Cipiloglu Yildiz
EURASIP Journal on Image and Video Processing (2023) 2023:8
https://doi.org/10.1186/s13640-023-00609-w

EURASIP Journal on Image and Video Processing

RESEARCH Open Access

Learning a crowd-powered perceptual distance metric for facial blendshapes

Zeynep Cipiloglu Yildiz[1*]

*Correspondence:
zeynep.cipiloglu@cbu.edu.tr

[1] Department of Computer Engineering, Manisa Celal Bayar University, Manisa, Turkey

Abstract

It is known that purely geometric distance metrics cannot reflect the human perception of facial expressions. A novel perceptually based distance metric designed for 3D facial blendshape models is proposed in this paper. To develop this metric, comparative evaluations of facial expressions were collected from a crowdsourcing experiment. Then, the weights of a distance metric, based on descriptive features of the models, were optimized to match the results with crowdsourced data, through a metric learning process. The method incorporates perceptual properties such as curvature and visual saliency. A formal analysis of the results proves the high correlation between the metric output and human perception. The effectiveness and success of the proposed metric were also compared to other distance alternatives. The proposed metric will enable intelligent processing of 3D facial blendshapes data in several ways. It will be possible to generate perceptually valid clustering and visualization of 3D facial blendshapes. It will help reduce storage and computational requirements by removing redundant expressions that are perceptually identical from the overall dataset. It can also be used to assist novice animators while creating plausible and expressive facial animations.

Keywords: Blendshapes, Animation, Facial expressions, Visual perception, Crowdsourcing, Metric learning

1 Introduction

Generating realistic facial models and animation has always attracted the computer graphics and animation community due to the demand in video games and the film industry. Interested readers are referred to the survey papers [10, 12, 16] for a detailed investigation of facial animation techniques. Among a variety of facial modeling and animation methods the most common are *physics-based methods*, *motion capture*, and *blendshape model*.

Physics-based methods aim to simulate the mechanics of muscles, bones, tissues, etc. [8]. Such methods are powerful for producing realism, but they are computationally expensive on account of the complexity of the facial mechanics. In motion capture techniques, actors control the movement of the digital character through the capture devices and markers. These methods have evolved rapidly to make real-time high-fidelity capturing of facial performance possible [3, 5].

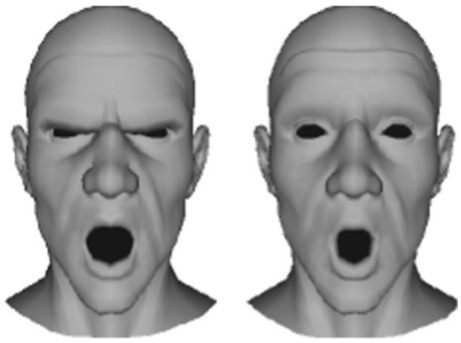


Fig. 1 Two facial blendshapes whose geometric distance is very small

Blendshapes also referred to as morph targets, are very popular as a simple linear model of facial expressions as well. [16]. In this model, a set of morph targets defines a linear space of facial expressions and blending these targets linearly with different weights enables the creation of new expressions. Although it is claimed that using the pre-made morph targets limits the animator's creativity, this property also allows even novice users to produce valid animations by disabling arbitrary deformations.

To improve the realism and expressivity of facial blendshape models, there is a need to quantify the similarity between two faces. The similarity between two facial expressions can be measured in different ways. One basic approach is to use a purely geometric measure such as Euclidian or Hausdorff distance. However, it is obvious that purely geometric measures are not sufficient to mimic perceptual similarity between facial expressions. Figure 1 exemplifies this situation. The Euclidian distance between these two faces is very small, but they perceptually correspond to very different expressions: angry and surprised.

It is possible to calculate the distance between two faces using the parametric weights of the morph targets if they are available. This will be more sensitive to perceptual differences; however, these weights may not be available at hand for a specific facial model. General-purpose visual quality metrics such as [25, 26] will not be sufficient either, since the perception of a face is different from arbitrary objects [4] and the point here is not only the quality. For that reason, a perceptual distance measure tailored for 3D facial models, which captures the semantic (i.e., emotional state) difference is required to quantify the facial expressivity in facial blendshape models.

The main goal of this paper is to develop such a perceptual distance metric for 3D facial blendshape models. The proposed method utilizes crowdsourcing and metric learning tools. The suggested metric is developed by learning the weights of a metric by maximizing the likelihood of obtaining human-evaluated distances between blendshapes. Experimental results demonstrate that the metric is well correlated with human perception. A list of the main contributions of the paper is given below:

- A novel perceptual distance metric which relies on crowdsourcing and metric learning techniques and is tailored specifically for 3D facial blendshape models has been proposed.
- A formal experimental analysis of the performance of the proposed metric compared to several distance alternatives has been performed.

– Possible applications and usage areas of the proposed metric have been described.

The rest of the paper is organized as follows: first, an overview of the related literature is provided in Sect. 2. In the third section, the data and methodology of this study are explained in detail. Quantitative results are analyzed and discussed in the fourth section. Several applications of the developed metric are suggested in Sect. 5, and the last section summarizes and wraps up the work.

2 Related work

In this section after giving a brief overview of the studies on general blendshape models, perceptual studies on facial expressions are elucidated in more detail.

2.1 Studies on blendshapes

There are several studies to improve the expressivity of blendshapes and alleviate their limitations. First of all, the direct manipulation blendshapes approach [2] solves a minimization problem to calculate the values of many degrees of freedom after the artist's editing of control points and allows intuitive control over the animation. Another study [22] aims to facilitate the tedious process of artistic sculpting of individual morph targets by automatically propagating the animator's edit on a specific expression to the rest of the sequence. Neumann et al. [19] propose a method to extract sparse and localized deformation modes from an input animation by means of a sparse matrix decomposition scheme, to enable localized artistic editing. Another branch of research on blendshapes is conducted for expression cloning, i.e., retargeting a source animation to a destination model [18, 21, 24]. In a recent study [7], a nonlinear semantic blending of parametric faces is enabled by deep face networks to enhance expressivity and creativity.

2.2 Perceptual studies on facial expressions

There are several attempts in the literature to incorporate perceptual mechanisms during the evaluation and synthesis of facial expressions. Wallraven and colleagues [23] evaluate the perceptual quality of animated facial expressions by means of a series of psychophysical experiments investigating the effect of different animation techniques and spatiotemporal features such as shape, texture, and motion. Their experimental results can give insight into the design of realistic-looking facial animations. Another study explores the effect of shape and material stylization on perceived realism, appeal, and expressivity of computer-generated faces [29]. Based on the psychophysical experiments conducted with varying levels of artistic stylization effects, they found that shape is the dominant factor when evaluating realism and expression intensity, while the material is more effective for appeal. Likewise, the influence of several factors on the perception of facial laughter expression is analyzed [20].

