Vision Research 100 (2014) 78–87

Contents lists available at ScienceDirect

Vision Research

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

Quantifying human sensitivity to spatio-temporal information in dynamic faces

Katharina Dobs [a,*], Isabelle Bülthoff [a,b], Martin Breidt [a], Quoc C. Vuong [c], Cristóbal Curio [a], Johannes Schultz [a,d]

[a] Department Human Perception, Cognition and Action, Max Planck Institute for Biological Cybernetics, Spemannstr. 38, 72076 Tübingen, Germany
[b] Department of Brain and Cognitive Engineering, Korea University, Anam-dong Seongbuk-Gu, Seoul 136-701, Republic of Korea
[c] Institute of Neuroscience, Newcastle University, Framlington Place, Newcastle upon Tyne NE2 4HH, United Kingdom
[d] Department of Psychology, Durham University, South Road, Durham DH1 3LE, United Kingdom

ARTICLE INFO

Article history:
Received 3 November 2013
Received in revised form 4 March 2014
Available online 28 April 2014

Keywords:
Face perception
Motion perception
Facial motion
Psychophysics
Motion retargeting
Animation

ABSTRACT

A great deal of perceptual and social information is conveyed by facial motion. Here, we investigated observers' sensitivity to the complex spatio-temporal information in facial expressions and what cues they use to judge the similarity of these movements. We motion-captured four facial expressions and decomposed them into time courses of semantically meaningful local facial actions (e.g., eyebrow raise). We then generated approximations of the time courses which differed in the amount of information about the natural facial motion they contained, and used these and the original time courses to animate an avatar head. Observers chose which of two animations based on approximations was more similar to the animation based on the original time course. We found that observers preferred animations containing more information about the natural facial motion dynamics. To explain observers' similarity judgments, we developed and used several measures of objective stimulus similarity. The time course of facial actions (e.g., onset and peak of eyebrow raise) explained observers' behavioral choices better than image-based measures (e.g., optic flow). Our results thus revealed observers' sensitivity to changes of natural facial dynamics. Importantly, our method allows a quantitative explanation of the perceived similarity of dynamic facial expressions, which suggests that sparse but meaningful spatio-temporal cues are used to process facial motion.

1. Introduction

Most of the faces we encounter and interact with everyday move. Dynamic faces are highly ecological stimuli from which we can extract various cues such as the affective states of others (e.g., Ambadar, Schooler, & Cohn, 2005; Cunningham & Wallraven, 2009; Kaulard et al., 2012; Krumhuber, Kappas, & Manstead, 2013), the intensity of emotions (e.g., Jack et al., 2012; Kamachi et al., 2001) or speech movements (e.g., Bernstein, Demorest, & Tucker, 2000; Rosenblum et al., 2002). Given the social relevance of facial motion, it is of great interest to study which face motion cues are used by observers during perceptual tasks. However, dynamic face information is complex, which makes it difficult to isolate and quantify meaningful cues. Such quantification would for example allow testing human sensitivity to various aspects of this spatio-temporal information (e.g., onset or acceleration of movements) using dynamic face stimuli with controlled information content. Here we first measured the perceived similarity of computer generated facial expressions. This similarity was then correlated with different cues in the animations to test observers' sensitivity to natural facial movements and explore the cues they used for face perception.

One common method to quantify the spatio-temporal information in complex facial movements is to use a coding scheme for facial expressions called Facial Action Coding System (FACS; Ekman & Friesen, 1978; Ekman, Friesen, & Hager, 2002). This system defines a number of discrete face movements - termed Action Units - as an intuitive and accurate description of the basic constituents of facial expressions (e.g., eyebrow raising). Each Action Unit can be represented as a time course which captures the magnitude of activation of a "local" facial region (e.g., eyebrow) over time. This magnitude can vary from no activation to some maximum

[*] Corresponding author.
E-mail addresses: katharina.dobs@tuebingen.mpg.de (K. Dobs), isabelle.buelthoff@tuebingen.mpg.de (I. Bülthoff), martin.breidt@tuebingen.mpg.de (M. Breidt), quoc.vuong@newcastle.ac.uk (Q.C. Vuong), cristobal.curio@tuebingen.mpg.de (C. Curio), j.w.r.schultz@durham.ac.uk (J. Schultz).

