

From Images To Animation Models

P. Fua
Computer Graphics Lab (LIG), EPFL
CH-1015 Lausanne, Switzerland
Pascal.Fua@epfl.ch*

Abstract

We show that we can effectively fit complex animation models to noisy image data. Thus, complete head models can be acquired with a cheap and entirely passive sensor. The shape and motion parameters of limbs can be similarly captured and fed to existing animation software to produce synthetic sequences.

In this paper, we show that we can effectively fit complex facial and body animation models to calibrated or uncalibrated sets of images and video sequences. We need minimal human intervention and neither targets nor structured light nor any other active device. We first introduce our animation models and then describe the fitting techniques we have developed.

Animation Models The modeling and animation of Virtual Humans has traditionally been decomposed into two subproblems: facial animation and body animation.

In both cases, muscular deformations must be taken into account, but their roles and importance differ. Facial animation primarily involves deformations due to muscular activity. Body animation, on the other hand, is a matter of modeling a hierarchical skeleton with deformable primitives attached to it so as to simulate soft tissues. It is possible to model bodies in motion while ignoring muscular deformations, whereas to do so for facial animation is highly unrealistic.

For heads, we use the facial animation model that has been developed at University of Geneva and EPFL and is depicted by Figure 1(a). It can produce the different facial expressions arising from speech and emotions.

The body model we use has been developed at EPFL and is depicted by Figure 1(b,c,d,e). It incorporates a highly effective multi-layered approach for constructing and animating realistic human bodies. Ellipsoidal metaballs are used to simulate the overall behavior of bone, muscle, and fat tissue; they are attached to the skeleton and arranged in an anatomically-based approximation. Skin construction is a three step process: First, the implicit surface resulting from the combination of the metaballs influence is automatically sampled along cross-sections. Second, the sampled points become control points of a B-spline patch for each body part. Third, a polygonal surface representation is constructed by tessellating those B-spline patches for seamless joining of different skin pieces and final rendering. This simple and intuitive method combines the advantages of implicit, parametric and polygonal surface representation, producing very realistic and robust body deformations.

*The work reported here was funded in part by the Swiss National Science Foundation

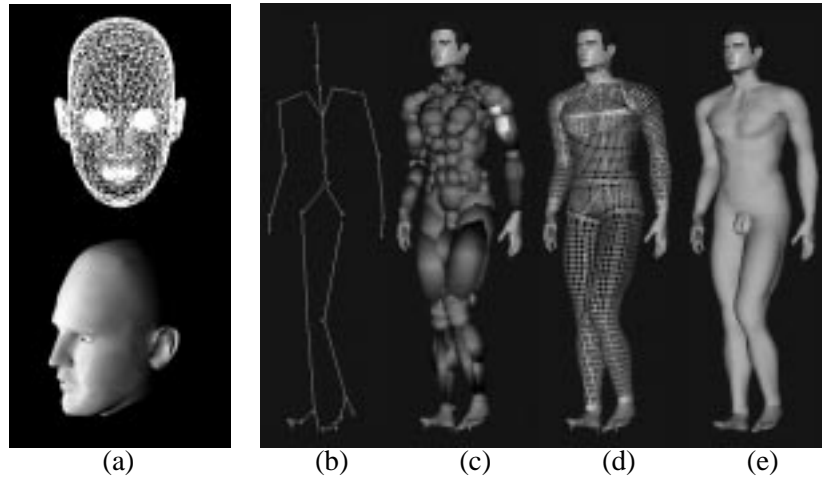


Figure 1: The layered human model: (a) Model used to animate heads, shown as a wireframe at the top and as a shaded surface at the bottom. (b) Skeleton. (c) Ellipsoidal metaballs used to simulate muscles and fat tissue. (d) Polygonal surface representation of the skin. (e) Shaded rendering.

