A Preliminary Investigation into the Impact of Training for Example-Based Facial Blendshape Creation

Emma Carrigan[1], Ludovic Hoyet[2], Rachel McDonnell[1] and Quentin Avril[3]

[1]Graphics Vision and Visualisation Group, Trinity College Dublin, Ireland
[2]Inria Rennes, France [3] Technicolor

Abstract

Our work is a perceptual study into the effects of training poses on the Example-Based Facial Rigging (EBFR) method. We analyse the output of EBFR given a set of training poses to see how well the results reproduced our ground truth actor scans compared to a Deformation Transfer approach. While EBFR produced better results overall, there were cases that did not see any improvement. While some of these results may be explained by lack of sufficient training poses for the area of the face in question, we found that certain lip poses were not improved by training, despite a large number of mouth training poses supplied.

Categories and Subject Descriptors (according to ACM CCS): I.3.3 [Computer Graphics]: Picture/Image Generation—Line and curve generation

CCS Concepts

•***Computing methodologies*** → ***Mesh models;*** *Animation;*

1. Introduction

The process of creating blendshapes is still very reliant on artists. Although there is a lot of research done on scanning and rigging faces, any facial blendshapes or rigs that are created using these methods require extensive editing for use in games and movies. Additionally, some methods require a large amount of facial scans or motion capture, which can be costly depending on the technology required or the amount of time for which an actor must be hired. Currently, these are necessary costs for any AAA game or blockbuster.

One method which is used in production is Example-Based Facial Rigging [LWP10], which is an extension of Deformation Transfer [SP04]. This method uses a generic blendshape rig template, i.e. a neutral face with a number of blendshape target faces, as well as the neutral face of the character which you want to create the rig for, and a number of facial poses of this character. The algorithm recreates each of the blendshapes of the generic rig for the desired target face, while also incorporating facial details which it learns from the supplied facial poses.

While this method is used in professional pipelines, companies still need to hire an actor to create numerous poses and 3D modelling artists to clean the final blendshapes. However, we believe that we can improve the blendshape creation process to cut down the dependency on actors by finding the optimal types of poses to supply to the system.

Our contribution is a preliminary perceptual study into the effect of a set of input scans on the EBFR system. From this, we can see what areas of the face were most improved by the algorithm and which areas were least improved. This gives us an idea of the impact of our supplied training poses and is a basis for further research into reducing the number of scans needed to attain suitable blendshape rigs using EBFR.

2. Related Work

Sumner and Popović define a method of deformation transfer for triangle meshes. This method deforms two meshes similarly given a source mesh, a deformation of the source mesh, a target mesh, and correspondence between the two [SP04]. The transfer is achieved through solving a constrained optimisation for the target mesh topology that matches the source deformations as closely as possible, while maintaining consistency constraints. Expanding on this, Ben-Chen et al. describe a spatial deformation transfer technique that allows deformation transfer to be applied to more than just single component manifold triangle meshes [BCWG09].

Li et al. propose a method for generating facial blendshapes given a generic facial rig and a neutral pose for the target mesh [LWP10]. This method improves upon Sumner and Popović's Deformation Transfer technique by supplying example poses of the target mesh to train the blendshape generation. This method uses a

generic template model with all of the blendshapes that are to be created for the target model.

Xu et al. propose a method for facial animation transfer that transfers detailed animations and allows for quick user-editing of the spatial-temporal domain [XCLT14]. This approach splits the high-fidelity facial performance into high-level facial feature lines, large-scale facial deformation, and fine-scale motion details. It then transfers and reconstructs them to create the retargeted animation.

Ribero et al. propose a method of facial retargeting that takes into account the range of motion of the source and target characters in order to allow retargeting between characters of significantly different styles and proportions [RZL*17].

3. Method

Our idea was to use all the common training poses available across a number of actors, which would theoretically create the best possible trained rigs for our data using EBFR. We then ran an experiment comparing these trained rigs, as well as untrained rigs created using the Deformation Transfer method, to a ground truth facial scan. The participants chose which of DT or EBFR faces best resembled the ground truth, then described how close their chosen pose was to the ground truth on a 5-point Likert scale. This showed us which parts of the face were most improved or disimproved by EBFR. An example of the stimuli can be seen in Figure 1.

