Contents lists available at ScienceDirect

Computers & Graphics

journal homepage: www.elsevier.com/locate/cag

Model for predicting perception of facial action unit activation using virtual humans

Rachel McDonnell[a,], Katja Zibrek[b], Emma Carrigan[a], Rozenn Dahyot[c]

[a] *Trinity College Dublin, Ireland*
[b] *Inria, Univ Rennes, CNRS, IRISA, France*
[c] *Maynooth University, Ireland*

ARTICLE INFO

Article history:
Received July 5, 2021

Keywords: Computers and Graphics, Formatting, Guidelines

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 (blendshapes) is costly in terms of memory and computation for gaming applications. Blendshape rigs are comprised of sets of semantically-meaningful expressions, which govern how expressive the character will be, often based on Action Units from the Facial Action Coding System (FACS). However, 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. Therefore, we believe that perception of blendshape visibility will not be reliably predicted by numerical calculations of the difference between the expression and the neutral mesh. 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.

1. Introduction

Virtual humans are becoming extremely popular in recent years for a range of diverse applications, such as video games, human-computer interfaces [1], live streaming, virtual reality entertainment, and personalized training [2]. With the increase in interactions with virtual humans comes the need for a greater understanding of how users perceive them, in particular their faces.

The perception of human faces and facial expressions is a much studied area in Psychology research. For virtual characters, facial expressions are generally created by animating blendshape rigs [3] based on FACS action units (AUs) [4], however these rigs are computationally expensive for real-time applications. 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 [5], or prioritising which blendshapes to include in expressions for example-based blendshape rig creation algorithms [6, 7], or to ensure facial expressions in rigs are being activated enough to be perceived clearly by the viewer. Additionally, algorithms that create or alter facial geometry are usually evaluated against ground-truth facial meshes using standard geometry error metrics [7], 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 nature of how facial perception it is a special form

e-mail: ramcdonn@tcd.ie (Rachel McDonnell), katja.zibrek@inria.fr (Katja Zibrek), rozenn.dahyot@adaptcentre.ie (Rozenn Dahyot)

of perception that humans are particularly attuned to [8, 9, 10], 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 the standard geometric and image error metrics. We also expect that due to social conditioning, sex and race will affect the perception of facial action units. It appears that female and male faces are observed differently, because the type and expressivity of particular emotions were found to be sex specific [11, 12, 13]. In addition, it has been shown that people perceive faces of their own race differently than other races in certain tasks such as facial recognition [14], so it is possible that perception of action units will differ across different race groups.

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. Geometric and image-based error metrics for triangle meshes are traditionally used for predicting mesh errors such as watermarking, simplification or lossy compression. However, we aim to determine if our question of perceived action unit importance can be predicted by simply calculating the *error* between the neutral pose and the expression blendshape, using common image and geometry error metrics. We investigate both standard and perceptually-based metrics, calculated from either the 3D geometry or the rendered 2D image, and perform linear regression analysis to determine if any of them can predict facial expression importance well, or if a new perceptual metric specific to facial expressions should be developed.

We address a number of questions, such as:

- Are certain facial action units more perceptually noticeable than others?
- Does a linear increase in activation of expressions (geometry alterations) result in a linear perceptual response for all action units equally?
- Are the same facial action units consistently noticeable across faces of different sex and race?
- Can we predict the saliency of facial action units using numerical error metrics, and is there a benefit to using existing perceptually based metrics?
- If metrics can predict saliency of facial action units, are 3D geometry metrics better than 2D image-based metrics?

Additionally, we tested several Generalised Linear Models in order to describe the relationship between our perceptual results and calculated errors. Our findings could be used for optimisation of blendshape rigs through blendshape reduction for facial animation in games. By identifying and removing blendshapes of lower visual saliency, we can save both memory and computation required. Additionally, our perceptual model could be used to guide real-time facial animation systems to ensure virtual agents are expressing perceptible expressions to a precise level (e.g., medium-level smile, etc.).

In this paper, we extend Carrigan et al. [15] with an online experiment with a more diverse pool of participants in terms of gender and race (Section 5) and a cross-validation test to assess how accurate our models are for prediction of unseen data (Section 8).

2. Related Work

Our interdisciplinary research relates to work in the areas of Psychology, Computer Vision and Computer Graphics, which we will discuss in this section.

Face perception is a very active area of study in **Psychology**, as humans have been shown to perceive faces in a different way to regular perception [8, 9, 10]. Work by Schwaninger et al. shows that faces are processed both in terms of their components as well as the configuration of those components [16, 17] rather than purely holistically.

As well, the different areas of the face have been shown to be important in terms of speech and emotion perception [18, 19]. 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 [20].

Another interesting property of face perception is that people perceive faces of their own race differently to faces of other races, with studies showing an own-race recognition memory advantage [21], as well as an own-race encoding advantage [22]. One explanation for this phenomenon is that people have more exposure to people of their own race, and there is evidence that experience can mitigate these other-group effects even if the experience is acquired during adulthood [23]. There is also a neurological basis for perceptual differences of faces based on both shape, pigment and internal features [24, 14]. Social conditioning appears to play a role in face perception of different sexes as well. There are sex differences in the readiness to express certain emotions - males tend to more readily express anger [11], while females more frequently express fear and sadness [12]. Therefore, a female expression of anger can actually be perceived as more intense than a male expressing the same intensity of anger, due to the violation of viewers' expectations [25, 13]. For these reasons, we include a diverse set of characters in our experiment, ranging in race and sex, to generalise our results.

In terms of perception of emotion, it has been shown that not all emotions are perceived equally. Happiness is most quickly recognised and least often confused with other emotions [26, 27, 28], while angry faces are more easily detected within a crowd [29]. For each emotional expression, specific parts of the expression appear to be more important for the classification of an emotion [30]. Since particular areas of the face are important for the recognition of emotion, 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 [31]. In addition, particular action units are responsible for the correct recognition of an emotion [32]: for happiness, this is the lip corner puller and parting of lips; for disgust, the most important are the raising and plucking of the lip. For fear, surprise, anger and sadness the regions around the eyes have the highest weights, with the lid raiser (exposing the sclera of the eyes) important for fear, and the lid tightener significantly most important for anger. Brows are important for sadness and both eyes and mouth contribute significantly to the recognition of surprise.

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 [33]. Reverse engineering expressions based on perceptual relevance helped with improved facial recognition in artificial faces [34]. 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 [35, 36] 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 [37], and can alter the perception of the eyes [38], however they are important in their own right for face recognition [39] and not just in relation to how they change the perception of eyes.

