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

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

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

4. Results

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

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

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

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

EBFR worsened the results: 11, 13, 16

4.1. Excluded Results

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

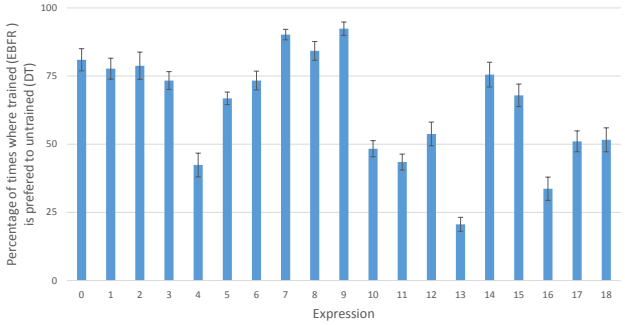

Figure 3: *Main effect of Expression on preference of EBFR over DT.*

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

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

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

(a) *Texture artifact example*

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

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

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

4.2. Analysis

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