CS191W Research Summary
Improving Facial Rig Semantics for Tracking and Retargeting: FLAME-Based Models

Allise Thurman
Department of Computer Science
Stanford University
allise@stanford.edu

June 12, 2025

1 Introduction

Beginning in October, I have had the opportunity to research in the lab of Professor Ron Fedkiw under the guidance of PhD student Dalton Omens working on improved semantics in facial rigging. This paper posits that performing identity-based calibration for a set of standard expressions will improve the semantic meaning of both tracked and retargeted rig solves using FLAME-based trackers. This ties in with Dalton's broader thesis on improved facial semantics across a variety of rig models. [3]

2 Background and Motivations

Modeling and animating 3-dimensional faces requires a deep understanding and accurate recreation of not only the geometry of the human face, but also the particular motions and facial movements used to effectively communicate the intended meaning of said expressions. Even more challenging is developing a model with consistent, reproducible results across a variety of different facial identities and expressions. Unlike manual approaches to animation like keyframe, performance-driven animation uses the captured motions of a live subject and retargets these actions to the animated model. This provides animators another layer of freedom to change the identity of their characters without needing to reanimate the entire action sequence. [4]

A facial rig is a mathematical basis which can be applied to a set of expression controls. Paired with shape information for the intended identity, this matrix operation between the controls and rig produces the geometry of the face being modeled. To recreate the expressions of a subject from a video, a rig solver requires tracked geometry reconstruction from the input video. FLAME (Faces Learned with an Articulated Model and Expressions) is a commonly used model for facial rigging.[2] Several common facial tracking and reconstruction techniques use the FLAME model as the underlying rig. However, FLAME has limitations which subsequently impact the quality of all technology built upon it. One significant limitation of FLAME is a non-semantic model, meaning its controls do not directly correlate to meaningful expressions. In order to better calibrate the rig, the controls must map to the Range of Motion (ROM) expressions we are using as benchmarks.

The second major limitation to the FLAME model is the lack of identity-based calibration. Though geometry reconstructions can produce a realistic representation of the input performance when using the original identity, as the controls are based entirely on the tracked geometry of the input video, rather

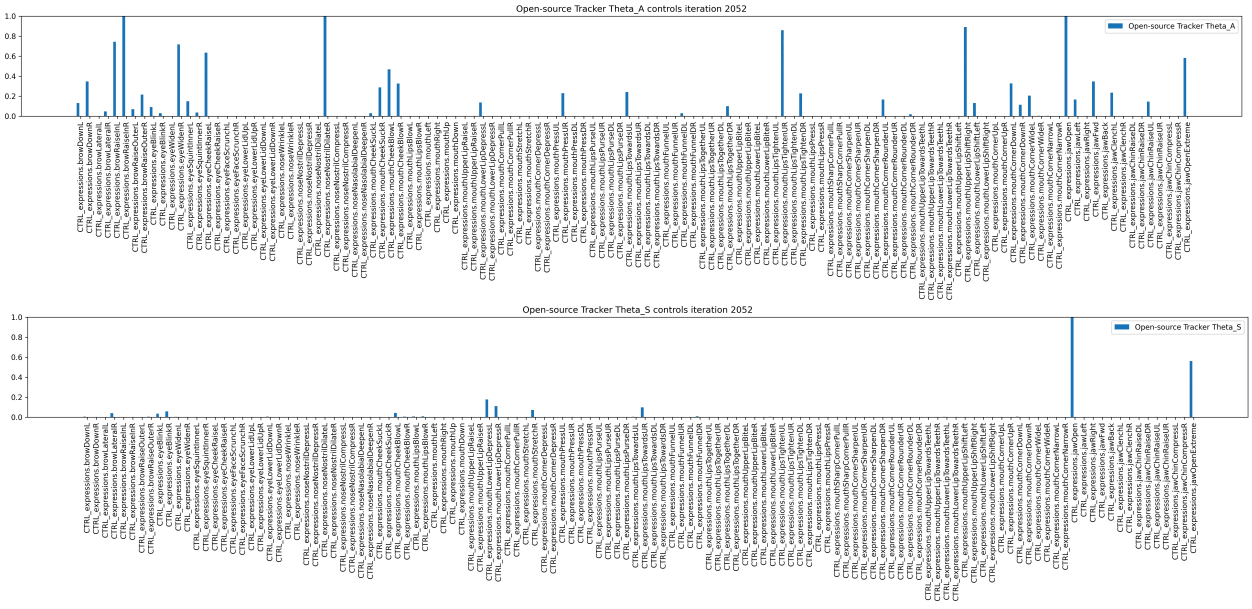

Figure 1: Metahuman raw control values captured in tracking the uncalibrated (top) and calibrated (bottom) FLAME rigs on a ROM frame where the performer is intending to make the "Jaw Open Extreme" expression.

