from the planar blobs that partitioned the visual hull of the captured subject and aligned model. Using the camera calibration parameters, all blobs are projected into the available images, and the detected regions of interest are used to find a point in the silhouettes that can be considered as a body feature, which divides the images into body parts (Figure 10). The quality of the body features is fundamental for this step and images may have to be adjusted in order to simplify the 2D body part mapping.

The process is based on knowledge about how 3D points of the synthetic model project into the available segmented input images, provided by the camera calibration parameters. The transformation of a 3D point $(X,Y,Z,1)^T$ in world coordinates into its corresponding 2D projected point $(u,v,1)^T$ can be expressed in homogeneous coordinates as:

$$\begin{pmatrix}u\\v\\1\end{pmatrix} = K\begin{bmatrix}1&0&0&0\\0&1&0&0\\0&0&0&1\end{bmatrix}\begin{bmatrix}R&\vec{t}\\\vec{0}^T&1\end{bmatrix}\begin{pmatrix}X\\Y\\Z\\1\end{pmatrix}$$

where K is the camera calibration matrix that contains the camera intrinsic parameters.

$$K = \begin{bmatrix}f_x&0&p_x\\0&f_y&p_y\\0&0&1\end{bmatrix}$$

Parameters R and t (Figure 10) give the view transformation between a point $(X,Y,Z,1)^T$ expressed in world coordinates into a point in camera coordinates $(X_c,Y_c,Z_c,1)^T$.

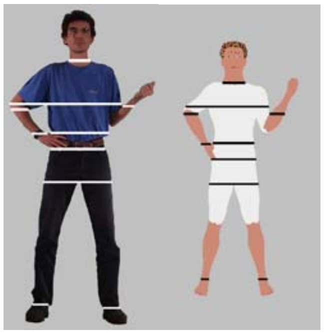

Figure 10. Body Part division.

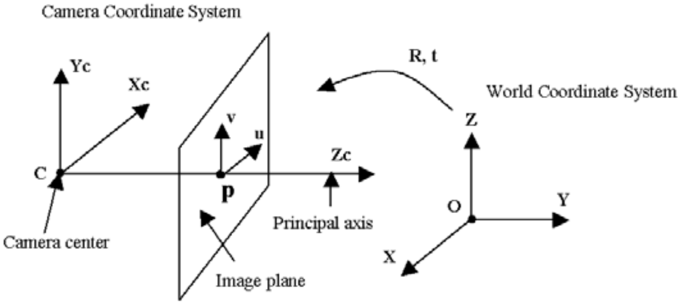

Figure 11. World and camera coordinate systems.

Once all the vertices of the aligned synthetic model are projected into the initial synthetic model images using the camera calibration parameters, 2D correspondences between images in a per body part basis can be established. First, we decide to which body part a vertex belongs according to the current camera viewpoint. Once this is determined, a 2D mapping function can be defined for each body part based on its range in screen coordinates (Figure 10). For a given $(u,v)^T$ 2D point in the synthetic reference model image, we know that (1) it maps to the coordinates $(u',v')^T$ in the captured subject image, (2) it belongs to a specific body part, (3) the limits of this body part along the v axis are v_{min} and v_{max} for the synthetic reference model image, and v'_{min} and v'_{max} for the captured subject image, and (4) the limits in the u axis for the given v value are u_{min} and u_{max} for the synthetic reference model image, and for the v' value, u'_{min} and u'_{max} in the captured subject image. The equivalent $(u',v')^T$ point in the subject image that corresponds to the $(u, v)^T$ point in the synthetic reference model image is given by:

$$s_v = \frac{v'_{\max} - v'_{\min} + 1}{v_{\max} - v_{\min} + 1} \qquad v' = s_v\left(v - v_{\min}\right) + v'_{\min}$$

$$s_u(v') = \frac{u'_{\max}(v') - u'_{\min}(v') + 1}{u_{\max}(v) - u_{\min}(v) + 1} \qquad u' = s_u(v)\left(u(v) - u_{\min}(v)\right) + u'_{\min}(v')$$

where s_u and s_v are scale factors.

