Simplified facial capture with head mounted cameras

Jose Serra*
jserra@d2.com
Digital Domain
Canada

Lucio Moser*
lmoser@d2.com
Digital Domain
Canada

David McLean
mclean@d2.com
Digital Domain
United States of America

Doug Roble
droble@acm.org
Digital Domain
United States of America

Figure 1: Two steps of solving for blendshapes (with/without eye landmarks constraints) and then doing Laplacian refinement.

ABSTRACT

We present a unified pipeline for high-resolution facial capture that replaces the initial traditional seated capture with a head-mounted camera setup. At its core, our approach relies on improving roughly personalized blendshapes by fitting handle vertices, in a Laplacian framework, to depth and image data. Thus, refining the geometry. This pipeline has been used in production to generate high quality animation to train our proprietary marker-based solution, leading to large time and cost savings.

CCS CONCEPTS

• **Computing methodologies** → **Motion capture**.

KEYWORDS

Face Capture, Animation, Optimization, Head Mounted Camera

ACM Reference Format:
Jose Serra, Lucio Moser, David McLean, and Doug Roble. 2021. Simplified facial capture with head mounted cameras. In *Special Interest Group on Computer Graphics and Interactive Techniques Conference Talks (SIGGRAPH '21 Talks), August 09-13, 2021.* ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/3450623.3464637

*Both authors contributed equally to this research.

SIGGRAPH '21 Talks, August 09-13, 2021, Virtual Event, USA

ACM ISBN 978-1-4503-8373-8/21/08.
https://doi.org/10.1145/3450623.3464637

1 INTRODUCTION AND OVERVIEW

Facial motion capture is a complex process that involves numerous steps. Amongst these, there is commonly a session, traditionally seated, to capture high quality actor specific poses, via e.g. ICT [ICT 2021] or Di4D [DI4D 2021]. This data can be used to track shots directly or to tailor generic face rigs, later used for tracking. This can cause delays before shots are processed and introduce errors when transferring motion between topologies. We present an approach that uses a range-of-motion (ROM), captured in the same 2 head-mounted cameras (HMC) as the core performances, to extract high quality consistent poses that are then used to train our marker-based system [Moser et al. 2017]. By avoiding the seated session, we significantly reduce costs and configuration time.

Our approach has two main steps: 1) **blendshape optimization**, where a rough approximation of the animation is computed by fitting blendshapes to video data, which creates an initial guess for 2), a **Laplacian refinement** step, i.e. a non-linear optimization of handle vertices in a Laplacian framework [Sorkine 2005]. We apply these 2-steps on a ROM and, from the results, create a PCA basis that replaces the blendshapes of step 1. We can then repeat the two steps as desired, each time replacing the previous initial shapes, and producing higher quality results. We found it generally only takes 2 full iterations to produce results with the desired quality.

The approach closest to ours is presented by [Fyffe et al. 2017]. It also uses a template that is iteratively improved. It produces high quality results, but is computationally heavy and requires complex hardware. Our method trades quality for speed and ease of use.

The main advantages of our pipeline are: 1) avoiding seated capture session; 2) working with a single stereo pair of cameras; 3) working with marker and markerless video.

2 METHOD & IMPLEMENTATION

Our pipeline starts with: a ROM captured with the HMC; eyelid 2D landmarks, tracked with [Park et al. 2018]; a depth map computed for each frame, by rasterizing the 3D reconstruction; and personalized blendshapes, obtained by interpolating multiple identities and blendshapes in our proprietary database. Finally, we define a set of weights/masks to control the importance of the optimization terms. These tend to be consistent for different people, unless there is a large change of camera placement or facial proportions.

2.1 Blendshape Optimization

With this data, we perform a non-linear optimization of the blendshapes and the head transform (translation and rotation) to match each frame's depth map and video. The constraints include: depth (and respective normals), optical flow between consecutive frames and anchor frames [Beeler et al. 2011], and an animation prior to prevent significant deviations from realistic poses. The prior is obtained by applying the blendshape weights of a generic ROM to the personalised blendshapes. Depth and optical flow influences are controlled by per-vertex maps that specify per-vertex weights and confidence. The eye/eyelid landmarks act as soft constraints, as the depth map/optical flow can be noisy around these areas. Minimization occurs iteratively: first attempting to fit the rigid head transform; then the blendshape weights, while allowing transforms to be tuned; and finally, adding 2D/3D constraints to the cost function. To improve the temporal consistency and speed up the solve, each frame is seeded using the results of the previous. The resulting animation roughly approximates the motion of the actor. Issues arise mostly from the blendshapes rig that was, itself, an approximation and, thus, does not accurately represent all actor's poses. And that is made worse by its linear combination nature. The animation is, however, a great starting point for the next step.

