Investigating perceptually based models to predict importance of facial blendshapes

Emma Carrigan
Trinity College Dublin

Katja Zibrek
INRIA, Rennes

Rozenn Dahyot
Trinity College Dublin

Rachel McDonnell
Trinity College Dublin
ramcdonn@tcd.ie

ABSTRACT

Blendshape facial rigs are used extensively in the industry for facial animation of virtual humans. However, storing and manipulating large numbers of facial meshes is costly in terms of memory and computation for gaming applications, yet the relative perceptual importance of blendshapes has not yet been investigated. Research in Psychology and Neuroscience has shown that our brains process faces differently than other objects, so we postulate that the perception of facial expressions will be feature-dependent rather than based purely on the amount of movement required to make the expression. In this paper, we explore the noticeability of blendshapes under different activation levels, and present new perceptually based models to predict perceptual importance of blendshapes. The models predict visibility based on commonly-used geometry and image-based metrics.

CCS CONCEPTS

• **Applied computing** → **Psychology**; • **Computing methodologies** → **Mesh geometry models**; • **Mathematics of computing** → *Equational models.*

KEYWORDS

action units, perception, linear model, blendshapes

ACM Reference Format:
Emma Carrigan, Katja Zibrek, Rozenn Dahyot, and Rachel McDonnell. 2020. Investigating perceptually based models to predict importance of facial blendshapes. In *Motion, Interaction and Games (MIG '20), October 16–18, 2020, Virtual Event, SC, USA.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3424636.3426904

1 INTRODUCTION

Virtual human expressions are generally created by animating blendshape rigs [Lewis et al. 2014] based on the Facial Action Coding System (FACS) [Ekman and Friesen 1978b]. However, these rigs are computationally expensive for real-time applications due to the large amounts of geometry processing. The question of importance of blendshapes is therefore of great interest to computer games and other real-time applications, with the aim of reducing the number of blendshapes needed for animating a rig [Costigan et al. 2016], or prioritising which blendshapes to include in expressions for example-based rig creation algorithms [Carrigan et al. 2020]. Additionally, algorithms that create or alter facial geometry are usually evaluated against ground-truth facial meshes using standard geometry error metrics, however, we postulate that standard error-metrics may not be sufficient to determine how perceptually different the results are to the ground-truth.

Due to the fact that humans are particularly attuned to face perception [Bruce and Young 2013], we expect that differences in perception of facial action units will not align with the magnitude of displacement on the mesh caused by the expression. We hypothesise that small displacements in salient regions (e.g., eyelids) will be more perceptually noticeable than larger displacements in less salient regions (e.g., puffing of cheeks), which may not be accurately reflected by standard geometric and image error metrics. We also expect that the sex and race of the face will affect the perception of action units on that face, due to social conditioning [Hess et al. 2004; Lindsay et al. 1991].

In this paper, we investigate the perceptual impact of a carefully selected range of expressive action units at varying activation levels across a number of characters of different race and sex. We then compare our qualitative perceptual results to quantitative metrics in order to determine whether the perceptual effect can be predicted directly. We aim to determine if our question of perceived action unit importance can be predicted by calculating the *error* between the neutral pose and the expression blendshape, using both standard and perceptually based metrics, calculated from either the 3D geometry or the rendered 2D image. We then perform linear regression analysis to determine the best model describing the relationship between perception and error metrics. We address a number of questions, such as:

- Are certain facial action units more perceptually noticeable than others?
- Can we predict the saliency of facial action units using numerical error metrics, and is there a benefit to using existing perceptually based metrics?
- Are 3D geometry or 2D image-based metrics better at predicting the saliency of facial action units?

2 RELATED WORK

Face perception is a very active area of study in **Psychology**, as humans have been shown to have specialised sensory and interpretative processes related to faces [Bruce and Young 2013; Farah et al.

MIG '20, October 16–18, 2020, Virtual Event, SC, USA

ACM ISBN 978-1-4503-8171-0/20/10.
https://doi.org/10.1145/3424636.3426904

1998; Kanwisher et al. 1997]. Work by Schwaninger et al. [2009; 2006] shows that faces are processed both in terms of their components, as well as the configuration of those components, rather than purely holistically. A great deal of research is ongoing in the areas of face recognition, detection, memory, the other-race effect and the effect of experience on face perception, critical features for recognition, and social evaluation of faces [Oruc et al. 2019].

Since particular areas of the face are important for the recognition of emotion [Smith et al. 2005], different action units could potentially be more salient than others. The evidence supporting this suggests that specialised areas exist in the brain (region pSTS) for the perception of action units. This could indicate that action units are a necessary precursor to categorization of emotion [Srinivasan et al. 2016]. In addition, particular action units are responsible for the correct recognition of an emotion [Wegrzyn et al. 2017].

There were also studies which used the information about individual action units to generate synthetic expressions. A gradual activation of specific action units resulted in detection of an expression [Yu et al. 2012]. Reverse engineering expressions based on perceptual relevance helped with improved facial recognition in artificial faces [Chen et al. 2018]. There is enough evidence to suggest that action units alone have a perceptually significant impact on emotion categorisation. However, it is unknown if certain action units are more salient than others because they are associated with a particular emotional expression.