In their experimental study, Dobs et al. [11] intend to quantify human sensitivity to spatiotemporal information in dynamic facial expressions by inspecting from what cues observers benefit to judge the similarity of facial movements. They work on motion-captured facial animations and generate several approximations of the original animation. The observers' task is then to choose which of the given two approximated animations is more like the original. They define several objective measures for

stimulus similarity, based on frame accumulation of facial action activation distance, optic flow, and Gabor similarity. For each of these similarity measures, they calculate the probability of choosing one approximation over the other and perform a regression analysis to test if the calculated choice probabilities could predict the observers' behavior well. Their results show that facial action time courses reflect human behavior better than image-based measurements, revealing the importance of high-level cues in the processing of facial motion.

Deng and Ma [9] learn a statistical perceptual prediction model called FacePEM which measures the perceptual expressivity of facial motion sequences using Support Vector Machines. Yu et al. [28] aim at developing a perception-driven platform for synthesizing arbitrary valid facial expressions from existing ones. They seek a parametric relationship between the temporal morph weights of facial muscles and the perception of a specific expression. The morph weights are estimated from observer judgments using linear regression.

In a recent study, Carrigan et al. [6] explore the relationship between the human perception of facial blendshapes and their action units (AUs), based on controlled user experiments. They investigate the perceptually salient AUs and whether they can be predicted accurately by common 2D or 3D error metrics such as mean squared error, structural similarity index, etc. They put forward the need for a perceptually based error metric tailored for facial blendshape models.

In summary, studies on visual attention show that human faces are subject to high attention, which leads researchers to investigate ways of augmenting the realism and expressivity of computer-generated facial expressions. But perfect physical realism may not be required most of the time when human perception comes into play. The studies in the literature reveal the necessity for a perceptual distance metric for 3D facial blendshapes. This is one of the few studies that put forward such a metric customized for facial blendshapes and powered by crowdsourcing. The metric only requires 3D geometry, it does not even require AUs.

3 Materials and methods

In this section, data, algorithms, tools, and methodologies used in this study are explained in detail.

3.1 Background on blendshapes

The blendshape model is expressed as in Eq. 1 [16]. It is basically a weighted linear combination of the base expressions. Assume that N is the number of vertices and n is the number of morph targets. In this equation, b_0 corresponds to the neutral face model, b_ks are the individual blendshapes, and w_ks are the blendshape weights where $w_k \in [0, 1]$ and $\sum_{k=1}^{n} w_k = 1$. There are $3N$ coordinate values (x, y, z for each vertex) in each blendshape. f is the resulting facial model which has also $3N$ coordinate values. Figure 2 illustrates the idea with an example. Three morph targets are mixed with different weights to generate a new face model:

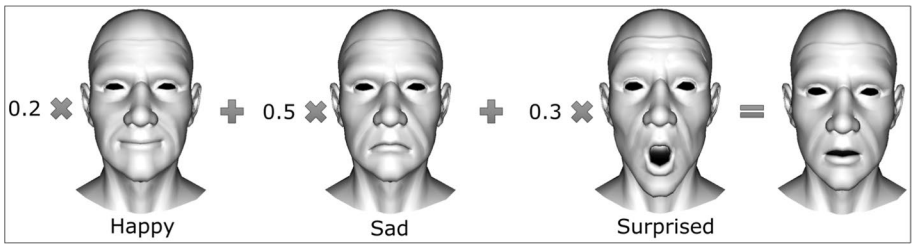


Fig. 2 Blendshape model illustration

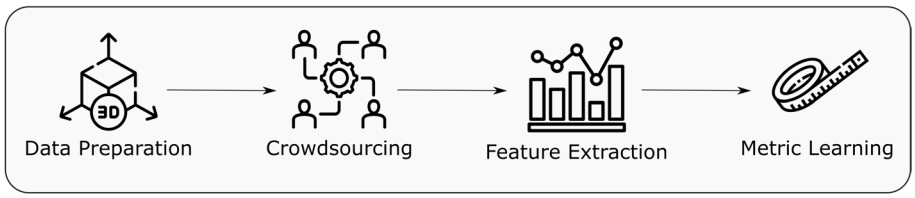


Fig. 3 Method overview

$$f = b_0 + \sum_{k=1}^{n} w_k (b_k - b_0). \tag{1}$$

3.2 Method overview

An overview of the methodology is illustrated in Fig. 3. First, facial expression blendshapes were obtained and processed for the user experiment. Secondly, a crowdsourcing experiment was conducted to collect data that represent the human perception of facial models. Different descriptive features were then extracted for each 3D facial model. Lastly, an optimization model was applied to learn a metric by identifying the mapping between the descriptive features of 3D models and human preferences obtained from a crowdsourcing experiment.

3.3 Data preparation

The original blendshape dataset is obtained from Turbosquid.[1] This dataset was chosen for experiments for two main reasons: (1) it contains essential emotional expressions in a clearly perceptible view, and (2) the target audience of the proposed method, novice animators, often prefer affordable and simple models. The original dataset contains 27 morph targets including emotional, eye blinking, and morphological expressions. In compliance with the purpose of this study, only basic emotional expressions were selected. The base facial expression morph targets are shown in Fig. 4. The neutral face was not used as a morph target, it is the base expression (b_0 in Eq. 1). Each 3D model has 2126 vertices and 4184 faces.

[1] https://www.turbosquid.com/3d-models/3d-model-male-head-morph-targets/261694.

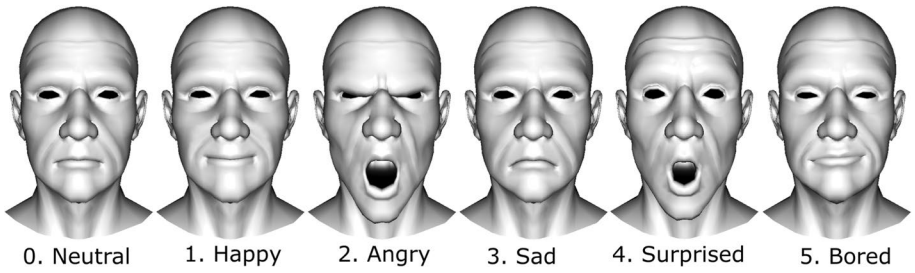


Fig. 4 Base expressions that are used in this study

Note that the original dataset also provides some skin textures; however, they are not used in this study. The reason behind this is twofold: first, skin textures, eye pupils, color, etc., may have a great impact on the perceived expression. This will complicate the problem and require an extensive user experiment. Furthermore, the modeller/animator may prefer to apply a different skin texture on the same model at different times. Hence, we only focus on the geometry of the models and develop a texture-independent metric in this study.

Five base expressions were combined in different weights to generate new facial expressions. Hence, $N = 2126$ and $n = 5$ in Eq. 1 for this study. The weight of each blendshape was changed between 0 and 1 with a step size of 0.1 in a uniform grid manner, without exceeding the limit of 1 for the sum of the weights. In this way, 877 new facial expressions were generated. However, the difference between many of these expressions was imperceptible and this number of expressions is quite much for a feasible and affordable user experiment. Therefore, a small and plausible subset of these 877 expressions should be selected. For this purpose, the models were sorted with respect to their Euclidian distances to the neutral model in ascending order and a uniform sampling based on these distances was done. This process produced 28 facial blendshape models which are given in the supplementary material along with their blendshape weights.

