DOI:10.1145/3292037

To view the accompanying paper, visit doi.acm.org/10.1145/3292039

Technical Perspective
Photorealistic Facial Digitization and Manipulation

By Hao Li

FOR MORE THAN a decade, computer graphics (CG) researchers and visual effects experts have been fascinated with bringing photorealistic digital actors to the screen. Crossing the well-known "uncanny valley" in CG humans has been one of the most difficult and crucial challenges, due to hypersensitivity to synthetic humans lacking even the slightest and most subtle features of genuine human faces. Given sufficient resources and time, photorealistic renderings of digital characters have been achieved in recent years. Some of the most memorable cases are seen in blockbuster movies, such as *The Curious Case of Benjamin Button*, *Furious 7*, and *Rogue One: A Star Wars Story*, in which large teams of highly skilled digital artists use cutting-edge digitization technologies. Despite the progress of 3D-scanning solutions, facial animation systems, and advanced rendering techniques, weeks of manual work are still needed to produce even just a few seconds of animation.

When depth cameras, such as structured light systems or time-of-flight sensors, were introduced, the 3D acquisition of highly deformable surfaces became possible. Graphics and vision researchers started to investigate the possibility of directly capturing complex facial performances, instead of manually key-framing them or applying complex simulations. While marker-based motion capture technologies are already widely adopted in industry, massive amounts of hand-tweaking and post-processing are still needed to generate lifelike facial movements. On the other hand, markerless solutions based on real-time RGB-D sensors provide dense and accurate facial shape measurements and were poised to automate and scale animation production.

The release of the mainstream Kinect depth sensor from Microsoft sparked a great deal of interest in real-time facial animation in the consumer space, most notably through several seminal SIGGRAPH publications between 2010 and 2013, as well as the popular facial animation software, Faceshift, later acquired by Apple. While computer vision-based facial landmark detectors are suitable for puppeteering CG faces using conventional RGB cameras, they do not capture nuanced facial expressions, as only sparse features are tracked. However, when dense depth measurements are available, an accurate 3D face model can be computed by refining the shape of a statistical face model to fit a dense input depth map. Not only can this face-fitting problem be solved in real time using efficient numerical optimization, but the shape and expression parameters of the face can be fully recovered and used for retargeting purposes. If facial performance capture is possible for conventional RGB videos in real time, then believable facial expressions can be transferred effortlessly from one person to another in a live-action scenario. This capability is demonstrated by the Face2Face system of Thies et al. detailed in the following paper.

As opposed to animating a CG character in a virtual environment, the key challenge is to produce a photorealistic video of a target subject whose facial performance matches the source actor. In addition to being able to track and transfer dense facial movements at the pixel level, the facial albedo and lighting environment also must be estimated on the target video, in order to ensure a consistent shading with the original footage. The solution consists of a real-time GPU implementation of a photometric consistency optimization that solves for parameters of a morphable face model originally introduced by Blanz and Vetter, extended with linear facial expression blendshapes. The authors also introduce an important data-driven technique to handle the non-linear appearance deformations of the mouth, in which plausible textures are retrieved instead of being rendered using a parametric model. Such an approach is particularly effective in producing a photorealistic output, as it bypasses the traditional and more complex rendering pipeline. While some limitations remain, such as the inability to control the head pose in the target video sequence, very convincing photorealistic facial reenactments are demonstrated on footages of celebrities and politicians obtained from YouTube.

While the original intent of performance-driven video was to advance immersive communication, teleconferencing, and visual effects, the ease and speed with which believable manipulations can be created with such technology has garnered widespread media attention, and raised concerns about the authenticity and ethical aspects of artificially generated videos.

Recent progress in artificial intelligence, such as deep generative models, is further accelerating these capabilities and making them even easier for ordinary people to use. For instance, Pinscreen's photorealistic avatar creation technology requires only a single input picture and can be used to create compelling video game characters at scale, but face replacement technologies, such as DeepFake, have been exploited to create inappropriate and misleading video content. I highly recommend the following paper, as it is one of the first that promotes awareness of modern technology's capability to manipulate videos, at a time in which social media is susceptible to the spread of doctored videos and fake news. C

Hao Li (hao@hao-li.com) is assistant professor of computer science at the University of Southern California, director of the Vision and Graphics Lab of the USC Institute for Creative technologies, and CEO of Pinscreen.