The computed 2D point $(u',v')^T$ is then unprojected into its corresponding $(X',Y',Z')^T$ 3D point, under the assumption that the distance between the viewer and this point is the same as the one that can be calculated between the viewer and the $(X,Y,Z)^T$ that projects to the coordinates $(u,v)^T$. This assumption turns out to give good visual results since the location of the viewer is far enough from the captured subject. As this process is repeated for each pair of views <captured subject, synthetic model>, the vertices in the final avatar are computed as the average position of all the vertices obtained for each of the available views.

4. EXPERIMENTS

In this section we present results of the avatar reconstruction pipeline for images of both synthetic and real posing actors. The acquisition system used to capture real actors consists of a single digital video camera recorder Sony DCR-TRV20 and a turn table. This turn table was used to position the test subject while helping the subjects pose unchanged. All tests were run in a 1.8GHz Pentium 4 with 512MB of RAM and a NVIDIA Quadro2 Pro graphics card. The acquired images were then cropped to 256 x 512 pixels to match the texture mapping capabilities of the videocard.

The reference model used in the avatar geometry estimation stage as the basis for performing the reconstruction of the captured subject was an HANIM 1.0 compliant VRML97 humanoid model called Hiro. The overall model mesh resolution was modified to 8554 faces.

Different datasets were used to evaluate the system. Table 1 provides an overview of the image model complexity.

Model	Type	Images	Images Resolution	Visual Hull Resolution
Wolf	Synthetic	8	256x512	64x64x64
Roberto	Real	4	256x512	64x64x64
Miki	Real	4	256x512	128x128x128

Table 1. Input datasets characteristics.

The first dataset is based on a synthetic character and eight input images were generated at 45 degree intervals in the horizontal plane. The camera calibration parameters were obtained from the 3D model visualization tool. Figure 12 shows the views used as input for the reconstruction process.

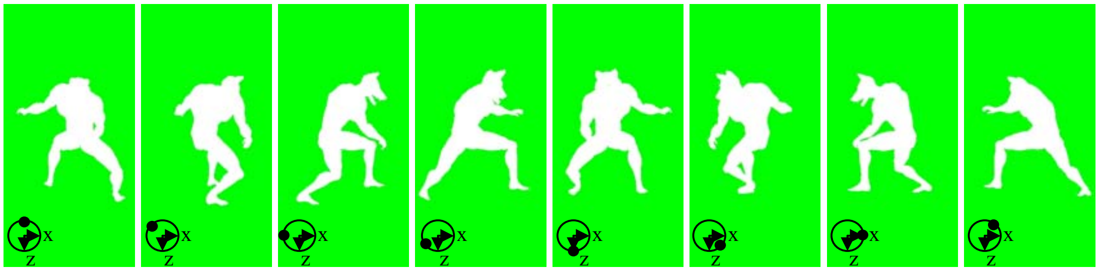

Figure 12. Input images of the synthetic character Wolf.

The second and third datasets consist of real images for a human test subject. In both cases 4 images were acquired at 90 degrees intervals in the horizontal plane. The subject was standing on a rotating platform while the images were taken. Figure 13 show some of the input views with segmented background.

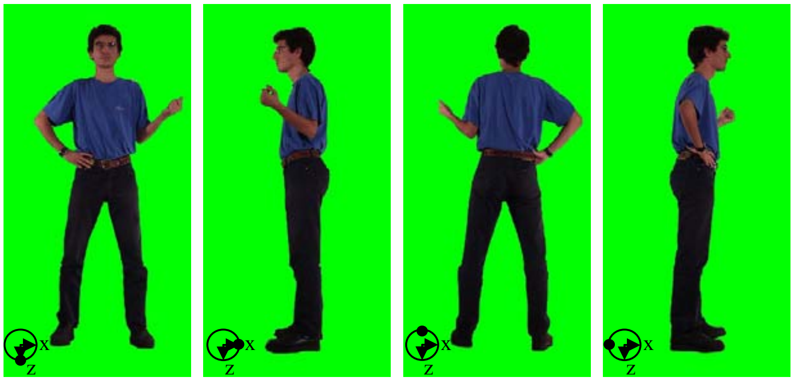

Figure 13. Input images of the Roberto dataset.