3.4 Crowdsourcing experiment

With the aim of developing a perceptual distance metric for facial blendshapes, we first need to collect user evaluations for the similarity of facial expressions. Crowdsourcing platforms provide a principled and effective means for this purpose. As a widespread crowdsourcing platform, Amazon Mechanical Turk (AMT)[2] was utilized in this study. In the AMT platform, simple user experiments are called *Human Intelligence Task (HIT)*.

Making comparative evaluations is known to be easier than providing absolute judgements for human observers [14]. Therefore, an experimental design shown in Fig. 5 was conducted in AMT. According to this design, the reference model was displayed in panel B and the task of the observer was to choose the facial expression that is closer to the reference face semantically. At the beginning of the HIT, some guidelines for the subject were listed. For the sake of computational efficiency, the interaction between the models (rotation, zooming, etc.) was not allowed since the facial expressions are better judged from the front view.

[2] https://www.mturk.com/mturk/welcome.

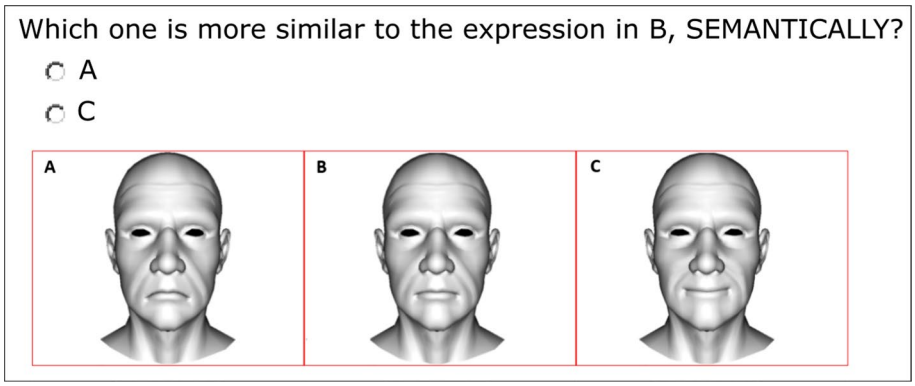


Fig. 5 Sample screen from the AMT experiment

As stated before, there are 28 models for comparison, which produces $28 * C(27, 2) = 9828$ query triplets (when each model is set as a reference, 2 of the remaining 27 models were chosen in 351 different ways). However, in some triplets, the Euclidian distance between the two models was very small that cannot be perceived. A threshold on the Euclidian distance was used to eliminate such kind of triplets that contain imperceptible differences. More specifically, if the Euclidian distance between any pair is less than 0.001 in a triplet, it was excluded from the dataset. This process reduced the number of query triplets to 2905.

Each HIT contained 12 questions (query triplets), two of which were control questions. Each query was evaluated by at least 20 users. $0.03 was paid per HIT and each HIT took approximately 4 min.

To ensure the reliability of the crowdsourced data, we should adopt some precautions [13]. Firstly, the subjects must answer all the questions in the qualification test correctly, to proceed to the actual test. Secondly, if a user's response time for a HIT is less than the standard deviation of the response times, that HIT is rejected. Furthermore, each HIT contains two control questions (in reversed order). If both control questions have inconsistent answers, the HIT is rejected again. Lastly, if the number of rejected HITs of a subject is three or more, that subject's responses are not included in the data since they are unreliable. 173 subjects participated in the experiments and about 3.2% of the HITs were rejected.

3.5 Feature extraction

The purpose of this step is to extract the features that can describe the semantics and perception of a 3D facial model. For this purpose, mean, variance, skewness, kurtosis, and median values of the common geometric attributes that are known to be related to human perception of 3D models are employed. The features and the procedure for extracting these features are explained below.

First, 3D models were centered at the origin and scaled to the unit bounding box, as a preprocessing step. Then, landmark vertices on a face were pinpointed, which highlight the important locations on a face. The modeller generally annotates these landmarks and they are used as control points in the animation. There are also different approaches for automatically determining facial landmarks. One heuristic approach is

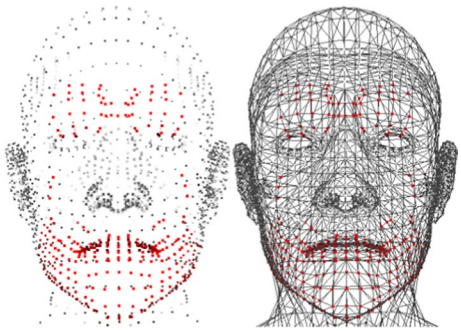


Fig. 6 Landmarks of the blendshape model (left: point cloud, right: wireframe)

to set a vertex as a landmark if its variance of displacement through the morph targets exceeds some threshold value.

All the per-vertex attributes described below were only calculated for the landmarks, not for all the vertices. Using all the vertices in computations is also possible but it is not recommended since it will increase the computational complexity and presumably detriment the success by incorporating non-relevant parts in the feature vector.

382 landmarks in total were identified according to the above heuristic. It can be replaced by any other state-of-the-art facial landmark extraction technique if needed. However, as shown in Fig. 6, this simple heuristic yields quite plausible landmarks. The extracted landmark points are located at the forehead, eyelids, eyebrows, mouth, chin, and cheeks and they correspond to the most dynamic parts of the face.

The following features of the landmarks were calculated and normalized to the [0–1] range. The final feature vector consists of 55 features:

- *Displacement.* Five descriptive statistics of x, y, and z displacements of the landmarks from the neutral face's corresponding landmarks were stacked in the feature vector (15 features).
- *Normal.* The descriptive statistics of the vertex normals in x, y, and z directions were also used in the feature vector, as the normals describe the surface geometry (15 features).
- *Curvature.* Surface curvature values are also known to be effective at describing the geometry of the model and their correlation to human perception is shown in the literature. *Minimum, maximum, mean,* and *Gaussian curvature* values were computed as described in [1] and Eq. 2. In this equation, T is the curvature tensor for each vertex, defined over a neighborhood of B, $\beta(e)$ is the signed angle between the normals of the faces incident to edge e, and $\bar{e}$ is a unit vector in the same direction of e. The eigenvalues of T are used as estimators for the minimum and maximum principle curvatures. Their five descriptive statistics were added to the feature vector too (20 features):

$$T(v) = \frac{1}{|B|} \sum_{edges\ e} \beta(e)\ |e \cap B|\ \bar{e}\ \bar{e}^t. \tag{2}$$

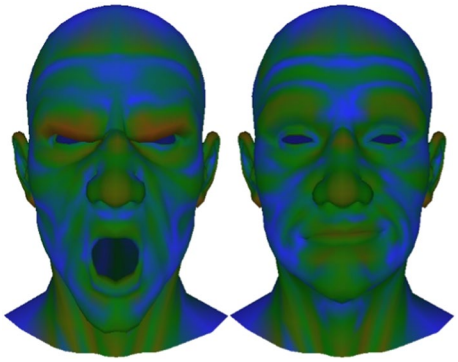


Fig. 7 Saliency maps for sample blendshapes (saliency increases from blue to red)

– *Saliency.* Saliency is a significant visual perception concept that emanates from low-level human visual system mechanisms. It is used to measure the visual importance of a region compared to neighboring regions. More salient regions attract more attention in general. Therefore, five descriptive statistics of the saliency values of the landmark points on the face were also included in the feature vector (5 features).