While the mouth is understandably a significantly attended to area due to its importance for emotional expression and communication [Nusseck et al. 2008] and its size relative to other facial features, the eyes and eyebrows can also be considered highly important despite their considerably smaller size. Eyebrows are integral for emotional and conversational signals [Ekman 1979], and can alter the perception of the eyes [Matsushita et al. 2015], however they are important in their own right for face recognition [Sadr et al. 2003] and not just in relation to how they change the perception of eyes.

While there has been much research in the area of Psychology on perception of the human face, these results are rarely utilized in **Computer Graphics** to improve the quality or computation of facial animation for real-time applications where resources are limited. The current state of the art for high quality real-time facial animation is blendshape animation [Lewis et al. 2014]. There is currently no consensus on what blendshapes a rig should contain, with the decision being left entirely to the artist. One solution is to use the Action Units from the Facial Action Coding System [Ekman and Friesen 1978a]. In theory, FACS breaks down facial expressions to their most basic components, making it a useful guideline for blendshape creation.

Optimisation of blendshape animation can be done in a few ways. Reducing mesh complexity is one method [Garland and Heckbert 1997], however this causes correspondence issues between shapes. The animation itself can be optimised by passing blendshapes [Lorach 2007] and animation to the GPU, and using GPGPU methods [Costigan et al. 2016]. The most relevant optimisation method for this paper would be blendshape reduction, either removing blendshapes from a rig or from an animation. Naturally, this would reduce the expressivity of a rig and reduce the quality of animations, so identifying salient blendshapes as we attempt to do in this work is important.

Mesh optimisations in graphics have traditionally been assessed using error metrics, which are used to measure dissimilarity between ground-truth geometry and geometry after undergoing simplification, watermarking, or lossy compression, with the goal of avoiding perceptible differences. The types of metrics used are *view-dependent* and *view-independent*, or image-based and geometry-based (see overview by Corsini et al. [2013]). We are interested if these metrics can be used in face-geometry perception. In the context of facial animation, Deng et al. [2008] used a similar approach to ours by bridging objective facial motion patterns with subjective perceptual outcomes to predict the emotional fidelity of expressive facial animations.

3 STIMULI CREATION

We first acquired a high-end photogrammetry-scanned *template model*, created by Eisko[1], a leading Digital Double company. The character had over 200 blendshapes, inspired by the FACS [Ekman and Friesen 1978b] with additional shapes for emotion and speech. Our *experiment characters* were a set of 6 high-resolution scanned neutral faces from 3D Scan Store[2], including 2 characters of each Asian, Black, and White race. Within each race group, there was 1 female and 1 male character.

3.1 Blendshape Transfer

In order to obtain a range of expressions for each of our experiment characters, we used the Russian 3D Scanner[3] Wrap 3.4 to transfer the topology of our template model to each of the neutral characters, using some feature points as guidance so that the semantics of the topology remained the same. We then used this wrapped mesh to warp the blendshapes of our template model to the experiment characters, thereby creating 6 new character rigs with equal topology and blendshapes. We chose not to include any hair on the characters as we are exclusively interested in facial features and wanted to avoid distracting elements.

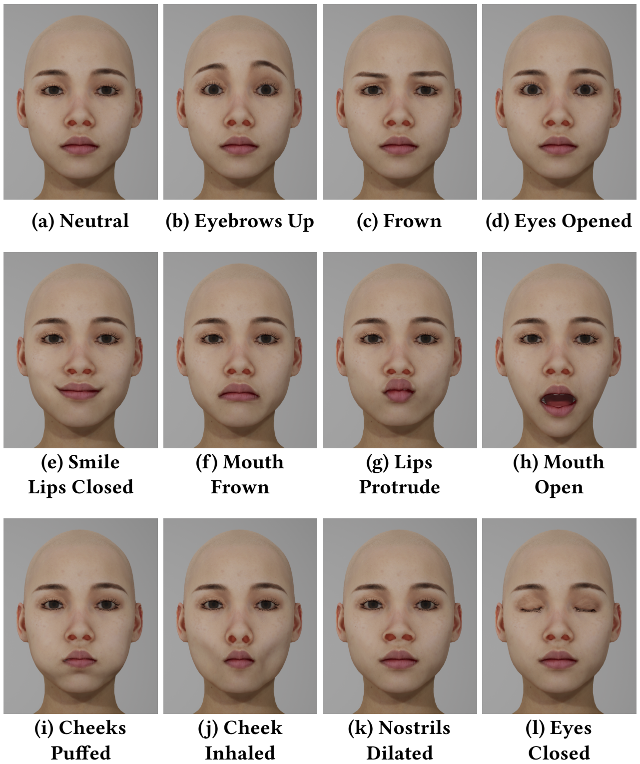
3.2 Action Unit Selection

We carefully chose 11 blendshapes from the character's set of 200 for the experiment (see Figure 1). The AUs chosen are those which are important for emotion (AUs 2, 4, 5, 12, 15, 26, 38) [Wegrzyn et al. 2017], speech (AUs 18, 26) [Meng et al. 2018], and those necessary for realistic and natural motion (AU 43) [Itti et al. 2003]. The cheeks have also been found to be important for facial recognition [Busso et al. 2004], so in order to fully cover potentially important features we also included cheek AUs 34 and 35. We also attempted to include opposite movements in each area, e.g., smile and frown.

