VIDEO-DRIVEN ANIMATION OF HUMAN BODY SCANS

Edilson de Aguiar, Rhaleb Zayer, Christian Theobalt, Marcus Magnor and Hans-Peter Seidel

MPI Informatik, Saarbrücken, Germany

ABSTRACT

We present a versatile, fast and simple framework to generate animations of scanned human characters from input multi-view video sequences. Our method is purely mesh-based and requires only a minimum of manual interaction. The proposed algorithm implicitly generates realistic body deformations and can easily transfer motions between human subjects of completely different shape and proportions. We feature a working prototype system that demonstrates that our method can generate convincing lifelike character animations from marker-less optical motion capture data.

Index Terms— Image motion analysis, Computer graphics, Animation.

1. INTRODUCTION

In recent years, photo-realistic computer-generated animations of humans have become the most important visual effect in motion pictures and computer games. To generate virtual people, animators make use of a well-established but often inflexible set of tools (see also Sect. 2) that makes a high amount of manual interaction unavoidable.

First, the geometry of the human body is hand-crafted in a modeling software or obtained from a laser scan of a real individual [1]. In a second step, a kinematic skeleton model is implanted into the body by means of, at best, a semi-automatic procedure [2]. In order to couple the skeleton with the surface mesh, an appropriate representation of pose-dependent skin deformation has to be found [3]. Finally, a description of body motion in terms of joint parameters of the skeleton is required. It can either be designed in a computer or learned from a real person by means of motion capture [4, 5]. Although the interplay of all these steps delivers animations of stunning naturalness, the whole process is very labor-intensive and does not easily allow for the interchange of animation descriptions between different virtual persons.

In this paper, we present a versatile, fast and simple mesh-based approach to animate human scans that completely integrates into the animator's traditional animation workflow. Our system produces realistic pose-dependent body deformations implicitly by means of a harmonic field interpolation. Furthermore, it solves the motion transfer problem, i.e. it enables the animator to interchange motions between persons of even widely different body proportions with no additional effort.

The paper proceeds with a review of related work in Sect. 2. An overview of our approach is given in Sect. 3, and our shape deformation method is described in Sect. 4. We demonstrate that we can realistically animate human scans using marker-less motion capture data in Sect. 5. Finally, results and conclusions are presented in Sect. 6.

2. RELATED WORK

The first step in human character animation is the acquisition of a human body model comprising a surface mesh and an underlying animation skeleton [6]. Thereafter, mesh and skeleton have to be connected such that the surface deforms realistically with the body motion [3]. The virtual human is awakened by specifying motion parameters for the joints in the skeleton. The most authentic method to generate such motion descriptors is through optical marker-based [4] or marker-free motion capture [5]. Unfortunately, reusing motion capture data for subjects of different body proportions is not trivial, and requires computationally expensive motion editing [7] and motion retargetting techniques [8].

By extending ideas on mesh deformation techniques we propose a versatile and simple framework to animate human scans. In the mesh editing context, see [9, 10], differential coordinates are used to deform a mesh while preserving its geometric detail. The potential of such methods for mesh editing [9] and animation [11, 12] has already been stated in previous publications. Most recently, a multi-grid technique for efficient deformation of large meshes was presented [13] and a framework for performing constrained mesh deformation using gradient domain techniques has been developed in [14]. Both methods are conceptually related to our system. However, none of the papers provides a complete integration of the surface deformation approach with a marker-less motion acquisition system. On the other hand, we see potential use of these methods within our framework for enhancing the animation quality and the speed of our system.

Our system is most closely related to the SCAPE method [15]. The SCAPE model learns pose and shape variation across in-

Thanks to EC for supporting within FP6 under Grant 511568 with the acronym 3DTV

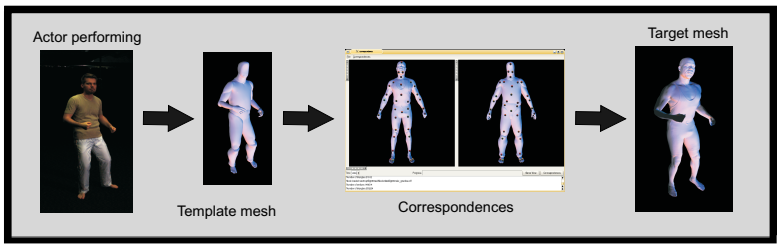


Fig. 1. Illustration of the workflow of our system.

dividuals from a database of body scans by solving a nonlinear optimization problem. Our approach addresses this problem by solving simple linear systems, hence delivering instantaneous results. By relying on the semantic similarities between characters it also provides an alternative solution to the retargetting problem.