Mesh saliency was computed according to the classical center-surround operator on the Gaussian-weighted mean curvature as in Eq. 3 [15]. In the equation, $\mathcal{S}_i(v)$ denotes the saliency of vertex v at scale level i, σ_i is the standard deviation at scale level i, and $G(\mathcal{C}(v), \sigma_i)$ is the Gaussian-weighted average of the mean curvature. Five scales are used for σ_i as $(2\epsilon, 3\epsilon, 4\epsilon, 5\epsilon, 6\epsilon)$, where ϵ is 0.3% of the diagonal of the bounding box of the model. Then the saliency maps of different scales are combined using a nonlinear suppression operator:

$$\mathcal{S}_i(v) = |G(\mathcal{C}(v), \sigma_i) - G(\mathcal{C}(v), 2\sigma_i)|. \tag{3}$$

In Fig. 7, saliency values for each point on sample faces are visualized. The reddish regions indicate highly salient parts on the mesh. Eyebrows have the highest saliency on the angry face (left image) as expected, while the nose and lips seem more salient on the bored face (right image).

Blendshape weights used for blending the morph targets in Eq. 1 were not included in the feature vector since they may be unknown to the user for specific applications. However, the effect of adding those parameters to the feature vector was also analyzed in the results section.

3.6 Metric learning

At this step, the aim is to find a mapping between the user evaluation data, gathered through a crowdsourcing experiment, and extracted features of the blendshapes. For such kind of problems that include comparative evaluations, the metric learning approach is adopted in the literature [17, 27]. In the metric learning model built for this study, the goal is to minimize the objective function in Eq. 4 by *Maximum A Posteriori (MAP)* estimation:

$$-\sum_{T} log(P_{AC}^{B}) + \lambda||w||_1. \tag{4}$$

In this equation, the second term is used as L_1 regularization to obtain a sparse feature vector. T is the training set, λ is the regularization weight (empirically set to 0.1), and w is the diagonal of the weight matrix W. P_{AC}^{B} is the probability that the user selects A as more like B than C, given a query triplet of blendshapes $< A, B, C >$. This probability is formulated as a sigmoid function of the form:

$$P_{AC}^{B} = \frac{1}{1 + e^{D(B,A) - D(B,C)}}, \tag{5}$$

where $D(X, Y)$ is the perceptual distance between two facial blendshapes X and Y. It is calculated as the simple-weighted Euclidian distance between their feature vectors f_X and f_Y:

$$D(X, Y) = \sqrt{(f_X - f_Y)^T W (f_X - f_Y)}. \tag{6}$$

The weights in the diagonal of the W matrix are learned by minimizing the objective function in Eq. 4 which in turn maximizes the likelihood of observing the training data. This nonlinear optimization problem is solved by *Limited Memory Broyden–Fletcher–Goldfarb–Shanno (L-BFGS)* optimization in Matlab.

4 Results and discussion

In this section, quantitative results are presented along with their elaboration.

4.1 Optimization process

The data collected from the crowdsourcing experiment were processed in the following way: let tuple t be in the form of $< F_r, F_1, F_2, r >$, where F_r is the reference face model, and F_1 and F_2 are the face models that are compared, and r is the binary response variable determined as below. The majority vote of 20 observers for a query triplet determines the r value (label) of that query:

$$r = \begin{cases} 1, & \text{if } F_1 \text{ is selected by majority} \\ 0, & \text{if } F_2 \text{ is selected by majority} \end{cases}.$$

The dataset of tuples was first randomly divided into training (70%) and test (30%) splits. Then fivefold cross-validation was performed on the training set and the learned metric was tested on the test set to predict the r values of the tuples. The weights in Eq. 4 were initialized with random values. Feature extraction part (per model) took about 3 s and the optimization process took about 22 s on a 2.6-GHz PC. 19 of the 55 features have non-zero weights after convergence.

The highest weight features are displayed in Fig. 8. The weight of the skewness of the minimum curvature seems very high, compared to other features. Because of the high impact of the minimum curvature, its distribution for different expressions is also displayed in Fig. 9. The shape of the distribution varies according to the expression types and it is seen that the magnitudes of skewness values are higher for the angry and surprised faces in which the movement of the vertices of the mouth and eyes are high.

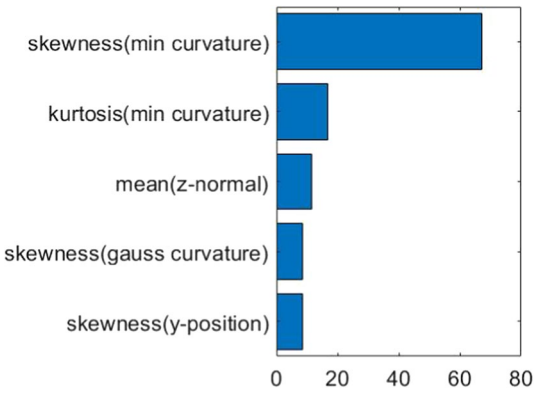


Fig. 8 Top-5 non-zero weights at the end of the optimization procedure

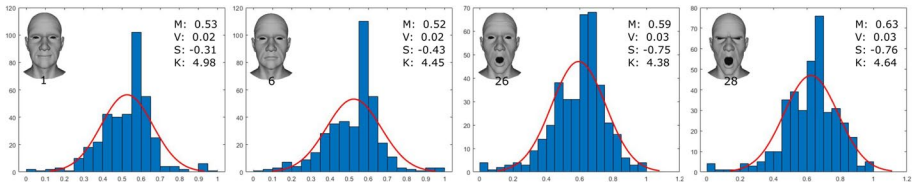


Fig. 9 Minimum curvature distribution for different expressions. (Red curve is the Gaussian function fitting. M: mean, V: variance, S: skewness, K: kurtosis.)

Table 1 Prediction accuracy (P.A.) on the test tuples, using different metrics as a decision-maker

Metric	P.A. (%)
Euclidian	78.38
Parametric	80.59
Dominant target	81.33
Uniform-weight	69.67
Perceptual	83.06
Perceptual + parametric	85.78

Therefore, it is reasonable that minimum curvature values are so effective at characterizing the expressions. Each attribute category (position, normal, different types of curvature, and saliency) contributes to the result somehow. Different combinations may produce similar results.