3.3 Activation Levels

We are interested in whether the increase in onset of an AU linearly affects its perceptual importance, or whether there is a point at which the AU becomes more noticeable. For this reason, we investigate each AU at a number of different levels of activation. For

[1]https://www.eisko.com/
[2]https://www.3dscanstore.com/3d-head-models/
[3]https://www.russian3dscanner.com/



(a) Neutral (b) Eyebrows Up (c) Frown (d) Eyes Opened

(e) Smile Lips Closed (f) Mouth Frown (g) Lips Protrude (h) Mouth Open

(i) Cheeks Puffed (j) Cheek Inhaled (k) Nostrils Dilated (l) Eyes Closed

Figure 1: The blendshape set used in our experiment, shown on the Asian Female character at full activation (1.0).

each of these expressions, we show 5 activation levels: 0.2, 0.4, 0.6, 0.8, 1.0, with 1.0 being the maximum activation of that expression performed by the actor during the scanning process. In terms of blendshapes, this is simply a linear interpolation from the neutral face to the blendshape, with 1.0 being the fully activated expression (e.g., eyes fully closed) and each intermediate step being a transition from neutral to that expression.

4 PERCEPTUAL EXPERIMENT

We chose to develop a real-time experiment system in Unreal Engine 4 for flexibility and so that we could utilize pre-built advanced lighting and shading for realistic virtual character visualisation. For each trial of the experiment, we displayed the Neutral expression on the left and the stimulus on the right, and asked the participants to answer "How different are the expressions?" using a slider. The slider ranged from 1 defined as "No Difference" to 9 defined as "Extremely Different". Participants were aware that the left image was always neutral. After each trial, a 1 second focus cross was displayed. We chose the Likert scale instead of a two-alternative forced-choice paradigm, in order to determine the relative saliency of AUs and activation levels, rather than simply whether the activation levels were noticed or not. Participants were asked to answer as quickly and accurately as possible.

At the beginning of the experiment, participants conducted a training session, where they completed 11 trials showing the full activated blendshapes on the template character, which was not used in the main experiment in order to calibrate participants to the most extreme examples of each AU. Three hundred and sixty trials were shown to participants in random order, 12 AUs x 5 activation levels × 6 characters.

4.1 Participants

Twenty participants volunteered for the experiment (3 female, 16 male, 1 prefer not to answer; 8 were in the age range 18-27, 10 in 28-37, and 2 in 38-47). All reported medium or high familiarity with computer graphics and video games. As the experiment characters varied in race, and there is a perceptual effect of one's own race and perception of other races [Lindsay et al. 1991], we asked the participants to disclose their race (5 Asian, 13 White, 0 Black, 2 Other).

4.2 Results

We ran a 4-way repeated measures ANOVA on the perceptual difference results with the within factors Sex, Race, Action Unit, and Activation Level. Due to the imbalance between participant race and sex groups, we did not include these between-groups factors in the analysis. The ANOVA results can be seen in Table 1. We ran post-hoc analysis using Tukey's HSD tests throughout.

Table 1: ANOVA interactions with dependent variable "Difference" from the perceptual results. (AU = Action Unit, * represents significant p-values, F* stand for Greenhouse-Geisser correction for violations of sphericity). Effects sizes are reported in the last column (η_p^2).

Factor	F(DFn, DFd) = F-value	p-value	η_p^2
Sex	F(1, 19) = 1.727	0.2	0.08
Race	F(2, 38) = 4.192	0.02*	0.18
Action Unit	F*(2.93, 55.58) = 123.8	0.00*	0.86
Activation	F*(1.21, 22.90) = 158.2	0.00*	0.89
Sex-Race	F(2,38) = 7.826	0.001*	0.29
Sex-AU	F(11, 209) = 2.99	0.001*	0.14
Race-AU	F(22, 418) = 6.885	0.00*	0.27
Sex-Activation	F(4, 76) = 2.887	0.03*	0.13
Race-Activation	F(8, 152) = 1.581	0.14	0.08
AU-Activation	F(44, 836) = 19.29	0.00*	0.50
Sex-Race-AU	F*(6.73, 127.86) = 5.301	0.00*	0.22
Sex-Race-Activation	F(8, 152) = 2.031	0.046*	0.10
Sex-AU-Activation	F(44, 836) = 0.979	0.5	0.05
Race-AU-Activation	F(88, 1672) = 1.592	0.001*	0.07
Sex-Race-AU-Activation	F(88, 1672) = 1.68	0.00*	0.08

4.2.1 Character Sex & Race. There was no main effect of the character Sex. However, we found a main effect of character Race, where shape differences were less perceptible for Black characters overall than for Asian characters ($p < 0.02$). An interaction between Race and Sex gave further insight that shape differences were more perceptible for the Asian Female character than other characters except for the White Male ($p < 0.03$ for all). There was an interaction between Race and AU, which showed the Frown and Cheeks Puffed ($p < 0.02$) were the main AUs that were less perceptible on

Table 2: The AUs ordered by average perceptual difference.

AU Name	Difference	AU Name	Difference
Mouth Open	5.97	Eyes Opened	3.15
Eyes Closed	5.2	Cheeks Puffed	2.77
Smile Lips Closed	4.18	Mouth Frown	2.56
Eyebrows Up	3.56	Frown	2.22
Lips Protude	3.55	Nostrils Dilated	1.78
Cheek Inhaled	3.24	Neutral	1.42

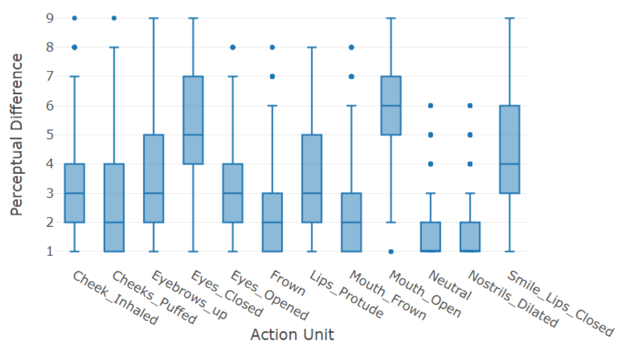


Figure 2: Main effect of AU from our experiment.