3. OVERVIEW

The workflow of our algorithm is shown in Fig. 1. The input to our framework is a multi-view video sequence of the subject performing. By using a marker-less optical motion estimation method, the subject's motion data is captured. As motion description, our marker-free motion capture method outputs a sequence of template mesh poses, henceforth termed *template mesh*. Thereafter, the user specifies a sparse set of correspondences between triangles of the *template mesh* and triangles of the mesh to be animated, henceforth termed the *target mesh*. This is easily and intuitively done using our prototype interface, Fig. 1. Using the correspondences, we transfer the motion from the moving template mesh onto the target mesh using a Laplacian mesh deformation technique, Sect. 4. As a result, the human scan performs the same motion as its real-world counterpart in the images, Sect. 5.

4. MESH DEFORMATION

The algorithmic core of our approach is a mesh deformation method that transfers motion from the template mesh onto the target mesh. We regard motion transfer as a pure deformation interpolation problem. This way, we put aside all difficulties relating to the dissimilarities between the template and the target, e.g. anatomical disparity (body proportions), and take advantage of their semantic similarities, e.g. the fact that both mesh representations have knees and elbows.

For this purpose, the user is asked to specify a set of *correspondence triangles* between the two meshes. In practice, this means that the user marks a set of triangles on the template and assigns to each of them a corresponding triangle on the target. This can be interactively done using our prototype interface tool. We resort to this interactive step since there exists no viable automatic approach that can identify body segments on meshes standing in general poses.

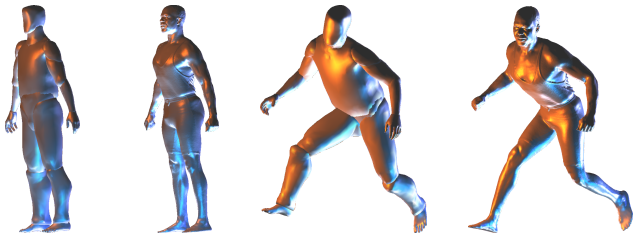
The motion of the template mesh from its reference pose (e.g. Fig. 2a) into another pose (e.g. Fig. 2c) can be captured by the deformation applied to a set of marked triangles. A correct interpolation of this deformation applied over the corresponding triangles of the target mesh would bring it from its own reference pose (e.g. Fig. 2b) into the template's pose (e.g. Fig. 2d). To this end, both reference poses are roughly aligned a priori.

After specifying per-triangle rotations for all marked triangles using quaternions, we follow an idea proposed in [16] and regard each component of a quaternion $Q = [w\ q_1\ q_2\ q_3]$ as a scalar field defined over the entire mesh. Hence, given the values of these components at the marked triangles, we interpolate each scalar field independently. In order to guarantee a smooth interpolation we regard these scalar fields as harmonic fields defined over the mesh. The interpolation can then be performed efficiently by solving the Laplace equation over the whole mesh with constraints at the correspondence triangles: $\nabla^2 S = 0$, where S is a scalar field which alternatively represents each of the quaternion components w, q_1, q_2 and q_3.

Once the rotational components are computed, we average the quaternion rotations of the vertices to obtain a quaternion rotation for each triangle. This way we establish a geometric transformation for each triangle of the target mesh M. However, this last step destroys its original connectivity and yields a new fragmented mesh M'. In order to recover the original geometry of the mesh while satisfying the new rotations, we have to solve the problem in a least square sense. The problem can be rephrased as finding a new tight mesh having the same topology as the original target mesh, such that its differential coordinates encode the same geometric detail as the ones of the fragmented mesh M'. This can be achieved by satisfying the following equation in terms of the coordinates x of M and u of M':

$$\nabla^2_M x = \nabla^2_{M'} u\,. \tag{1}$$

In order to carry out this discretization correctly the topological difference between both meshes should be addressed. Technically, the differential coordinates of the fragmented mesh are computed by deriving the Laplacian operator for the fragmented mesh and then applying it to its coordinates. This, in fact, yields a vector of size $3 \times nT$, where nT is the number of triangles. We sum the components of this vector accord-



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

Fig. 2. A template model (a) and a high-resolution body scan (b) in their respective reference poses. The template in a pose obtained via motion capture (c) and its pose transferred to the human scan (d).

ing to the connectivity of the original mesh M. This yields a new vector $U_{reduced}$ of size nV, where nV is the number of vertices in M, and the discrete form of Eq. 1 reads as simple as

$$LX = U_{reduced} \,, \tag{2}$$

where the matrix L is the discrete Laplace operator. During the processing of an animation sequence, the differential operator matrix does not change. Furthermore, since it is symmetric positive definite we can perform a sparse Cholesky decomposition as preprocessing step and perform only back substitution for each frame. This enables us to compute novel poses of the target mesh at interactive rates for meshes of the order of 30 to 50 thousand triangles.

