A Theory of Stabilization by Skull Carving

Mathieu Lamarre
mlamarre@ea.com
SEED Electronic Arts
Montréal, Québec, Canada

Patrick Anderson
panderson@ea.com
SEED Electronic Arts
Montréal, Québec, Canada

Étienne Danvoye
edanvoyer@ea.com
SEED Electronic Arts
Montréal, Québec, Canada

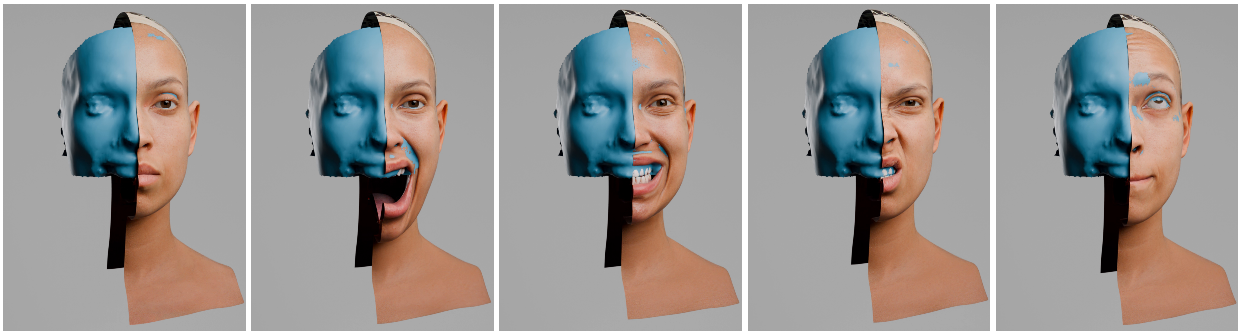

Figure 1: We present the skull carving algorithm for skull stabilization in facial animation. Above a split-view of reference neutral and four expressions with combined action units (AU), with the stable hull in blue on the left side of each render.

ABSTRACT

Accurate stabilization of facial motion is essential for applications in photoreal avatar construction for 3D games, virtual reality, movies and training data collection. For the last case, stabilization needs to work automatically for the general population with people of varying morphology. Distinguishing rigid skull motion from facial expressions is critical since misalignment between skull motion and facial expressions can lead to an animation model that is hard to control and can not fit natural motion. Existing methods struggle to work with sparse sets of very different expressions, such as when combining multiple units from Facial Action Coding System (FACS). Others are not robust enough, some depend on motion data to find stable points and other make one-for-all invalid physiological assumptions. In this paper, we leverage recent advances in neural signed distance fields and differentiable isosurface meshing to compute skull stabilization rigid transforms directly on unstructured triangle meshes or point clouds, significantly enhancing accuracy and robustness. We introduce the concept of a stable hull as the surface of the boolean intersection of stabilized scans, analogous to the visual hull in shape-from-silhouette and the photo hull from space carving. This hull resembles a skull overlaid with minimal soft tissue thickness, upper teeth are automatically included. Our skull carving algorithm simultaneously optimizes the stable hull shape and rigid transforms to get accurate stabilization of complex expressions for large diverse set of people, outperforming existing methods.

CCS CONCEPTS

• **Computing methodologies** → **Animation**; **Computer vision**; *Shape modeling*; *Neural networks.*

KEYWORDS

computer vision, avatars, animation

ACM Reference Format:
Mathieu Lamarre, Patrick Anderson, and Étienne Danvoye. 2024. A Theory of Stabilization by Skull Carving. In . ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

High resolution photorealistic avatar likeness capture for hero characters in 3D games or virtual reality face-to-face conversation or digital double for movies is a mature but complex process that still requires signification manual effort. On the other hand, major progress in the field of generative AI promise to make avatar creation as simple as generating images from text. However, there is still a some way to go before generated photoreal 3D facial animations become convincing enough for use the previously mentioned applications. See [Zollhöfer et al. 2018] for a review of the field, [Saito et al. 2024] for a recent example of a machine learned photoreal animated avatar and [Wu et al. 2024] for an example of text to avatar generation. In any case, one of the best way to improve both photoreal 3D avatar likeness capture and generative AI training data is to automate as many steps as possible in the creation pipeline while maintaining or improving the quality of the final result, especially for animations.