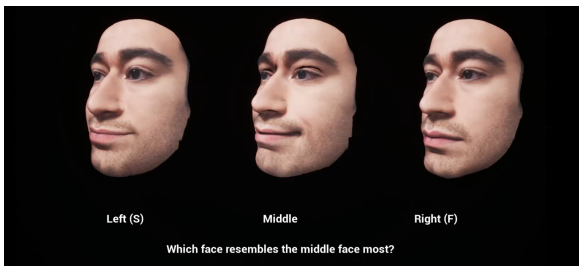

Figure 1: *An example of stimuli shown to participants comparing the facial rigs created without training, and those with training.*

3.1. Stimuli

We used the Bosphorus Database to get data for our experiment [SAD*08]. The data is provided as point clouds with textures. We meshed, cleaned, normalized and registered this data to create meshes with consistent topology for ease of use for our experiment. After this preprocessing step, we had a large number of facial expressions from different actors which we could use both as ground truth and as input to EBFR. An example of the cleaned data we used can be seen in Figure 2.

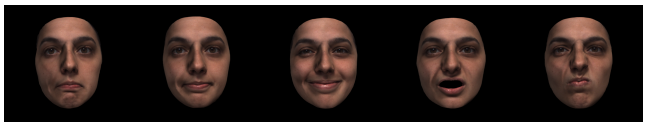

Figure 2: *Examples of scanned poses from the Bosphorus database.*

The Bosphorus database is a database of facial scans of over a hundred actors attempting to recreate the Action Units (AUs) as described in Ekman et al.'s Facial Action Coding System (FACS) [EF78], however due to the difficulty of activating certain facial muscles in isolation, most scans are actually a combination of AUs. Fortunately, each scan in the database has been annotated by a FACS expert. Using this information, we can attempt to recreate the scans using our facial rigs created using DT and EBFR, as the blendshapes of the rigs we used were based on FACS.

AU No.	FACS Name	Expression No.
9	Nose Wrinkler	0
10	Upper Lip Raiser	1
12	Lip Corner Puller	2
14	Dimpler	3
15	Lip Corner Depressor	4
16	Lower Lip Depressor	5
17	Chin Raiser	6
18	Lip Pucker	7
22	Lip Funneler	8
23	Lip Tightener	9
24	Lip Pressor	10
25	Lips Part	11
26	Jaw Drop	12
27	Mouth Stretch	13
28	Lip Suck	14
34	Cheek Puff	15
2	Outer Brow Raiser	16
4	Brow Lowerer	17
43	Eyes Closed	18

Table 1: *The Action Units we used in our experiment with their FACS names. The third column shows the numbers we used for them in our experiment. Expressions 0-15 are lower face expressions, 16-18 are upper face.*

We selected 4 female and 4 male actors from this database, and created trained and untrained rigs for each selected actor. The untrained rig was created using Deformation Transfer using a generic facial animation rig whose blendshapes were based on FACS AUs. The trained rig was created using Example-Based Facial Rigging, using the same generic model and neutral pose as the untrained rig, but including 19 additional facial scans of the actor with different expressions as training poses.

The expressions in the scans used as training were the same across all actors. These expressions consisted of the 19 Action Units as detailed in Table 1. These expressions were chosen because they were the most commonly represented expressions in the database and we required a common set of expressions across all actors. They were also chosen to convey information from all the different areas of the face, e.g., mouth, nose, eyes.

3.2. Participants and Procedure

Participants were presented with 152 trials: 2 Actor Sex (Female, Male) $\times$ 4 Actors $\times$ 19 Expressions. In each trial, participants were presented with the ground truth face (scan) in the middle of the screen, and both the trained and untrained faces randomly presented on the left or right side of the screen (Figure 1). Participants

could rotate the faces simultaneously using the arrow keys on the keyboard, to a maximum of 30 degrees in each direction. For each stimulus they were asked "Which face resembles the middle face most?", and answered using the S and F keyboard keys. They saw the faces for a maximum of 10s, after which they were forced to provide an answer. Then they were asked to rate how close the face they selected was to the middle face on a scale from 1 (Not at all) to 5 (Identical) using the keyboard. The trials in the experiment were presented in blocks: each actor of one gender was presented in a random order, then the actors of the other gender. The genders were presented in a randomized order. All the expressions for one actor were presented in a random order before moving to the next actor.

We included training stimuli at the beginning of the experiment, identical across participants and using an actor who did not appear in the experiment. The participants used these stimuli to become familiar with the experiment and the buttons needed to answer our questions. Responses for these stimuli were not recorded. A screen was shown between the training and real experiment to warn the participants that their responses would begin to be recorded.

Twenty-three participants took part in our experiment (5 female, 17 male and 1 other, aged 23-61 years). They viewed the experiment on a 24" display of resolution 1920x1200. Each participant was given an information sheet and consent form to sign. The information was repeated on the screen at the beginning of the experiment. The participant was then asked to input some demographics information before they began the experiment.