http://dx.doi.org/10.1016/j.visres.2014.04.009
0042-6989/© 2014 The Authors. Published by Elsevier Ltd.

intensity. As exemplified in Fig. 1 (red line), the eyebrow can naturally rise and lower from a resting, neutral position over time as an actor makes a facial expression. These time courses thus capture spatio-temporal properties of local facial movements (e.g., onset, acceleration of eyebrow raising). Curio and colleagues developed a novel 3D facial animation approach inspired by FACS to decompose motion-capture data recorded from actors into time courses of local facial movements termed facial actions (Curio et al., 2006). Like Action Units of the FACS system, facial actions are semantically meaningful. In their study, Curio and colleagues showed that using a set of local facial actions to approximate the facial motion led to more natural animations than using a global approximation in which the whole face is deformed at the same time.

Recently, a series of studies have used synthesized time courses for FACS Action Units to generate animations of facial expressions in the absence of actor data (e.g., Jack et al., 2012; Roesch et al., 2011; Yu, Garrod, & Schyns, 2012). Without real-data recorded from a performing actor, the particular shape of an Action Unit's time course is arbitrary, and various methods can be used to generate it. In its simplest form, an Action Unit's activation can increase linearly over time from no activation to some level of activation (see blue line in Fig. 1). When applying this linear interpolation to all Action Units of a facial expression, the resulting stimulus is very similar to an image sequence made by gradually morphing between two images (e.g., neutral and peak of the facial expression). Given the simplicity and ease of control, such techniques have been used in many studies investigating facial motion perception (e.g., Furl et al., 2010; Ku et al., 2005; LaBar et al., 2003; Sarkheil et al., 2012; Sato & Yoshikawa, 2007). More recent studies have combined spline interpolation (see green line in Fig. 1) with advanced reverse-correlation methods and found that observers used fine-grained spatio-temporal cues to categorize facial expressions (e.g., Jack et al., 2012; Yu, Garrod, & Schyns, 2012). In line with these findings, other studies showed that advanced spatio-temporal interpolations are perceived as more natural than linear or global interpolations of facial motion (e.g., Cosker, Krumhuber, & Hilton, 2010; Curio et al., 2006). However, how sensitive humans are to spatio-temporal cues in facial motion has not been investigated quantitatively so far.

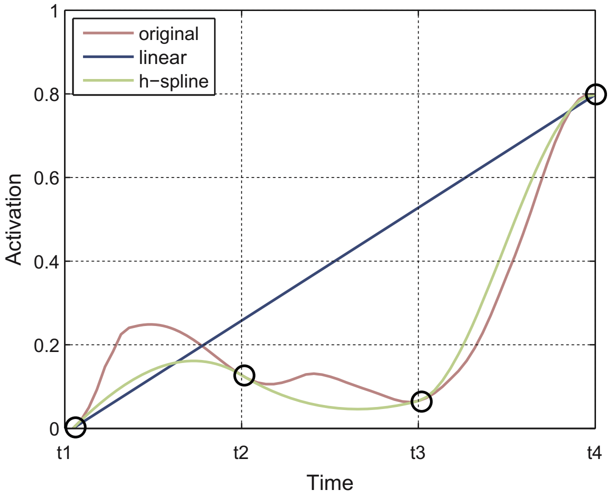

Fig. 1. Exemplary time course of activation for one local facial action. The "original" time course derived from facial motion tracking is shown in red. The simplest kind of approximation is a "linear interpolation" from t1 to t4 (outer left and outer right black circles) of the original time course, and is shown in blue. A more sophisticated approximation method is to use a Hermite spline interpolation ("h-spline"), based on four control points (t1, t2, t3 and t4, black circles), which is shown in green. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