In the field of **Computer Vision**, the recognition of Action Units from FACS has been explored using facial component models, with AUs being recognised with greater than 95% accuracy [40]. Computer recognition of AUs is interesting to our work as it allows us to see the similarities and differences between human perception and computer vision. Most AUs were recognised correctly, with incorrect recognition being attributed to either an additional similar AU being recognised (e.g. both Inner and Outer Brow Raiser being recognised when only one was present), or a similar AU being incorrectly recognised (e.g. Jaw Drop being recognised instead of Lips Part). It is noted that one of the pairs of AUs that were confused, Cheek Raiser and Lid Tightener, are confused by humans as well [41]. Recognition of AUs, as well as automatic recognition of intensity of AUs, has also been accomplished using pure deep learning methods [42]. The relative importance of facial features for recognition of emotion has been investigated by Kumar et al. [43] who automatically recognized the six basic emotions viewed at several different angle using a multi-level classification model and only extracting features from relevant parts of the face and then separating facial expressions into three categories: lip, lip-eye, lip-eye-forehead. This method allowed for a recognition rate of 95.51%, outperforming state-of-the-art multi-view learning methods, showing the benefit of a segmented rather than holistic view of facial perception.

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 [3]. A blendshape is a mesh representing a certain shape, typically a simple movement like an eye blink or mouth open shape. Animation is achieved by linearly combining a number of these blendshapes with the neutral face to create an expression. 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 [44]. In theory, FACS breaks down facial expressions to their most basic components, making it a useful guideline for blendshape creation.

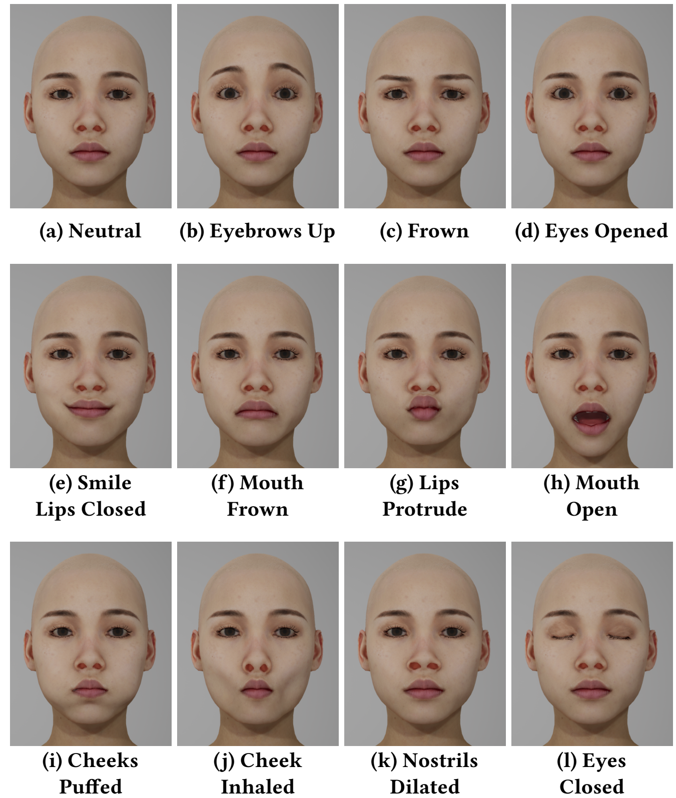


Fig. 1: The blendshapes set used in our experiment, shown on the Asian Female character at full activation (1.0).

Blendshapes can be costly to create, however, they can be transferred from a template rig containing the desired shapes to a target character rig using Deformation Transfer [45]. The quality and personalisation of these blendshapes can be improved by providing examples of the target character face [46]. Similar to the question of which blendshapes should be included in a rig, there is no consensus on which examples should be provided to best improve a rig. Initial perceptual research has been done in this area [6] as well as a first attempt at creating an example suggestion algorithm [7]. Another method for personalising rigs is to use an actor's performance to train an existing set of blendshapes to better match the actor's face [47].

Optimisation of blendshape animation can be done in a few ways. Reducing mesh complexity is one method [48], however this causes correspondence issues between shapes. The animation itself can be optimised by passing blendshapes [49] and animation [50] to the GPU, and using GPGPU methods [5]. 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. One area in particular where this optimisation method is applicable is optimisation for level of detail, where distance obscures the detail of the face so reduced quality is less perceptible.

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. [51]). We are interested if these metrics can be used in face-geometry perception.

Root-mean-square error (RMS) is a commonly used model-based error metric. Similarly, mean-squared-error (MSE) is used for image quality measurement. However these metrics are quite simplistic as they do not take into account the way in which a model is deformed, only measure the overall difference. This can lead to models with the same error but wildly different perceptual difference. To account for this, error metrics based on the human vision system have been proposed.

Of special interest to our work is the Structural Similarity Index Metric (SSIM) [52], which is a preferred image-based perceptual metric since it incorporates important perceptual phenomena such as contrast and luminance and also takes into account the structure of objects in the scene. Also of interest is the Spatio-temporal Edge Difference (STED) [53], which is a perceptual metric for meshes that works on edges as basic primitives as opposed to vertices. In our work, we investigate error metrics typically used for measuring mesh optimisation for the purpose of identifying importance of facial blendshapes, with the aim of reducing computation for facial animation in games.

3. Stimuli Creation

We explored acquiring a range of high-resolution full-head meshes with semantically-matching AUs and diversity of facial features from open-source databases. However, to our knowledge, no such set exists, therefore we created our own data-set.

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 [4] with additional shapes for emotion and speech. Our *experiment characters* were a set of 6 neutral faces (Fig. 2) created utilising high resolution scan data, from 3D Scan Store[2].

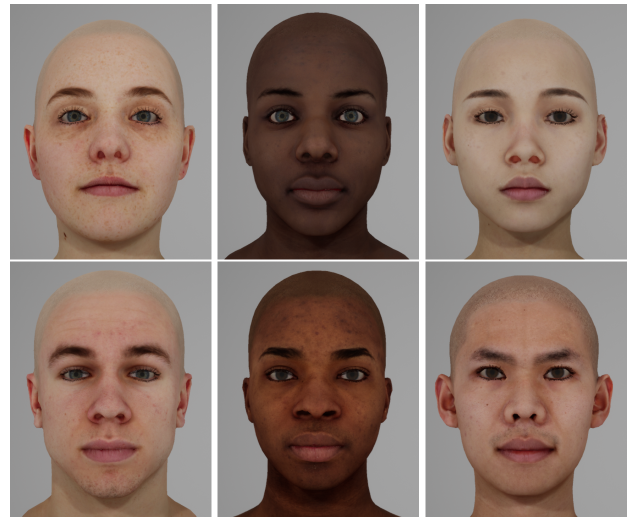


Fig. 2: Neutral faces of the characters used in our experiment. Left: white, middle: black, right: asian faces. Top row shows the female faces, while the bottom row shows the male faces.

