Automatic Generation of Personalized Human Avatars from Multi-view Video

Naveed Ahmed, Edilson de Aguiar, Christian Theobalt,
Marcus Magnor, Hans-Peter Seidel

MPI Informatik
Saarbrücken, Germany

{nahmed,edeaguia,theobalt,magnor,hpseidel@mpi-sb.mpg.de}

ABSTRACT

In multi-user virtual environments real-world people interact via digital avatars. In order to make the step from the real world onto the virtual stage convincing the digital equivalent of the user has to be personalized. It should reflect the shape and proportions, the kinematic properties, as well as the textural appearance of its real-world equivalent. In this paper, we present a novel spatio-temporal approach to create a personalized avatar from multi-view video data of a moving person. The avatar's geometry is generated by shape-adapting a template human body model. Its surface texture is assembled from multi-view video frames showing arbitrary different body poses.

Categories and Subject Descriptors

I.3.7 [**Computer Graphics**]: Three-Dimensional Graphics and Realism—*Animation, Virtual Reality*; I.4.8 [**Image Processing and Computer Vision**]: Scene Analysis—*Motion, Tracking, Time-varying Imagery*

General Terms

Algorithms, Measurement.

Keywords

Avatar creation, virtual reality, texturing, shape deformation

1. INTRODUCTION

In recent years, virtual environments in which real-world people can interact through controllable digital characters, so-called avatars, have become accessible even to the user at home. In order to make their appearance on the virtual stage convincing, many users want to give their digital equivalent a personal touch. Unfortunately, in most online games or 3D chat rooms, the degree to which a user can personalize his avatar is very restricted. At best, he can manually modify a body shape taken from a database of template geometries, and texture the face of the virtual puppet with a digital photograph. It is obvious that, in order to make the personal touch fully convincing, the animatable human model should reflect the complete shape and textural appearance of the real-world human that it represents.

In order to serve this purpose, we have developed a novel fully-automatic method to build a customized digital human from easy-to-capture input data. The inputs to our method are multiple synchronized video streams that show only a handful of frames of a human performing arbitrary body motion (Sect. 3). Our approach is based on a template human body model consisting of a triangle mesh surface representation and an underlying kinematic skeleton (Sect. 4). This body representation is automatically deformed until it matches both the shape and the skeletal structure of its real-world counterpart captured in the video footage. The realistic appearance of the digital human is completed by reconstructing a consistent surface texture from the multi-view video footage. By simultaneously employing images from multiple camera views and multiple time steps of video, it is made sure that even temporarily invisible parts of the body surface are faithfully captured in the texture (Sect. 5). With our novel method we quickly generate photo-realistic digital equivalents of real-world people (Sect. 6 and Sect. 7).

2. RELATED WORK

Acquisition of visually convincing models of humans from images has been a long standing problem in computer graphics and virtual reality. In order to generate a realistic human avatar, the kinematics, shape and appearance have to be captured simultaneously.

Full-body range scanning systems exist that can quickly acquire the full surface geometry of a human body. However, they are highly expensive and don't straightforwardly enable to estimate a skeleton of the human [11]

Alternatively, image- or video-based methods can be used to reconstruct body models. In one line of research, it is the primary goal to derive the kinematic structure and a simple surface geometry from image data [5, 10, 3]. A surface texture, however, is not reconstructed.

In 3D video, novel views of a real person are rendered from multiple input video streams [9, 6, 8]. Unfortunately, these approaches do not reconstruct models that could be animated with arbitrary motion data.

VRST'05, November 7–9, 2005, Monterey, California, USA.

In contrast, we propose a model-based approach that creates a fully-animatable avatar comprising a customized geometry, a realistic surface texture, and an appropriately rescaled skeleton.

Our work is similar to the methods proposed by Hilton et al. [4] and Lee et al. [7], where a human template model is deformed until it aligns with multiple silhouette images. Surface textures are created by mapping photographs back onto the body representation.

We introduce an improved reconstruction method that employs multiple time steps of multi-view video footage to capture shape and texture at higher accuracy. As opposed to many previous approaches, our method is fully-automatic and even enables the extraction of multiple face textures depicting different facial expressions.

3. MULTI-VIEW VIDEO RECORDING

We employ eight frame-synchronized video cameras, each of which features a 1-megapixel CCD sensor and 25 fps of sustained acquisition frame rate, to capture multi-view video (MVV) footage. The imaging devices are arranged in a convergent setup around the center of the scene, and they are geometrically and photometrically calibrated. In each MVV sequence the person first strikes an initialization pose for short moment, and thereafter is free to move arbitrarily. The silhouette of the person in each frame is found via color-based background subtraction.