In the current study, we investigated observers' sensitivity to changes in facial motion, and studied what cues observers extract and interpret when making judgments about facial motion. Identifying these cues would provide clues about the importance of different aspects of facial motion for perception, also in comparison to static faces, and thus have implications for theories of mental representations of facial motion. Given the importance of motion for facial expressions, we focused on this aspect of face perception but it should be noted that identity and expressions may be processed by different pathways in the brain (Bruce & Young, 1986; Haxby, Hoffman, & Gobbini, 2002; but see also Calder & Young, 2005). With static faces, one widely used approach to determine cues important for face perception is to correlate objective measures of similarity (e.g., Gabor jets, principal components) with perceived similarities between facial expressions (e.g., Lyons et al., 1998; Susskind et al., 2007) and facial identities (e.g., Rhodes, 1988; Steyvers & Busey, 2000; Yue et al., 2012). Here we adopted a similar approach for dynamic facial expressions to assess whether objective measures of facial motion similarity could explain the perceived similarity of facial motion. As discussed below, the different measures captured both low-level and high-level cues in our dynamic faces. We used the system developed by Curio et al. (2006) to generate high quality animations based on natural facial motion, which we will refer to as "original animations". We then created additional animations based on different approximations of the facial action time courses obtained from the actors' motion, which we will call "approximations". To this end, we chose interpolation techniques such that the approximations systematically varied in the amount of information they contained about the natural motion dynamics. Observers judged which of two approximations was more similar to the original animation. If observers were sensitive to differences between the approximations, they would consistently judge one animation of the pair to be more similar to the original. We captured facial expressions with different dynamics (e.g., including speech movements) to investigate whether the goodness of an approximation varied with the type of facial expression. The pattern of choices served as a measure of the perceived similarity between approximations and the original animation, and allowed us to directly compare perceived similarities between stimuli with objective measures of similarity. Here, we calculated these objective similarity measures based on three kinds of information: (1) time courses of facial action activation, (2) optic flow, and (3) Gabor-jet filters. Importantly, facial action time courses capture semantically meaningful high-level changes to a sparse set of local facial regions (e.g., eyebrow) whereas optic flow and Gabor-jets capture detailed low-level image changes (e.g., movement direction of one pixel). To anticipate our results: We found that high-level cues about spatio-temporal characteristics of facial motion best explained observers' choice pattern.

2. Material and methods

2.1. Participants

Fourteen participants (6 female; mean age: 28.6 ± 5.2 years) were recruited from the subject database of the Max Planck Institute for Biological Cybernetics, Tübingen, Germany. They were naive to the purpose of the experiment and had normal or corrected-to-normal vision. All participants provided informed written consent prior to the experiment and filled out a post-questionnaire after the experiment was finished. The study was conducted in accordance to the Declaration of Helsinki.

2.2. Stimuli

To create highly controllable and accurate animations of facial expressions, we used a system that decomposes recorded motion data into time courses of facial actions (e.g., eyebrow raising) which are used to animate a 3D head model with corresponding facial actions (Curio et al., 2006). This facial animation procedure is schematically shown in Fig. 2 and is explained in detail in Appendix A.

2.2.1. Constructing approximations of natural facial expressions

We created approximations of the original time courses for each facial action obtained from the motion decomposition (see Appendix A for details). We did not attempt to find the optimal technique to approximate natural facial motion, but focused on different approximation techniques (linear and spline interpolations) previously used to investigate perception of facial motion (e.g., Jack et al., 2012; Sarkheil et al., 2012). We selected different time points at fixed intervals of the original time courses as control points to create our approximations. The start and the end of the time course were always included as control points but the number of points in between was varied to create approximations that preserved different aspects of the original time course. We selected a subset of four approximation techniques to span a range of possible techniques. Many more techniques could have been used to reveal a more fine-grained pattern of results; however, restrictions imposed by the experimental design (mainly the total number of trials) meant that this would have gone beyond the scope of the present study.