One of the goals of this experiment was to obtain results that could be generalisable across different character faces, therefore we attempted to create a diverse set of stimuli by 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. These characters can be seen in Fig. 2. 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

In order to keep the experiment from having too many variables, we carefully chose 11 blendshapes from the character's set of 200 for the experiment (see Fig. 1). Since our work is aimed at character animation, we selected AUs that were particularly relevant for conversing virtual humans. AUs were chosen that were previously shown to be important for emotion (AUs 2, 4, 5, 12, 15, 26, 38 [32, 54]), speech (AUs 18, 26 [55], and those necessary for realistic and natural motion (AU 43 [56]). The cheeks have also been found to be important for facial recognition [57], 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.

[1] https://www.eisko.com/

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

[3] https://www.russian3dscanner.com/

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 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, e.g., 0.4 of the eyes closed expression would be eyes almost half closed.

4. Experiment 1: Laboratory

We chose to develop a real-time experiment system in Unreal Engine 4 for flexibility, and the fact that adjustments could be made easily to all characters without having to re-render a large set of images. Additionally, 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. The amount of time given to view each stimulus was not limited, although 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. The idea of the training session was 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 (including Neutral) x 5 activation levels × 6 characters. To avoid the experiment becoming too long, we used only one repetition of each character.

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 [21, 22], we asked the participants to disclose their race (5 Asian, 13 White, 0 Black, 2 Other). Due to the fact that this was an in-laboratory experiment, recent restrictions related to the COVID-19 pandemic meant that we were unable to recruit a larger or more diverse sample of participants. However, we address this shortcoming in our Online Experiment (Section 5).

4.2. Perceptual Experiment 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. In order to meet the assumptions for ANOVA, we analysed the data for sphericity violations and applied Greenhouse-Geisser corrections to the degrees of freedom (see Table 1). We also conducted the Kolmogorov–Smirnov analysis for the normality of residuals per each level of the factors and found that not all residuals were distributed normally, however, we assumed sufficient robustness of ANOVA for these violations. The ANOVA results can be seen in Table 1. We ran post-hoc analysis using Tukey's HSD tests throughout.

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

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).

4.2.1. Character Sex & Race

There was no main effect of the Sex of the character. There were some smaller interactions showing some individual differences in the models, but no interesting trends.

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 Activation Level, which showed the Frown and Cheeks Puffed ($p < 0.02$) were the main AUs affected. This implies that differences in the cheek and frown expressions were less perceptible on Black characters.

4.2.2. Activation Level

A main effect of Activation Level showed a significant increase in perceived differences as the activation increased, as expected.

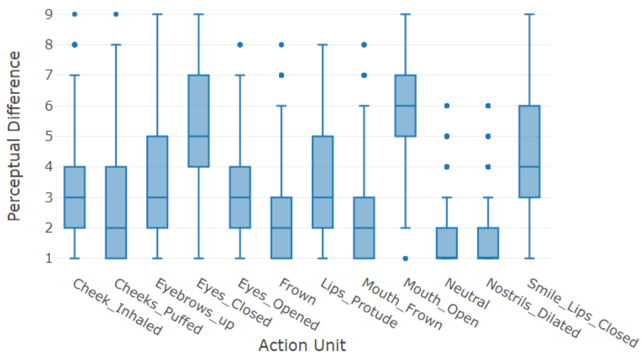


Fig. 3: Main effect of AU from our experiment.

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

Table 2: The AUs ordered by average perceptual difference.

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 Fig. 3 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) [13]. 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. Experiment 2: Online

Our online experiment was devised to investigate the effect of participant race on perception of model race (i.e., the other-race effect [21, 22]) with a larger and more diverse pool of participants.

We rendered out images of the stimuli and used an online form for presentation. To make the experiment shorter, we used only the full activation level (1.0) for AUs. One hundred and forty-four trials were shown to participants in random order, 12 AUs (including Neutral) × 6 characters x 2 sides (right, left).

For each trial of the experiment, we displayed the neutral expression side-by-side with the stimulus, and counterbalanced whether the stimulus was displayed on the left or right hand side. Participants were asked to answer "How different are the expressions?" by selecting a radio-button. The radio buttons ranged from 1 defined as "Not Different" to 5 defined as "Extremely Different".

5.1. Participants

In order to reject participants that were not concentrating on the experiment, we checked our data where 'No Difference' was not selected above a chosen threshold for the 12 trials where the neutral face was displayed on both the left and right.

After removal of 24 users that failed our attention test, 120 participants completed the experiment (40 White, 40 Black, 40 Asian, with 20 Male and 20 Females in each race group).

Since the experiment was conducted online, we did not have control of screensize so we included a question on the form for participants to report their monitor screen-size. 17 participants viewed the stimuli on a screen size of 8"-12", 63 on 13-17", 24 on 18-23", 15 on 24-26", and 1 on screen of 27" and above.

5.2. Results

In order to evaluate if smaller screen sizes made perceiving geometric differences more difficult, we first conducted an ANOVA with between factor Screen Size and within factor AU. The normality assumption for our data was tested using Shapiro-Wilk test and found that none of the residuals were normally distributed. Therefore, a non-parametric analysis with Aligned Rank Transformation (ART) was used, since it allows interaction effects to be analysed (unlike the non-parametric Friedman's test alternative) and does not require assumptions for ANOVA to be met. Post-hoc tests ($\alpha = .05$) with Tukey's adjustment were conducted to check significance for pairwise comparisons.

We did not find a main effect of Screen Size or an interaction with AU, confirming that the size of participants' screen did not affect their judgments.

5.2.1. Race

A mixed model non-parametric ANOVA was then conducted to determine if there was an interaction between participant race and character race, considering the within-group factors character AU, and Race and between-groups factor participant Race. A main effect of participant Race was found ($F(2, 117) = 10.17, p < 0.0001$), where White participants rated differences overall lower than Asian or Black participants ($p < 0.04$ in both cases). An interaction between participant Race and AU ($F(2, 4095) = 5.70, p = 0.000$) was found but a closer look at the post-hoc comparisons did not reveal many significant differences, except for White participants giving significantly lower ratings for the Neutral AU compared to Black and Asian participants ($p < 0.05$).

A main effect of character Race also occurred ($F(2, 4095) = 4.94, p = 0.008$), where differences shown on Asian characters were rated higher than differences shown on Black characters, as before. A main effect of AU ($F(2, 4095) = 773.09, p = 0.000$), and interactions between AU and character Race occurred ($F(22, 4095) = 11.63, p = 0.000$), which followed the same trends as before - differences were rated higher for Neutral AU and lower for Smile Lips Closed and Frown for White characters compared to other two races. Higher differences were found for Eyebrows Up, Eyes Open AUs and lower for Mouth Frown and Cheeks Puffed for Black characters compared to the same AUs of other races ($p < 0.05$, for all).

Importantly, we found no interactions between the participant Race and character Race, implying that an 'other-race' effect did not occur, and results on character race were consistent across participants.

