Constructing A Realistic Head Animation Mesh for a Specific Person

Ulrich Neumann, Jiankun Li, Reyes Enciso, Jun-Yong Noh, Douglas Fidaleo, and Tae-Yong Kim

Integrated Media Systems Center, Computer Science Department, USC

Lucien Nocera

Eyematic Interfaces, Inc

Abstract

This paper addresses the problem of constructing an realistic and complete animation mesh that portrays a specific person's head geometry and texture. Our approach deforms a prototype mesh containing vector-based delineated muscles to fit one or more geometric models obtained from stereo image pairs of a specific person's head. The resulting personalized mesh facilitates animation with the same realism and predictability as the original prototype mesh. The model construction requires some manual interaction, however automatic refinement methods reduce the need for precision. The sensing process is passive and no physical markers are needed on the person's face. Models produced by our method are suited to realistic animations of specific individuals for applications in special effects, games, and 3D teleconferencing.

1. Introduction

Ever since the pioneering work of Frederic I. Parke [1] in 1972, researchers have attempted to generate realistic facial models and animation. Recent interest in facial modeling and animation is spurred by the increasing appearance of virtual characters in film and video, inexpensive desktop processing power, and the potential for new 3D immersive communication metaphors for human-computer interaction.

The complexity of human facial anatomy and our natural sensitivity to facial appearance increase the difficulty of modeling human facial appearance and subtle expressions. Although some recent work [2-5] produces realistic results with relatively fast performance, the process of generating a specific person's head model suitable for facial animation often entails extensive human intervention, physical markers on the face, or tedious tuning.

We create realistic 3D models of specific people from multiple stereo image pairs. The approach requires some interactive user interaction, however automatic refinement allows the user input to be approximate. Our target application is 3D teleconferencing or tele-immersion, an emerging technology for interpersonal communication in distributed virtual environments [6]. This application requires real-time animation of photo-realistic 3D models of specific people.

Figure 1 illustrates our method. Volume morphing fits a prototype mesh to one or more reconstructed models of a specific person. These models are reconstructed from stereo image pairs. To obtain a complete model of the head, we may acquire frontal and side views of the person's face, each producing separate reconstructions of the visible head regions. Our system facilitates the merging of multiple models during the fitting of the prototype mesh. Our prototype mesh contains only 1700 polygons (for speed purposes) compared to about 20-40K triangles per reconstruction, therefore the

Fig. 1 - Our approach morphs a prototype mesh (a) to fit one or more reconstructed models (b) to produce a complete personalized animation mesh (c).
fitting process is implicitly also a simplification or compression process. The prototype mesh contains the animation mechanism, (in our case muscle-based information per vertex), so after the fitting is complete, we can animate the personalized complete head animation model with realism and predictability.

The novelty of our approach is in its integration of high-quality stereo reconstruction and texture mapping; semi-automatic volume morphing and refinement; and texture blending. One advantage of our approach is a decrease in the usual number of correspondences needed to build a specific face model from hundreds to 12 points per mesh.

The remainder of paper is organized as follows. In section 2 we introduce the two prototype meshes and the animation method we use in our experiments. For comparison, our initial experiments with 3D scanners for head modeling are reviewed in section 3, and our new approach with stereo images is detailed in section 4. Section 5 describes the volume morphing and refinement method for constructing the final model.

2. The Animation System

Our target application is a performance-driven animation system with unobtrusive sensing (i.e, no markings on the face). Our modeling approach does not depend on any particular animation method or even a polygonal mesh, but our implementation is built around a prototype mesh that incorporates a muscle-based animation engine. Based on work by Waters and Terzopoulos [7, 8], our real-time implementation mimics facial gestures like blinking, smiling, jaw motion, and head pose changes [9]. The system also contains a simple backup control system to maintain reasonable parameters in the case of network or sensor failure. Muscle-based animation is independent of the mesh resolution so in addition to the Waters model, we selected a Viewpoint mesh of a complete head that enables our system to run in real time while maintaining visual realism (Fig. 2).

