Blendshapes from Commodity RGB-D Sensors

Dan Casas[1], Oleg Alexander*[1], Andrew W. Feng†[1], Graham Fyffe‡[1], Ryosuke Ichikari[1], Paul Debevec§[1], Rhuizhe Wang¶[2], Evan Suma‖[1], and Ari Shapiro**[1]

[1]Institute for Creative Technologies , University of Southern California
[2]University of Southern California

CR Categories: I.3.6 [Computer Graphics]: Methodology and Techniques—Interaction Techniques; I.3.7 [Computer Graphics]: Three-Dimensional Graphics and Realism—Animation;

Keywords: face animation, RGB-D, depth sensors, blendshapes

1 Exposition

Creating and animating a realistic 3D human face is an important task in computer graphics. The capability of capturing the 3D face of a human subject and reanimate it quickly will find many applications in games, training simulations, and interactive 3D graphics. We demonstrate a system to capture photorealistic 3D faces and generate the blendshape models automatically using only a single commodity RGB-D sensor. Our method can rapidly generate a set of expressive facial poses from a single depth sensor, such as a Microsoft Kinect version 1, and requires no artistic expertise in order to process those scans. The system takes only a matter of seconds to capture and produce a 3D facial pose and only requires a few minutes of processing time to transform it into a blendshape-compatible model. Our main contributions include an end-to-end pipeline for capturing and generating face blendshape models automatically, and a registration method that solves dense correspondences between two face scans by utilizing facial landmarks detection and optical flows. We demonstrate the effectiveness of the proposed method by capturing different human subjects and puppeteering their 3D faces in an animation system with real-time facial performance retargeting.

Since our method is low cost, fast, and requires little or no expertise on the part of the operator, this method has the potential to expand the number of photoreal, individualized faces that are available for simulation. Our method is data agnostic, and can utilize scan input from depth-sensors; as the scan quality of such componenets improved, the quality of the resulting blendshapes will also improve. For example, we demonstrate photorealistic results suitable for close-up encounters when using the Occipital Structure Sensor instead of the Kinect v1.

A key to our method is the proper alignment of both texture and geometry. Rather than storing our scans as geometry and textures, we choose instead to store our scans as images. Each one of our scans is stored as a 32 bit float EXR texture map image, and a high resolution point cloud. Our texture and geometry alignment algorithm has three steps and runs in 2D. First, we construct a Delaunay triangulation between the user supplied points and apply affine triangles to roughly pre-warp the source diffuse texture to the target. Second, we use GPU-accelerated optical flow to compute a dense warp field from the pre-warped source diffuse texture to the target. Finally, we apply the dense warp to each one of our source texture maps, including diffuse, specular, specular normals, and point cloud. The result is the source scan warped to the target UV space. The sub-millimeter correspondence is able to align individual pores across the majority of the face.

Figure 1: *Blendshape results using our method with data from a commodity RGB-D sensor (Occipital Structure Sensor)*

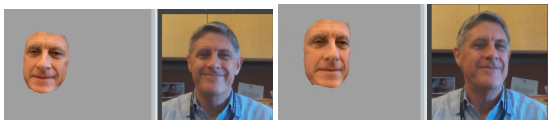

Figure 2: *Real-time puppeteering on digital image. Our shapes preserve their photorealistic facial quality generated from the original scans.*

Recent template-based face scanning technologies deform a generic template model toward raw face scans, resulting face model represents only an approximation of the original geometry instead of an exact reconstruction. Moreover, since the template model is deformed under geometric constraints, there is no guarantees on exact texture alignments between different facial expressions. By contrast, our method starts from raw face scans and directly extract consistent mesh from the scans. Therefore the geometric representations are exact to the original shapes. By contrast, our method unwraps both geometric and texture information into 2D images and solves for alignments in the UV space using facial feature detection and optical flow. Thus, template-based methods tend to produce results that are similar across scan subjects, since their template is shared, while our method preserves the unique appearance of each face, resulting in greater photorealism.

*oalexander@ict.usc.edu
†e-mail:feng@ict.usc.edu
‡e-mail:fyffe@ict.usc.edu
§e-mail:debevec@ict.usc.edu
¶e-mail:rhuizewa@usc.edu
‖e-mail:suma@ict.usc.edu
**e-mail:shapiro@ict.usc.edu

SIGGRAPH 2015 Talks, August 09 – 13, 2015, Los Angeles, CA.
ACM 978-1-4503-3636-9/15/08.
http://dx.doi.org/10.1145/2775280.2792540