5. MARKER-LESS ANIMATION

Marker-less tracking methods estimate motion parameters from image features in the raw video footage showing a moving person. Using such motion capture technique as front-end to our algorithm, we are able to create two intriguing applications: video-driven animation and 3D video.

5.1. Video-driven Animation

For non-intrusively estimating motion parameters, we make use of the passive optical motion capture approach proposed in [17]. To this end, we record a moving person with eight static video cameras that are roughly placed in a circle around the center of the scene. From the frame-synchronized video streams, the shape and the motion parameters of the human are estimated. To achieve this purpose, a template model, (see Fig. 2a), comprising of a kinematic skeleton and sixteen separate closed surface segments is fitted to each time step of video by means of silhouette-matching. The output of the method conveniently represents the captured motion as a sequence in which the template model subsequently strikes the estimated body poses.

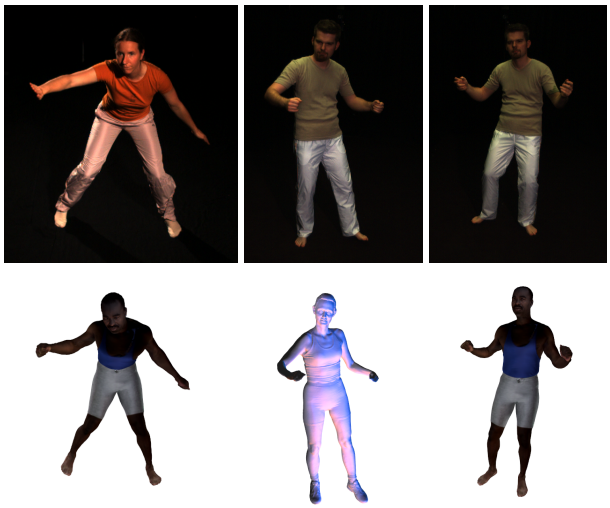


Fig. 3. Video-driven animation: Motion parameters are extracted from raw video footage of human performances (top row). By this means, body poses of a video-taped individual can easily be mapped to body scans of other human subjects (bottom row).

This output format can be directly used as input to our pipeline. The animator specifies triangle correspondences between the template and the scanned mesh that shall be animated. Finally, our algorithm makes the human scan mimic the motion that we have captured in video. Realistic surface deformations of the output mesh are implicitly generated. In order to demonstrate the performance of video-driven animation, we animate our female (264K triangle) and male (294K triangle) Cyberware scans with two very different captured motion sequences. The first sequence contains 156 frames and shows a female subject performing a capoeira move. The second sequence is 330 frames long and shows a dancing male subject. Fig. 3 shows a comparison between actual input video frames and human scans striking similar poses. It illustrates that body poses recorded on video can be faithfully transfered to 3D models of arbitrary human subjects. Differences in body shape and skeletal proportions can be completely neglected.

5.2. 3D Video

By means of video-driven animation, Sect. 5.1, we can also generate 3D videos of moving characters. In the traditional model-based approach to 3D video a simplified body model is used to carry out a passive optical motion estimation from multiple video streams [17]. During rendering, the same simplified shape template is displayed in the sequence of captured body poses and textured from the input videos. Although these methods deliver realistic free-viewpoint renditions of virtual actors, we expect that a more accurate underlying geometry increases realism even further.

Fig. 4. Our approach enables the creation of 3D videos with high-quality geometry models.

To demonstrate the feasibility of this approach in practice, we have acquired full-body surface scans of several individuals in our studio. To this end, we merged several partial body scans performed with our MINOLTA VI-910 which is best suited for scanning small objects. Thus the quality of our scans is far below the quality of scans acquired using full-body scanners. For each scanned individual, we also recorded several motion sequences with multiple synchronized video cameras. We use the method from Sect. 5.1 to animate the scans from the captured motion data. During 3D video display, the animated scan is projectively textured with the captured video frames.

Fig. 4 shows two free-viewpoint renditions of a dynamically textured animated scan in comparison to input images of the test subject. The free-viewpoint renditions reflect the true appearance of the actor. Since we are given a better surface geometry, texture blending artefacts are hardly observed. Remaining artefacts in the rendering can be clearly attributed to the non-optimal scanning apparatus we used.

6. RESULTS AND CONCLUSION

To demonstrate the potential of our method we conducted several experiments. Due to their high resolution, we used the Cyberware models in most of our experiments. Marker-less motion acquisition enables us to perform video-driven animation. Both of our models in Fig. 3 authentically mimic the human performances captured on video. This also allows for producing 3D video, Fig. 4. The substantiated results and the accompanying video (that can be downloaded from [18]) confirm that our method is capable of animating human scans at a low interaction cost.