4. PERSONALIZING THE SHAPE OF THE AVATAR

Our virtual human is represented by means of a template body model comprising 16 individual body segments whose shape is modeled via closed triangle meshes, and an underlying kinematic skeleton providing 35 pose parameters. Two sets of anthropomorphic parameters enable shape customization. The first set controls uniform scaling of individual segments, the second set controls free-form deformation of each body segment by means of B-spline scaling curves.

The shape of the model is adapted until it matches the look of its real world equivalent by searching for an optimal set of anthropomorphic parameters. To this end, we employ a silhouette-based analysis-by-synthesis approach [2]. It performs an optimization search in both the shape and the pose parameters in a pre-defined sequence in order to maximize the overlap between the silhouette of the reprojected model and the image silhouette in all camera views. The energy function EF that numerically assesses this overlap sums up the number of set pixels in binary XOR images between image and model silhouettes from all camera perspectives. We further extend this robust and fast optimization scheme such that we can personalize the skeleton dimension (Sect. 4.1), as well as the segment geometries (Sect. 4.2) by not only looking at one but several time steps of multi-view video.

4.1 Adapting the Skeleton

To customize the bone lengths of the default skeleton we only employ the eight input frames depicting the person in the initialization pose. Our skeleton rescaling method is an iterative procedure that alternates between an optimization of pose parameters and an estimation of uniform scaling parameters.

In the first step of each iteration the scaling parameters of all body segments are adjusted. The second step of each iteration uses the rescaled body model and computes an estimate of the body pose parameters. These two steps are repeated several times (Fig. 1b).

4.2 Spatio-temporal Free-form Deformation

Via simple uniform scaling the shape of the real actor can not be realistically mimicked. We have thus developed a novel spatio-temporal free-form deformation scheme. It deforms the individual segments' geometries until they are in accordance with the actor's body in multiple body poses. The deformation of each individual segment is controlled by the anthropomorphic B-spline parameters. For each of the 16 triangle meshes, four local B-spline curves are defined. The curves scale the geometry in directions parallel to the x- and z-planes of the local coordinate frame.

The geometry of one individual segment is deformed by simultaneously finding four optimal sets of N local control values. Each set of control values specifies one of the local scaling curves. The criterion that guides the optimization search is the previously mentioned silhouette-XOR energy function, EF. For numerical minimization we employ the LBFGS-B method [1].

Through experiments we have found out that $N = 4$ control values per curve represent the best compromise between deformation flexibility and fitting speed.

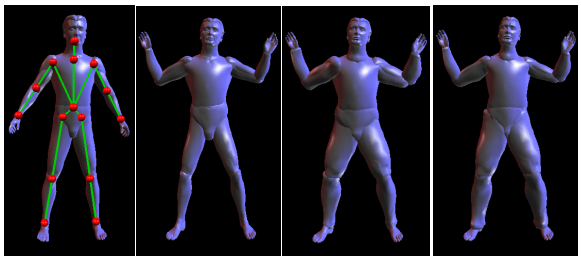
We have developed a spatio-temporal optimization procedure that employs the previously described principle to shape-adapt the geometry of all body segments. It allows us to robustly infer deformation values that correctly reproduce the geometry of an actor not only in one but in several body poses. Since the stance of the skeleton changes over time, we apply a two-step iterative procedure that alternates between pose determination and segment deformation.

In the first step of each iteration, the pose parameters of the model at each time step of video are estimated using the silhouette-based analysis-by-synthesis approach.

In the second step, the B-spline control values for each of the 16 segments are computed by means of the previously described optimization scheme. To this end, K time steps of video are automatically selected out of the M time steps that the input video sequence contains. We find scaling parameters that optimally reproduce the shape of the segments in all of these K body poses simultaneously. A modified energy function EF_R sums over the silhouette-XOR contributions EF_I at each of these K time steps, $EF_R = \sum_{I=1}^{K} EF_I$. Optionally, the two-step optimization procedure can be iterated. The final model possesses a spatio-temporally silhouette-consistent shape (Fig. 1d).

5. RECONSTRUCTING A PERSONALIZED SURFACE TEXTURE

Previous approaches to avatar creation reconstructed a static surface texture from multiple photographs showing the person in a single pose. Although the so-created virtual humans look authentic if they strike the same pose as the person in the images, very disturbing artifacts may occur when they are animated. The main reason for these artifacts is texture undersampling due to non-optimal camera placement or due to visibility problems at mutually occluding body segments (Fig. 4a).



(a) (b) (c) (d)

Figure 1: (a) Adaptable generic human body model; (b) initial model after skeleton rescaling; (c) model after one (d) and several iteration of the spatio-temporal free-form deformation scheme.