than the intended expressions of the performer. This presents a challenge as different people often produce different performances of the same expression. For example, when instructed to perform "Jaw Open" one subject may perform an action which is, geometrically, more similar to another subject's "Jaw Open Extreme". Therefore, when a set of controls are retargeted to another identity from a FLAME-based tracker the result often loses some of the semantic significance of the original capture. In our research, Dalton and I set out to improve upon the existing FLAME model by creating an equivalent model which maps to meaningful expressions, calibrated to the basic ROM expressions of the performers. By enhancing the FLAME model, we were thus also able to improve the results of FLAME-dependent tools which are farther down in the facial animation pipeline.

3 Approach and Methodology

Improvements were made to FLAME-based retargeting methods in four main steps. Firstly, a semantic version of the FLAME face model was created, mapping the 100 FLAME expression components as well as the six jaw and neck pose parameters to semantically-meaningful MetaHuman raw controls by solving for and applying a basis vector to transform the raw controls into the FLAME parameter space. [3, 1]

From here, the FLAME rig could then be calibrated on the range of motion expressions provided by the performer. Since the project was focused on improving expression accuracy, we were able to run an iterative LBFGS solver to optimize the expression basis used by the FLAME model. To do so, we identified which frames in the training ROM corresponded to the 20 primary controls facial controls. The optimized model should then, in theory, output those target controls when those or similar frames/expressions

are run through a tracker. To ensure that the model was not overfit to the training video, we also added a regularizing factor to the solve.

With the optimized expression basis, we could then apply our tracker using an identity-calibrated FLAME model to any video or image of the original performer to produce identity-specific semantic control values. In addition to improving the tracker's ability to identify when the subject was presenting the 20 ROM expressions, we wanted to minimize both the number of active controls and the average value of tweaker controls in each control solve. Any controls beyond those required to produce the ROM expression were identified as tweaker controls. Thus, we added penalties to the overall loss used in the LBFGS solver to discourage the model from turning on less common controls or too many controls at once. Figure 1 displays the uncalibrated and calibrated controls for frame 2052 of a training range of motion video in which the performer is executing the "Jaw Open Extreme" ROM expression which correlates to the full activation of the "Jaw Open" and "Jaw Open Extreme" Metahuman raw controls. For this particular trial, we used an under-regularized rig and tracker to validate the efficacy of the calibration on the training dataset. In comparing the two figures, we can verify that using the calibrated model with the aforementioned penalties allows the tracker to isolate and maximize the controls which map to the original performer's ROM expressions.

After tracking a performance using the calibrated FLAME model, we could then use the identity-based controls to retarget that performance onto any other character using the uncalibrated semantic FLAME model with a higher level of semantic similarity to the original video.

4 Results

To evaluate the success of our calibration process, we collected visual and numerical data comparing the controls and retargeted performances produced by the tracker using the calibrated and uncalibrated FLAME rigs. Figure 3 provides a side-by-side comparisons of the uncalibrated and calibrated tracker controls applied to the same randomly-generated virtual avatar. Table 1 shows the average controls sparsity before and after calibration as well as average tweaker control value across three different performances tested using two different FLAME-based open source trackers.

	Tracker	Sparsity (↓) Baseline	Sparsity (↓) Ours	Tweakers (↓) Baseline	Tweakers (↓) Ours
ROM 2	OS1	0.1566	**0.1313**	0.0425	**0.0152**
	OS2	0.1736	**0.1053**	0.0673	**0.0116**
ROM 1	OS1	0.1628	**0.1237**	0.0538	**0.0173**
	OS2	0.1681	**0.0868**	0.0562	**0.0077**
Pangram 1	OS1	0.1782	**0.1695**	0.0695	**0.0315**
	OS2	**0.1568**	0.1799	0.0472	**0.0228**

Table 1: *left column* Sparsity values, defined as the proportion of all controls which are turned on (set to a value over 0.05 out of 1) across all frames for the uncalibrated (baseline) and calibrated (ours) controls. *right column* Average value of all tweaker controls—any controls not required to produce the 20 basic ROM expression—across all frames. Both models were tested using two different common open-source trackers (OS1 and OS2).