For the first two subjects, the visual hull was computed using a resolution of 64x64x64 voxels which provides enough detail for the pose estimation algorithm. For the third one, the resolution was 128x128x128. Figure 14 shows the volume representation of the synthetic and real datasets. The visual hull voxels are shown as light grey dots, while the centroids of the different slices used to compute the skeleton are shown in black.

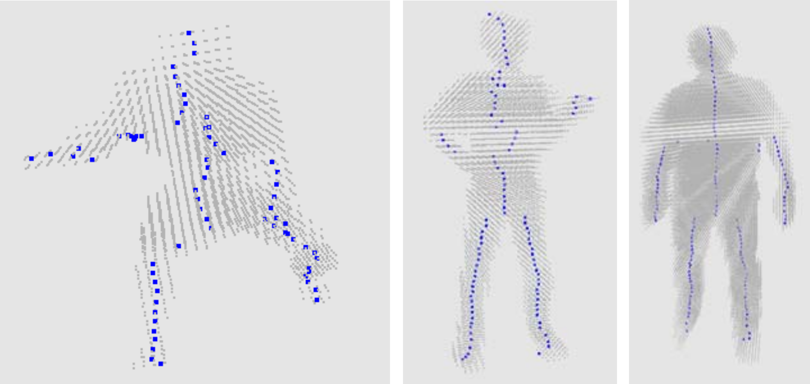

Figure 14. Visual hull and estimated centroids (Wolf, Roberto, Miki datasets).

Once the skeleton pose has been obtained from the volumetric model, the 3D synthetic reference models are aligned and used for the geometry estimation stage. Figure 15 show the captured subjects and reference models matching the pose.

The next step is to extract the silhouettes for each reference view. Using the skeleton information and the discontinuities projected into the input 2D images we calculate the boundaries of the different body parts. Figure 10 shows a pair of <captured subject, reference model> images segmented into body parts.

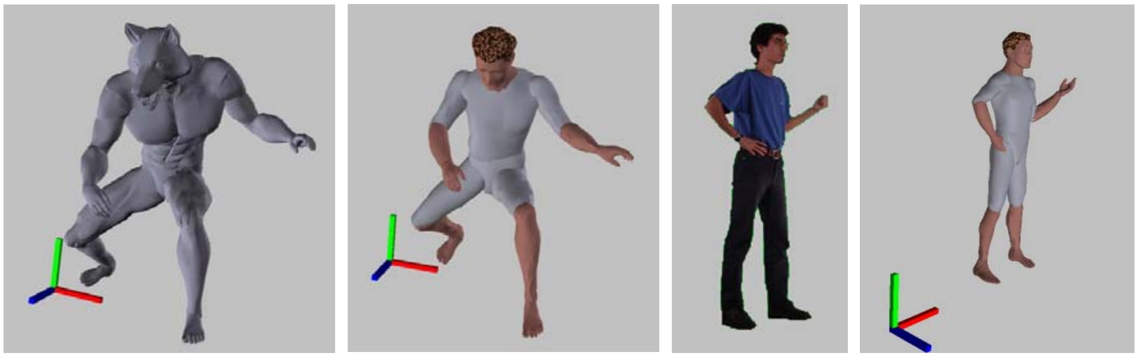

Figure 15. Aligned synthetic reference models.

After performing the 2D mapping that relates the body parts of the captured subject with the ones of the reference model, the 3D displacements are estimated and combined to obtain the reconstructed 3D avatars in Figure 16. The quality of the results for the pose estimation stage depends on several factors, with the number of available views of the subject being the most important. This is primarily due to the fact that the convergence of the visual hull to the correct shape of the subject is improved with a higher number of reference views, directly affecting pose extraction and body part identification.

The variety of poses that the presented approach can recover is constrained by, (1) how much self occlusion is present in the input images and (2) if it can be resolved with the combination of available views. This is not considered a limitation, since the system is intended for avatars reconstruction from multiple synchronized video streams. The principal issue is to determine a robust criterion to decide when an input set of images is good enough to be reconstructed, a step currently performed with user guidance. The alignment of the reference skeleton is also very critical and it depends on the captured input pose, since certain assumptions have to be made, such as a vertical pose of the actor, with the head located in the upper part of the volume and feet in the lower part.

