High-quality, cost-effective facial motion capture pipeline with 3D Regression

Lucio Moser
Digital Domain

Mark Williams
Digital Domain

Darren Hendler
Digital Domain

Doug Roble
Digital Domain

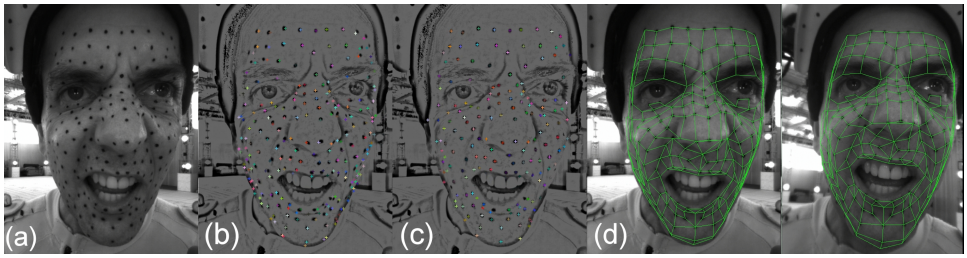

Figure 1: (a) input, (b) 3D regression, (c) automatic corrections and (d) results in top and bottom views.

ABSTRACT

We present our improved marker-based facial motion capture pipeline that leverages on 3D regression from head-mounted camera (HMC) images to speed up and reduce the cost of high quality 3D marker tracking. We use machine learning to boost productivity by training regressors on traditionally tracked performances and applying those models to the remaining performances. Our specialized regressor for HMC marker-based tracking shows improvements in quality and robustness for marker tracks. The regressor results are automatically refined by a simple blob detection tool and then imported back into the tracking tool such that manual correction can be applied as needed and subsequently included as additional training data. This iterative approach reduces 70% the amount of artist time required for traditional tracking methods and does not add much setup time nor planning as alternative techniques.

CCS CONCEPTS

• **Computing methodologies** → **Motion capture**; **Supervised learning by regression**;

KEYWORDS

Facial capture, head-mounted cameras, shape regression

ACM Reference Format:
Lucio Moser, Mark Williams, Darren Hendler, and Doug Roble. 2018. High-quality, cost-effective facial motion capture pipeline with 3D Regression. In *Proceedings of SIGGRAPH '18 Talks*. ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/3214745.3214755

SIGGRAPH '18 Talks, August 12-16, 2018, Vancouver, BC, Canada

ACM ISBN 978-1-4503-5820-0/18/08.
https://doi.org/10.1145/3214745.3214755

1 OVERVIEW

The facial animation pipeline at Digital Domain relies on precise 3D marker data to generate high resolution actor animation (see [Moser et al. 2017]) that subsequently drives virtual characters whilst maintaining the subtleties of the original performance.

HMC facial motion capture for visual effects is primarily performed using specialized software to track markers on the actor's face. The process requires extensive operator input making it time consuming and expensive, and usually only justified for main characters. Film production as well as new media presents a growing need for facial tracking in terms of volume and shot length, which can be 1-2 orders of magnitude longer, making processes such as these prohibitive. In this work we aimed to overcome this limitation and provide assisted high-fidelity tracking for large frame counts.

We investigated modern alternatives based on machine learning and formulated a new workflow that uses a regression model designed around the particularities of HMC imagery and marker-based tracking, followed by a simple automated correction step to iteratively improve and help the tracking.

2 RELATED WORK

Recent work by [Laine et al. 2017] achieved production level markerless capture results using convolutional neural networks, but they require significant quantities of expensive high-resolution seated capture performance training data including range of motion, pangrams and in-character material.

Cascaded regression ([Cao et al. 2012],[Cao et al. 2013]) has been used in markerless motion capture in the context of real-time consumer-level applications. Most recently there have been attempts to adapt it to professional HMC configurations, both as single view capture [McDonagh et al. 2016] and multi-view capture [Klaudiny et al. 2017] but both works also require high resolution seated capture as training data, followed by creation of a blendshape rig, and finally synthesis of photo-real training images under varying lighting and transformations for improved robustness. We believe that besides their attempt to reduce the number of training frames, considerable complexity remains in their setup prior to

tracking new material. Their reported average errors can also be as large as 5mm which does not meet our quality criteria.

3 OUR APPROACH

We decided to supplement, rather than replace, the standard costly marker tracking approach with machine learning for the situations that it can handle effectively. In a production environment it is important that a fallback solution exists when automation fails, and in this scenario manual tracking and fixes can be used so as not to delay progress. With that in mind, we focused on the architecture from [Cao et al. 2013], which uses Cascaded Fern regression and typically requires significantly less training data and training time than neural networks. It also does not require blendshape extraction.