Black characters. These small differences across Race could have been due to our predominantly European and Asian participant pool [Lindsay et al. 1991; Walker and Tanaka 2003] or possibly due to differences in perception due to stand-out or distracting features in the rendering of their faces.

4.2.2 Activation Level. A main effect of Activation Level showed a significant increase in perceived differences as the activation increased, as expected. There was no difference across all characters and AUs at the lowest Activation Level of 0.2. However, some characters were rated as relatively more different at higher Activation Levels. Specifically, Asian Female at 0.6 Activation Level was rated similarly to the AUs of some characters at the 0.8 level.

4.2.3 Action Units. Mouth Open, Eyes Closed, and Smile Lips Closed appeared to have a higher perceptual effect since the perceived differences were significantly higher when compared to all other AUs ($p < 0.02$). Nostrils Dilated had the smallest effect since it was not significantly different from the Neutral. See Figure 2 and Table 2.

Further interactions showed that Mouth Open was significantly more different than most other shapes ($p < 0.005$). Eyes Closed were also prominent on some characters, while Nostrils Dilated and Frown were not different from Neutral, for some characters.

Mouth Frown was the only AU to be rated significantly differently between the sexes ($p < 0.05$), with the female characters being rated as more different. This could potentially be related to the inverse effect of gender stereotyping increasing saliency of unexpected emotions seen in previous work (i.e., that females are perceived as more angry than males) [Hess et al. 2004]. We also found interactions with Race, as well as interactions with Race and Sex (see Table 1). While we observed many significant differences from post-hoc tests, we did not observe any meaningful patterns.

5 ERROR METRICS

We now move to investigating the relationship between numerical error metrics and perception. We first record the error metric scores for each activation level of each AU, for each character. Each metric is calculated between the neutral face and the activated AU.

5.1 Geometric Error Metrics

Root-Mean-Square. We calculate the RMS error between two meshes by getting the sum across all vertices $n \in N$ of the square root of the average of the square of each (x, y, z) component of each delta vertex $\delta\vec{v}_n$, i.e. the difference between that vertex position in the blendshape mesh and the same vertex in the neutral mesh, as described in Equation 1.

$$\delta_{RMS} = \sum_{n=1}^{N} \sqrt{\frac{1}{3}\delta\vec{v}_n^T \delta\vec{v}_n} = \frac{1}{\sqrt{3}} \sum_{n=1}^{N} \|\delta\vec{v}_n\| \tag{1}$$

Spatio-Temporal Edge Difference. STED is a perceptual metric for dynamic meshes which focuses on local and relative changes of edge length by measuring the standard deviation of relative edge length around each vertex, rather than global mesh difference. The model parameters have been tuned such that its results best match those of the perceptual experiment, described in the paper [Vasa and Skala 2010].

5.2 Image Error Metrics

To calculate our image metric results, we took screenshots of each stimulus and cropped out a large amount of the empty space surrounding each head. MSE and SSIM were calculated using scikit-image [van der Walt et al. 2014].

Mean-Squared-Error. We calculate MSE by getting the per-pixel average error between images A and B, where N is the total number of pixels in the image, and $\vec{x}_n^A$ is the n^{th} pixel of image A.

$$\delta_{MSE} = \frac{1}{N} \sum_{n=1}^{N} \vec{x}_n^A - \vec{x}_n^B \tag{2}$$

Structural Similarity Index Metric. SSIM is calculated as defined by Wang et al. [2004] and using the default suggested parameters. It is designed to model the response of the human vision system and should correlate better to our perceptual results than standard MSE. As this metric measures similarity rather than difference, we invert this metric (1-SSIM) for better comparison with our other metrics where appropriate.

6 MODEL FIT

To find the best model describing the relationship between perceptual results and the calculated errors, several Generalised Linear Models were tested and compared using Akaike Information Criterion (AIC) that combines the log-likehood (best fit) penalised by the model complexity (as measured by the number of parameters to estimate in the model) for selection of the best model [Dobson and Barnett 2008]. The model with the lowest AIC is deemed the best model (amongst those tested) for explaining the observations. A χ^2 test for the deviance is then used to assess if this selected 'best' model is actually a good model for explaining the data [Dobson and Barnett 2008]. Poisson and Gaussian distributions were tested in combination with several link functions (identity, log, and square root). We found that the Poisson distribution captures the discrete nature of the perceived difference best and provides lower AICs than with the Gaussian distribution in the many models tested including the ones shown in Table 3.

Table 3: Model comparison with AIC↓ to explain the perceived difference (columns 2 and 3). Best link function reported between Identity (Id), log and sqrt. The lowest AIC for each metric are displayed in bold. Deviances (all with Poisson distribution and best link function) for models are shown column 4. A good model has a deviance in the interval [0; $\chi^2_{0.95}$] with $\chi^2_{0.95}$ reported in column 5.