4.2 Performance evaluation

There are no publicly available distance metrics that can be used as benchmarks, designed for 3D facial blendshapes to our knowledge. Nevertheless, since the ultimate benchmark is the human perception for a perceptual metric, it is convenient to evaluate its performance in terms of correlation with the human-evaluated data. Still, the proposed metric should be compared to several common baseline methods. Thus, as the performance measure, prediction accuracy which is the ratio of correctly predicted tuples over all tuples was calculated on the test set, using different distance heuristics as a decision-maker. The results are presented in Table 1.

In the table, *Euclidian*, *parametric*, and *dominant target* distances were computed as baseline approaches. Euclidian distance is calculated as the mean Euclidian distance between the corresponding vertices of two models. Parametric distance is calculated based on the Euclidian distance between the weights of the morph targets used to generate respective models. Dominant target distance is developed based on the idea that the morph target with the highest weight would be effective on the perceived expression. For instance, if a blendshape is constructed as a mixture of <Happy, Angry, Sad, Surprised, Bored> morph targets with weights of $< 0, 0.8, 0.1, 0, 0.1 >$, one can expect that the final model will be similar to angry face model. This idea is employed as a heuristic decision-maker by choosing the model with the closest weight in the dimension where the reference model has the maximum weight. In other words, while calculating parametric distance, all the weights except the weight of the morph target for which the reference model has the maximum weight are assumed to be zero.

Perceptual distance is the proposed distance metric whose weights are learned through the described optimization procedure. *Uniform-weight* metric uses the same feature vector as the proposed distance metric without applying the learning procedure (the weights of the features are equal). In this study, we assume that the parametric weights of the morph targets are not available. Therefore, these parameters were not included in the feature descriptors. However, the effect of augmenting the feature vector described in Sect. 3.5 with these parametric weights, was also examined by computing a new distance metric which is named *perceptual+parametric* distance. To calculate this metric, in addition to the features in Sect. 3.5 five blending parameters were added to the feature vector. Then the same optimization procedure was applied to learn the weights of the distance metric.

4.3 Discussion

The results in Table 1 validate the foresight that purely geometric distance metrics (i.e., Euclidian) do not reflect human perception. It is also seen that the extracted features are not sufficiently descriptive without optimizing their weights, as the learning procedure dramatically improved the success of the proposed distance metric. Another remark from the results is that the dominant target heuristic correlates with human perception slightly better than parametric distance. The proposed metric performs better than parametric and dominant target distance metrics, while it does not require morph target weights to be known in advance. An important observation from the results is that augmenting the feature vector with blending parameters significantly improves the performance compared to both purely parametric and purely perceptual distances. This result suggests that even if the blending parameters are known, applying this technique better captures the human perception of facial blendshapes while measuring distance.

As another remark, the same optimization procedure was applied using the feature vector which is calculated over all the vertices (not just landmarks). This time the prediction accuracy could only reach up to 79%. This result implicates the importance of landmarks. Performing the feature extraction over the landmarks improves both the prediction performance and computational efficiency.

Overall, the proposed perceptual metric's performance is quite well. Note that for perceptual issues, it is unrealistic to expect full accuracy rates since human

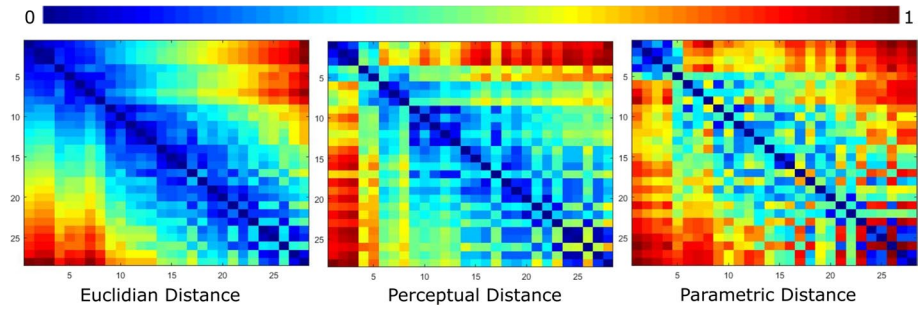


Fig. 10 Heat maps depicting the pairwise distances between 28 facial models used in the study, according to different distance metrics

Table 2 Correlation coefficients between different metrics

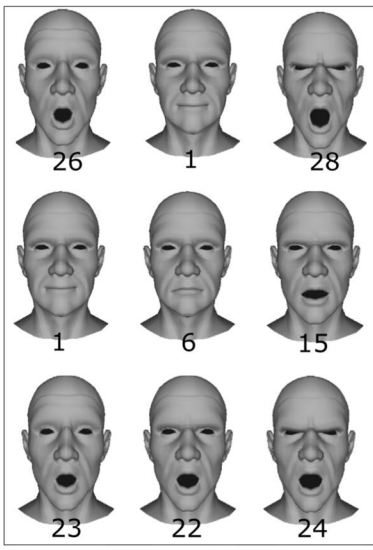
Metrics	PCC (%)	SROCC (%)
Euclidian vs. perceptual	77.15	76.11
Euclidian vs. parametric	73.47	73.07
Parametric vs. perceptual	82.25	81.72

perception varies from person to person. Therefore, the methodology here uses a heuristic method to model average human response.

Figure 10 visualizes the pairwise distances between 28 face models (see Additional file 1) used in this study, according to three distance metrics, in the form of a heat map. All the distance values were normalized to the range of [0,1] for each metric. The distance matrices are symmetric, and the diagonals are 0 as expected. The Euclidian distance matrix exhibits a regular pattern (the distance increases from model 1 through model 28) unsurprisingly. On the other hand, parametric and learned distances show a different distribution, revealing the lack of full correlation between geometric and perceived distances. In the perceptual distance heat map, an interesting pattern draws attention; some consecutive models (e.g., 1–3, 9–11, 14–16) construct a band of almost the same color, and then it suddenly jumps to another color. These bands may correspond to perceptual bands where the human visual system cannot perceive the difference.

The lack of correlation between geometric and perceptual distances was also quantified by calculating the pairwise *Pearson Correlation Coefficient (PCC)* and *Spearman Rank Order Correlation Coefficient (SROCC)* between the results of three metrics, which are listed in Table 2. According to these results, the proposed distance metric highly correlates with parametric distance, while the correlation between the proposed metric and Euclidian distance is higher than the correlation between parametric and Euclidian distances. One can deduce that human perception of facial expressions lies somewhere in-between geometric and parametric distance. This is also observable on heat maps.

In Fig. 11, several interesting cases which characterize the behavior of each metric are demonstrated. From the first sample, we see that the maximum distance is between the first (1) and last (28) expressions, regardless of the metric type. With



	$D_m(1,26)$	$D_m(1,28)$
m = Euclidian	0.84	1.00
m = Parametric	1.00	1.00
m = Perceptual	0.88	1.00

	$D_m(6,1)$	$D_m(6,15)$
m = Euclidian	0.16	0.40
m = Parametric	0.77	0.51
m = Perceptual	0.64	0.39

	$D_m(22,23)$	$D_m(22,24)$
m = Euclidian	0.17	0.17
m = Parametric	0.44	0.44
m = Perceptual	0.33	0.23

Fig. 11 Distance measurements for sample cases

respect to the parametric distance, expression 1 has the same distance as both expressions 26 and 28. But Euclidian and perceptual distances find expression 1 closer to expression 26 than 28. From the human perception perspective, this can be explained by the fact that expression 28 has a negative meaning (angry), whereas expression 26 has not that much negative meaning. The parametric distance cannot capture this nuance since it regards each parameter dimension with the same weight.