Fig. 3 illustrates the four approximation techniques we used, with the facial action “mouth open” from the facial expression “fear” serving as example. The red line represents the time course from the motion decomposition (*orig*, Fig. 3A and B). For the linear approximation *lin1* (magenta, Fig. 3A), three equidistant control points were chosen from the original time course (frame 1, 50 and 100, black circles) and used to linearly interpolate the original time course. The second approximation *lin2* (blue, Fig. 3B) is another linear interpolation of the original time course based on four equidistant control points (frame 1, 33, 67 and 100, black circles). Lin2 contains more information about the original time courses than lin1, and the animation made on its basis should thus be perceived as more similar to the animation based on the original time course. Linear interpolations may contain sharp changes in the time courses at the position of the control points (see frame 67 of lin2, black circle in Fig. 3B). The visual system may be sensitive to these changes. Thus, for another set of approximations, we created spline interpolations of the original time courses, which are very smooth at the control points and should thus appear more similar to the original motion than lin1 and lin2. For the first spline approximation *spl1* (green in Fig. 3A) we used a cubic spline interpolation based on the same three control points as lin1 (frame 1, 50 and 100, black circles). While this approximation is smoother than the linear approximations, splines tend to exceed (overshoot) the interpolated time course at extremes, resulting in a large difference from the original time course. We reduced this in the next approximation by using cubic Hermite splines *hspl* (yellow, Fig. 3B), in which the spline interpolation not only goes through the same four control points as lin2 (frame 1, 33, 67 and 100, black circles) but also preserves the slope of the time course at the given control points. This approximation contains the most information about the original time courses of facial actions.

2.2.2. Animation stimuli

For each facial expression, we sub-sampled the number of frames by a factor of three to 34 frames to ensure fluid video display during the experiments (*Note:* this down-sampling did not affect the smoothness or other characteristics of the time courses and was thus not perceptible in the final stimuli). We then loaded the original and approximated time courses of facial action activation into 3ds Max to produce 20 Quicktime animation movies with a resolution of 480 × 640 pixels, a duration of about 1 s (34 frames at 30 Hz), and scene and rendering settings optimized for facial animations.

To assess whether our stimuli could be correctly recognized, we performed a preliminary experiment with a different set of participants ($N = 10$). In a 4 alternative-forced-choice task, participants were able to correctly identify the four expressions from the animations based on the original time courses (chance = 25%). Recognition was perfect for happiness (mean and standard error of the mean: 100 ± 0%) and good for anger and surprise (80 ± 13% for both). Performance for the expression fear was lower but still clearly above chance level (60 ± 16%).

Videos that matched the animations frame-by-frame were recorded by the scene camera during the motion capture, and after scaling to match the visual angle subtended by the size of the face in the animations (approximately 8° × 13°), they were saved at the