Model	Gaussian	Poisson	Deviance	$\chi^2_{.95}$
STED	28704 (Id)	27346 (Id)	6551	7396
STED*AU	24850 (sqrt)	24680 (sqrt)	3844	7374
STED*AU+Race:Sex	24832 (sqrt)	**24670** (sqrt)	3823	7368
STED*AU*Sex*Race	24781 (sqrt)	24742 (sqrt)	3676	7252
RMS	27965 (Id)	26903 (Id)	6109	7396
RMS*AU	24878 (Id)	24688 (Id)	3852	7374
RMS*AU+Race:Sex	24853 (Id)	**24673** (Id)	3827	7368
RMS*AU*Sex*Race	24810 (Id)	24749 (Id)	3683	7252
SSIM	29534 (Id)	28075 (Id)	7280	7396
SSIM*AU	26287 (Id)	25591 (Id)	4753	7374
SSIM*AU+Race:Sex	25505 (sqrt)	25112 (log)	4264	7368
SSIM*AU*Sex*Race	24799 (sqrt)	**24758** (log)	3680	7252
MSE	29920 (Id)	28678 (Id)	7884	7396
MSE*AU	26940 (log)	26120 (log)	5277	7374
MSE*AU+Race:Sex	25933 (Id)	25384 (Id)	4536	7368
MSE*AU*Sex*Race	24839 (Id)	**24776** (Id)	3698	7252

6.1 Variable selection with ANOVA

The variables used to design the models shown in Table 3 have been chosen using ANOVA. Table 4 in the Supplemental Material shows the ANOVA when fitting a linear regression to explain the Perceived Difference (response variable) with explanatory variables RMS, AU and the 6 experiment characters (captured with variables Sex and Race). As can be seen by the high values for Sum Sq., most of the perceived difference is explained using RMS and AU with their interactions (variable highlighted in green in Table 4). Similarly, Table 5 in the Supplemental Material shows the ANOVA with explanatory variables STED, AU and the 6 virtual characters used and the high values for Sum Sq. imply that most of the perceived difference is explained using STED and AU with their interactions (variable highlighted in green Table 5). These results imply the relationship between the perceived values and the geometry metrics are AU-specific, and using an AU-specific model is necessary for prediction. ANOVA tables for image metrics are shown likewise in Table 6 and 7 in the Supplemental Material. We note that MSE and SSIM alone have less explanatory power than RMS and STED variables (see lower Sum Sq. in the tables). These ANOVA tables explain the comparison shown in Table 3 where AICs of models shown are either using only the metrics (*Metric*=STED/RMS/SSIM/MSE), the full models (*Metric**AU*Sex*Race), the ones considering interactions between metrics and blendshapes (*Metric**AU), and the models that include Sex and Race as additional variables.

6.2 Best Metric?

In the geometry domain, all fitted models are good models as per their deviance reported in Table 3 [Dobson and Barnett 2008]. However, we note that the perceptual metric STED achieves a lower AIC (marginally) in comparison to the standard metric RMS (see Table 3). Similarly, in the image domain, the perceptual metric SSIM achieves a lower AIC (marginally) in comparison to the standard metric MSE (Table 3). All fitted models are good models as per their deviance reported in Table 3 with the exception of the simplest one using only MSE. This shows that MSE has less explanatory power than SSIM for explaining the perceived difference, which is not surprising since it does not account for structural fidelity of the image. Figure 3 in the Supplemental Material illustrates the different models used.

We found that the perceptual image metric SSIM (measured in a 2D projective space) is not as powerful as even the standard geometry metric RMS (measuring the deformation in 3D) for explaining the perceived difference.

This is interesting, as our participants viewed the stimuli as a 2D projection, however their recorded perceived difference is better explained by geometric metrics computed from 3D meshes. A potential explanation may be that because faces are very familiar objects, a 3D representation is automatically imagined or inferred by participants when viewing 2D facial images. Despite this, having a model fitted using image metrics can be useful for prediction of perceived difference when geometry metrics are not available (e.g., for facial photograph comparisons).

7 DISCUSSION

In this paper, we presented the first experiment on perceptibility of facial action units, and the relationship with numerical metrics describing the displacements. Our main contribution is our perceptual models, which will provide a starting point for the development of a universal perceptual error metric suitable for human faces[4].

Additionally, we found that some facial action units were more perceptually noticeable than others, and provide a table showing the order of importance (Table 2). This perceptual ordering will be useful for game developers for tasks that require an order of blendshapes, such as level-of-detail blendshape reduction methods [Costigan et al. 2016], or example creation for blendshape transfer [Carrigan et al. 2020]. By identifying and removing blendshapes of lower visual saliency, which equates to simply removing rows from the blendshape matrix, we can save both memory and computation required.

In general, we found an equally-spaced linear relationship between perceptual difference and activation level. This implies that future experiments can focus solely on the highest activation level (1.0) of an AU.

[4]Data and R code used for this study are shared at https://github.com/Roznn/facial-blendshapes allowing others to build on our data investigating a larger range of faces, viewpoints, and facial action units.

In terms of the error metrics, one of our aims was to determine if existing standard and perceptual geometry- and image-based error metrics could predict the perceptibility of facial action units when compared to the neutral face. We did not find this to be the case, as our statistical models showed that each AU had a different relationship with the metrics (see different slopes in Supplemental Material, Figure 3). It is interesting how no single geometric or image-based metric alone was able to capture well the perceptual difference as seen by a human, even by the metrics which are based on the human vision system. This implies an AU-specific perception that must be taken into account when measuring the perceptual impact of changes in facial expressions, which is not catered for by the current metrics used alone in the models.