Current avatar animation models whether they are based on blendshapes or linear blend skinning or more advanced machine

Conference'17, July 2017, Washington, DC, USA

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM
https://doi.org/10.1145/nnnnnnn.nnnnnnn

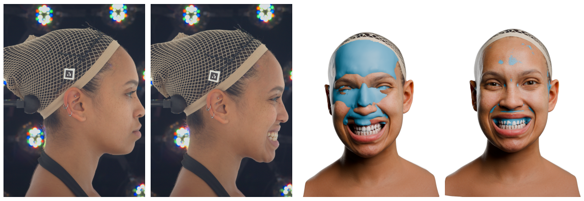

Figure 2: Even with a headrest, the whole head moves when performing expressions (left). Stabilization is the process of estimating the rigid transform to remove the skull motion from non-rigid expression deformation (right).

learning methods first need the captured data to be normalized. The goal of likeness capture is to obtain a controllable model that is independent of the capturing modalities, for example, the head position and angle with respect to the rest of the body. A small but crucial step is to estimate the position and orientation of the skull. The skull must stay fixed when other control variables change, for example, opening the lips should not cause the upper teeth and eyes to move. Strong facial expressions are often correlated with some head motion. Removing as much correlation between control variables from the training data is desirable for next generation neural network based facial animation models.

If the goal of likeness capture is creating a database of head representative of the general population and resources are limited, using static expressions with combined FACS units for a larger group of people instead of using 4D data from a smaller group is a good compromise. This leads to the challenge of estimating the skull pose of various complex expressions from static captures. Existing methods require either 4D temporal data to find stable features or make simplifying assumptions about the skull shape and skin thickness, hence they are not appropriate for stabilizing an expression database including a variety of ethnicity and age, with large difference in morphology and skin thickness.

2 RELATED WORK

[Bouaziz et al. 2013b] uses Iterated Closest Point (ICP) with a custom template that focus on the face's most rigid parts, like the forehead and the nose. This method works relatively well for all expressions that do not involve eyebrows motion. However, it is not very accurate because it relies on a single fixed template and the rest positions of points on the surface of the face have some variance.

[Beeler and Bradley 2014] propose a non-linear optimization method that requires customizing a template skull model on each person by making several assumptions, for instance that the skin thickness at five specific points on the face is constant between individuals. This assumption doesn't hold at all. For example, for the nose radix point between the eyes, Rhinoplasty literature [Dey et al. 2021] shows that the standard deviation of the skin thickness at this point is 1.7 mm for a 6.7 mm average and that this value is correlated with the body mass index. Other assumptions include a fixed single one-for-all texture mask that describe how the skin thickness varies on the face and a very simplified physic model for how the nose should bend.

The FLAME parametric facial animation model [Li et al. 2017] is trained on thousands of low resolution laser scans and thousands of frames of 4D data. It includes a linear blend skinning model with a bone for the skull that may be used for stabilization. However, the model is trained on raw scan data and the minimized energy function has no skull specific term, so the fitting error is distributed randomly between shape, expression and pose parameters and the skull pose will carry some inaccuracy.

[Lamarre et al. 2018] observe that in the skull coordinate frame, the zero distance-to-neutral mode of the head mesh vertex histogram is maximized. Their mode pursuit method searches for the skull pose trajectory that minimizes a custom loss function that gradually approximates the ℓ_0 norm of vertex positions and velocities. This is equivalent to maximizing the zero mode of the distance-to-neutral and velocity histograms. Since mode-pursuit is carried on both positions and velocities of the head mesh vertices, without temporal data, on static expressions, this method lacks motion information to find stable regions. It reduces to an approximate ℓ_0 norm ICP, similar to [Bouaziz et al. 2013a] but with a better behaved penalty function for gradient descent. Mode pursuit fails if the zero maximum mode hypothesis fails, which occurs when all points on the upper portion of the head move at the same time which occur when many FACS action units are combined.

3 METHOD

Our skull carving method works on a set of 3D facial static expressions scans. It doesn't depend on a specific capture method and only requires a 3D point cloud or unstructured triangle mesh per expression and a template mesh aligned to the neutral to be used as the reference coordinate frame; see [Zollhöfer et al. 2018] for a survey of methods to create 3D facial scans.

To explain our algorithm we first need to explain the stable hull concept, which is analogous to two well known ideas in computer vision : the visual and photo hulls. The visual hull introduced by [Laurentini 1994] is the intersection of the binary silhouettes perspective cones of a set of cameras. [Kutulakos and Seitz 2000] invented the photo hull concept, which is the remaining isosurface after carving voxels with a photo-consistency criterion. Their algorithm is named space carving, by analogy we named ours skull carving.