Fitting Head Models Given a set of uncalibrated images or a video sequence, our goal is to fit the animation mask of Figure 1(a) with minimal manual intervention. To this end, our procedure takes the following steps:

- **Estimation of relative head motion:** We match regularly sampled points in the images and use a bundle-adjustment approach to estimate the motion. We introduce model-based regularization constraints that allow a reliable recovery of the motion parameters, even though the quality of the points matches cannot be expected to very good.
- **Computation of reliable 3-D surface information:** We compute disparity maps for each stereo pair or each consecutive pair in the video sequences, fit local surface patches to the corresponding 3-D points, and use these patches to compute a central 3-D point and a normal vector.
- **Modeling of the Face:** We attach a coarse control mesh to the face of the animation model and perform a least squares adjustment of this control mesh so that the model matches the previously computed data. We weigh the data points according to how close—in the least squares sense—they are to the model and use an iterative reweighting technique to eliminate the outliers. We then subdivide the control mesh and repeat the procedure to refine the result.
- **Modeling of the Hair:** We repeat the same procedure for the hair. Because hair is where stereo tends to fail, we typically rely more heavily on the use of silhouettes to derive an estimate of the shape than we do for the face.
- **Generation of a texture map:** We use the original images to compute a cylindrical texture map that allows realistic rendering. This is achieved by first generating a cylindrical projection of the head model and then, for each projected point, finding the images in which it is visible and averaging the corresponding gray-levels.

Figure 2 depicts the output of our modeling procedure. In all cases, the head models can be animated. To ensure that some of the key elements of the face—corners of the eyes, mouth and hairline—project

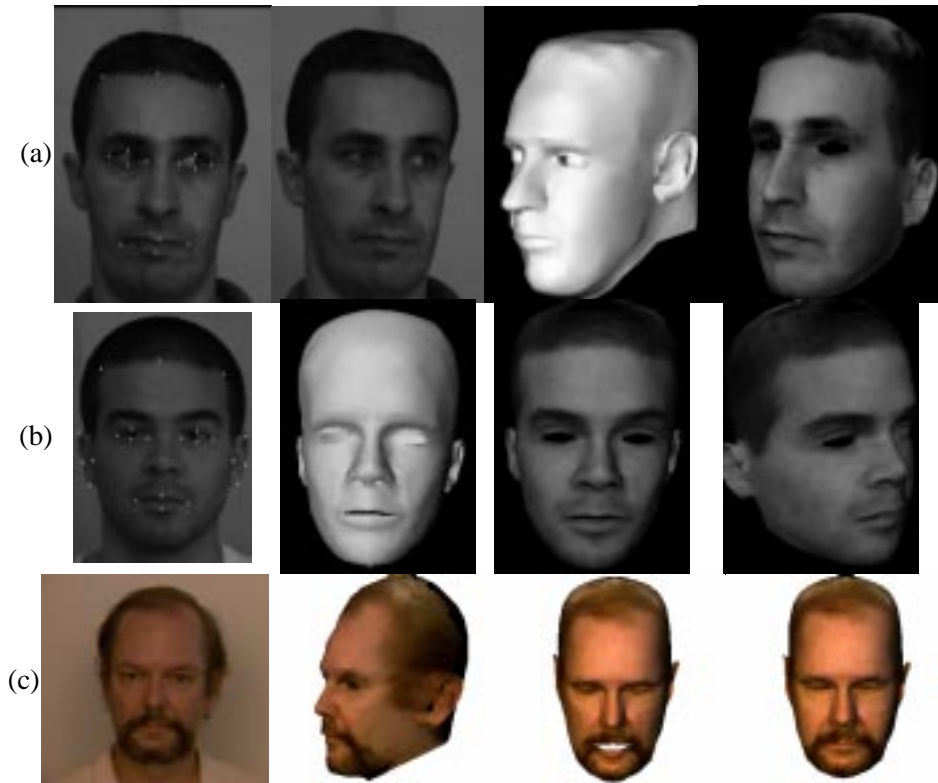


Figure 2: Facial reconstruction and animation. (a) Two images out of a sequence of nine; shaded view of the complete head model; and, textured model. (b) For a second person, a shaded and two textured views of the head model. (c) For a third person, a texture-mapped view of the corresponding head model; and two synthetic expressions, opening of the mouth and closing of the eyes.

at the right places, we have manually supplied the location of the projection in one single image of a few feature points such as the ones shown in the first column of Figure 2. We add observation equations that force the projection of the generic mask’s corresponding vertices to be close to them. Among these points, only five are required—corners of the eyes and mouth and tip of the nose—and the other are optional. To produce these face models, the required manual intervention therefore reduces to supplying these few points by clicking on their approximate locations in only one image, which can be done quickly.