2.2 Laplacian Refinement

This step tunes the surface via handle vertices with Laplacian-based deformation [Sorkine 2005]. The optimization builds on the previous terms, with the addition of a Laplacian term to enforce smoothness (Laplacian matrix calculated on a per-frame basis), and the removal of 2D/3D constraints (to avoid sharp edges). A novel aspect involves solving frames in batches, instead of one at a time. This produces a more temporally consistent result by reducing noise from the depth and optical flow. We use frames from solved neighbouring batches to enforce temporal consistency. The processed mesh is closer to the actor's performance than the blendshape guess (Fig. 2).

2.3 Improve and feed back into the pipeline

After each step, the artist tunes the results to prevent propagation of errors to the remaining pipeline. Tuning happens primarily in two forms: improving the default input maps; and applying example-based correctives [Hendler et al. 2018]. Once that happens, we extract the principal components of the sequence (with PCA) and replace the initial shapes with this new basis. We then repeat the full process, which now starts from much improved shapes. Once the artist is happy with the results, the ROM is used as training data for Masquerade [Moser et al. 2017]. The PCA shapes can be further refined with additional shots, if the initial training data for Masquerade does not suffice.

3 RESULTS/VALIDATION

Figure 2: Results of blendshape optimization (2nd col.); with 2D eye landmarks as soft constraints (3rd col.); followed by Laplacian Refinement (4th col.). The first row output does not need manual adjustment, while the second still needs example-based corrections.

Results were captured using a HMC with 2 Ximea MQ042MG-CM cameras (48fps). Processing each frame takes, on average, 45 seconds on a dual Intel Xeon Silver 4210/Nvidia Quadro RTX5000. This process is highly parallelizable using the anchor frames as break points. Anchor frames are also not limited to neutral poses.

The presented approach produces high quality results from a single pair of cameras. It allows generating data for Masquerade faster and with less errors than using the traditional seated capture session. This system has already been used on an upcoming show. There is still room for improvements: manual intervention is needed for more complex poses that can have issues either due to coverage limits or are particularly challenging such as contours of eye or lips; high frequency noise and occasional pops are still present as drift accumulates, although we found a PCA reduction step helps.

REFERENCES

T. Beeler, F. Hahn, D. Bradley, B. Bickel, P. Beardsley, C. Gotsman, R. Sumner, and M. Gross. 2011. High-Quality Passive Facial Performance Capture Using Anchor Frames. *ACM Trans. Graph.* 30, 4, Article 75 (July 2011), 10 pages.

DI4D. 2021. DI4D Pro. Retrieved Feb 10, 2021 from https://www.di4d.com/di4d-pro/

G. Fyffe, K. Nagano, L. Huynh, S. Saito, J. Busch, A. Jones, H. Li, and P. Debevec. 2017. Multi-View Stereo on Consistent Face Topology. *CGF* 36 (2017), 295–309.

D. Hendler, L. Moser, R. Battulwar, D. Corral, P. Cramer, R. Miller, R. Cloudsdale, and D. Roble. 2018. Avengers: Capturing Thanos's Complex Face. In *ACM SIGGRAPH 2018 Talks*. Article 58, 2 pages.

ICT. 2021. ICT - Light Stages. Retrieved Feb 10, 2021 from https://vgl.ict.usc.edu/LightStages/

L. Moser, D. Hendler, and D. Roble. 2017. Masquerade: Fine-Scale Details for Head-Mounted Camera Motion Capture Data. In *ACM SIGGRAPH 2017 Talks*. New York, NY, USA, Article 18, 2 pages.

S. Park, X. Zhang, A. Bulling, and O. Hilliges. 2018. Learning to Find Eye Region Landmarks for Remote Gaze Estimation in Unconstrained Settings. In *ACM Symposium on Eye Tracking Research and Applications* (Warsaw, Poland). ACM.

O. Sorkine. 2005. Laplacian Mesh Processing. In *Eurographics 2005 - State of the Art Reports*. The Eurographics Association.