We attack these problems by means of a spatio-temporal texture reconstruction scheme that samples from multiple time steps of the MVV sequence.

We lay out the surface texture of our avatar in the 2D plane by means of a patch-based surface parameterization.

Since we know the exact body pose of the model in each time step of multi-view video we can incorporate image data of multiple body poses into one consistent surface texture. This, in turn, enables us to fill-in color information for surface areas that are invisible in one body pose from images of the model in another body pose. There are two main reasons for why a surface point may not be visible from any input camera view: mutual geometry occlusion at segment boundaries or non-optimal camera placement. Unfortunately, there is no simple way of deciding which of the two cases applies to a specific invisible surface point. Thus, we employ the following two-step procedure which implicitly handles both cases:

Before texture reconstruction commences, U time steps of the input MVV sequence from which the color information for the final texture is assembled are automatically selected.

In step 1, the *single-time-step texture assembly*, we create U individual consistent surface textures, $stex_i$, with $i \in \{1,\ldots,U\}$. Each $stex_i$ is only reconstructed from the multi-view video images of time step i. The color of a texel is computed by weightedly blending the colors at its projected locations in each of the camera views [2].

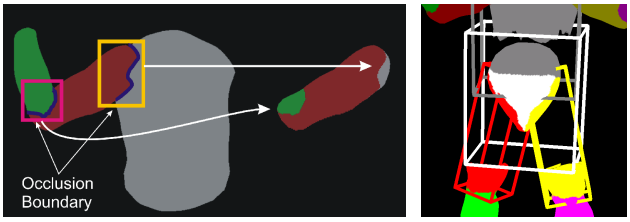
In order to compute the visibility of each surface point in all of the camera views, we have developed a scheme which looks at each of the 16 body segments separately. Using the pelvis as an example, the scheme works as follows: First, a slightly enlarged bounding box of the pelvis is generated. All triangle vertices on directly adjacent segments (i.e. torso and upper legs) that are inside that bounding box are trimmed. For each input camera view, the visibility of each vertex in the pelvis is determined from the trimmed version of the model. In Fig. 2b the trimming procedure for the pelvis segment is visualized. The white regions on the pelvis illustrate those parts of the geometry that have been visible in input camera 2 even in the untrimmed model. All pelvis areas with another color were occluded by one of the directly adjacent segments. By this means, we implicitly create texture information for parts of the surface geometry that are invisible due to mutual occlusion between neighboring triangle meshes. Our visibility computation scheme makes sure that occluded texture parts are filled-in from those parts of the occluding geometry that are spatially close to the occlusion boundary on the 3D surface (Fig. 2a). Texture parts of the occluder that are further distant from the occlusion boundary do not contribute to the occluded texture area.

In step 2, the *texture combination* step, we merge all single-time-step textures, $stex_i$, into one final texture (Fig. 3a). The texture generated from the model in the initialization pose, $stex_1$, is considered as the reference texture. For every texel in $stex_1$ whose color is not known we make a look-up, in ascending order, into all remaining single-time-step textures $stex_i, i \in \{2,\ldots,U\}$. The color of the texel is copied from the first texture in which it is visible. Fig. 3b illustrates from what time steps of the multi-view video sequence the colors in the final texture of the left upper leg were taken. Color discontinuities in the final texture are locally smoothed. As an additional feature, the user can manually select individual time steps of video that show interesting facial expressions. These face subtextures are assembled into one packed texture which can be used to change the avatar's appearance on-the-fly according to his mood.

6. RESULTS

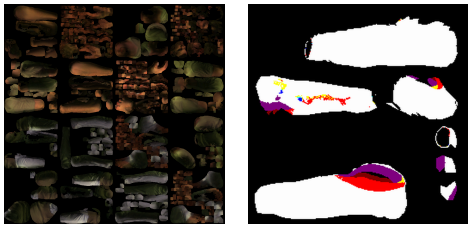
We have tested our method using a few seconds of video footage of a male and a female test subject. Through experiments we have found out that it is sufficient to employ around 5 non-subsequent frames of an MVV sequence for shape and texture reconstruction. On a PC featuring a Pentium™ 4 CPU and an Nvidia GeForce 6800 GPU one iteration of the skeleton rescaling method on average takes around 1 minute. If 5 MVV time steps are considered, free-form deformation takes around 15 minutes. Under the same circumstances spatio-temporal texture reconstruction takes around 40 seconds.