To illustrate the potential of this technique, we have asked a graphics designer to replace the heads of crew members of the original Star Trek TV series by the synthetic heads of Figure 2 to produce Figure 3. In this case, this was achieved by manually fitting heads on synthetic bodies and recreating a completely synthetic environment. In the future, we intend to extend our approach to automate this process.

Fitting Body Models We now turn to the modeling of the motion and shape of a limb. Here, we use a simplified version of the body animation model of Figure 1—with only three metaballs per limb—to approximate the shape and derive joint angles values for the underlying skeleton. In future work, we will use this knowledge to initialize the complete model and further optimize its shape using the same input data.

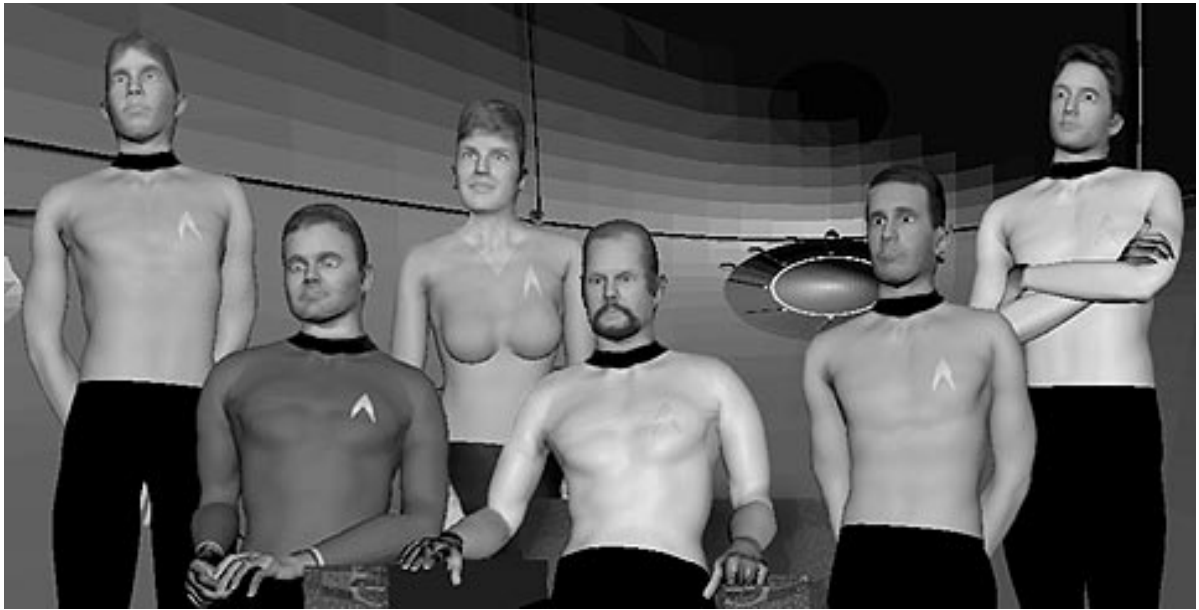


Figure 3: Star Trek revisited: Entirely synthetic crew members whose heads have been reconstructed from video sequences.

The model's state is defined by the Euler angles that encode the position and posture of the body with respect to its rest position and by the radii of the metaballs. In standard least-squares fashion, we use the image data to write observation equations. In the case of 3-D points derived from stereo, these observation equations are designed to minimize the distance of the reconstructed limb to all such points. We use again an iterative reweighted least squares algorithm to progressively discount outliers.

Using this approach and the noisy stereo data of Figure 4(b), we can reconstruct the positions and shapes of the arm as shown in Figure 4(c). The recovered joint angles can then be used to animate the virtual human of Figure 4(d).

Conclusion and Further Reading We have presented a technique that allows us to fit complex animation models for human faces and bodies to noisy image data with very limited manual intervention. As a result, these models can be produced cheaply and fast. Even though they were primarily designed for animation rather than fitting purposes, we have designed a framework that allows us to exploit them to resolve ambiguities in the image data.