We also conducted a user study where 17 participants were shown side-by-side videos of the original performance alongside the uncalibrated and calibrated tracker controls applied to the same virtual avatar. We found that 82% of participants preferred the calibrated videos, demonstrating improvements to animation quality from a consumer perspective [3].

5 Discussion

The results of our tests concluded that identity-specific calibration could produce more semantically-consistent retargeted performances in existing

Figure 2: *left figure* Model ROM captures (*top*) of "Teeth Grimace" and "Mouth Funnel" expressions compared against my captures (*bottom left*). *right figure* Resulting expression solves in test Pangram. Due to funneling in my input teeth grimace capture, we note that the resulting calibrated controls solve *(right column)* is less semantically similar to the Pangram captures *(left column)* than the uncalibrated geometry reconstruction *(middle column)*.

FLAME-based trackers. Given the wide use of FLAME and FLAME-based trackers like MICA, this research offers a method for achieving improved facial rigging semantics in existing industry-standard models without requiring any fundamental changes to the models themselves. Once the optimized expression basis has been found for a particular user, the user-calibrated FLAME model can then be inserted into any model or tracker which already consumes it [5].

However, we did identify some continued challenges with the method that are open to discussion. In Figure 2, we in my sample "Teeth Grimace" frame that my lips are more pursed than the model example, resulting in an expression that can easily be conflated with a partial "Mouth Funnel". As a result, the mouth funnel control was on far more often than intended, resulting in a set of solved controls producing overly funneled expressions when the mouth was open as shown in the resulting geometry reconstructions for those calibrated controls. This is because the quality of the calibrated expression solve is heavily dependent on the quality of the initial ROM capture as well as the regularization and penalties applied to the rig calibration and tracking. We've noted that some expressions are particularly difficult or unnatural for users to execute and to isolate particular parts of their face. Some of this can be mitigated by using binary masks when optimizing the rig to each expression, however clear instruction and supervision in the ROM capture is likely the best way to achieve the most accurate expressions for training.

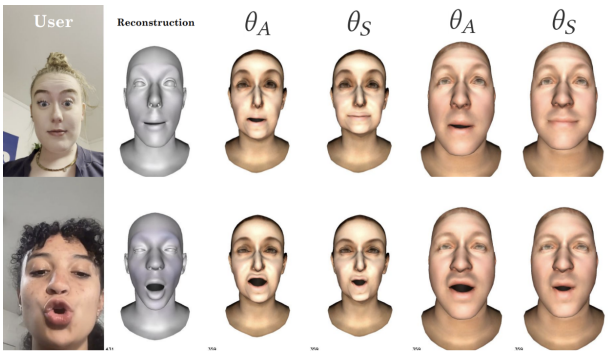

Figure 3: Uncalibrated θ_A and calibrated θ_S controls retargeted to two different randomly generated virtual avatars. Calibrated results better convey semantic meaning despite inaccuracies in the original reconstruction geometry.

References

[1] Epic Games. *MetaHuman Animator*. `https://www.unrealengine.com/en-US/digital-humans`. Accessed: 2025-06-10. 2025.

[2] Tianye Li et al. "Learning a model of facial shape and expression from 4D scans". In: *ACM Transactions on Graphics, (Proc. SIGGRAPH Asia)* 36.6 (2017), 194:1–194:17. URL: `https://doi.org/10.1145/3130800.3130813`.

[3] Dalton Omens. "Improving Facial Rig Semantics for Tracking and Retargeting". Unpublished doctoral dissertation. Ph.D. thesis. Stanford University, 2025.

[4] ChangAn Zhu and Chris Joslin. "A review of motion retargeting techniques for 3D character facial animation". In: *Computers & Graphics* 123 (2024), p. 104037. ISSN: 0097-8493. DOI: `https://doi.org/10.1016/j.cag.2024.104037`.

[5] *Towards Metrical Reconstruction of Human Faces*. European Conference on Computer Vision. 2022.