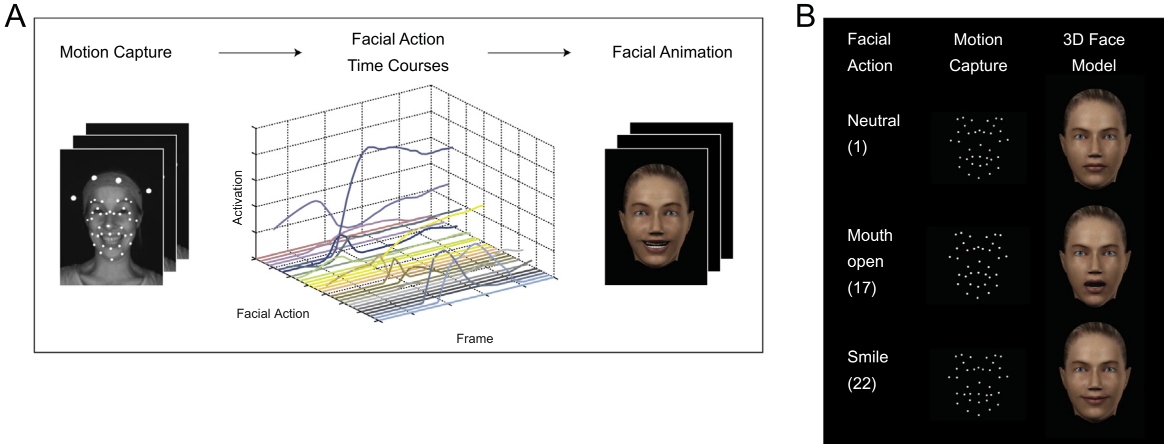

Fig. 2. (A) Schematic overview of the facial animation procedure, shown for the facial expression happiness. In the first step, motion capture data from a set of facial actions and the facial expression happiness is recorded from the actor (left). In the second step, the facial animation system decomposes the motion capture data of the facial expression into time courses of facial action activation (middle). In the last step, the time courses of facial action activation are used to animate a semantically matched 3D face model (right). (B) Three example facial actions “Neutral” (facial action 1), “Mouth open” (facial action 17) and “Smile” (facial action 22). The recorded facial marker positions for the facial actions (middle), and the semantically matched 3D facial action shapes created for the 3D face model (right).

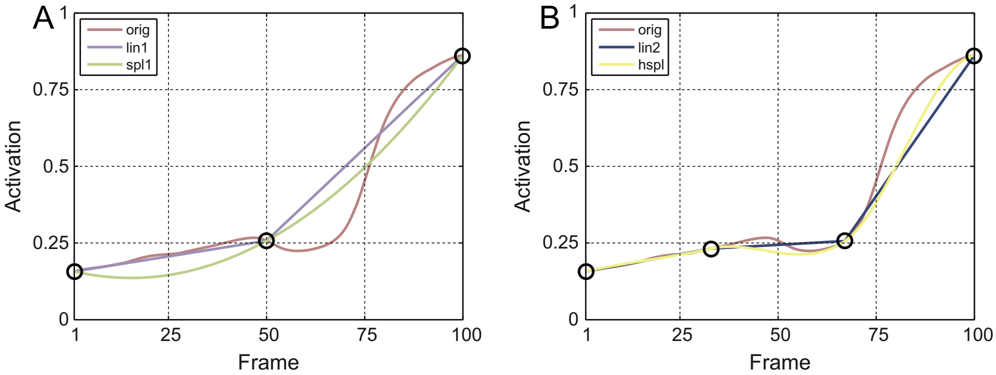

Fig. 3. Time course of activation for the facial action "Mouth open" during the facial expression "fear" directly resulting from the motion analysis (orig, shown in red) and interpolated using four approximation types. (A) Approximations lin1 (magenta) and spl1 (green) based on three control points (frame 1, 50 and 100, black circles). (B) Approximations lin2 (blue) and hspl (yellow) based on four control points (frame 1, 33, 67 and 100, black circles). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

same frame rate as the animations. We used these videos as stimuli in a second preliminary experiment which tested whether the original animation was perceptually the most similar to the corresponding expression video. The participants of our main study ($N = 14$, see Section 2.1) performed a delayed match-to-sample task using the video as sample and the animations as comparison stimuli. When paired with any approximation, the original animation was chosen in more than 50% for all expressions (anger: 89%, $t(13) = 18.97$, $p < 0.0001$; fear: 69%, $t(13) = 9.33$, $p < 0.0001$; happiness: 71%, $t(13) = 5.37$, $p < 0.001$; surprise: 72%, $t(13) = 7.11$, $p < 0.0001$). Thus, the original animations were perceived to be most similar to videos of the expressions.

2.3. Design and procedure

Perceptual sensitivity to the different approximations was tested in a delayed match-to-sample task: After watching the original animation driven by one of the four facial expressions (sample), observers were asked to indicate which of the two approximations (matching stimuli A and B) was most similar to the original animation. The six possible combinations of the four approximations were repeated ten times (60 trials) for each of the four facial expressions, for a total of 240 trials. Trials were run in random order and the presentation on the left or the right side of the screen was counter-balanced for each animation within a specific pair.

Fig. 4 depicts the trial procedure in the experiment. All trials began with a white fixation cross on a black background shown for 0.5 s at the center of the screen, followed by the original animation. After the animation, a black screen appeared for 0.5 s, followed by two matching animations presented side-by-side, 6.7° to the left and right of fixation. As the difference between animations was subtle, we decided to present the animations simultaneously to allow for detailed, continuous assessment of the facial motions without influence of memory load. The same presentation procedure had already been successfully used in the study by Curio et al. (2006). Observers could indicate their readiness to respond by pressing any key on a standard computer keyboard during the trial. The sequence of animation, black screen and two animations was repeated until a key was pressed or three presentations were reached. Each sequence was repeated 1.45 times per trial on average across observers with a standard deviation of 0.31. Then a response screen showing the question "Which of the two animations was most similar to the original?" appeared. Observers pressed the left or right cursor arrow key to choose the corresponding animation. The response screen remained until observers responded. No feedback was provided.