Another interesting case is shown in the second row of Fig. 11. In this case, expression 6 is geometrically closer to expression 1 than expression 15 since the mouth is closed. On the other hand, both parametric and perceptual distances find expression 6 closer to expression 15. This is perceptually plausible because expression 1 is a smiley face but expressions 6 and 15 give a negative impression emotionally. Perceptual distance gives moderate results compared to parametric distance.

In the third case of Fig. 11, expression 22 has the same distance to expressions 23 and 24, both geometrically and parametrically. However, perceptual distance detects it as slightly closer to expression 24. From the point of visual perception, we can explain it by the saliency of eyebrows. Although the shape of the mouth is more similar in expressions 22 and 23, eyebrows are salient regions in the face and more effective at the perception of the expression.

Limitations Besides the success of the proposed distance metric, there are several limitations. Firstly, it is a full-reference metric which gives the distance of a model with respect to a reference model. Another obvious restriction which was also stated in Sect. 3.3 is that the proposed distance metric does not incorporate color and texture properties. In addition, user experiments were conducted using a specific face model. Although a generalizability analysis is done as described in the next section, a small bias due to the specific model may exist in the results. Lastly, the metric requires a manual feature extraction process which is laborious.

Fig. 12 Sample morph targets from the validation dataset

4.4 Validation of the metric

4.4.1 Generalizability of the metric

With the intention of validating the generalizability of the metric, it was applied to a different dataset of facial blendshapes. The dataset was also purchased from Turbosquid. Sample morph targets from this dataset are displayed in Fig. 12. The same criteria, explained in Sect. 3.3, hold for choosing this dataset too, but a female character was chosen on purpose for the validation. These morph targets were blended to generate new models, as described in Sect. 3.3. Then, pairwise distances were calculated for 40 newly generated models, both using the proposed metric and parametric distance. SROCC between the learned and parametric distance was calculated as 83.48%, which is consistent with the findings in the previous section. This result verifies the generalizability of the proposed distance metric over different datasets.

4.4.2 Effectiveness of the metric on clustering performance

As another validation technique, the effectiveness of the proposed metric on clustering performance was measured in comparison to other alternatives. For this purpose, another crowdsourcing experiment was conducted. In this experiment, the subjects were provided four of the base expressions in Fig. 4 (happy, angry, surprised, and sad) as cluster seeds, and they were asked to assign 28 expressions (used throughout this study) into the most similar categories. They were informed that the number of faces in each category may be different. The labels of the categories were not written in the experiment. 50 subjects performed the experiment, and they were paid \$0.1 for this task. The average response time for this experiment was about 8 min. The majority response for each facial model was then used to determine the cluster of that model. The resulting clusters are given in Fig. 13. These clusters will be referred to as 'experimental clusters' henceforth.

On the other side, 28 models were embedded in the 2D plane, using the classical multi-dimensional scaling implementation in Matlab and simple k-means clustering with $k = 4$ was also applied on the same dataset using both parametric and proposed distance metrics. In Fig. 14, the embedding of 28 models in the 2D plane and generated clusters are displayed according to both metrics. The cluster distributions are also demonstrated with their details in Figs. 15 and 16.

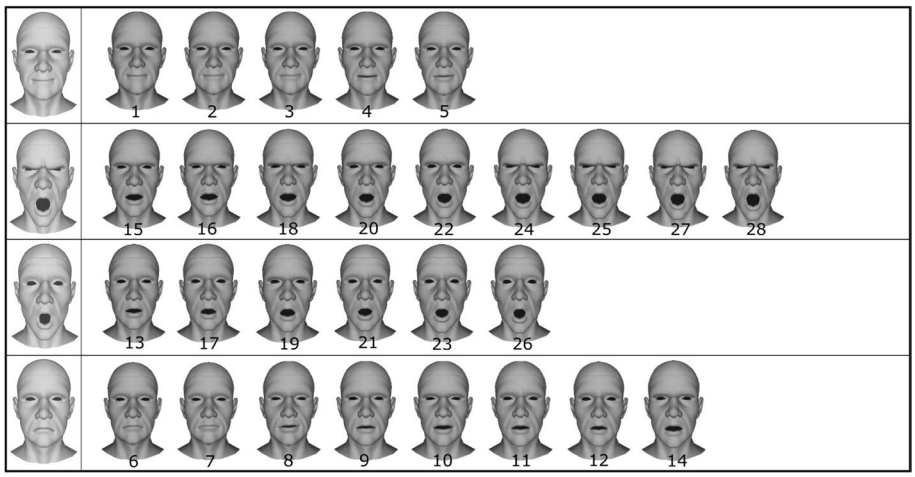


Fig. 13 Experimental clusters generated according to crowdsourcing experiment. The first column includes reference expressions for each cluster provided to the users

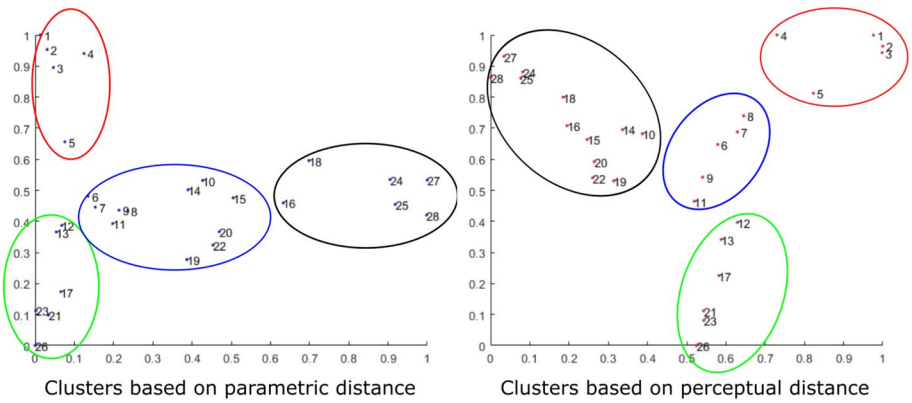


Fig. 14 Visualization of multi-dimensional scaling and k-means clustering with $k = 4$. (Different colors represent different clusters. Numbers correspond to the facial model index.)