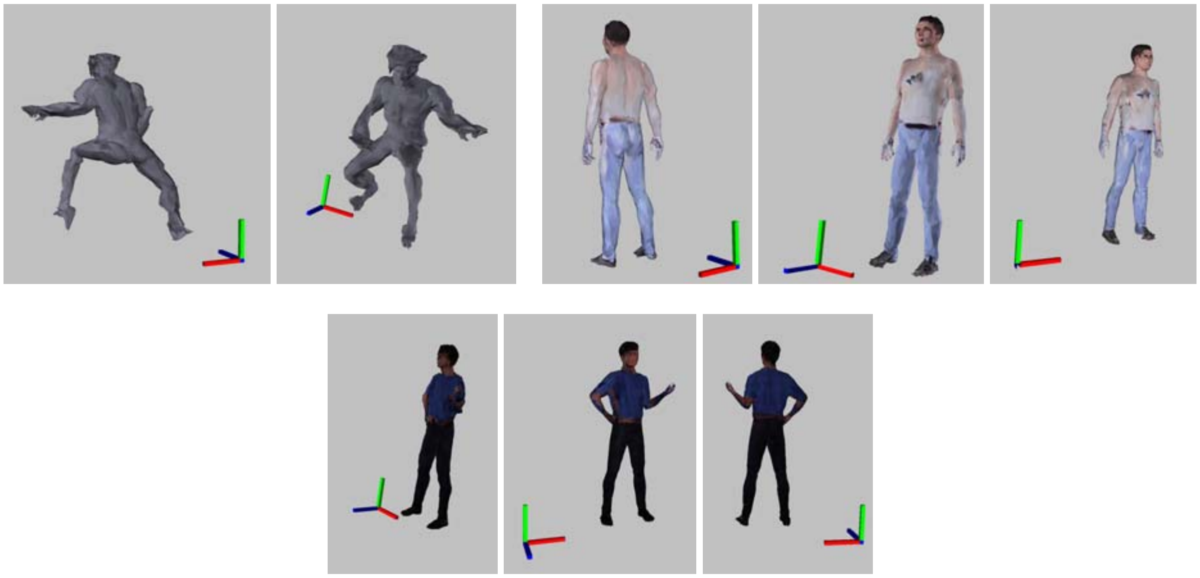

Figure 16. Reconstructed avatars

The body part identification and assignment between body parts of the captured subject and body parts of the aligned reference model is important. It is fundamental that the body parts relationship between posing actor and reference model is one-to-one, that is, all vertices of the aligned reference model map to a body part in the reference model input images and that a correspondence exists in the captured subject images. The mesh resolution of the 3D synthetic reference model determines how fast the avatar reconstruction is performed and its final accuracy.

5. CONCLUSIONS

This paper presents a framework for the construction of high-quality avatars from image data. A pipeline is presented that combines a set of hardware-accelerated stages into one seamless system. Primary stages in this pipeline include pose estimation, skeleton fitting, body part segmentation, geometry construction and coloring. Different techniques including visual hull reconstruction, 2D image moment calculation, least squares minimization were implemented, to recover an initial skeleton and to segment it into different body parts that in turn can be used to construct properly defined avatars in near real-time. The presented system removes traditional constraints on the initial pose of the captured subject by introducing silhouette-based techniques in combination with a reference model. The presented pose estimation technique reduces the need for user intervention significantly and high-quality results can be obtained at interactive speeds. Main contributions of this paper include hardware-accelerated computation for visual hull approximations in combination with fast 2D moment calculation algorithms used to estimate pose and location of the different body parts of the target subject. Using this information a synthetic reference model can be aligned and the silhouette information segmented to create the basis for realistic 3D avatars. A primary advantage of this approach is that it accepts arbitrary poses and an unconstrained number of different views of the captured subject. As shown in the experiments section, the visual quality of the reconstructed avatars is good, although the coloring stage can be further optimized to achieve photorealism.

6. ACKNOWLEDGMENTS

This research was supported in part by the National Science Foundation under contracts CCR-0083080, EIA-0203528 and by the Balsells Fellowship Program. Models used for this study were provided by http://www.3dcafe.com and http://www.cis.upenn.edu/~beitler/hanim1.0/. The above support is greatly appreciated.