Interestingly, we found that image metrics were worse at predicting perceived differences than geometry metrics, even though the viewers only viewed the 3D geometry from a single viewpoint. This implies that humans have a strong ability to infer 3D shape of faces from a 2D image, and that the pixel-based differences in the images do not capture these differences as well as 3D geometry comparisons. This is unlikely to hold true for different viewpoints besides the front view, but will be interesting to investigate in future work.

In this work, we focused on realistic virtual human faces, so it is not clear if our results would generalize to other character rigs or even real faces. It would be interesting to investigate stylized or cartoon character rigs in future work. Also, as this is an initial study into investigating the perceived importance of facial AUs, we limited our study to static expressions of single AUs. Naturally, perception of animated faces with combined expressions would be more complicated, particularly since specific AUs are important for the perception of emotions (e.g., AU 7 Lid Tightener for anger [Wegrzyn et al. 2017]). It might be the case that even if these AUs are not perceptually important according to our approach, removing them from a rig might alter the interpretation of emotion of a virtual human, which we will study in future work.

ACKNOWLEDGMENTS

This work was funded by Science Foundation Ireland under the ADAPT Centre for Digital Content Technology (Grant 13/RC/2106) and the Game Face project (Grant 13/CDA/2135).

REFERENCES

Vicki Bruce and Andy Young. 2013. *Face perception.* Psychology Press.

Carlos Busso, Zhigang Deng, Serdar Yildirim, Murtaza Bulut, Chul Min Lee, Abe Kazemzadeh, Sungbok Lee, Ulrich Neumann, and Shrikanth Narayanan. 2004. Analysis of emotion recognition using facial expressions, speech and multimodal information. In *Proceedings of the 6th international conference on Multimodal interfaces.* 205–211.

Emma Carrigan, Eduard Zell, and Rachel McDonnell. 2020. Expression Packing: As-Few-As-Possible Training Expressions for Blendshape Transfer. In *Proceedings of the 41st Annual European Association for Computer Graphics Conference*, Vol. 39. Eurographics Association, 219–233.

Chaona Chen, Oliver GB Garrod, Jiayu Zhan, Jonas Beskow, Philippe G Schyns, and Rachael E Jack. 2018. Reverse Engineering Psychologically Valid Facial Expressions of Emotion into Social Robots. In *Int. Conf. on Automatic Face & Gesture Recognition.* 448–452.

Massimiliano Corsini, Mohamed-Chaker Larabi, Guillaume Lavoué, Oldřich Petřík, Libor Váša, and Kai Wang. 2013. Perceptual metrics for static and dynamic triangle meshes. In *Computer Graphics Forum*, Vol. 32. Wiley Online Library, 101–125.

Timothy Costigan, Anton Gerdelan, Emma Carrigan, and Rachel McDonnell. 2016. Improving blendshape performance for crowds with GPU and GPGPU techniques. In *Proceedings of the 9th International Conference on Motion in Games.* ACM, 73–78.

Zhigang Deng and Xiaohan Ma. 2008. Perceptually Guided Expressive Facial Animation. In *Proceedings of the 2008 ACM SIGGRAPH/Eurographics Symposium on Computer Animation* (Dublin, Ireland) *(SCA '08)*. Eurographics Association, Goslar, DEU, 67–76.

A. J. Dobson and A. G. Barnett. 2008. *An Introduction to Generalized Linear Models.* CRC Press, Third Edition.

Paul Ekman. 1979. About brows: Emotional and conversational signals. *Human Ethology* (1979).

Paul Ekman and Wallace V Friesen. 1978a. *Facial action coding system.* Consulting Psychologists Press, Stanford University.

Paul Ekman and Wallace V. Friesen. 1978b. *Facial Action Coding System: A Technique for the Measurement of Facial Movement.* Consulting Psychologists Press.

Martha J Farah, Kevin D Wilson, Maxwell Drain, and James N Tanaka. 1998. What is "special" about face perception? *Psychological review* 105, 3 (1998), 482.

Michael Garland and Paul S Heckbert. 1997. Surface simplification using quadric error metrics. In *Proceedings of the 24th annual conference on Computer graphics and interactive techniques.* 209–216.

Ursula Hess, Reginald B Adams Jr, and Robert E Kleck. 2004. Facial appearance, gender, and emotion expression. *Emotion* 4, 4 (2004), 378.

Laurent Itti, Nitin Dhavale, and Frederic Pighin. 2003. Realistic avatar eye and head animation using a neurobiological model of visual attention. In *Applications and Science of Neural Networks, Fuzzy Systems, and Evolutionary Computation VI*, Vol. 5200. International Society for Optics and Photonics, 64–78.

Nancy Kanwisher, Josh McDermott, and Marvin M Chun. 1997. The fusiform face area: a module in human extrastriate cortex specialized for face perception. *Journal of neuroscience* 17, 11 (1997), 4302–4311.

John P Lewis, Ken Anjyo, Taehyun Rhee, Mengjie Zhang, Frederic H Pighin, and Zhigang Deng. 2014. Practice and Theory of Blendshape Facial Models. *Eurographics (State of the Art Reports)* 1, 8 (2014), 2.

D Stephen Lindsay, Philip C Jack, and Marcus A Christian. 1991. Other-race face perception. *Journal of applied psychology* 76, 4 (1991), 587.