Given that each expression scan is converted to a signed distance field (SDF), with a defined interior (negative) and exterior (positive), When each SDF is rigidly transformed to the stabilized skull coordinate frame, the surface of the intersection of their interior volume is the stable hull. The optimal stable hull of a sufficiently large set of expressions should be the skull layered with the minimum observable soft tissue thickness of the person. If an expression show visible upper teeth, the stable hull will wrap them accurately.

Our main hypothesis is that *for each expression, there is always a large enough region of the stable hull that is close enough to the scan surface to maximize the zero distance mode in the skull coordinate frame.*

It is true for all expressions where upper teeth are sufficiently visible on the scan surface since they are part of stable hull and also on the scan surface. For the other expressions, we assume that muscle and skin moves around the face and can never totally cover the stable hull. We validated that this hypothesis empirically produces production quality results on a group of 32 persons. This hypothesis is enough to build an energy function that can be minimized to solve simultaneously for the skull poses and stable hull using gradient descent. Using the mode as the main loss function makes the method robust, support regions may be sparse and small.

3.1 Raw 3D Scan to SDF

To convert raw 3D scans to SDF, we first compute the bounding cube of all scans and create voxel grids for each scan with the same world coordinate extent. Next we determine if each voxel is inside or outside the scan mesh using the Fast Winding Number method of [Barill et al. 2018]. In practice, we only tested with unstructured triangle mesh input, but this algorithm also works on 3D point cloud. SDF are then computed with the Eikonal equation solver of [Vicini et al. 2022] which only works on cubic voxel grids. To implement an efficient stabilization process scan SDF must be compact and fast to evaluate on the graphical processing unit (GPU). The expression sets being stabilized may have between 10 and 100 scans. Voxel grid evaluation are very fast but even at low bit depth they require too much memory to stabilize 100 expressions. We propose to use a very simple tri-plane based neural SDF model similar to [Wang et al. 2023] but even simpler. For our purpose, feeding the output of the 128x128x3 tri-plane features to a single multi-layer perceptron (MLP) with 2 hidden layers of 196 neurons each is enough to get sub-millimeter accuracy close to the scan surface. We train the distinct parameters θ_i of this model to approximate each expression SDF_i over the whole voxel domain Ω.

$$\phi_{\theta_i}(x) \approx SDF_i(x), \forall x \in \Omega \tag{1}$$

This model evaluates fast and takes 7 MB of GPU memory per scan.

3.2 Skull Carving Optimization

Skull carving is a non-linear optimization problem solved with gradient descent. To model rigid transformations, we use unit dual quaternion which are well suited for numerical optimization. Carving is implemented by taking the maximum distance over all stabilized expressions. If for a point in stabilized space, the maximum distance to any expression is positive, this point is considered outside the stable volume. Let γ be the differentiable isosurface extraction function of Flexicube [Shen et al. 2023]. The function γ takes a scalar field as input and outputs the vertex positions of the stable hull triangle mesh. In practice, Flexicube also outputs the triangle vertex indices which are useful for visualization but are not used during optimization. Let $Q = \{q_1 = I, q_i \mid i \in [2..N]\}$ be the set of stabilization dual quaternions and X_r be an array of voxel grid points in the neutral reference frame. The first transform is identity and is matched with the neutral reference SDF. The stable hull function $\mathcal{S}$ is

$$\mathcal{S}(Q) = \gamma\Big(\max_{i \in [N]} \phi_{\theta_i}(\mathbf{q}_i X_r \overline{\mathbf{q}_i})\Big) \tag{2}$$

With ψ as the ℓ_0 norm approximating penalty function of [Lamarre et al. 2018], the optimization process is

$$\arg\min_{Q} \frac{1}{N} \sum_{i=1}^{N} \psi\left(\phi_{\theta_i}\Big(\mathbf{q}_i \mathcal{S}(Q) \overline{\mathbf{q}_i}\Big)\right) \tag{3}$$

These equations are implemented in Pytorch and optimized using the Adam optimizer. We implement a two-step mode pursuit schedule, first optimizing with a histogram bin size of 2 mm and then 1 mm. The voxel grid point set X_r size is 60^3 but we use a mask to ignore voxels very far from inside or outside all surface at the initialization stage (+/- 4 mm) to accelerate computations. The masked voxel have fixed signed distance and do not influence the stable hull mesh. Each step run for 2000 iterations with learning rates $1e-3$. For 25 expressions, the process takes 5 minutes on a GeForce GTX 3090 and requires 15 GB of GPU memory.