5.2.2. Sex

A mixed model non-parametric ANOVA was conducted, considering the within-group factors AU, and character Sex and between-groups factor participant Sex. There was no main effect of participant Sex, or character Sex, or interaction between them. An interaction between participant Sex and AU ($F(11, 2714) = 9.72, p = 0.000$) showed that some AUs were perceived differently by male and female participants. Male participants perceived greater differences for Neutral, Mouth Open, and Eyes Closed, while female participants rated Mouth Frown higher ($p < 0.05$ for all). An interaction between AU and character Sex ($F(11, 2714) = 4.37, p = 0.000$) showed similar effects as in Experiment 1. For females, the differences were higher for Neutral and Mouth Frown AUs, while differences were higher for males compared to female characters for Smile Lips Closed ($p < 0.05$ for all). There was also a 3-way interaction between AU, participant Sex and character Sex ($F(11, 2714) = 1.85, p = 0.041$), where only one difference was found in post-hoc tests for the Neutral AU - male participants rated female characters higher than female participants ($p < 0.04$).

5.3. Discussion

Our online experiment confirmed our findings from the laboratory study on a larger sample size, and added the fact that our results are generally consistent across participants, regardless of sex or race. Since our laboratory experiment was conducted in a more controlled environment and tested more variables than the online experiment, we use this data for our subsequent model fit (Section 7).

6. Error Metrics

We investigate here the relationship between numerical error metrics and perception. We calculate each metric for each Activation Level of each AU, for each character. Each metric is calculated between the neutral face and the activated AU.

6.1. Geometric Error Metrics

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

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

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 perceptual data. For details and implementation, please refer to the paper by Vasa and Skala [53], and an overview by Corsini et al. [51].

6.2. Image Error Metrics

To calculate our image metric results, we took screenshots of each stimulus during the experiment and cropped out a large amount of the empty space surrounding each head. An example of the crop can be seen in Fig. 2. MSE and SSIM were calculated using scikit-image [58].

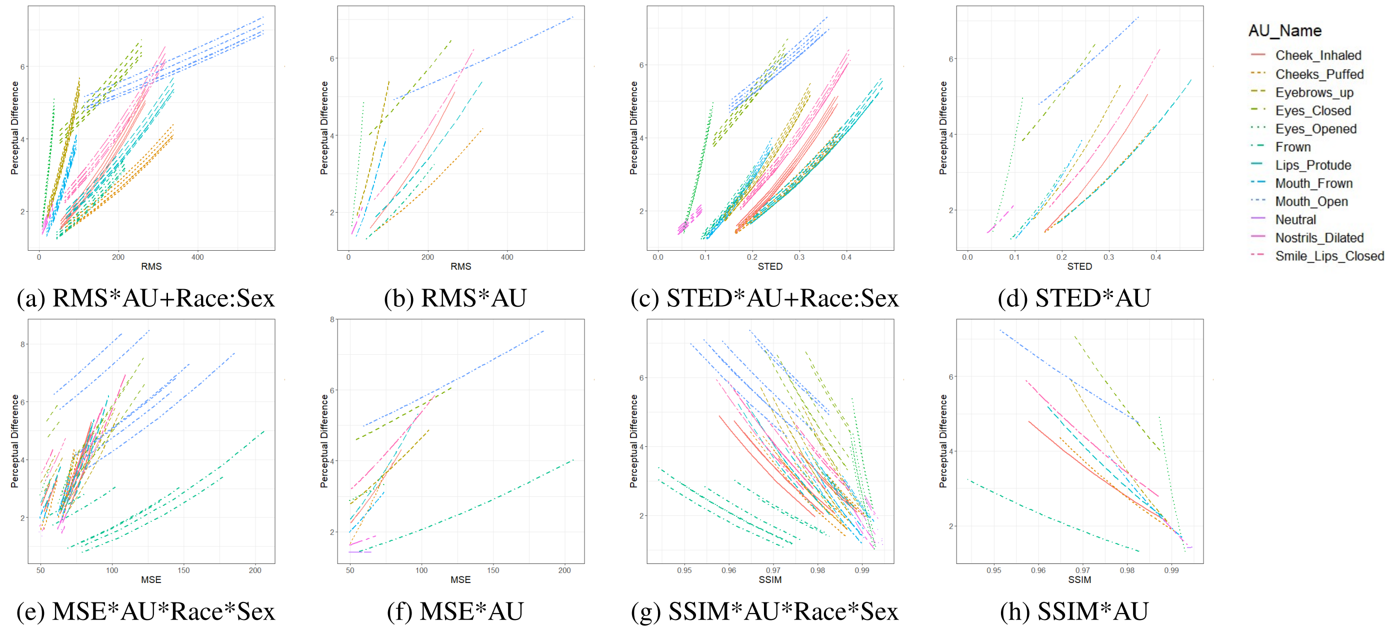
Mean-Squared-Error. We calculate MSE by getting the per-pixel average error between images A and B, where N is the total number of pixels in the image, and $\vec{x}_n^A$ is the n^{th} pixel of image A.

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

Structural Similarity Index Metric. SSIM is calculated as defined by Wang et al. [52] 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. SSIM measures similarity between 0 and 1 rather than dissimilarity: we invert this metric (i.e. 1-SSIM → SSIM) for better comparison with our other metrics where appropriate.

7. 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 [59]. 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 [59]. Poisson and Gaussian distributions were tested in combination with several link functions (identity, log, and



(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

Fig. 4: 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.

square root) [59]. 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.

	AIC ↓		is D ∈ [0; $\chi^2_{0.95}$] ?	
Model	Gaussian	Poisson	Deviance D	χ^2_{95}
Activ	29700 (Id)	28440 (Id)	7650	7396
Activ *AU	24880 (Id)	24690 (Id)	3852	7374
Activ*AU+Sex:Race	24860 (Id)	**24680** (Id)	3827	7368
Activ*AU*Sex*Race	24820 (Id)	24760 (Id)	3681	7252
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

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.

7.1. Variable selection with ANOVA

Tables 4, 5, 6 and 7 shows ANOVA results for the models `Metric*AU*Sex*Race` with `Metric` corresponding to RMS, STED, SSIM and MSE respectively (see also Appendix A). These tables show the importance of each variable and their interactions when fitting a model with Gaussian distribution and identity link function with perceived difference as the dependent variable (some of these models have their AICs reported in Table 3). As can be seen by the high values for Sum Sq., a large amount of the perceived difference is explained using the Metric and the blendshapes (AU) along with their interactions `Metric*AU`. These results imply the relationship between the perceived difference and the metrics (geometric or image based) are AU-specific, and using an AU-specific model is necessary for good prediction. 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 contributing variables. Note that when these two variables (e.g. terms `Sex:Race` or `Sex*Race`) appear in the fitted models, the models become character specific for our experiment (c.f. the 6 characters used shown Fig. 2 for which individual fitted lines appears and overlaps at times in Fig. 4 (a), (c), (e) and (g)).