These two geometric meshes become prototype animation meshes with the addition of muscle-based information per vertex. The Waters delineated muscle models are defined by the vector field direction, an origin, and an insertion point (Fig 3). The field extent is defined by cosine functions and fall off factors that produce a cone shape when visualized as a height field. Waters also models the mouth sphincter muscles as a simplified parametric ellipsoid. The sphincter muscle contracts around the center of the ellipsoid and is primarily responsible for the deformation of the mouth region.

3. Models from a 3D Scanner

A 3D scanner (e.g., Cyberware) can acquire a 3D model of a person's head. These scanners often produce artifacts when digitizing heads. Hair, for example is generally not digitized unless the hair is artificially colored white, but that destroys the texture. The scanners covers the full 360 degrees around the head in 20-30 seconds, during which user motion causes artifacts. Our experiments also required hand-tuning of the geometry to remove outliers, smooth geometry, and add details like the top of the head.

Figure 4 (a) shows a 3D scanner in operation. The output for a dark-haired person is shown in (b), illustrating the loss
of hair during digitizing and some outlier artifacts. Figure 4 (c) and (d) show a post-processed model without and with texture, respectively. The hair was painted white during the scan to ensure digitization. There are approximately 40K polygons in the output scans. After post-processing, the scanner data becomes the target for fitting a prototype mesh. The fitting process is described in greater detail for the stereo reconstruction models.

4. Models from Stereo Reconstruction

Stereo reconstruction has been an active area of research in computer vision. The stereo algorithm we use is an area-based correlation approach [11] based on [12, 13]. This class of algorithms provides accurate and dense 3D information that may in turn define higher-level object descriptions.

We calibrate the stereo camera pair with a method based on [14] developed by T. Padjila and R. Sara from the University of Praha [12]. Calibration computes the matrix relating a 3D point in space to its projection on the image plane. To compute this projection we need the 3D coordinates of a known point in a world coordinate frame with high accuracy and the projection of these points on the image. One can use a calibrated fixture with two planes, but such objects require precision manufacturing. The other possibility is to use a known planar pattern (Fig. 5), and introduce the third dimension by displacing the pattern with a calibrated mechanical linear motion platform. General-purpose platforms provide controlled precise displacements.

The stereo cameras are two color CCD Sony XC-999 cameras with 12 mm lenses. We calibrate the system at less than one meter to use full resolution for calibration and operation. The digital image size is 512 x 512. The cameras are in a fixed configuration with a baseline of approximately 8-cm., verging at approximately 40-degrees.

To ensure accurate geometry over the entire face and head, we momentarily project a noise texture onto the scene with a slide projector when acquiring a stereo image pair (Fig. 6). A high-resolution color image for texturing the final model is acquired immediately after disabling the noise texture. The sequence of stereo and texture images may be acquired in consecutive video frames.

To recover range data from (polynocular) stereo, the corresponding projections of the spatial 3D points have to be found in the left and right images. This is known as the correspondence (matching) problem. To reduce the search-space dimension from a 2D region to a 1D line, we rectify the images so that all corresponding points lie on the same scan lines [15]. By definition, corresponding points have coordinates \((u, v)\) and \((u-d, v)\), in left and right rectified images, and \(d\), the distance in pixels, is the disparity.

A modified normalized cross-correlation measures the degree of correspondence (see [11] for a review of common correlation functions). The correlation score \(c\) measures the similarity between the selected windows in the left and right images (7x7 or 11x11 are usual window sizes).
where \( I_l \) and \( I_r \) are the (left/right) rectified images over the selected correlation windows, and the covariance \( \text{cov}(I_l, I_r) \) and variance \( \text{var}(I) \) are defined over these square windows assuming a Gaussian distribution. For each pixel \((u, v)\) in the left image, the matching produces a correlation profile \( c(u, v, d) \), where \( d \) ranges over acceptable integer disparities. The range \( \{d_{\text{min}}, d_{\text{max}}\} \) is fixed in advance, depending on the camera configuration and the scene distance.