Observers could take up to seven self-timed breaks, one every 30 trials. The experiment lasted approximately 60–70 min and was programmed using PsychToolbox 3 for Matlab (http://www.psychtoolbox.org) (Kleiner, 2010). Observers were seated approximately 68 cm from a Dell 2407WFP monitor (24 in. screen diagonal size, 1920 × 1200 pixel resolution; 60 Hz refresh rate).

2.4. Calculating objective similarity measures

One aim of this experiment was to determine the extent to which observers' choice behavior correlated with objective measures of similarity between the animations. Each stimulus consists of a sequence of images (34 frames). Given how each animation was generated, an animation stimulus can also be conceived as a set of time courses (with each time course representing the activation of a facial action over time). Various cues can thus be extracted from either the image sequences or the time courses and used to measure the similarity between two animations.

2.4.1. Similarity based on facial action activation

First, we calculated the similarity of animations based on the time courses of facial action activation. Each frame of the original animation and the approximations can be described in terms of

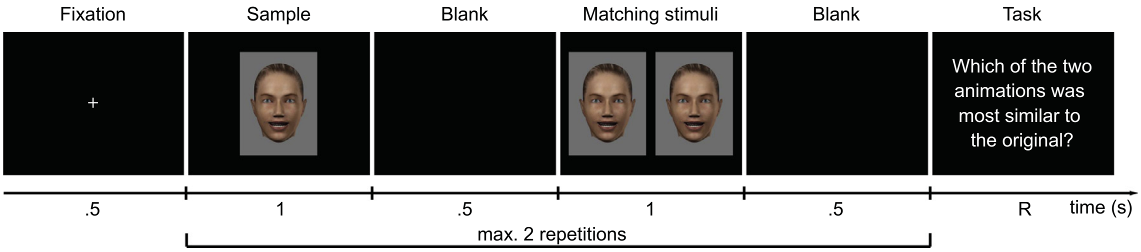

Fig. 4. The trial procedure of the experiment.

the activation of facial actions used to construct this frame. As those values cannot be retrieved in a straightforward way using image analysis, we consider these facial action activations to be high-level cues. To calculate the similarity between two animations, we carried out the following steps. First, we interpreted each frame of an animation as coordinates in an n-dimensional facial action space, where n represents the number of facial actions used to generate a frame of the stimulus. For a single frame, we computed the distance between two animations in this space as the Euclidean distance between the facial action activations. We then summed the distances across all frames. This procedure was implemented in the following equation, for two animations a and b:

$$d_{FA}(a,b) = \sum_{i=1}^{m} \sqrt{\sum_{j=1}^{n} (a_{i,j} - b_{i,j})^2}, \tag{1}$$

where $m = 34$ is the total number of frames per animation, $n = 30$ is the number of facial actions and $a_{i,j}$ represents the activation level of facial action j at time i. Note that the same equation can be applied to calculate the similarity between two animations based on a subset of facial actions, or even based on a single facial action (e.g., for facial action “eyebrows raised” with $n = 1$).

2.4.2. Similarity based on optic flow

Second, we calculated the similarity between two animations based on optic flow. In the context of an image sequence, optic flow is defined as the spatial displacement of pixels from one image of the animation to the next image (Horn & Schunck, 1981). We used 3ds Max to directly output the pixel motion of our animations (called “velocity render element” in the software). The motion output of one time frame consists of q 3-dimensional vectors of pixel space motion (x-, y- and z-motion), where q represents the number of pixels in the animation. For simplicity, we ignored motion in the z-axis (depth) as the stimuli rendered from our animations were 2-dimensional. We calculated the distance in pixel motion between two animations a and b as follows:

$$d_{OF}(a,b) = \sum_{i=1}^{m-1} \sum_{p=1}^{q} \sqrt{\sum_{c=1}^{2} (a_{i,p,c} - b_{i,p,c})^2}, \tag{2}$$

where c represents the motion direction (in horizontal and vertical dimensions), $q = 480 \times 640$ is the number of pixels p per frame, summed over m (the total number of frames per animation) $-1 = 33$ consecutive pairs of frames. (Note that the sum is over $m - 1$ frames as the motion vectors represent pixel motion from two consecutive frames, so there is no motion information about the first frame of the animation.)