Tristan Lorach. 2007. Directx 10 blend shapes: Breaking the limits. *GPU Gems* 3 (2007), 53–67.

Soyogu Matsushita, Kazunori Morikawa, Saya Mitsuzane, and Haruna Yamanami. 2015. Eye shape illusions induced by eyebrow positions. *Perception* 44, 5 (2015), 529–540.

Zibo Meng, Shizhong Han, Ping Liu, and Yan Tong. 2018. Improving speech related facial action unit recognition by audiovisual information fusion. *IEEE transactions on cybernetics* 49, 9 (2018), 3293–3306.

Manfred Nusseck, Douglas W Cunningham, Christian Wallraven, and Heinrich H Bülthoff. 2008. The contribution of different facial regions to the recognition of conversational expressions. *Journal of vision* 8, 8 (2008), 1–1.

Ipek Oruc, Benjamin Balas, and Michael S Landy. 2019. Face perception: A brief journey through recent discoveries and current directions. *Vision Research* 157 (2019), 1–9.

Javid Sadr, Izzat Jarudi, and Pawan Sinha. 2003. The role of eyebrows in face recognition. *Perception* 32, 3 (2003), 285–293.

Adrian Schwaninger, Janek S Lobmaier, Christian Wallraven, and Stephan Collishaw. 2009. Two routes to face perception: evidence from psychophysics and computational modeling. *Cognitive science* 33, 8 (2009), 1413–1440.

Adrian Schwaninger, Christian Wallraven, Douglas W Cunningham, and Sarah D Chiller-Glaus. 2006. Processing of facial identity and expression: a psychophysical, physiological, and computational perspective. *Progress in Brain Research* 156 (2006), 321–343.

Marie L Smith, Garrison W Cottrell, FrédéAric Gosselin, and Philippe G Schyns. 2005. Transmitting and decoding facial expressions. *Psychological science* 16, 3 (2005), 184–189.

Ramprakash Srinivasan, Julie D Golomb, and Aleix M Martinez. 2016. A neural basis of facial action recognition in humans. *Journal of Neuroscience* 36, 16 (2016), 4434–4442.

Stéfan van der Walt, Johannes L. Schönberger, Juan Nunez-Iglesias, François Boulogne, Joshua D. Warner, Neil Yager, Emmanuelle Gouillart, Tony Yu, and the scikit-image contributors. 2014. scikit-image: image processing in Python. *PeerJ* 2 (2014), e453.

Libor Vasa and Vaclav Skala. 2010. A perception correlated comparison method for dynamic meshes. *IEEE transactions on visualization and computer graphics* 17, 2 (2010), 220–230.

Pamela M Walker and James W Tanaka. 2003. An encoding advantage for own-race versus other-race faces. *Perception* 32, 9 (2003), 1117–1125.

Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. 2004. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing* 13, 4 (2004), 600–612.

Martin Wegrzyn, Maria Vogt, Berna Kireclioglu, Julia Schneider, and Johanna Kissler. 2017. Mapping the emotional face. How individual face parts contribute to successful emotion recognition. *PloS one* 12, 5 (2017).

Hui Yu, Oliver GB Garrod, and Philippe G Schyns. 2012. Perception-driven facial expression synthesis. *Computers & Graphics* 36, 3 (2012), 152–162.

SUPPLEMENTAL MATERIAL

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
RMS	1	11125.30	11125.30	6174.53	0.00
AU	11	5114.03	464.91	258.02	0.00
Sex	1	3.56	3.55	1.97	0.16
Race	2	24.91	12.45	6.91	0.001
RMS:AU	10	2103.48	210.35	116.74	0.00
RMS:Sex	1	7.98	7.98	4.43	0.035
AU:Sex	11	30.06	2.73	1.52	0.118
RMS:Race	2	21.93	10.96	6.09	0.002
AU:Race	22	132.01	6.00	3.33	0.00
Sex:Race	2	36.30	18.15	10.07	0.00
RMS:AU:Sex	10	10.55	1.05	0.59	0.827
RMS:AU:Race	20	63.22	3.16	1.75	0.02
RMS:Sex:Race	2	6.87	3.43	1.91	0.149
AU:Sex:Race	22	140.94	6.41	3.56	0.00
RMS:AU:Sex:Race	20	58.88	2.94	1.63	0.037
Residuals	7062	12724.35	1.80	NA	NA

Table 4: ANOVA interactions with dependent variable "Difference" and within factors RMS, Sex, Race and AU.

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
STED	1	8909.20	8909.20	4959.53	0.00
AU	11	8879.79	807.25	449.38	0.00
Sex	1	0.6	0.6	0.33	0.563
Race	2	27.51	13.76	7.66	0.00
STED:AU	10	591.92	59.19	32.95	0.00
STED:Sex	1	10.49	10.49	5.84	0.016
AU:Sex	11	34.24	3.11	1.73	0.060
STED:Race	2	16.42	8.21	4.57	0.010
AU:Race	22	116.68	5.30	2.95	0.00
Sex:Race	2	26.00	13.00	7.24	0.00
STED:AU:Sex	10	9.72	0.97	0.54	0.862
STED:AU:Race	20	65.08	3.25	1.81	0.015
STED:Sex:Race	2	5.76	2.88	1.60	0.201
AU:Sex:Race	22	165.56	7.53	4.19	0.00
STED:AU:Sex:Race	20	59.37	2.97	1.65	0.034
Residuals	7062	12686.04	1.80	NA	NA