7.2. Best Metric?

In the geometry domain, all fitted models are good models as per their deviance reported in Table 3 [59]. 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

	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	12758.19	1.80	NA	NA

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

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 [59]. 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.

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
MSE	1	4736.38	4736.387	2620.34	0.00
AU	11	8523.97	774.906	428.71	0.00
Race	2	599.75	299.875	165.90	0.00
Sex	1	33.00	0.330	0.18	0.67
MSE:AU	11	2162.63	196.603	108.77	0.00
MSE:Race	2	133.16	66.580	36.83	0.00
AU:Race	22	1098.38	49.926	27.62	0.00
MSE:Sex	1	137.96	137.969	76.33	0.00
AU:Sex	11	126.47	11.497	6.36	0.00
Race:Sex	2	266.21	133.105	73.64	0.00
MSE:AU:Race	22	547.68	24.894	13.77	0.00
MSE:AU:Sex	11	136.11	12.373	6.85	0.00
MSE:Race:Sex	2	25.38	12.690	7.02	0.00
AU:Race:Sex	22	123.35	5.607	3.10	0.00
MSE:AU:Race:Sex	22	232.52	10.569	5.85	0.00
Residuals	7056	12754.04	1.807	NA	NA

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

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).

8. Model Prediction

One application of our models can be to predict the viewer's perceived difference for a given character's deformation (as measured by geometric or image metrics) from its neutral pose. We note y an actual perceptual difference (data point) and $\hat{y}$ its prediction by one of our models. Prediction errors are computed with formula error= $\hat{y} - y$ for each N data point and these are expected to be centered on 0. The RMSE= $\sqrt{\frac{\sum_i error_i^2}{N}}$ is a global score that we use here for assessing our models.

8.1. RMSE & Cross validation

A K-fold cross-validation test (K=10) was conducted to assess how accurate the models are for prediction on unseen data. We report in Table 8 RMSE values with this cross validation strategy (RMSE.CV) as well as the RMSE of the model when fitted to the whole data (Column RMSE) as a baseline. Table 8 shows that the predictive precision are about identical for models using geometric metrics (STED or RMS) in combination or not with factors Sex and Race. On the other hand, models using image metrics (MSE or SSIM) perform better with these additional factors that help to compensate for the image metrics lacks of explanatory power in the models. We note that the models Metrics*AU*Sex*Race fitted with all the data slightly

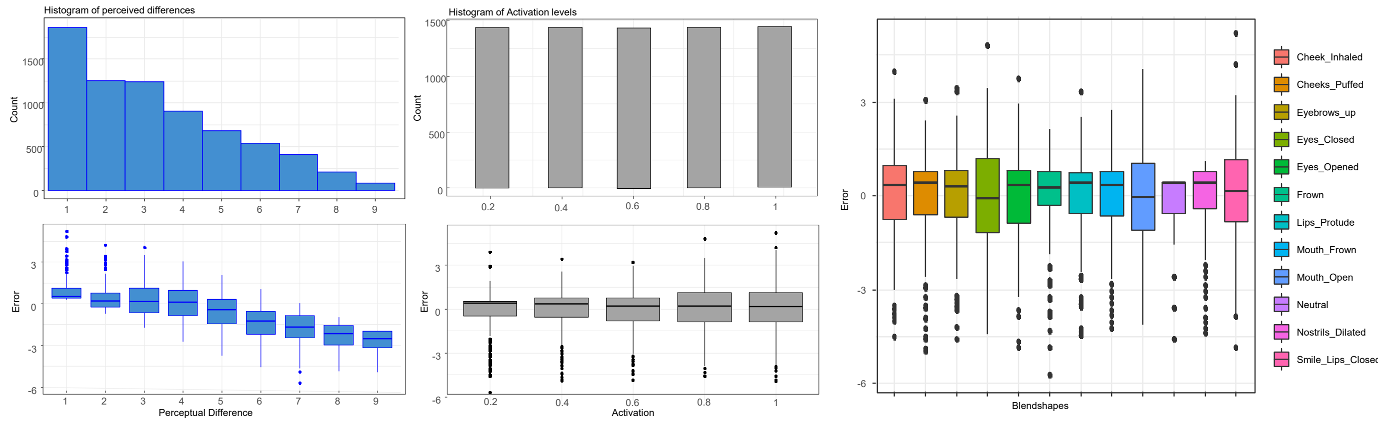


Fig. 5: For model RMS*AU, boxplots of errors= $\hat{y} - y$ are represented from left to right w.r.t. Perceptual difference y, Activation and blendshapes. Histogram of perceived differences from all collected responses from participants is also shown (top left). Histogram of collected responses per Activation level is also shown (top middle) for comparison and this flat distribution is also observed when counting responses w.r.t. AU (as per our experiment design explained in Sec. 3.3 and 4).

over-fit (i.e. RMSE.CV is systematically higher than RMSE by about 0.03).

Model	RMSE	RMSE.CV
STED*AU	1.355	1.359
STED*AU+Race:Sex	1.353	1.358
STED*AU*Sex*Race	1.328	1.354
RMS*AU	1.358	1.362
RMS*AU+Race:Sex	1.355	1.359
RMS*AU*Sex*Race	1.330	1.358
SSIM*AU	1.497	1.501
SSIM*AU+Race:Sex	1.425	1.431
SSIM*AU*Sex*Race	1.328	1.356
MSE*AU	1.566	1.571
MSE*AU+Race:Sex	1.497	1.503
MSE*AU*Sex*Race	1.332	1.361

Table 8: Comparing RMSE (full dataset) and K-fold cross-validation prediction error (measured with RMSE.CV averaged over 5 replications reported with standard error of less than 10^{-3})).

8.2. *Error analysis*

We analyse these prediction errors in more detail to check their distribution. As a representative result, Fig. 5 shows the box plots of these prediction errors for the model RMS*AU for each level of perceived difference, Activation level, and for each blendshape. We note that boxplots have median value close to 0 for these errors when shown w.r.t. Activation and blendshapes. Fig. 5 (left) shows the box plots of these prediction errors for each perceived difference level as reported by participants (x-axis). In this case, we note that for low level of perceived difference at 1, the model provides a slightly systematic over-estimated prediction ($\hat{y} > y$). On the other hand, for high levels of perceived difference between level 5 to 9, the model provides a under-estimated prediction ($\hat{y} < y$). Participants are not using evenly the Likert Scale for rating their perceived difference (cf. histogram in Fig. 5 (top left)) and 82% of collected perceived difference data is in fact on the levels 1 to 5. Our models provide mainly good performance for reported differences on levels 1 to 5 where most of the data is.