2.4.3. Gabor similarity

As a third similarity measure, we computed the Gabor similarity between two animations. Gabor similarity is a biologically-inspired physical similarity measure that emulates the responses of simple and complex cells (see Lades et al., 1993). In early visual cortex (V1), both simple and complex cells are organized into hypercolumns that respond to different spatial frequencies at different orientations. Importantly, this similarity measure is highly correlated with the perceived similarity of the identity of static faces (e.g., Yue et al., 2012) and has been successfully applied as similarity measure for facial expressions (Xu & Biederman, 2010). We computed the Gabor similarity between two animations as follows. First, we took all corresponding frames from each animation and converted them into grayscale images (256 levels). Second, we placed a Gabor jet at the intersections of a uniform grid (11 × 14) covering the entire image. Each jet consisted of five spatial scales and eight equidistant orientations (i.e., 22.5° differences in angle; for details see Yue et al., 2012). Third, we convolved the image with each jet to get its response to an image. The responses from all the Gabor jets thus form a high-dimensional feature vector (5 scales × 8 orientations × (11 × 14) jets = 6160 features) for each frame. Lastly, the Gabor similarity between corresponding images was computed as the Euclidean distance between the two feature vectors, J_a and J_b, and summed across all corresponding pairs of frames in the two animations a and b:

$$d_{GS}(a,b) = \sum_{i=1}^{m} \sqrt{\sum_{j=1}^{n} (J_{a_{i,j}} - J_{b_{i,j}})^2}, \tag{3}$$

where $m = 34$ represents the number of frames and n is the number of features in each feature vector.

2.5. Calculating choice probabilities

In order to compare the objective similarity measures to observers’ choice behavior, we computed choice probabilities based on the three objective similarity measures (Luce, 1959). Observers had to choose which of two approximations was most similar to the original animation. For each similarity measure, we used Luce’s choice rule to calculate the probability of choosing one approximation over the other. The probability of choosing which of two approximations, a and b, is most similar to the original animation can be expressed as the conditional probability of selecting a (response r_a) given an original animation o:

$$P(r_a|O) = 1 - \frac{d(a,O)}{d(a,O) + d(b,O)}, \tag{4}$$

where $d(a,o)$ is the similarity between the approximation a and the original animation o in terms of facial action activation, optic flow or Gabor similarity, and $d(b,o)$ is the corresponding similarity between approximation b and the original o. Note that the choice probability is given as 1-fraction because the similarity is represented as distance, such that large distances indicate low similarity and small distances indicate high similarity. If two approximations are equally similar to the original animation o (i.e., $d(a,o) = d(b,o)$), the probability of choosing a is 0.5. If the approximation a is very similar to the original animation o (e.g., the distance $d(a,o) = 0.1$) and the approximation b is very dissimilar (e.g., the distance $d(b,o) = 0.9$), the probability of choosing a is very high (e.g., $P(r_a|o) = 0.9$).

