Journal of Multimedia Information System VOL. 12, NO. 2, June 2025 (pp. 67-70): ISSN 2383-7632 (Online)
http://doi.org/10.33851/JMIS.2025.12.2.67

Brief Paper:
Transform-Invariant Facial Expression Editing for 4D Mesh Sequences

Jaewon Song[1*], Minyeong Jeong[1], Jaeho Im[1]

Abstract: Conventional editing of 4D facial mesh sequences often involves mesh stabilization, which removes head motion to align facial expressions across frames. While effective for maintaining consistent vertex correspondence, this process discards valuable performance cues embedded in natural head movements. In this paper, we present a transform-invariant facial expression editing framework for 4D mesh sequences that retains the original head transformations while enabling precise, non-rigid facial modifications. Our method decouples global rigid motion from local facial deformation using a hybrid approach that combines alignment tracking with localized expression modeling. This allows intuitive per-frame editing while maintaining temporal coherence and preserving the subject's identity. Experimental results on real 4D capture sequences demonstrate stable and realistic edits, making the method suitable for applications in facial retargeting, performance-driven animation, and digital human production.

Key Words: 4D Facial Capture, Dynamic Face Mesh Sequence, Facial Expression Editing, Head Transform Preservation.

I. INTRODUCTION

4D facial capture, which produces a time series of 3D face meshes by recording performances at high frame rates (typically 24–120 fps), has become an essential technique in digital human modeling, visual effects (VFX), and performance-driven animation. In this context, the term "4D" refers to a time series of 3D face meshes that capture both spatial geometry and temporal dynamics of facial performances. Various approaches have been developed to capture and manipulate facial geometry, including deformation transfer for mesh animations [1], high-quality multi-view facial capture techniques [2], and interactive face editing using deformable models [3]. However, raw 4D facial scan data inherently contains both rigid head motions and non-rigid facial deformations, making it difficult to isolate and adjust specific expressions without unintentionally discarding meaningful motion cues.

A common practice in processing such data involves mesh stabilization removing global head transformations based on rigid facial regions (e.g., the forehead and nose) to enable consistent topology and temporal alignment across frames. While effective for retargeting and expression manipulation, this approach also eliminates natural head movements, which are often crucial for preserving the actor's performance, emotion, and intent. As a result, stabilized data may lose authenticity and appear artificially constrained in downstream applications. Recognizing this limitation, some recent methods have aimed to preserve the head motion during facial performance capture or retargeting [4,5]. However, to our knowledge, no prior work has specifically addressed the editing of captured 4D sequences in a way that retains the original head movement.

To address this issue, we propose a transform-invariant facial expression editing framework for 4D mesh sequences. Our method retains the global head transform embedded in the original scans while allowing precise local shape editing. We achieve this by registering raw mesh frames into a common topology *without* applying stabilization, and by introducing a transform-aware editing scheme that applies user-defined corrective shapes (which we call *fix shapes*) in a manner that respects each frame's pose. The contributions of this work are threefold:

Head-Motion-Preserving 4D Capture Pipeline: A high-fidelity 4D data acquisition and preprocessing pipeline that preserves the subject's natural head motion throughout the sequence.

Pose-Invariant Editing Mechanism: A practical editing mechanism that enables frame-wise expression correction

Manuscript received May 14, 2025; Revised June 09, 2025; Accepted June 13, 2025. (ID No. JMIS-25M-02-002)
Corresponding Author (*): Jaewon Song, +82-2-6391-7652, jaewon.song@dexterstudios.com
[1]R&D Department, Dexter Studios, Seoul, Korea, jaewon.song@dexterstudios.com, minyeong.chung@dexterstudios.com, jaeho.im@dexterstudios.com

using artist-defined fix shapes applied in a pose-invariant manner.

Temporally Consistent Smoothing Strategy: A linear interpolation strategy for propagating edits to neighboring frames, ensuring temporal smoothness of corrections across the sequence.

Experimental results demonstrate that our method supports high-quality facial edits without sacrificing natural motion, making it applicable to performance retargeting, digital human production, and subtle emotion refinement tasks.