9. 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 for perceptibility of facial action units which we demonstrated through cross-validation could predict perceptual results from unseen data. Our model will provide a starting point for the development of a universal perceptual error metric suitable for human faces. Our GitHub repository[4] is provided (data and models in R-code), allowing others to build on our data investigating a larger range of faces, viewpoints, and facial action units.

Our other contribution is the results of our experiments which answer our questions from before. Firstly, 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 [5], or example creation for blendshape transfer [7]. By removing blendshapes of lower saliency, game developers can reduce memory usage and computation time.

We noted that diversity is missing from much of the psychology and computer vision research on recognition and perception of faces. Therefore, we included Asian, Black, and White characters with various skin tones to determine if our model could generalize across characters with different appearances. In general, there were no large differences at a per-Race or per-Sex level, implying that our results were generally consistent across characters. However, we did find an effect of Race (see Section 4.2.1), which showed that certain expressions were less perceptible on our Black characters. We felt that this result may

[4]https://roznn.github.io/facial-blendshapes/

have been due to our predominantly European and Asian participant pool in the Laboratory experiment, indicating that differences in perception of Black characters could be caused by the other-race effect [21, 22]. However, we tested a more diverse participant pool in our Online Experiment, which showed that the result was not due to the other-race effect.

We also hypothesized that male and female faces would be observed differently, but did not find much evidence for this, except that the Mouth Frown AU was more noticed on the female than on the male faces in our laboratory experiment. We believe this could be related to the inverse effect of gender stereotyping increasing saliency of unexpected emotions, in this case the Mouth Frown could have been perceived as anger. Our online experiment confirmed this effect and additionally found that male smiles (associated with happiness) were rated as more salient than female smiles, which is consistent with previous work by Hess et al. [13]. Interestingly, this was not affected by the sex of the participant.

With regard to activation level, we found an equally-spaced linear relationship between perceptual difference and activation level for most AUs. Additionally, we found that almost all AUs were not perceptibly different from the Neutral at our lowest activation level (0.2), with the only exceptions being Eyes Closed and Mouth Open, which were the AUs with the highest perceived difference overall. However, there were some AUs that remained imperceptibly different from Neutral at higher activation levels. For example, Cheeks Puffed and Mouth Frown only became significantly different at 0.6 activation, Frown at 0.8, and Nostrils Dilated at 1.0.

None of our image or geometric metrics used alone provided us with good statistical models. On the other hand, the perceived difference is well explained by metrics for each AU taken independently (as seen with the different slopes in Fig. 4).

Lower AICs have been measured using more complex GLM models (not reported here) using two metrics in combination with AU and we believe that non linear models such as neural networks may be able to learn more informative metrics computed directly from vertices or pixels for predicting the perceived difference more accurately (e.g. for removing the bias of predictive errors shown in Fig. 5).

Image metrics were shown to be worse at predicting perceived differences than geometry metrics, even though the viewers only viewed the 3D geometry from a single viewpoint (i.e., they were not allowed to interact with the geometry). 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.

Additionally, we found that eye AUs (Eyes Closed and Eyes Opened) were rated high in terms of perceptual difference (Table 2) despite their low error metric values, showing that humans are relatively more sensitive to eye expressions than other areas of the face. Additionally, Frown was one of the least perceptually different AUs, however it had medium-level geometric error values compared to other AUs, and had either the highest or second-highest error using image-based metrics. These results further highlight the need for a perceptually AU-based error metric for describing facial geometry alterations.

10. Limitations and Future Work

In this paper, we limited our study to static expressions of individual AUs to avoid confounds and to establish baseline models. However, it must be noted that perception of animated faces with combined expressions is more complicated, particularly since specific AUs are important for the perception of emotion (e.g., AU 7 Lid Tightener for anger [32]). It is possible that activation of AUs that are considered unimportant according to our model, could be extremely important for the interpretation of emotion of a virtual human, which we will study in future work. Additionally, we plan to broaden our investigation to the full range of AUs from FACS in future work.

We used only two characters to represent White, Black, and Asian races, for the purposes of creating material variation in the character models. We found some small effects of character race, however, more character models would be needed to generalize our results. Similarly, while we found few differences across our sample of female and male Black, White and Asian participants, it is possible that other factors might affect results such as participant age, etc.

In the future, our perceptual experiment could be replicated and new models fit for individuals that have more difficulty perceiving facial expressions than the general population (e.g., those with Autism Spectrum Disorder [60]). Results would allow us to create custom virtual agent systems that can increase or decrease blendshape activation levels to ensure clear perception of action units.

Acknowledgments

This research was funded by Science Foundation Ireland under the ADAPT Centre for Digital Content Technology (Grant 13/RC/2016) and the Game Face project (Grant 13/CDA/2135).

References

[1] Swartout, W, Traum, D, Artstein, R, Noren, D, Debevec, P, Bronnenkant, K, et al. Ada and grace: Toward realistic and engaging virtual museum guides. In: International Conference on Intelligent Virtual Agents. Springer; 2010, p. 286–300.

[2] Hubal, RC, Kizakevich, PN, Guinn, CI, Merino, KD, West, SL. The virtual standardized patient. In: Medicine Meets Virtual Reality. 2000, p. 133–138.

[3] Lewis, JP, Anjyo, K, Rhee, T, Zhang, M, Pighin, FH, Deng, Z. Practice and theory of blendshape facial models. Eurographics (State of the Art Reports) 2014;1(8):2.

[4] Ekman, P, Friesen, WV. Facial Action Coding System: A Technique for the Measurement of Facial Movement. Consulting Psychologists Press; 1978.

[5] Costigan, T, Gerdelan, A, Carrigan, E, McDonnell, R. Improving blendshape performance for crowds with gpu and gpgpu techniques. In: Proceedings of the 9th International Conference on Motion in Games. ACM; 2016, p. 73–78.

[6] Carrigan, E, Hoyet, L, McDonnell, R, Avril, Q. A preliminary investigation into the impact of training for example-based facial blendshape creation. 2018,.

[7] Carrigan, E, Zell, E, McDonnell, R. 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; 2020, p. 219–233.

[8] Bruce, V, Young, A. Face perception. Psychology Press; 2013.

[9] Farah, MJ, Wilson, KD, Drain, M, Tanaka, JN. What is “special” about face perception? Psychological review 1998;105(3):482.

[10] Kanwisher, N, McDermott, J, Chun, MM. The fusiform face area: a module in human extrastriate cortex specialized for face perception. Journal of neuroscience 1997;17(11):4302–4311.

[11] Plant, EA, Hyde, JS, Keltner, D, Devine, PG. The gender stereotyping of emotions. Psychology of Women Quarterly 2000;24(1):81–92.