The concern here is not the performance of a clustering algorithm on the ability to separate the data. The aim is to identify whether different distance metrics reflect human-generated clustering performance. Thus, classical clustering efficiency measurements such as silhouette score, partition coefficient, etc., are not feasible for this problem. As a substitute, clustering accuracy was calculated using Eq. 7, in which experimental clusters were used as ground truth labels:

$$\text{Accuracy} = 100 * \frac{\text{Consistent assignments}}{28}. \tag{7}$$

Inconsistent cluster assignments are also marked in Figs. 15 and 16. Five and four of the total instances were inconsistent with the experimental clustering assignments for parametric and perceptual distance-based clustering, respectively. Thus, the clustering accuracies are 82% and 86% for parametric and perceptual distance metrics, respectively. When inconsistent assignments are inspected, it is seen that the inter-observer variability is high and hence they fall into the grey region of perception (see Table 2 of the

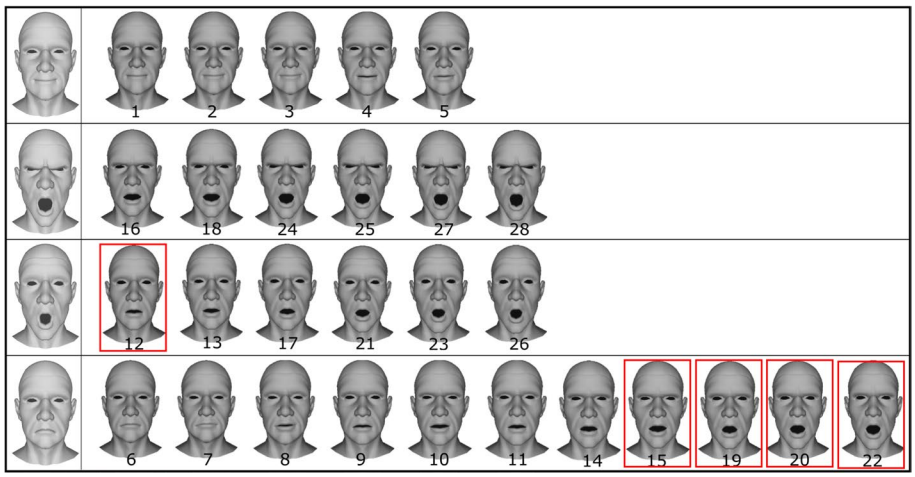


Fig. 15 Clusters generated according to the parametric distance metric. The first column includes reference expressions for each cluster provided to the users, they were not used in automatic clustering. K-means clustering was used with $k = 4$. Faces with red borders show assignments that are different from experimental clusters

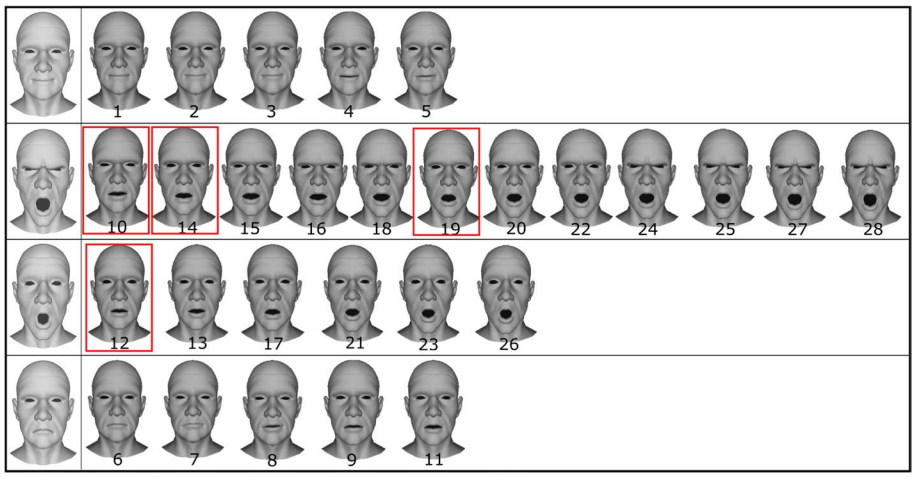


Fig. 16 Clusters generated according to the proposed distance metric. The first column includes reference expressions for each cluster provided to the users, they were not used in automatic clustering. K-means clustering was used with $k = 4$. Faces with red borders show assignments that are different from experimental clusters

supplementary material). From the results, we can see that both the experimental and perceptual clustering assign a high number of expressions to the angry face category. This is also consistent with the success of the dominant target distance heuristic and it can be explained by the fact that the angry face is perceptually more salient, and hence it dominates in the perception.

5 Applications

In this section, several applications in which the proposed metric can be integrated are suggested.

5.1 Embedding and clustering applications

Distance metrics are generally used for the visualization of data in a lower dimension when the number of elements is high. There are several different techniques to embed high-dimensional data in two or three dimensions to make the data comprehensible. The most common techniques for this purpose can be listed as *multi-dimensional scaling (MDS)* and *t-Distributed Stochastic Neighbor Embedding (t-SNE)*. These techniques take a pairwise distance matrix between the elements and return its low-dimensional embedding.

They generally use common distance measures such as Euclidian, city-block, cosine similarity, etc. In addition, unsupervised learning methods such as clustering also employ a distance metric to discover the natural clusters in the data. Utilizing the proposed perceptual distance metric in such an embedding or clustering application would produce a perceptually correlated visualization of face models.

Such a kind of embedding is also helpful for storage purposes. If the perceived distance is too small between some blendshape models and they fall into the same perceptual space, there is no need to generate and store all of them.

5.2 Assistance for novice animators

For especially novice users, it may be difficult to determine the blendshape weights which generate a desired facial expression. The proposed metric can be used while finding the optimum weight combination of the morph targets that minimize the perceptual distance of the blended model to a reference face model. Consider the following scenario: An apprentice user has one or several blendshapes that were previously created by an artist. The user has also generated some morph targets and desires to blend these targets in an animation in such a way that the blended model resembles the reference models created by the artist. This is an optimization problem in which the objective function in Eq. 8 should be minimized; where D is the proposed perceptual distance function, F_{novice} is the blendshape obtained by Eq. 1, w is the weight vector to blend morph targets in Eq. 1, and F_{ref} is the reference face model created by the artist. Some regularization terms and constraints based on user preferences could also be added to the objective function. For instance, the L_1 norm of the weight vector w could be added to make the vector sparse and remove the perceptually ineffective targets from further processing:

$$\min_{w} D(F_{\text{novice}}, F_{\text{ref}}). \tag{8}$$

6 Conclusion

A perceptual distance metric specified for 3D facial blendshapes is proposed in this paper. The proposed metric was learnt by optimizing the match between human perception data obtained from a crowdsourcing experiment and the estimated distance. The analysis of the results shows that purely geometric distance metrics do not suffice to reflect human perception whereas the proposed metric exhibits a high correlation between human perception. Several applications in which the proposed distance metric can be employed are also suggested. The metric can be used to assist novice users while creating plausible facial expression models and reducing the storage size of blendshapes. Using this metric, it is even possible to compare blendshapes with different geometry

and topology, as the features are obtained from the distribution statistics of some properties of the landmarks. In this way, the optimum level of detail of a model can be found for trading off between expression perception and computational complexity.