II. METHODS

2.1. 4D Facial Data Acquisition and Preprocessing

2.1.1. Data Acquisition

To capture high-resolution dynamic facial expressions, we constructed a custom multi-view facial scanning system (Fig. 1) consisting of eight industrial 6-megapixel machine-vision cameras arranged in a semi-circular arc to maximize facial coverage while maintaining a compact setup and synchronized capture. The system captured image sequences at 24 frames per second under uniform lighting conditions. Each subject's performance was recorded in real time while they delivered various expressions or lines of speech. The captured multi-view frames were temporally synchronized and used to reconstruct a 3D mesh for each frame via multi-view stereo techniques. The result was a raw 4D mesh sequence, denoted as S_{raw}, as defined in Eq. (1):

$$S_{\text{raw}} = \{M_i \mid i = 1,2,\dots,N\}, \tag{1}$$

where each M_i represents a frame-wise unregistered mesh with its own unique topology and vertex count. To manage storage and enable fast data handling, each mesh frame was compressed using the Google Draco geometry compression format. This compression preserves the surface shape while significantly reducing file size in our case, the size of a single frame's mesh was reduced from approximately 102.21 MB to 0.84 MB (a compression ratio of about 100:1) without introducing visible degradation.

2.1.2. Post-Processing

Although each frame of the raw sequence contains detailed facial expression geometry, the sequence suffers from inconsistent mesh topology from frame to frame. This inconsistency impedes both temporal analysis and geometric editing of the data. To resolve this, we performed non-rigid mesh registration using a predefined template mesh as a reference. This process, commonly known as mesh re-topology or remeshing, resulted in a sequence of meshes with consistent topology, as shown in Eq. (2):

$$S_{\text{reg}} = \{\widetilde{M_i} \mid i = 1,2,\dots,N\}, \tag{2}$$

where each $\boldsymbol{m_i}$ shares the same vertex connectivity and vertex count (topology) as the template.

Unlike conventional pipelines that include a stabilization step to remove head motion (typically by rigidly aligning all frames to a reference pose based on facial zones such as the T-zone), we deliberately preserve the global head movement by skipping stabilization. This decision reflects the core motivation of our work: to preserve actor-driven motion cues such as nods, emphasis, or expressive head tilts that are often lost when all frames are forced into a common head pose. As a result, the processed sequence retains both the intended facial deformations and the natural head transforms across time. Fig. 2 shows an example of a 4D facial scan sequence in which the subject's head motion is preserved along with facial expressions.

2.2. Transform-Invariant Editing Method

2.2.1. Frame-Wise Editing with Fix Shape

In 4D facial performance editing, users often need to correct specific frames that exhibit flawed or incomplete expressions for example, a blinking gesture where the eyelid does not fully close. To address such issues while preserving the frame's natural head pose, we introduce artist-authored fix shapes.

Let frame f be the target frame to edit. The fix shape $\boldsymbol{m}^*$ is manually created in a canonical frontal pose and encodes the desired facial expression, such as fully closed eyes or a corrected mouth shape. Because the original 4D frame $\boldsymbol{m_f}$ includes head rotation, we cannot directly apply $\boldsymbol{m}^*$. Instead, we compute a rigid transformation $\mathbf{T}_f$ that

Fig. 1. (left) Our custom 4D facial-scanning rig; (right) synchronized multi-view images captured simultaneously.

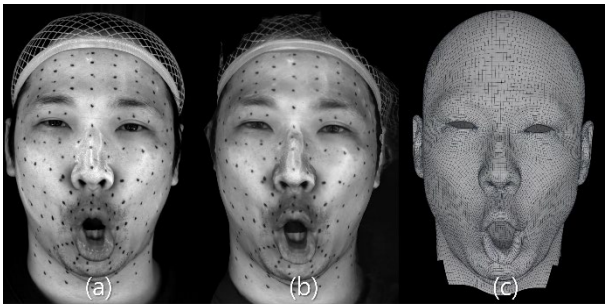

Fig. 2. (a) Captured image; (b) raw reconstructed mesh; (c) registered mesh with template mesh.

aligns $\boldsymbol{m}^*$ to $\boldsymbol{m_f}$, based on sparse anatomical landmarks. Typically, 6 to 8 landmarks are used, including the outer eye corners, nose tip, ear tragus, and chin. These landmarks are chosen for their stability and visibility across expressions and frames. The alignment is computed using the Iterative Closest Point (ICP) algorithm, initialized by these landmarks. In rare cases where ICP fails to converge due to extreme pose differences or occlusions, we fall back to a constrained alignment using a predefined frontal pose as a reference.