Our approach consists of first applying standard marker tracking to a contiguous range of representative frames containing the most expressive moments, to use as training data. We have had success tracking as little as 5s of dialogue when aiming to solve for the full dialogue, but we found that if we track a range of motion performance of about 1min it can be used to solve a wide range of performances. These tracked markers and their corresponding HMC images are used as training data for our regressor, as detailed in section 3.1, which is then applied to the remaining frames and other performances. We found that the regression results can fail to track exactly the position of the markers, especially in challenging situations, those which differ too much - either in the actor pose, helmet placement or illumination conditions. We apply a simple and automated blob detection algorithm to snap the regression results to the center of the markers, as described in section 3.2 and the results can be imported back into the tracking tool for final fixes by the artists. The results of these manual fixes are fed back as training data for subsequent model iterations, adding robustness where needed.

3.1 3D regression

We made several improvements to the model by [Cao et al. 2013] to aid robustness to variable lighting and helmet positions as well as to exploit the fact that the markers in our HMC imagery are very salient features in the face to track. The main improvements are:

(1) Multiple cameras: Our HMC setup consisted in four cameras: top, bottom, left and right. We found that using the top and bottom perspectives as independent training samples increased robustness to different helmet placements which inevitably arise during production.
(2) Image filtering: We apply a filter when sampling the images which returns the ratio between the pixel intensities and the local average. That representation greatly improves illumination invariance for overall changes in the light intensity as well as situations where the face is unevenly lit.
(3) Sampling strategy: We adopt a coarse-to-fine approach during sampling such that initial layers in the cascade use larger landmark-relative offsets whilst subsequent regressors use smaller offsets. This reflects the idea that more local information needs to be used as confidence in predictions increases, and, conversely global information becomes less relevant. We also decrease the maximum allowable distance between these samples to build localized feature pairs.
(4) Data augmentation: We augment training data by applying rotation as well as translation. We adopt a strategy similar to [Klaudiny et al. 2017] by using regularly-spaced variations. However, because prediction uses stabilized guessed poses, we apply the transformations to the target shapes (and their respective images) instead of the initial guess shapes.
(5) Random perturbations: We found that model robustness could be increased by duplicating training data and introducing random flips to the binary fern decision boundaries as they compute the residuals. This forces subsequent layers in the cascade to learn how to correct for such corruptions, and can also be considered another form of augmentation as each flip produces variation in the predicted poses.

3.2 Correction by blob detection

This tool first applies the same image filtering as the regressor, to improve robustness to illumination changes, then it leverages on the OpenCV[1] library for fast implementations of blob detection and sparse Lucas-Kanade optical flow to create the 2D blob tracks over time. Lastly, the 3D regression results are projected to each camera view point to assign marker ids to the blobs, based on their proximity to the tracks. The successful assignments are exported to a file so artists can validate the assignments, apply minor corrections and track additional markers if required.

4 CONCLUSIONS AND FUTURE WORK

The new facial motion capture pipeline allows artists to track the minimum amount of footage, taking advantage of shape regression as necessary, avoiding expensive high resolution seated capture data, planning and setup phases required by previous works. Time savings of around 70% for final production quality results were reported. As future work we intend to further reduce artist time by improving the correction tool to minimize misassignment as well as to apply multi-view regression.

REFERENCES

Chen Cao, Yanlin Weng, Stephen Lin, and Kun Zhou. 2013. 3D Shape Regression for Real-time Facial Animation. *ACM Trans. Graph.* 32, 4, Article 41 (July 2013), 10 pages. https://doi.org/10.1145/2461912.2462012

X. Cao, Y. Wei, F. Wen, and J. Sun. 2012. Face alignment by Explicit Shape Regression. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*. 2887–2894. https://doi.org/10.1109/CVPR.2012.6248015

Martin Klaudiny, Steven McDonagh, Derek Bradley, Thabo Beeler, and Kenny Mitchell. 2017. Real-Time Multi-View Facial Capture with Synthetic Training. *Computer Graphics Forum* (2017). https://doi.org/10.1111/cgf.13129

Samuli Laine, Tero Karras, Timo Aila, Antti Herva, Shunsuke Saito, Ronald Yu, Hao Li, and Jaakko Lehtinen. 2017. Production-level Facial Performance Capture Using Deep Convolutional Neural Networks. In *Proceedings of the ACM SIGGRAPH / Eurographics Symposium on Computer Animation (SCA '17)*. ACM, New York, NY, USA, Article 10, 10 pages. https://doi.org/10.1145/3099564.3099581

S. McDonagh, M. Klaudiny, D. Bradley, T. Beeler, I. Matthews, and K. Mitchell. 2016. Synthetic Prior Design for Real-Time Face Tracking. In *2016 Fourth International Conference on 3D Vision (3DV)*. 639–648. https://doi.org/10.1109/3DV.2016.72

Lucio Moser, Darren Hendler, and Doug Roble. 2017. Masquerade: Fine-scale Details for Head-mounted Camera Motion Capture Data. In *ACM SIGGRAPH 2017 Talks (SIGGRAPH '17)*. ACM, New York, NY, USA, Article 18, 2 pages. https://doi.org/10.1145/3084363.3085086

[1] https://opencv.org