The limitations explained in the Results section also construct future research directions. To dispose of the manual feature extraction process, deep learning methods should be investigated. The effects of skin color, texture, eye, teeth, etc., and race and gender differences should also be elaborated. The method should be tested in more practical cases such as during animations and using a more diverse dataset. No-reference metric development could be another research problem.

Abbreviations

2D	Two-dimensional
3D	Three-dimensional
AMT	Amazon Mechanical Turk
AU	Action unit
HIT	Human intelligence task
L-BFGS	Limited Memory Broyden–Fletcher–Goldfarb–Shanno
MAP	Maximum a posteriori
MDS	Multi-dimensional scaling
PCC	Pearson correlation coefficient
SROCC	Spearman rank order correlation coefficient
t-SNE	T-Distributed Stochastic Neighbor Embedding

Acknowledgements
Not applicable.

Author contributions
All authors read and approved the final manuscript.

Funding
This work was supported by Scientific Research Project Office of Manisa Celal Bayar University. Project Number: 2022-053.

Availability of data and materials
The data presented in this study are available on request from the corresponding author.

Declarations

Competing interests
The authors declare that they have no competing interest.

Received: 26 January 2023 Accepted: 9 May 2023
Published online: 15 May 2023

References

1. P. Alliez , D. Cohen-Steiner, O. Devillers, B. Lévy, M. Desbrun. Anisotropic polygonal remeshing. In: ACM SIGGRAPH 2003 Papers, pp 485–493 (2003)
2. K. Anjyo, H. Todo, J.P. Lewis, A practical approach to direct manipulation blendshapes. J. Graphics Tools **16**(3), 160–176 (2012)
3. T. Beeler, B. Bickel, P. Beardsley, B. Sumner, M. Gross. High-quality single-shot capture of facial geometry. In: ACM SIGGRAPH 2010 papers, pp 1–9 (2010)
4. V. Bruce, A.W. Young, *Face perception* (Psychology Press, 2012)
5. C. Cao, D. Bradley, K. Zhou, T. Beeler, Real-time high-fidelity facial performance capture. ACM Trans. Graphics (ToG) **34**(4), 1–9 (2015)
6. E.Carrigan, K. Zibrek, R. Dahyot, R. McDonnell, Investigating perceptually based models to predict importance of facial blendshapes. In: Motion, Interaction and Games, pp 1–6 (2020)
7. P. Chandran, D. Bradley, M.Gross, T. Beeler, Semantic deep face models. In: 2020 International Conference on 3D Vision (3DV), IEEE, pp 345–354 (2020)
8. M. Cong, M. Bao, E JL, Bhat KS, R. Fedkiw, Fully automatic generation of anatomical face simulation models. In: Proceedings of the 14th ACM SIGGRAPH/Eurographics Symposium on Computer Animation, pp 175–183 (2015)
9. Z. Deng, X. Ma, Perceptually guided expressive facial animation. In: Proceedings of the 2008 ACM SIGGRAPH/Eurographics Symposium on Computer Animation, pp 67–76 (2008)
10. Z. Deng, J. Noh, Computer facial animation: A survey. In: Data-driven 3D facial animation, Springer, pp 1–28 (2008)

11. K. Dobs, I. Bülthoff, M. Breidt, Q.C. Vuong, C. Curio, J. Schultz, Quantifying human sensitivity to spatio-temporal information in dynamic faces. Vision. Res. **100**, 78–87 (2014)
12. N. Ersotelos, F. Dong, Building highly realistic facial modeling and animation: a survey. Vis. Comput. **24**(1), 13–30 (2008)
13. Y. Gingold, A. Shamir, D. Cohen-Or, Micro perceptual human computation for visual tasks. ACM Trans. Graphics (TOG) **31**(5), 1–12 (2012)
14. M. Lagunas, E. Garces, D. Gutierrez, Learning icons appearance similarity. Multim. Tools Appl. **78**(8), 10733–10751 (2019)
15. CH. Lee, A. Varshney, DW. Jacobs, Mesh saliency. In: ACM SIGGRAPH 2005 Papers, pp 659–666 (2005)
16. J.P. Lewis, K. Anjyo, T. Rhee, M. Zhang, F.H. Pighin, Z. Deng, Practice and theory of blendshape facial models. Eurographics (State of the Art Reports) **1**(8), 2 (2014)
17. Z. Lun, E. Kalogerakis, A. Sheffer, Elements of style: learning perceptual shape style similarity. ACM Trans. Graphics (TOG) **34**(4), 1–14 (2015)
18. C. Mousas, CN. Anagnostopoulos, Structure-aware transfer of facial blendshapes. In: Proceedings of the 31st Spring Conference on Computer Graphics, pp 55–62 (2015)
19. T. Neumann, K. Varanasi, S. Wenger, M. Wacker, M. Magnor, C. Theobalt, Sparse localized deformation components. ACM Trans. Graphics (TOG) **32**(6), 1–10 (2013)
20. R. Niewiadomski, C. Pelachaud, The effect of wrinkles, presentation mode, and intensity on the perception of facial actions and full-face expressions of laughter. ACM Trans. Appl. Percept. (TAP) **12**(1), 1–21 (2015)
21. Y. Seol, J.P. Lewis, J. Seo, B. Choi, K. Anjyo, J. Noh, Spacetime expression cloning for blendshapes. ACM Trans. Graphics (TOG) **31**(2), 1–12 (2012)
22. Y. Seol, J. Seo, P.H. Kim, J.P. Lewis, J. Noh, Weighted pose space editing for facial animation. Vis. Comput. **28**(3), 319–327 (2012)
23. C. Wallraven, M. Breidt, D.W. Cunningham, H.H. Bülthoff, Evaluating the perceptual realism of animated facial expressions. ACM Trans. Appl. Percept. (TAP) **4**(4), 1–20 (2008)
24. F. Xu, J. Chai, Y. Liu, X. Tong, Controllable high-fidelity facial performance transfer. ACM Trans. Graphics (TOG) **33**(4), 1–11 (2014)
25. C. Yan, T. Teng, Y. Liu, Y. Zhang, H. Wang, X. Ji, Precise no-reference image quality evaluation based on distortion identification. ACM Trans. Multimed. Comput. Commun. Appl. (TOMM) **17**(3s), 1–21 (2021)
26. Z.C. Yildiz, T. Capin, A perceptual quality metric for dynamic triangle meshes. EURASIP J. Image Video Proc. **2017**, 1–18 (2017)
27. Z.C. Yildiz, A.C. Oztireli, T. Capin, A machine learning framework for full-reference 3d shape quality assessment. Vis. Comput. **36**(1), 127–139 (2020)
28. H. Yu, O.G. Garrod, P.G. Schyns, Perception-driven facial expression synthesis. Comput. Graphics **36**(3), 152–162 (2012)
29. E. Zell, C. Aliaga, A. Jarabo, K. Zibrek, D. Gutierrez, R. McDonnell, M. Botsch, To stylize or not to stylize? the effect of shape and material stylization on the perception of computer-generated faces. ACM Trans. Graphics (TOG) **34**(6), 1–12 (2015)

Publisher's Note