Driving High-Resolution Facial Blendshapes with Video Performance Capture

Graham Fyffe Andrew Jones Oleg Alexander Ryosuke Ichikari Paul Graham Koki Nagano Jay Busch Paul Debevec [*]
USC Institute for Creative Technologies

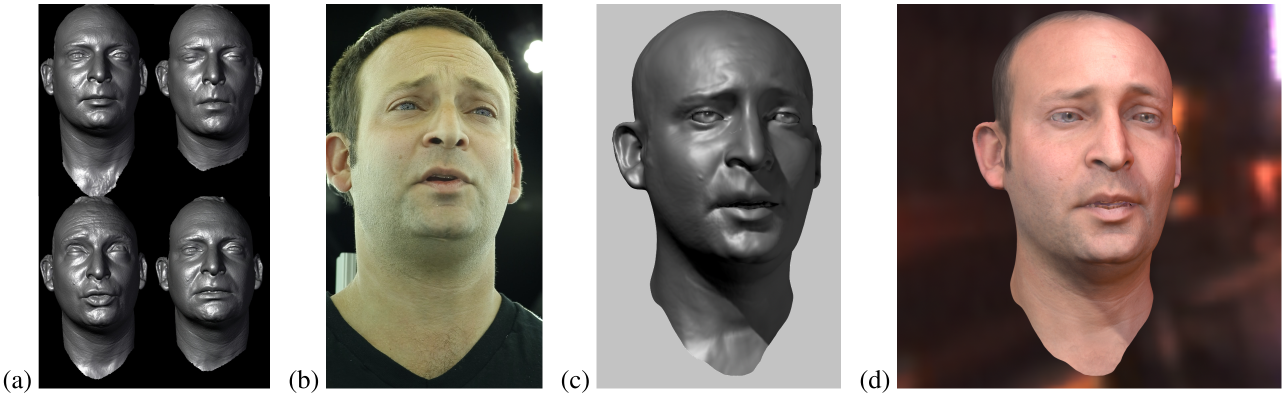

Figure 1: *(a) Geometry and reflectance data from multiple static scans are combined with (b) dynamic video frames to recover (c) animated high resolution geometry that can be (d) relit under novel illumination. This example is recovered using only a single camera viewpoint.*

We present a technique for creating realistic facial animation from a set of high-resolution static scans of an actor's face driven by passive video of the actor from one or more viewpoints. We capture high-resolution static geometry using multi-view stereo and gradient-based photometric stereo [Ghosh et al. 2011]. The scan set includes around 30 expressions largely inspired by the Facial Action Coding System (FACS). Examples of the input scan geometry can be seen in Figure 1 (a). The base topology is defined by an artist for the neutral scan of each subject. The dynamic performance can be shot under existing environmental illumination using one or more off-the shelf HD video cameras.

Our algorithm runs on a *performance graph* that represents dense correspondences between each dynamic frame and multiple static scans. An ideal performance graph will have edges that are sparse and well distributed as sequential frames are similar in appearance and any single dynamic frame only spans a small subset of facial expressions. Alternatively, facial deformation may be local (only affecting part of the face), for example the eyebrows may be raised while mouth remains neutral. Theoretically, all performance frames should lie within the domain of scanned FACS expressions allowing us to minimize temporal drift.

We developed an efficient scheme for selecting a subset of possible image pairs for reduced optical flow computation. We prune the graph based on a quarter-resolution GPU optical flow [Werlberger 2012] between each static frame and each dynamic frame on a single frontal camera view. If necessary, we can rerender the static frames to match the lighting and rough head pose in the dynamic scene. We then compute normalized cross correlation between the warped dynamic frame and each original static expression and average over twelve facial regions. We developed a iterative greedy voting algorithm based on per-region confidence measure to identify good edges. In each iteration we search, over all static expressions and all sequence frames, to identify the frame whose region has the highest confidence match to a static expression. We add that dynamic to static edge to the performance graph and compute high-resolution flow for that image pair. We then adjust the remaining confidence weights by subtracting a hat function scaled by the confidence of the selected expression in each region and centered around the last chosen frame. The hat function suppresses other facial regions that are satisfied by the new flow. We continue to iterate until the maximum confidence value falls below a threshold.

Based on the pruned performance graph, we apply a novel 2D drift-free vertex tracking scheme leveraging temporal optical flow and multiple texture targets. We track each mesh vertex as a 2D path in each camera image both forward and backwards in time. We combine flow vectors as a weighted sum of bidirectional temporal flows and nearby static-to-dynamic flows. Similar to the voting scheme, we weigh each static-to-dynamic flow vectors as a hat function that decays over temporal flows. We compute the 3D position of each vertex in the face mesh independently for each frame of the performance by triangulating the 2D projected vertex locations. Novel to our triangulation is a least-squares formulation penalizing the squared distance from the ray emitted from each camera at the 2D projected location. We regularize the triangulation using a small penalty on the distance along the ray to a hand-placed proxy face mesh. This regularization improves the stability of our solution, and at the same time enables the use of *monocular* data for triangulation. We simultaneously enforce a multi-target Laplacian shape regularization term within the same least-squares framework. In contrast to previous work, we consider the *three-dimensional coupling* between the shape term and triangulation terms, thereby obtaining a robust estimate that gracefully fills in missing or unreliable triangulation information using multiple target shape exemplars.

References

GHOSH, A., FYFFE, G., TUNWATTANAPONG, B., BUSCH, J., YU, X., AND DEBEVEC, P. 2011. Multiview face capture using polarized spherical gradient illumination. In *Proceedings of the 2011 SIGGRAPH Asia Conference*, ACM, New York, NY, USA, SA '11, 129:1–129:10.

WERLBERGER, M. 2012. *Convex Approaches for High Performance Video Processing*. PhD thesis, Institute for Computer Graphics and Vision, Graz University of Technology, Graz, Austria.

[*] e-mail:gl@usc.ict.edu