[12] Fischer, AH, Rodriguez Mosquera, PM, Van Vianen, AE, Manstead, AS. Gender and culture differences in emotion. Emotion 2004;4(1):87.

[13] Hess, U, Adams Jr, RB, Kleck, RE. Facial appearance, gender, and emotion expression. Emotion 2004;4(4):378.

[14] Wong, HK, Stephen, ID, Keeble, DRT. The own-race bias for face recognition in a multiracial society. Frontiers in Psychology 2020;11:208. URL: `https://www.frontiersin.org/article/10.3389/fpsyg.2020.00208`. doi:`10.3389/fpsyg.2020.00208`.

[15] Carrigan, E, Zibrek, K, Dahyot, R, McDonnell, R. Investigating perceptually based models to predict importance of facial blendshapes. MIG ’20; New York, NY, USA: Association for Computing Machinery. ISBN 9781450381710; 2020,URL: `https://doi.org/10.1145/3424636.3426904`. doi:`10.1145/3424636.3426904`.

[16] Schwaninger, A, Wallraven, C, Cunningham, DW, Chiller-Glaus, SD. Processing of facial identity and expression: a psychophysical, physiological, and computational perspective. Progress in Brain Research 2006;156:321–343.

[17] Schwaninger, A, Lobmaier, JS, Wallraven, C, Collishaw, S. Two routes to face perception: evidence from psychophysics and computational modeling. Cognitive science 2009;33(8):1413–1440.

[18] Bruce, V, Young, A. Understanding face recognition. British journal of psychology 1986;77(3):305–327.

[19] Adolphs, R. Perception and emotion: How we recognize facial expressions. Current directions in psychological science 2006;15(5):222–226.

[20] Oruc, I, Balas, B, Landy, MS. Face perception: A brief journey through recent discoveries and current directions. Vision Research 2019;157:1–9.

[21] Lindsay, DS, Jack, PC, Christian, MA. Other-race face perception. Journal of applied psychology 1991;76(4):587.

[22] Walker, PM, Tanaka, JW. An encoding advantage for own-race versus other-race faces. Perception 2003;32(9):1117–1125.

[23] Cassia, VM, Picozzi, M, Kuefner, D, Casati, M. Short article: Why mix-ups don’t happen in the nursery: Evidence for an experience-based interpretation of the other-age effect. Quarterly Journal of Experimental Psychology 2009;62(6):1099–1107.

[24] Balas, B, Nelson, CA. The role of face shape and pigmentation in other-race face perception: An electrophysiological study. Neuropsychologia 2010;48(2):498–506.

[25] Hess, U, Blairy, S, Kleck, RE. The intensity of emotional facial expressions and decoding accuracy. Journal of Nonverbal Behavior 1997;21(4):241–257.

[26] Calvo, MG, Nummenmaa, L. Detection of emotional faces: salient physical features guide effective visual search. Journal of Experimental Psychology: General 2008;137(3):471.

[27] Palermo, R, Coltheart, M. Photographs of facial expression: Accuracy, response times, and ratings of intensity. Behavior Research Methods, Instruments, & Computers 2004;36(4):634–638.

[28] Leppänen, JM, Hietanen, JK. Positive facial expressions are recognized faster than negative facial expressions, but why? Psychological research 2004;69(1-2):22–29.

[29] Ceccarini, F, Caudek, C. Anger superiority effect: The importance of dynamic emotional facial expressions. Visual Cognition 2013;21(4):498–540.

[30] Smith, ML, Cottrell, GW, Gosselin, F, Schyns, PG. Transmitting and decoding facial expressions. Psychological science 2005;16(3):184–189.

[31] Srinivasan, R, Golomb, JD, Martinez, AM. A neural basis of facial action recognition in humans. Journal of Neuroscience 2016;36(16):4434–4442.

[32] Wegrzyn, M, Vogt, M, Kireclioglu, B, Schneider, J, Kissler, J. Mapping the emotional face. how individual face parts contribute to successful emotion recognition. PloS one 2017;12(5).

[33] Yu, H, Garrod, OG, Schyns, PG. Perception-driven facial expression synthesis. Computers & Graphics 2012;36(3):152–162.

[34] Chen, C, Garrod, OG, Zhan, J, Beskow, J, Schyns, PG, Jack, RE. Reverse engineering psychologically valid facial expressions of emotion into social robots. In: Int. Conf. on Automatic Face & Gesture Recognition. 2018, p. 448–452.

[35] Nusseck, M, Cunningham, DW, Wallraven, C, Bülthoff, HH. The contribution of different facial regions to the recognition of conversational expressions. Journal of vision 2008;8(8):1–1.

[36] Eisenbarth, H, Alpers, GW. Happy mouth and sad eyes: scanning emotional facial expressions. Emotion 2011;11(4):860.

[37] Ekman, P. About brows: Emotional and conversational signals. Human Ethology 1979;.

[38] Matsushita, S, Morikawa, K, Mitsuzane, S, Yamanami, H. Eye shape illusions induced by eyebrow positions. Perception 2015;44(5):529–540.

[39] Sadr, J, Jarudi, I, Sinha, P. The role of eyebrows in face recognition. Perception 2003;32(3):285–293.

[40] Tian, YI, Kanade, T, Cohn, JF. Recognizing action units for facial expression analysis. IEEE Transactions on pattern analysis and machine intelligence 2001;23(2):97–115.

[41] Cohn, JF, Zlochower, AJ, Lien, J, Kanade, T. Automated face analysis by feature point tracking has high concurrent validity with manual facs coding. Psychophysiology 1999;36(1):35–43.

[42] Shao, Z, Liu, Z, Cai, J, Wu, Y, Ma, L. Facial action unit detection using attention and relation learning. IEEE Transactions on Affective Computing 2019;.

[43] Kumar, S, Bhuyan, M, Iwahori, Y. Multi-level uncorrelated discriminative shared gaussian process for multi-view facial expression recognition. The Visual Computer 2020;:1–17.

[44] Ekman, P, Friesen, WV. Facial action coding system. Consulting Psychologists Press, Stanford University; 1978.

[45] Sumner, RW, Popović, J. Deformation transfer for triangle meshes. ACM Transactions on graphics (TOG) 2004;23(3):399–405.

[46] Li, H, Weise, T, Pauly, M. Example-based facial rigging. In: ACM transactions on graphics (TOG); vol. 29. ACM; 2010, p. 32.

[47] Ma, WC, Lamarre, M, Danvoye, E, Ma, C, Ko, M, von der Pahlen, J, et al. Semantically-aware blendshape rigs from facial performance measurements. In: SIGGRAPH ASIA 2016 Technical Briefs. 2016, p. 1–4.

[48] Garland, M, Heckbert, PS. Surface simplification using quadric error metrics. In: Proceedings of the 24th annual conference on Computer graphics and interactive techniques. 1997, p. 209–216.