3.3 Initialization

Initialization is important has we found the energy function to be non-convex. If the head can move a lot in space, a coarse head alignment method using face feature Mediapipe [Lugaresi et al. 2019] should be used before computing the voxel array bounding cube to use the volumetric resolution wisely. Afterwards, we use the mode pursuit method of [Lamarre et al. 2018] using vertex distance from SDF instead of vertex-vertex match.

4 RESULTS

To evaluate the skull carving algorithm on a significant sample of the population, we build a database of 32 persons. We picked randomly from our larger internal head capture database to get an equal partition of male-female, asian-black-latino-white phenotype and young-old. The sample contains people of different BMI.

For comparison with our skull carving algorithm, we test five stabilization methods: (FLAME, ℓ_2-LBFGS-ICP, ℓ_1-LBFGS-ICP, GM-LBFGS-ICP, mode pursuit). For FLAME we use the official Chumpy based fitting software [Li et al. 2017]. We use the pose of the 2nd bone in the skinning hierarchy. We got the best result by first fitting the FLAME model on the neutral scan with expression parameter estimation disabled, and then fitting all expressions with the shape parameters estimation disabled and expression parameter enabled. Since we have neural SDFs available for which the gradient of the distance is available, we can use a gradient descent version of ICP which was shown in Levenberg-Marquardt ICP (LM-ICP) [Fitzgibbon 2003] to be more flexible and robust than the standard implementation. LM is not available in Pytorch but L-BFGS with strong Wolfe line search is, so we implemented L-BFGS ICP with three different loss functions: ℓ_2, ℓ_1 and Geman-McClure robust function as in [Li et al. 2017]. ICP and mode-pursuit use the mask illustrated Fig.3, which is also used as the basis to compute the skull carving bounding box. As an ablation study, we also test a version of skull carving where Flexicube gradients and grid deformation offsets are disabled. In this case, Flexicube becomes a fixed isosurface extraction function.

The quantitative evaluation process is manual and performed on expression with visible upper teeth. There is a total of 227 such expression scans, between 4 and 10 per person with an average of 7. The expression list varies slightly between captures, but even for the same action unit combo the visibility of the upper teeth may vary because of morphology. The results are compiled in Table 1. Skull carving is superior to all other methods. All ICP methods have similar results. Robust estimators do not work on this dataset because it is mostly composed of combined action units that deform

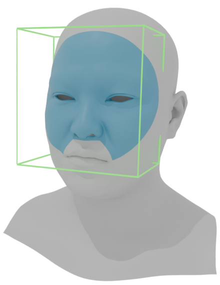

Figure 3: ICP and mode-pursuit use the mask region in blue. Skull carving is computed in the mask region rotated 3D bounding box with some extra room in vertical axis to cover the upper teeth.

all the upper part of the face, which mean the outlier percentage is too high for these estimators to work. Mode-pursuit often fails badly for the same reason. FLAME results are generally consistent but with a relatively high error for this task. It seems to indicate that training a similar model with a specific algorithm for the skull pose estimation could be beneficial. The bad result of skull carving without surface extraction gradients show that they are required to make the method optimize all expressions in a set globally. When the gradient are disabled, skull carving will converge to a local minima that focus on an expression subset which depends on the initialization and the shape of the initial stable hull.

Method	Max Upper Teeth Alignment Error			
	<= 1 mm (%)	<= 2 mm (%)	<= 3 mm (%)	> 3 mm (%)
ℓ_2-ICP	3	22	88	12
ℓ_1-ICP	3	28	63	37
GM-ICP	6	25	56	44
FLAME	9	50	94	6
Mode pursuit	13	38	53	47
Skull Carving	**78**	**97**	**97**	3
NG[1] *Skull Carving*	13	31	63	38

Table 1: In a 3D DCC software, for each of the 32 persons of the database, for the set of expression with visible upper teeth, we align an upper teeth template manually on the stabilized expressions scans and estimate the worst case translation error in the set and classify each person in the four error brackets.

5 DISCUSSION AND CONCLUSION

Our experiment show that differentiable isosurface extraction is a necessary component of the skull carving stabilization algorithm. We observed that the (3) does not converge from a coarse initialisation. It seems the stable hull must have already well defined 3D features like teeth stub and fairly complete nose arch for convergence. We suspect we could add some domain specific regularization on the stable hull shape, for example a distance to a learned subspace, or a statistical anthropometric distance, to improve convergence. Skull carving is not adapted to stabilize 4D data since even on a high-end 48GB GPU, it can process only about 100 frames at a time. More research is needed to use the stable hull concept on large number of sequential scans.