We then blend the original mesh and the transformed fix shape using the weighted interpolation in Eq. (3):

$$m'_f = (1 - \alpha) \cdot m_f + \alpha \cdot T_f(m^*), \tag{3}$$

where $\alpha \in [0,1]$ controls the intensity of the correction. This formulation ensures that the edit is applied in the local coordinate system of the original frame, maintaining the transform-invariant property.

2.2.2. Temporal Smoothing

To maintain temporal consistency, we propagate the keyframe edit to neighboring frames by linear interpolation. For each frame $f + k$ within a window of K frames around the edited frame f, the output mesh is computed as a weighted average using Eq. (4):

$$\boldsymbol{m'_{f+k}} = (1 - w_k) \cdot \boldsymbol{m_{f+k}} + w_k \boldsymbol{m'_f}. \tag{4}$$

To prevent visual popping artifacts and ensure temporal coherence, we propagate the edit to adjacent frames using a linear smoothing strategy. Let K denote the half-width of the temporal window used for interpolation. For each neighboring frame $f + k$, where $k \in [-K, K]$, we define a weight w_k for each offset frame, as in Eq. (5):

$$w_k = 1 - \frac{|k|}{K}, \quad \text{with } K = 5. \tag{5}$$

Here $w_k \in [0,1]$ is the interpolation weight for frame offset k, decreasing as k moves away from the keyframe. We empirically set $K = 5$, which provided visually acceptable smoothing results without noticeable artifacts. While other interpolation schemes such as Gaussian interpolation were considered, linear interpolation was adopted due to its computational simplicity and its effectiveness in producing smooth transitions over the short temporal window used in our method.

III. RESULT

We demonstrate the effectiveness of the proposed transform-invariant editing on a 4D facial scan sequence consisting of 120 frames. Within this sequence, one particular frame exhibited an artifact: the subject's eyes were captured in a semi-open state where they should have been closed (due to a momentary failure in the capture or detection process). To correct this, we used an artist-created *fix shape* mesh representing naturally closed eyes. Using our editing framework, this fix shape was transferred onto the problematic frame in a transform-invariant manner. In other words, the original head pose and orientation of that frame were preserved, and only the facial expression around the eyes was adjusted to match the intended closed-eye expression.

As shown in Fig. 3, the editing result successfully fixes the artifact while maintaining seamless continuity in the sequence. Figure 3 presents a visual comparison of the facial sequence before and after editing the target frame. The top row of Fig. 3 shows the original frames around the problematic moment (with the noticeable eye artifact), whereas the bottom row shows the corresponding frames after applying our edit. After the correction, the subject's eyes in the target frame appear properly closed, and the transition from the preceding frame through to the following frame remains smooth. The edit did not introduce any visible discontinuity or jitter in the sequence, indicating that the expression fix was seamlessly integrated with the original motion.

We primarily rely on qualitative visual inspection to evaluate the results, as there is no ground-truth or numeric error metric that directly applies to this kind of facial expression editing task. The side-by-side comparison in Fig. 3 clearly demonstrates the improvement in expression consistency (the eyes are correctly closed in the corrected frame) while showing that the actor's original head movements have been preserved. Overall, the edited sequence looks natural and maintains the realism of the performance despite the localized correction.

In addition, Fig. 4 visualizes the geometry of the fix shape used for the editing. The fix shape mesh (designed

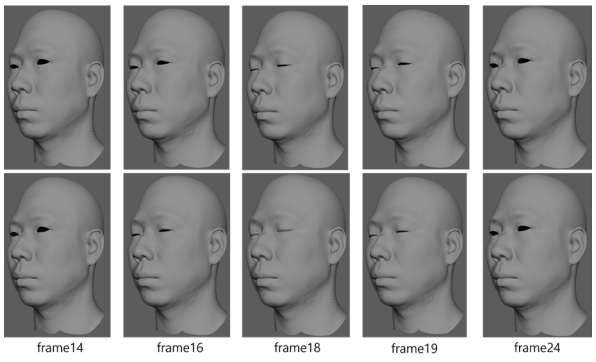

Fig. 3. (top) Original sequence with eye-closure artifact; (bottom) sequence after transform-invariant editing.