Figs. 5a,b show a comparison between the actor as it appears in one of the video frames, and the rendered avatar in a novel body pose. Our approach faithfully captures the shape and the textural appearance of the male and the female actor for different types of apparel. Even in body poses that are significantly different from any of the captured ones, ap-



(a) (b)

Figure 2: (a) At occlusion boundaries non-visible geometry is assigned the texture from nearby parts of the occluder. (b) Trimming procedure at torso: vertices of neighboring segments are trimmed. White: geometry visible in camera 2 prior to trimming: Remaining colors: denote neighboring occluding segments.



(a) (b)

Figure 3: (a) Complete body texture. (b) Texture patches for one segment; different colors indicate different time steps of video from which the texel color was taken.

(a) (b) (c)

Figure 4: Undersampling artifacts at segment boundaries (a) that are corrected by our method (b). (c) An avatar can be used to insert a real-world person into arbitrary virtual environments.

pearance artifacts due to texture undersampling are hardly visible.

Fig. 4a demonstrates that, if the surface texture is only reconstructed from a single time step, severe rendering artifacts may appear at segment boundaries which are not observed if our spatio-temporal method is applied (Fig. 4b).

The reconstructed realistic virtual humans can be used to realistically populate artificial virtual environments (Fig. 4c). Currently, our method can not handle wide apparel, and some undersampled regions may still be visible if the person did not attain sufficiently different body postures.

Nonetheless, our results show that the employment of image data of multiple body poses during shape adaptation and texture reconstruction enables us to reconstruct human avatars that exhibit a very high level of authenticity.

7. CONCLUSION

We have presented an automatic model-based approach to generate a personalized avatar from multi-view video streams showing a moving person. By employing dynamic multi-view image data for shape customization and texture reconstruction we obtain convincing virtual humans that exhibit a visual quality that would not have been achievable by reconstructing from single-pose photographs.

8. REFERENCES

[1] R. Byrd, P. Lu, J. Nocedal, and C. Zhu. A limited memory algorithm for bound constrained optimization. *SIAM J. Sci. Comp.*, 16(5):1190–1208, 1995.

[2] J. Carranza, C. Theobalt, M. A. Magnor, and H.-P. Seidel. Free-viewpoint video of human actors. *ACM Trans. Graph.*, 22(3):569–577, 2003.

[3] E. de Aguiar, C. Theobalt, M. Magnor, H. Theisel, and H.-P. Seidel. Marker-free model reconstruction and motion tracking from 3d voxel data. *Proc. IEEE Pacific Graphics 2003, Seoul, South Korea*, pages 101–110, Oct. 2004.

[4] A. Hilton, D. Beresford, T. Gentils, R. Smith, and W. Sun. Virtual people: Capturing human models to populate virtual worlds. In *CA '99: Proceedings of the Computer Animation*, page 174, Washington, DC, USA, 1999. IEEE Computer Society.

[5] I. A. Kakadiaris and D. Metaxas. 3d human body model acquisition from multiple views. In *ICCV '95: Proceedings of the Fifth International Conference on Computer Vision*, page 618, Washington, DC, USA, 1995. IEEE Computer Society.

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

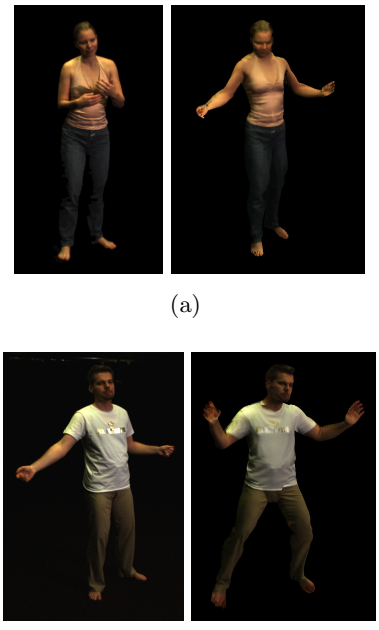
[7] W.-S. Lee, J. Gu, and N. Magnenat-Thalmann. Generating animatable 3D virtual humans from photographs. In M. Gross and F. R. A. Hopgood, editors, *Computer Graphics Forum (Eurographics 2000)*, volume 19(3), 2000.

[8] W. Matusik, C. Buehler, R. Raskar, S. Gortler, and L. McMillan. Image-based visual hulls. In *Proceedings of ACM SIGGRAPH 00*, pages 369–374, 2000.

[9] S. Moezzi, L.-C. Tai, and P. Gerard. Virtual view generation for 3d digital video. *IEEE MultiMedia*, 4(1):18–26, 1997.

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

[11] S. Paquette. 3d scanning in apparel design and human engineering. *IEEE Comput. Graph. Appl.*, 16(5):11–15, 1996.



(a)

(b)

Figure 5: (a),(b) Side-by-side comparison between the real human (left) and the reconstructed avatar in a novel pose.