In this paper, we introduced a new solution to the facial animation skull stabilization problem on a sparse set of combined FACS unit expressions. This problem is more challenging than stabilization on 4D capture since there is no motion information available. While there is room for improvement, the skull carving algorithm is more accurate than existing methods on our diverse database and we believe it should generalize well because it makes no morphological assumption.

[1] No gradient

REFERENCES

Gavin Barill, Nia Dickson, Ryan Schmidt, David I.W. Levin, and Alec Jacobson. 2018. Fast Winding Numbers for Soups and Clouds. *ACM Transactions on Graphics* (2018).

Thabo Beeler and Derek Bradley. 2014. Rigid Stabilization of Facial Expressions. *ACM Trans. Graph.* 33, 4, Article 44 (July 2014), 9 pages.

Sofien Bouaziz, Andrea Tagliasacchi, and Mark Pauly. 2013a. Sparse Iterative Closest Point. In *Eurographics/ACMSIGGRAPH SGP (SGP '13)*. 113–123.

Sofien Bouaziz, Yangang Wang, and Mark Pauly. 2013b. Online modeling for realtime facial animation. *ACM Transactions on Graphics (ToG)* 32, 4 (2013), 1–10.

Jacob K Dey, Chelsey A Recker, Michael D Olson, Andrew J Bowen, and Grant S Hamilton III. 2021. Predicting nasal soft tissue envelope thickness for rhinoplasty: a model based on visual examination of the nose. *Annals of Otology, Rhinology & Laryngology* 130, 1 (2021), 60–66.

Andrew W Fitzgibbon. 2003. Robust registration of 2D and 3D point sets. *Image and vision computing* 21, 13-14 (2003), 1145–1153.

Kiriakos N Kutulakos and Steven M Seitz. 2000. A theory of shape by space carving. *International journal of computer vision* 38 (2000), 199–218.

Mathieu Lamarre, John P Lewis, and Etienne Danvoye. 2018. Face stabilization by mode pursuit for avatar construction. In *2018 IVCNZ*. IEEE, 1–6.

Aldo Laurentini. 1994. The visual hull concept for silhouette-based image understanding. *IEEE PAMI* 16, 2 (1994), 150–162.

Tianye Li, Timo Bolkart, Michael. J. Black, Hao Li, and Javier Romero. 2017. Learning a model of facial shape and expression from 4D scans. *ACM Transactions on Graphics, (Proc. SIGGRAPH Asia)* 36, 6 (2017), 194:1–194:17.

Camillo Lugaresi, Jiuqiang Tang, and Hadon Nash. 2019. MediaPipe: A Framework for Perceiving and Processing Reality. In *Third Workshop on Computer Vision for AR/VR CVPR 2019*.

Shunsuke Saito, Gabriel Schwartz, Tomas Simon, Junxuan Li, and Giljoo Nam. 2024. Relightable Gaussian Codec Avatars. In *CVPR*.

Tianchang Shen, Jacob Munkberg, Jon Hasselgren, Kangxue Yin, Zian Wang, Wenzheng Chen, Zan Gojcic, Sanja Fidler, Nicholas Sharp, and Jun Gao. 2023. Flexible Isosurface Extraction for Gradient-Based Mesh Optimization. *ACM Trans. Graph.* 42, 4, Article 37 (jul 2023), 16 pages. https://doi.org/10.1145/3592430

Delio Vicini, Sébastien Speierer, and Wenzel Jakob. 2022. Differentiable Signed Distance Function Rendering. *Transactions on Graphics (Proceedings of SIGGRAPH)* 41, 4 (July 2022), 125:1–125:18. https://doi.org/10.1145/3528223.3530139

Yiqun Wang, Ivan Skorokhodov, and Peter Wonka. 2023. Pet-neus: Positional encoding tri-planes for neural surfaces. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 12598–12607.

Yunjie Wu, Yapeng Meng, Zhipeng Hu, Lincheng Li, Haoqian Wu, Kun Zhou, Weiwei Xu, and Xin Yu. 2024. Text-Guided 3D Face Synthesis-From Generation to Editing. In *Proceedings of the IEEE/CVF CVPR*. 1260–1269.

Michael Zollhöfer, Justus Thies, Pablo Garrido, Derek Bradley, Thabo Beeler, Patrick Pérez, Marc Stamminger, Matthias Nießner, and Christian Theobalt. 2018. State of the art on monocular 3D face reconstruction, tracking, and applications. In *Computer graphics forum*, Vol. 37. Wiley Online Library, 523–550.