As for any novel technique our method still has some limitations. For extreme deformation we note that there is generally some loss in volume due to the nature of our interpolation. We expect that using the volumetric approach proposed in [19] would reduce such artefacts. Another limitation is that our system can not enforce hard constraints. Our method satisfies the deformation constraints in a least-square sense. Although it is not possible to explicitly enforce hard constraints, they can be implicitly enforced by increasing the number of correspondences associated with a marker.

We nonetheless devised a powerful framework for animating human scans. The proposed method is easy and intuitive to use. By means of the same efficient methodology our approach simultaneously solves the animation, the surface deformation and the motion retargetting problem. Since our method relies only on setting up and solving linear systems, the implementation and the reproduction of our results are straightforward. As a direction for future work, we would like to combine our technique with an approach to learn per-time-step surface deformations from input video footage.

7. REFERENCES

[1] B. Allen, B. Curless, and Z. Popovic, "The space of human body shapes: reconstruction and parameterization from range scans," *ACM Trans. Graph.*, vol. 22, no. 3, pp. 587–594, 2003.

[2] L. Herda, P. Fua, R. Plänkers, R. Boulic, and D. Thalmann, "Skeleton-based motion capture for robust reconstruction of human motion," in *Proc. of CA '00*. 2000, p. 77ff, IEEE Computer Society.

[3] J. P. Lewis, M. Cordner, and N. Fong, "Pose space deformation: a unified approach to shape interpolation and skeleton-driven deformation," in *Proc. of ACM SIGGRAPH'00*, 2000, pp. 165–172.

[4] B. Bodenheimer, C. Rose, S. Rosenthal, and J. Pella, "The process of motion capture: Dealing with the data," in *Computer Animation and Simulation '97*, Sept. 1997, pp. 3–18.

[5] T. B. Moeslund and E. Granum, "A survey of computer vision-based human motion capture," *CVIU*, vol. 81, no. 3, pp. 231–268, 2001.

[6] N. Badler, D. Metaxas, and N. Magnenat Thalmann, *Virtual Humans*, Morgan Kaufmann, 1999.

[7] M. Gleicher, "Motion editing with space-time constraints," in *Proc. of 1997 Symposium on Interactive 3D Graphics*, 1997, p. 139ff.

[8] S. Tak and H.-S. Ko, "A physically-based motion retargeting filter," *ACM Trans. Graph.*, vol. 24, no. 1, pp. 98–117, 2005.

[9] Marc Alexa, Marie-Paule Cani, and Karan Singh, "Interactive shape modeling," *In Eurographics course notes (2005).*

[10] Olga Sorkine, "Differential representations for mesh processing," *Computer Graphics Forum*, vol. 25, no. 4, 2006.

[11] R. W. Sumner and J. Popovic, "Deformation transfer for triangle meshes," *ACM Trans. Graph.*, vol. 23, no. 3, pp. 399–405, 2004.

[12] R. W. Sumner, M. Zwicker, C. Gotsman, and J. Popovic, "Mesh-based inverse kinematics," *ACM Trans. Graph.*, vol. 24, no. 3, pp. 488–495, 2005.

[13] L. Shi, Y. Yu, N. Bell, and W.-W. Feng, "A fast multigrid algorithm for mesh deformation," *ACM Trans. Graph.*, vol. 25, no. 3, pp. 1108–1117, 2006.

[14] J. Huang, X. Shi, X. Liu, K. Zhou, L.-Y. Wei, S.-H. Teng, H. Bao, B. Guo, and H.-Y. Shum, "Subspace gradient domain mesh deformation," *ACM Trans. Graph.*, vol. 25, no. 3, pp. 1126–1134, 2006.

[15] D. Anguelov, P. Srinivasan, D. Koller, S. Thrun, J. Rodgers, and J. Davis, "Scape: shape completion and animation of people," *ACM Trans. Graph.*, vol. 24, no. 3, pp. 408–416, 2005.

[16] R. Zayer, C. Rössl, Z. Karni, and H.-P. Seidel, "Harmonic guidance for surface deformation," in *Proc. of Eurographics 2005*, 2005, vol. 24, pp. 601–609.

[17] J. Carranza, C. Theobalt, M. Magnor, and H.-P. Seidel, "Free-viewpoint video of human actors," *ACM Trans. Graph. (Proc. of SIGGRAPH'03)*, vol. 22, no. 3, pp. 569–577, 2003.

[18] ," http://www.mpi-inf.mpg.de/∼edeaguia/3dtvConVideo.avi .

[19] K. Zhou, J. Huang, J. Snyder, X. Liu, H. Bao, B. Guo, and H.-Y. Shum, "Large mesh deformation using the volumetric graph laplacian," *ACM Trans. Graph.*, vol. 24, no. 3, pp. 496–503, 2005.