[49] Lorach, T. Directx 10 blend shapes: Breaking the limits. GPU Gems 2007;3:53–67.

[50] Dudash, B. Skinned instancing. NVidia white paper 2007;.

[51] Corsini, M, Larabi, MC, Lavoué, G, Petřík, O, Váša, L, Wang, K. Perceptual metrics for static and dynamic triangle meshes. In: Computer Graphics Forum; vol. 32. Wiley Online Library; 2013, p. 101–125.

[52] Wang, Z, Bovik, AC, Sheikh, HR, Simoncelli, EP. Image quality assessment: from error visibility to structural similarity. IEEE transactions on image processing 2004;13(4):600–612.

[53] Vasa, L, Skala, V. A perception correlated comparison method for dynamic meshes. IEEE transactions on visualization and computer graphics 2010;17(2):220–230.

[54] Mortillaro, M, Mehu, M, Scherer, KR. Subtly different positive emotions can be distinguished by their facial expressions. Social Psychological and Personality Science 2011;2(3):262–271.

[55] Meng, Z, Han, S, Liu, P, Tong, Y. Improving speech related facial action unit recognition by audiovisual information fusion. IEEE transactions on cybernetics 2018;49(9):3293–3306.

[56] Itti, L, Dhavale, N, Pighin, F. 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; 2003, p. 64–78.

[57] Busso, C, Deng, Z, Yildirim, S, Bulut, M, Lee, CM, Kazemzadeh, A, et al. Analysis of emotion recognition using facial expressions, speech and multimodal information. In: Proceedings of the 6th international conference on Multimodal interfaces. 2004, p. 205–211.

[58] van der Walt, S, Schönberger, JL, Nunez-Iglesias, J, Boulogne, F, Warner, JD, Yager, N, et al. scikit-image: image processing in Python. PeerJ 2014;2:e453.

[59] Dobson, AJ, Barnett, AG. An Introduction to Generalized Linear Mod-

els. CRC Press, Third Edition; 2008.
[60] Kennedy, D, Cheng, B, Holcomb, C. Perception of emotions from facial expressions in high-functioning adults with autism. Neuropsychologia 2012;50(14):3313–3319.

Appendix A. Additional Analysis

ANOVA has been used as a preliminary analysis for selecting and understanding the role of the independent variables in our fitted models. Here, we show some additional analysis to further examine the ANOVA presented in the paper.

Table A.9 shows the results of the ANOVA analysis (Gaussian distribution with Identity link function): The dependent variable `Difference` is well explained (with significant level) using `Activation`, `AU`, `Race`, `Activation:AU`, the interaction `Activation:Race`, and to a lesser extent (cf. order of magnitude the `Sum Sq`) with interaction `AU:Race:Sex`. Note that this model for explaining dependent variable `Difference` is not the best suited (cf. AICs reported in the paper showing Poisson regression as performing best).

Table A.10 shows the results of the ANOVA analysis with Poisson regression model which is a better fit as reported in the paper (based on AIC). The dependent variable `Difference` is likewise well explained using `Activation`, `AU`, interaction `Activation:AU`, and `Race:Sex`, `AU:Race:Sex`.

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Activation	1	5.53e+03	5532.88	3069.03	0.00
AU	11	1.17e+04	1064.25	590.33	0.00
Race	2	2.49e+01	12.46	6.91	0.00
Sex	1	3.56e+00	3.56	1.97	0.16
Act.:AU	11	1.10e+03	100.29	55.63	0.00
Act.:Race	2	5.51e-01	0.27	0.15	0.86
AU:Race	22	1.53e+02	6.94	3.85	0.00
Act.:Sex	1	5.14e-01	0.51	0.28	0.59
AU:Sex	11	3.73e+01	3.39	1.88	0.04
Race:Sex	2	3.63e+01	18.15	10.07	0.00
Act.:AU:Race	22	6.40e+01	2.91	1.61	0.03
Act.:AU:Sex	11	1.22e+01	1.11	0.61	0.82
Act.:Race:Sex	2	4.54e+00	2.27	1.26	0.28
AU:Race:Sex	22	1.47e+02	6.68	3.70	0.00
Act.:AU:Race:Sex	22	5.74e+01	2.61	1.45	0.08
Residuals	7056	1.27e+04	1.80	NA	NA

Table A.9: ANOVA interactions with dependent variable "Difference" and within factors Activation, Sex, Race and AU.

	Df	Deviance	Res. Df	Res. Dev	F	Pr(>F)
NULL	NA	NA	7199	9388	NA	NA
Activation	1	1737.96	7198	7650.28	1737.96	0.00
AU	11	3227.60	7187	4422.68	293.42	0.00
Race	2	8.90	7185	4413.79	4.45	0.01
Sex	1	3.28	7184	4410.51	3.28	0.070
Act:AU	11	570.63	7173	3839.88	51.88	0.00
Act:Race	2	0.15	7171	3839.73	0.076	0.93
AU:Race	22	37.72	7149	3802.01	1.71	0.020
Act:Sex	1	0.01	7148	3802.00	0.0028	0.96
AU:Sex	11	10.01	7137	3791.995	0.91	0.53
Race:Sex	2	12.33	7135	3779.67	6.164	0.002
Act:AU:Race	22	26.54	7113	3753.13	1.21	0.23
Act:AU:Sex	11	4.33	7102	3748.799	0.39	0.96
Act:Race:Sex	2	2.87	7100	3745.93	1.44	0.24
AU:Race:Sex	22	43.89	7078	3702.047	1.995	0.0037
Act:AU:Race:Sex	22	21.04	7056	3681.008	0.96	0.518

Table A.10: Poisson ANOVA interactions with dependent variable "Difference" and within factors Activation, Sex, Race and AU.

Appendix B. Residuals

Fig.B.6 shows the QQplot for the Poisson model `RMS*AU`: KS (Kolmogorov–Smirnov test) fails indicating that simulated data from the model (i.e. predicted differences) does not have exactly the same distribution as the collected data (actual differences). Residual distributions shown as boxplots (Fig. 5) indicate that the model does not capture all deterministic patterns in the data: more complex models may provide a better fit.

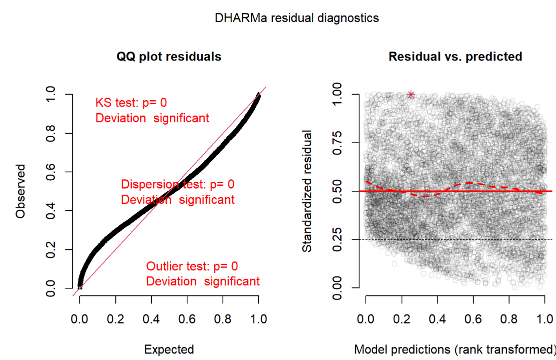


Fig. B.6: QQplot for model `RMS*AU`.