All peaks of the correlation profile are possible disparity hypotheses. The resulting list of hypotheses for all positions forms a disparity volume. The hypotheses in the disparity volume are pruned by a selection procedure based on a visibility constraint, ordering constraint, and disparity gradient constraint [16, 11]. The output of this procedure is an integer disparity map.

The reconstruction precision is proportional to the disparity error. A subpixel correction of the integer disparity map can be obtained by a simple interpolation of the correlation scores, or a more general approach described in [17], which takes into account the perspective distortion between the left and right correlation windows for a planar region of surface. The first approach is the fastest while the second gives a more reliable subpixel disparity estimate. Once the subpixel disparity map is computed, interpolation and filtering fill holes and cull outliers. From the disparity map, 3D points are computed by calibrated triangulation and tessellated to define quadrangular facets.

The meshes generated by this method are usually smooth and without significant holes (Fig. 7). Depending on the image size of the head, the mesh size is approximately 20-40K facets. The perspective projection matrix \( P \) is known for the cameras, one of which acquires the color texture image, so the texture mapping is defined by \( P \) and the rectification matrices.

5. Mesh Fitting by Volume Morphing

Scanned or reconstructed models have variable complexity and mesh arrangements. Although these capture the geometry and coloring of a person, they are not suitable for direct animation. The prototype mesh, on the other hand, has muscle parameters suited for animation, but its geometry and texture does not match a specific individual. By fitting the prototype mesh to the reconstruction model(s), the animation mesh takes on the shape and coloring of a specific individual.

There are several steps in a fitting process. The first step is a landmark-based volume morphing where the transformation and deformation of the prototype mesh is guided by the interpolation of a set of landmark points with a radial basis function [18-20]. The landmark volume morphing only guarantees that the morphing of the prototype mesh is accurate near the landmark points, and since these are sparse and scattered, a further optimization is needed to ensure the overall quality of the morph. The second step of surface optimization minimizes a cost function to further improve the overall similarity between the reconstructed model and the prototype mesh based on the Euclidean distance between vertices. A third step for texture extraction generates the spherical personal texture map from the morphed prototype and the original color images. When multiple reconstructed models are available, a fourth step for blending merges the corresponding spherical texture maps into one aggregate final texture map.

5.1. Landmark Volume Morphing

Given a set of landmark points \( P = \{p_i\} \), with \( i = \{1 \ldots n\} \) in the prototype and a corresponding landmark set \( Q = \{q_i\} \), of same size in the reconstructed model, the landmark based volume morph is defined as a mapping \( F: \mathbb{R}^3 \rightarrow \mathbb{R}^3 \) satisfying \( F(p_i) = q_i \).
Applying this mapping to the prototype transforms each landmark point $p_i$ to the position of its corresponding point $q_i$. The quality of the interpolation depends on the landmark point sets as well as the mapping function $F$.

We choose $F$ as a radial basis function, which is known to generate globally smooth interpolation [19]. A radial basis function is defined as a function whose value depends only on the distance from a point. A radial basis function interpolation method is simply the linear combination of such basis functions:

$$F(p) = \sum c_i R(p-p_i)$$

A disadvantage of the above equation is that it doesn’t have affine precision. When $Q$ is an affine transformation of $P$, $F$ can not recover it. A more generic solution is to add a linear term in the equation to absorb the linear part of the morphing. Therefore our interpolant has the form

$$F(p) = \sum c_i R(p-p_i) + Mp + t$$

Where $M$ is the rotation matrix, and $t$ is the vector for translation.

$R$ is selected to be the Hardy function $R(r) = \sqrt{r+d}$. The only parameter $d$ controls the stiffness of the deformation around the landmark points. The coefficients $c_i$, $M$, and $t$ are determined by the interpolation constraints $F(p_i) = q_i$, $i = 1 \ldots n$, and the linear constraints $\sum c_i = 0$ and $\sum c_i p_i^T = 0$.