In future work, we intend to extend the approach to capturing not only the static shapes of the body and face but also to characterizing their motion. We will also extend the self-calibration techniques used here for faces to the full body model so that we can model complete people using ordinary video sequences acquired with regular camcorders.

For further details, we refer the interested reader to the following publications:

- Facial Modeling:
 - P. Kalra, A. Mangili, N. Magnenat Thalmann, and D. Thalmann, “Simulation of Facial Muscle Actions Based on Rational Free Form Deformations,” in *Eurographics*, 1992.

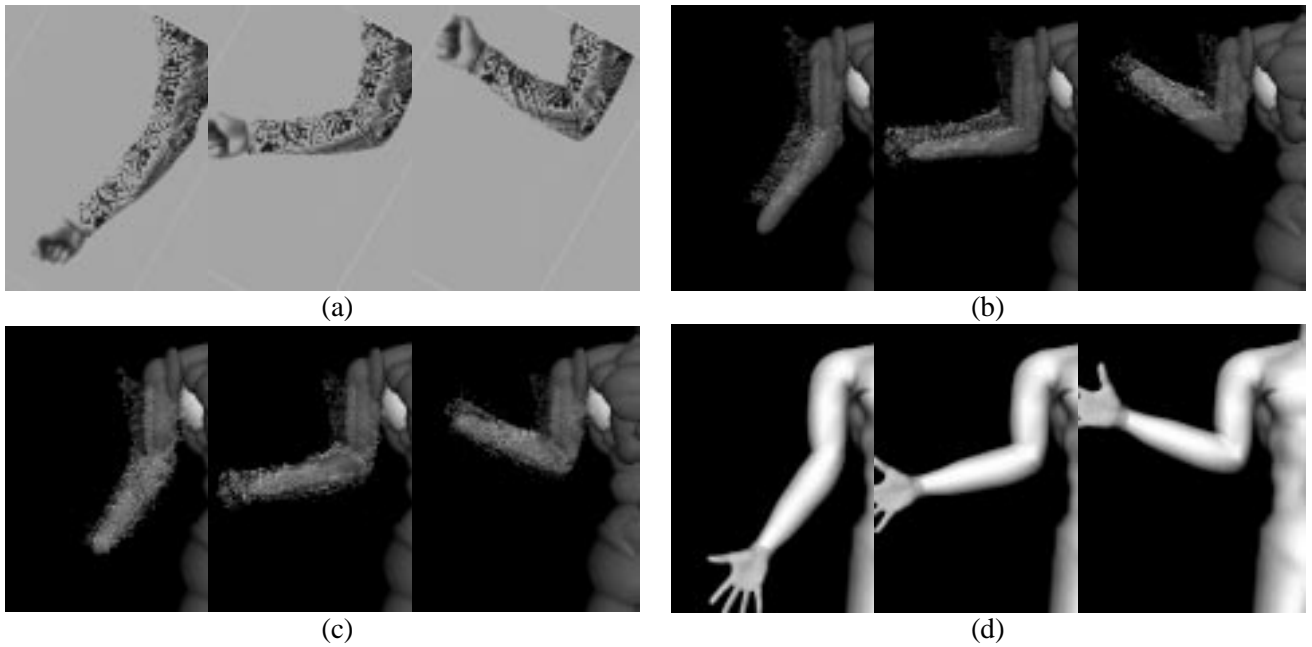


Figure 4: Fitting the model to stereo data. (a) Images taken by the first of two synchronized video cameras at three different times. (b) The 3-D points derived from stereo and the initial position of the model. (c) After fitting, the system has computed the joint angles. (d) An avatar can be made to reproduce the same gesture.

- P. Fua. Modeling Faces from Video Sequences. In *Electronic Imaging, SPIE Photonics West Symposium*, San Jose, CA, January 1999.

- Body Modeling:

- J. Shen and D. Thalmann, “Interactive shape design using metaballs and splines,” in *Implicit Surfaces*, April 1995.
- N. D’Appuzo, R. Plänkner, P. Fua, A. Grün, and D. Thalmann. Modeling Human Bodies from Video Sequences. In *Electronic Imaging, SPIE Photonics West Symposium*, San Jose, CA, January 1999.

These papers and additional examples are available on our web site: <http://ligwww.epfl.ch/fua/>