4. Results

To assess whether trained (EBFR) faces were preferred to untrained (DT) faces, as well as whether differences appear for different parts of the faces, we performed a one-way repeated measures Analysis of Variance (ANOVA) with within-subject factors *Expression* on the percentage of times EBFR was preferred over DT. To analyse these results, each participant's results were averaged across all the actors for each condition. All effects are reported at $p < 0.05$. When we found main or interaction effects, we further explored the cause of these effects using Newman-Keuls ($p < 0.05$) post-hoc tests for pairwise comparisons.

First, we found a main effect of *Expression* ($F_{18,396}$=41.65, $p \approx 0$), where post-hoc analysis showed that EBFR was clearly preferred for some expressions, and less for others (Figure 3). To further explore these effects, we conducted single t-tests against 50% to evaluate if preference was above chance level ($p < 0.05$). Results showed 3 categories of expression, which are listed below:

Improved by EBFR: 0, 1, 2, 3, 5, 6, 7, 8, 9, 14 and 15

No preference between EBFR and DT: 4, 10, 12, 17 and 18

EBFR worsened the results: 11, 13, 16

4.1. Excluded Results

We found that, for some of the expressions, participants preferred the untrained faces across all actors, which was unusual as we expected the trained faces to be equal or better in every case. In order

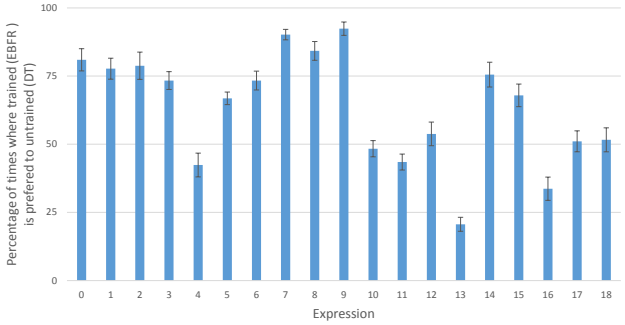

Figure 3: *Main effect of Expression on preference of EBFR over DT.*

to understand why, we manually examined the stimuli and found some artifacts across almost all actors for certain expressions.

There were texture artifacts for expressions 16 and 18, with FACS names Outer Brow Raiser and Eyes Closed, as can be seen on the left in Figure 4a. Although our interest was purely morphological and we asked participants to ignore texture artifacts to the best of their ability, we found these artifacts to be too noticeable to ignore. For this reason, we chose to exclude expressions 16 and 18 from our analysis.

While we could have avoided these issues by removing the textures on every stimulus, we found that the meshes with no texture were unnatural and might have affected the perception of participants, as they were too unlike real faces. As we are interested in human facial perception, we decided to include the textures to ensure the faces looked as human as possible.

(a) *Texture artifact example*

(b) *Left: Neutral scan, Centre: Expression 13, Right: Trained rig recreation of expression 13*

Figure 4: *(a) The texture artifact which affected expressions 16 and 18. (b) The artifact which affected expression 13.*

We also found that the trained stimulus for expression 13 (Mouth Stretch) was often unnatural looking, which we found to be caused by an error in scaling the scan from the database. In our data cleaning process, we scaled the faces to be of unit length. This had a strong negative effect on expression 13, as the actor opens their mouth as wide as possible, which causes the face to be a lot longer than when at rest. In the training process, we were essentially telling our algorithm to make the neutral actor scan (Figure 4b Left) shrink to match the scanned expression 13 (Figure 4b Centre). This resulted in an unnatural face (Figure 4b Right). For this reason, we excluded expression 13 from our results.

4.2. Analysis

After removing the results that were caused by artifacts, we can separate the results into groups as shown in Table 2.

AU	FACS Name	Exp.
9	Nose Wrinkler	0
10	Upper Lip Raiser	1
12	Lip Corner Puller	2
14	Dimpler	3
16	Lower Lip Depressor	5
17	Chin Raiser	6
18	Lip Pucker	7
22	Lip Funneler	8
23	Lip Tightener	9
28	Lip Suck	14
34	Cheek Puff	15

(a) *The expressions where EBFR was significantly preferred.*

AU	FACS Name	Exp.
15	Lip Corner Depressor	4
24	Lip Pressor	10
26	Jaw Drop	12
4	Brow Lowerer	17

(b) *The expressions where there was no significant difference between EBFR and DT.*

AU	FACS Name	Exp.
25	Lips Part	11

(c) *The expressions where DT was significantly preferred.*

Table 2: *Results grouped by ratio of trained to untrained responses.*

We excluded the Mouth Stretch expression, as our data cleaning algorithm scaled the faces so the meshes would be unit-length from top to bottom. This made Mouth Stretch smaller than it should have been. However, we did not exclude Jaw Drop as there were no obvious artifacts, although it appears a similar issue may have happened. Jaw Drop is a slightly longer than normal face, so the scaling should have affected this expression unfavourably as well.