5.2. Surface Optimization

Landmark points are typically scattered, so a further optimization is needed to ensure the overall quality of morphing. To this effect, we define a cost function as follows:

$$E = E_{dis} + E_{spring}$$

where $E_{dis}$ is the sum of the minimum distances from each point in the reconstructed model to the surface of the prototype mesh, and $E_{spring}$ is a term that measures the energy of the prototype by placing on each edge of the mesh a spring of rest length zero with a fixed spring constant. The positions of all the vertices of the prototype are iteratively adjusted to minimize $E$. The inclusion of the $E_{spring}$ term guarantees the existence of a minimum and regularizes the optimization to a desirable local minimum [21].

5.3. Texture Extraction

This step creates a mapping from the prototype mesh to the reconstructed model texture image(s). For the prototype mesh, we define the texture coordinates for each vertex by a spherical mapping $M_s$. First, we transform the mesh so that it faces the positive $z$-direction and its center is moved to the origin. Then each vertex $v_i$ of the mesh is projected to a point on the unit sphere with colatitude $\theta_i$ and azimuth $\phi_i$, where $\theta_i \in (0, \pi)$ and $\phi_i \in (0, 2\pi)$. The texture coordinates of vertex $v_i$ are defined as: $M_s(v_i) = (\theta_i/\pi, \phi_i/(2\pi))$. The resulting $u,v$ coordinates map to a suitable aspect and resolution image (1024x512 in our tests).

The stereo algorithm provides, for each reconstructed mesh, the mapping function $M_r$, that relates a 3D point to the color texture image. Through the volume morphing and surface optimization, a function $T$ is defined that maps a vertex of the prototype mesh (now

Fig. 8 - Texture mapping for the prototype mesh based on a spherical projection $M_s$. The rectangle in the center shows the region where only the frontal texture is applied.

Fig. 9 - Through morphing and refinement, the reconstruction-model textures map to the prototype texture.
deformed) to its closest neighbors in the reconstructed model(s). Finally, combining the projections as shown in Fig. 9 creates the required mapping between a reconstructed-model texture image and the prototype-mesh texture image.

5.4. Blending

This final stage combines multiple reconstruction-model texture maps into a single, comprehensive prototype-mesh texture map. To merge texture images from different views, we consider the following rules:

- The texture inside the rectangle of Fig. 8 comes only from the frontal view, since it is the most important region.
- The texture outside the rectangle is extracted by a weighted average, where the weighting coefficient is taken as the cosine of the angle between the normal of the corresponding point in the generic model and the viewing direction from which the stereo images are taken.

Figure 10 illustrates the data flow for the entire fitting process and Figure 11 shows the results of a fitting to a member of our research group.
6. Summary and Conclusions

We presented a method for creating person-specific head models for animation. Our method modifies a prototype animation mesh to a specific person with minimal interaction. We integrated the acquisition and use of multiple stereo reconstruction models to create a complete head mesh. We described the fitting process in the context of a muscle-based animation system to illustrate its utility. The manual interactions consist only of clicking a few correspondences on both the prototype mesh and each stereo model. In general, about 12 correspondences distributed among the eyes, nose, ears, and mouth are sufficient to fit the prototype mesh to a new stereo model set. We imagine that these correspondences could be automatically initialized in the near future using the information provided by a vision face recognition/tracking system.

The refinement step is a significant addition to volume morphing. A future goal is to augment the energy function with additional constraints to improve the robustness and automation of the fitting. We believe that “active snakes” can be initialized at the volume morphing stage to more precisely locate the outline of important human facial features such as the eyes and mouth. Since we have full knowledge of the prototype mesh, the additional correspondence information may enable fully automatic model fitting and more accurate models than currently feasible.

7. Acknowledgements

This research is funded in part by Advanced Network & Services [6], The Annenberg Center for Communication at USC, the Integrated Media Systems Center (IMSC) at USC, and Eyematic Interfaces, Inc.

References