7. REFERENCES

1. A. Hilton, D. Beresford, T. Gentils, R. Smith, and W. Sun, "Virtual People: Capturing human models to populate virtual worlds", in *Proceedings IEEE Computer Animation 1999*, pp. 174-185, 1999.
2. A. Hilton, D. Beresford, T. Gentils, R. Smith, W. Sun and J. Illingworth, "Whole-body modeling of people from multiview images to populate virtual worlds", in *The Visual Computer* **16**, pp. 411-436, Springer-Verlag 2000.
3. N. D'Apuzzo, R. Plänkers, P. Fua, A. Gruen and D. Thalmann, "Modeling human bodies from video sequences", in S.F. El-Hakim, A. Gruen, ed., *Videometrics VI, Proc. of SPIE* **3461**, pp. 36-47, 1999.
4. R. Plänkers, P. Fua, "Tracking and Modeling People in Video Sequences", in *Computer Vision and Image Understanding* **81**, pp. 285-302, 2001.
5. T. B. Moeslund, E. Granum, E., "A Survey of Computer Vision-Based Human Motion Capture", in *Computer Vision and Image Understanding* **81**, pp. 231-268, 2001.
6. I. A. Kakadiaris, Metaxas, D., "Three-Dimensional Human Body Model Acquisition from Multiple Views", in *International Journal of Computer Vision* **30**, pp. 191-218, 1998.
7. R. Rosales and S. Sclaroff, "Learning and Synthesizing Human Body Motion and Posture", *Proceedings of the 4th IEEE International Conference on Automatic Face and Gesture Recognition 2000*, pp. 506-511, 2000.
8. I. Mikic, M. Triverdi, E. Hunter, and P. Cosman, "Articulated body posture estimation from multicamera voxel data". in *Proc. of the IEEE Computer Vision and Pattern Recognition Conference* **1**, pp. 455-460, 2001.
9. C. Barron, I.A. Kakadiaris, "Estimating Anthropometry and Pose from Single Uncalibrated Image", in *Computer Vision and Image Understanding* **81**, pp. 269-284, 2001.
10. K.M. Cheung, T. Kanade, J. Bouguet, and M. Holler, "A real time system for robust 3D voxel reconstruction of human motions", in *Proceedings of the 2000 IEEE Conference on Computer Vision and Pattern Recognition* **2**, pp. 714–720, 2000.
11. Y. Yang, X. Wang, and J. X. Chen, "Rendering avatars in virtual reality: Integrating a 3D model with 2D images", in *Computing in Science & Engineering* **4**, pp. 86-91, IEEE Computer Society, 2002.
12. I. Cohen, G. Medioni, H. Gu, "Inference of 3D human body posture from multiple cameras for vision-based user interfaces", *5th World Multi-Conference on Systemics, Cybernetics and Informatics, Orlando, Florida*, 2001.
13. J. Starck, A. Hilton, and J. Illingworth. "Reconstruction of animated models from images using constrained deformable surfaces" in *10th Int. Conf. on Discrete Geometry for Computer Imagery (DGCI)*, *Lecture Notes in Computer Science* **2301**, pp. 382-391, Springer-Verlag, 2002.
14. A. Laurentini, "The Visual Hull Concept for Silhouette-Based Image Understanding", in *IEEE Transactions on Pattern Analysis and Machine Intelligence* **16**, pp. 150-162, 1994.
15. Philips, W. "A new fast algorithm for moment computation", in *Pattern Recognition* **26**, pp. 1619-1621, 1993.
16. B. Lok, "Online Model Reconstruction for Interactive Virtual Environments", in *Proc. of Symposium on Interactive 3D graphics*, pp. 69-72, 2001.
17. M. Sainz, N. Bagherzadeh, A. Susin, "Hardware Accelerated Voxel Carving". *1st Ibero-American Symposium in Computer Graphics*, pp. 289-297, 2002.
18. M. K. Hu, "Visual pattern recognition by moment invariants", *IRE Trans. Inf. Theory* **8**, pp. 179-187, 1962.
19. P. J. Schneider, D. H. Eberly, *Geometric Tools for Computer Graphics*, Morgan Kaufmann Publishers, 2002.