Brow Lowerer was the only upper face expression that remained after we excluded results. We had a noticeable lack of upper face expressions to choose from, and two of the three expressions we had were excluded due to artifacts. Brow Lowerer's neutral result may be caused by not having enough upper face training poses.

Interestingly, for Lip Corner Depressor and Lip Pressor, expressions which cause the lips to be pushed together and stretched, it seems that the algorithm simply has a hard time recreating these. For Lip Pressor, the lips become quite thin, which confused our training algorithm and seemed to accentuate some sharp edges around the lip contour, and sometimes caused overlapping faces. For Lip Corner Depressor, the downward movement seemed to make the mouth open slightly in some cases, and stretch the bottom lip to make it look slightly larger in other cases. Lips Part had a similar issue in that it often made the contour of the lips slightly sharper. These small errors seem to be enough to affect the perception of these expressions.

5. Discussion

We created a number of facial rigs using EBFR and showed that EBFR produces perceptually better facial rigs than Deformation Transfer. We found that artifacts caused by the algorithm that affected the contour of the lips were more noticeable than artifacts that affected the other areas of the face. Our results indicate that the lip area is important when creating facial rigs. However, it is possible this was caused by the lack of an internal mouth structure. This caused any opening of the mouth to be very apparent. Future work will investigate the importance of the internal structure.

More interesting is the fact that the Lips Part expression was noticeably affected. This expression had the mouth slightly open, so we see that it is not the difference between an open and closed mouth that is noticeable, but the actual shape of the lips, specifically the edge between the lip and the inside of the mouth.

Our main limitation came from the database of facial scans we used. We chose it because we already had a pipeline for processing meshes from this database, however we found it difficult to get a wide sample of facial expressions across many actors. This was the reason we were lacking in upper-face expressions in our study, we simply did not find a subset of actors in the database that had scans of the same upper-face expressions.

Our initial goal for this project was to identify what facial expressions are important to use as training when using Example-Based Facial Rigging to create facial rigs. While our work here has indicated certain parts of the face that might require more attention when automatically creating blendshapes, there is room for more investigation into this topic. We would like to be able to specify a subset of facial expressions that would be considered the "ideal" subset to use for training the EBFR algorithm. To do this, we would need to create multiple rigs with separate subsets and compare them. This fell out of the scope of our study, but we feel it is an important next step.

Acknowledgements

This research was partially funded by Science Foundation Ireland as part of the Game Face (13/CDA/2135) project.

References

[BCWG09] BEN-CHEN M., WEBER O., GOTSMAN C.: Spatial deformation transfer. In *Proceedings of the 2009 ACM SIGGRAPH/Eurographics Symposium on Computer Animation* (2009), ACM, pp. 67–74. 1

[EF78] EKMAN P., FRIESEN W. V.: *Facial action coding system*. Consulting Psychologists Press, Stanford University, 1978. 2

[LWP10] LI H., WEISE T., PAULY M.: Example-based facial rigging. In *ACM Transactions on Graphics (TOG)* (2010), vol. 29, p. 32. 1

[RZL*17] RIBERA R. B. I., ZELL E., LEWIS J. P., NOH J., BOTSCH M.: Facial retargeting with automatic range of motion alignment. *ACM Transactions on Graphics (TOG) 36*, 4 (July 2017), 154:1–154:12. 2

[SAD*08] SAVRAN A., ALYÜZ N., DIBEKLIOĞLU H., ÇELIKTUTAN O., GÖKBERK B., SANKUR B., AKARUN L.: Bosphorus database for 3d face analysis. *Biometrics and identity management* (2008), 47–56. 2

[SP04] SUMNER R. W., POPOVIĆ J.: Deformation transfer for triangle meshes. *ACM Transactions on Graphics (TOG) 23*, 3 (2004), 399–405. 1

[XCLT14] XU F., CHAI J., LIU Y., TONG X.: Controllable high-fidelity facial performance transfer. *ACM Transactions on Graphics (TOG) 33*, 4 (2014), 42. 2