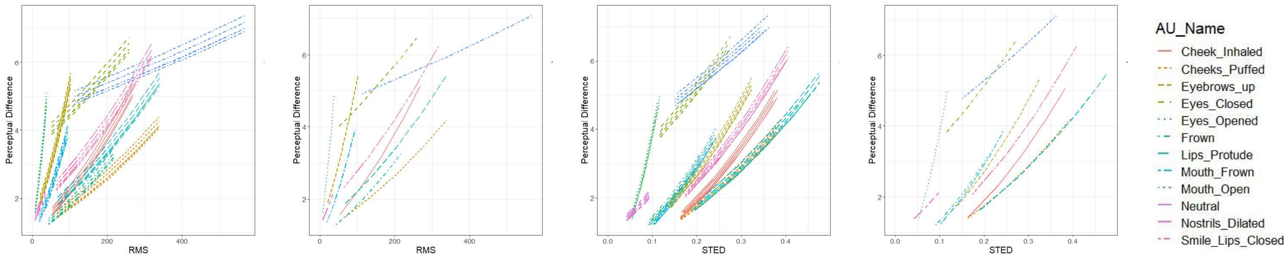
Table 5: ANOVA interactions with dependent variable "Difference" and within factors STED, Sex, Race and AU.

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
SSIM	1	6135.88	6135.88	3393.49	0.00
AU	11	8500.78	772.8	427.4	0.00
Sex	1	204.14	204.14	112.9	0.00
Race	2	615.25	307.62	170.13	0.00
SSIM:AU	11	924.91	84.08	46.5	0.00
SSIM:Sex	1	23.57	23.57	13.03	0.00
AU:Sex	11	405.87	36.9	20.41	0.00
SSIM:Race	2	113.56	56.78	31.4	0.00
AU:Race	22	484.94	22.04	12.19	0.00
Sex:Race	2	645.41	322.7	178.47	0.00
SSIM:AU:Sex	11	97.8	8.89	4.92	0.00
SSIM:AU:Race	22	148.01	6.78	3.72	0.00
SSIM:Sex:Race	2	148.00	74.00	40.93	0.00
AU:Sex:Race	22	297.97	13.54	7.49	0.00
SSIM:AU:Sex:Race	22	100.09	4.55	2.52	0.00
Residuals	7056	12758.19241	1.808134	NA	NA

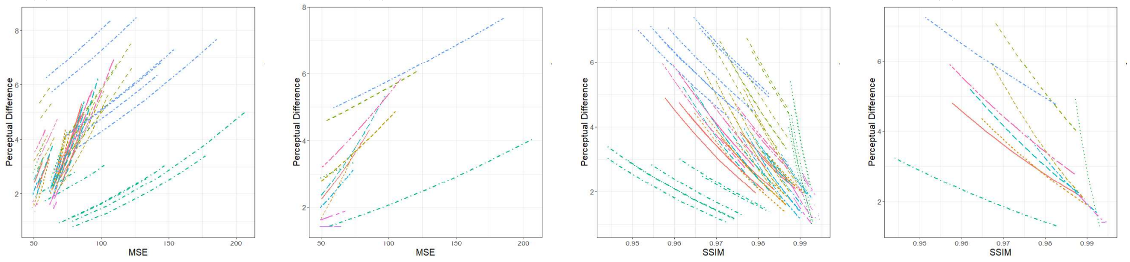
Table 6: ANOVA interactions with dependent variable "Difference" and within factors SSIM, Sex, Race and AU.

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
MSE	1	4736.38	4736.3879475	2620.34	0.00
AU	11	8523.97	774.9068976	428.71	0.00
Race	2	599.75	299.8756275	165.90	0.00
Sex	1	33.00	0.3300257	0.18	0.67
MSE:AU	11	2162.63	196.6032101	108.77	0.00
MSE:Race	2	133.16	66.5801372	36.83	0.00
AU:Race	22	1098.38	49.9265635	27.62	0.00
MSE:Sex	1	137.96	137.9690345	76.33	0.00
AU:Sex	11	126.47	11.4976458	6.36	0.00
Race:Sex	2	266.21	133.1057120	73.64	0.00
MSE:AU:Race	22	547.68	24.8945964	13.77	0.00
MSE:AU:Sex	11	136.11	12.3736730	6.85	0.00
MSE:Race:Sex	2	25.38	12.6901141	7.02	0.00
AU:Race:Sex	22	123.35	5.6070706	3.10	0.00
MSE:AU:Race:Sex	22	232.52	10.5693524	5.85	0.00
Residuals	7056	12754.04	1.8075454	NA	NA

Table 7: ANOVA interactions with dependent variable "Difference" and within factors MSE, Sex, Race and AU.



(a) RMS*AU+Race:Sex (b) RMS*AU (c) STED*AU+Race:Sex (d) STED*AU



(e) MSE*AU*Race*Sex (f) MSE*AU (g) SSIM*AU*Race*Sex (h) SSIM*AU

Figure 3: Model-fit for perceived difference using geometry metrics RMS (a-b), STED (c-d), and image metrics MSE (e-f) and SSIM (g-h) as per models listed in Table 3. The 6 virtual characters behave in a similar fashion when using STED (c) and are well captured with the simpler model (d) corresponding to the average model fit across the 6 virtual characters for each AU.