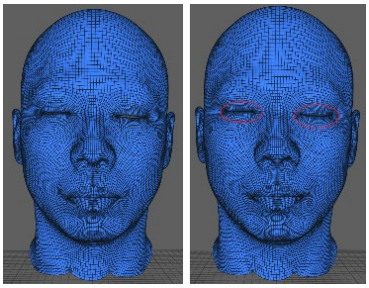

Fig. 4. (left) Target frame with artifact (frontal view shown for comparison); (right) artist-authored fix shape with closed eyes.

with closed eyes in a neutral head pose) is shown in Fig. 4 as a reference. Notably, although this fix shape was designed in a frontal pose, our transform-invariant editing approach allows it to be applied convincingly even when the character's head in the sequence is in a different orientation. The fix shape aligns to the target frame without any misalignment or distortion, and the desired expression change is achieved regardless of the global head rotation of that frame.

IV. CONCLUSION

In this paper, we presented a framework for transform-invariant facial expression editing of 4D mesh sequences. The key contribution of our work is the ability to modify or correct facial expressions in a captured sequence while *preserving the original head transform* that is, the global head pose and motion remain unchanged. This approach allows an artist or an automated pipeline to fix localized expression errors (such as an improperly captured eye closure in a single frame) without disrupting the temporal coherence or altering the pose consistency of the sequence. Such a capability is highly valuable for applications like facial performance retargeting and visual effects, where one often needs to adjust or enhance facial expressions in recorded performances while ensuring the character's head movements and overall appearance stay faithful to the original capture.

However, our work also has certain limitations. First, the current method relies on a manually created fix shape provided by an artist. Creating an appropriate fix shape for each new subject or type of expression can be time-consuming and requires artistic expertise, limiting the scalability of the approach. In some cases, however, fix shapes can be reused across subjects with similar facial structure or expression context, partially mitigating the manual effort. Second, our evaluation was limited to qualitative visual comparison, without quantitative metrics or user studies to objectively measure the improvement in realism or viewer perception. In future work, we plan to address these limitations by exploring ways to automate or assist the creation of fix shapes—possibly through learning-based techniques such as generative models trained on facial expression deltas, or by using example-based synthesis from a curated library of facial priors and common corrective shapes. We also intend to incorporate quantitative evaluation criteria and user feedback to more rigorously validate the effectiveness and perceptual impact of the edits. Additionally, extending the framework to handle a wider range of facial editing scenarios (including other types of expression corrections or stylistic modifications) and testing it on more diverse 4D facial datasets would further demonstrate the generality and robustness of our approach.

ACKNOWLEDGEMENT

This work was supported by Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No. RS-2023-00229451, Interoperable Digital Human (Avatar) Interlocking Technology Between Heterogeneous Platforms).

REFERENCES

[1] R. W. Sumner and J. Popović, "Deformation transfer for triangle meshes," *ACM Transactions on Graphics*, vol. 23, no. 3, pp. 399-405, Aug. 2004.

[2] T. Beeler, B. Bickel, P. Beardsley, R. Sumner, and M. Gross, "High-quality single-shot capture of facial geometry," in *Proceedings of ACM SIGGRAPH 2010*, Los Angeles, CA, Jul. 2010, pp. 1-9.

[3] J. Yu, M. Zollhöfer, H. Kim, J. Huang, V. Koltun, and C. Theobalt, "Interactive face editing using a deformable model," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Boston, MA, Jun. 2015, pp. 5783-5791.

[4] O. Ait-Aider, N. Berthouzoz, R. M. Boulic, and J. Thiran, "Pose-preserving dynamic facial expression retargeting," in *Proceedings of the International Conference on 3D Vision*, Lyon, France, Nov. 2015, pp. 313-321.

[5] S. Li, Y. Weng, H. Zhao, and K. Ji, "Robust head-motion-aware facial performance capture using a single RGB camera," *Computer Graphics Forum*, vol. 38, no. 7, pp. 105-116, Oct. 2019.