2.6. Regression analyses

We investigated whether the calculated choice probabilities based on objective similarity measures could predict observers’ choice behavior. After assessment of the normality of the data (quantile–quantile plot of data against a normal distribution), we ran three separate linear regression analyses to assess the contribution of each similarity measure to the behavioral choices. A Kolmogorov–Smirnov test on the residuals was not significant and thus confirmed the suitability of parametric analyses. Second, we investigated the best fitting model combining the three similarity measures to explain the behavioral choices. As facial action time courses are used to create the animation stimuli which image-based measures are based on, we expected the similarity measures to be correlated. In case of multicollinearity, the prediction accuracy of ordinary least squares regression can be reduced and the results are difficult to interpret (e.g., see Hoerl & Kennard, 1970). Compared to ordinary least squares regression, regularized regression obtains higher prediction accuracy (in particular if multicollinearity exists) and provides a simpler model by selecting the most informative predictors. Regularized regression methods include an additional regularization term in the cost function (e.g., an l_1-norm regularizer as in the “Lasso method”; see Tibshirani, 1996; or an l_2-norm as in “ridge regression”; see Hoerl & Kennard, 1970) for which a regularization parameter λ defines the degree to which coefficients are penalized. While ridge regression performs a shrinking of all coefficients, Lasso additionally selects variables by setting small coefficients to zero. Recently, a regularized regression method was proposed which combines Lasso and ridge regression (called “Elastic net”; see Zou & Hastie, 2005). Elastic net selects the most important predictors under consideration of multicollinearity (Zou & Hastie, 2005) where an additional parameter α ($0 < \alpha \leqslant 1$) defines the weight of lasso ($\alpha = 1$) versus ridge regression ($\alpha = 0$). Here, we applied regularized regression to investigate which similarity measure could explain most of the variance in the behavioral choices.

3. Results

3.1. Observers' perceptual choices

Fig. 5 shows which approximation the observers judged to be perceptually closest to the original animation in all possible pairs (Fig. 5A) and separated for facial expressions (Fig. 5B). The *y*-axis represents the mean proportion of trials in which observers chose a specific approximation and the *x*-axis represents the six possible approximation pairs. For trials in which approximation A was paired with approximation B, 0 indicates that B (bottom label on *x*-axis) was chosen on 100% of the trials, 1 indicates that A (top label of *x*-axis) was chosen on 100% of the trials, and 0.5 indicates that both approximations were chosen equally often.

As can be seen in Fig. 5A, choice proportion was different from chance in all pairs (lin2 > lin1: 70%, $t(13) = 9.91$, $p < 0.0001$, $d = 2.65$; hspl > lin1: 76%, $t(13) = 9.96$, $p < 0.0001$, $d = 2.53$; lin2 > spl1: 69%, $t(13) = 9.46$, $p < 0.0001$, $d = 2.46$; hspl > lin2: 66%, $t(13) = 9.26$, $p < 0.0001$, $d = 2.66$; hspl > spl1: 69%, $t(13) = 9.19$, $p < 0.0001$, $d = 2.48$) except for the pair lin1-spl1 (lin1 > spl1: 56%, $t(13) = 1.08$, $p > 0.1$, $d = 0.29$). This result suggests that observers were sensitive to differences between approximations because they consistently chose one approximation over another, with the exception of lin1 and spl1 (chosen equally often). The data further show that the four animations can be ranked in terms of observers' decreasing choice proportion: hspl > lin2 > lin1 = spl1.

A 6 approximation pairs × 4 expressions ANOVA revealed a main effect of approximation pair on choice proportions ($F(5, 13) = 81.49$; $p < 0.0001$, $\eta^2 = 0.474$). Choices also varied as a function of expression ($F(3, 13) = 8.3$; $p < 0.001$, $\eta^2 = 0.026$). Lastly, an interaction between the two factors ($F(15, 195) = 14.93$; $p < 0.0001$, $\eta^2 = 0.192$; see Fig. 5B) revealed that observers' choices were not consistent across different facial expressions, suggesting that there was not one specific approximation that was perceived to be most similar to the original animation for all expressions.

3.2. Explaining observers' perceptual choices

We investigated whether the calculated choice probabilities based on the three objective measures of similarity could explain the behavioral choice pattern. First, we assessed their separate contribution using linear regression. To this end, we calculated three separate linear regression analyses, in each of which only one measure was used as predictor and the choice behavior was the predicted measure. All predictors could significantly explain the variance of the behavioral choices. The choice probabilities based on facial action activation were highly predictive and explained 59% of the variance ($r = 0.77$, $p < 0.0001$), indicating that the semantically meaningful facial actions capture spatio-temporal properties which are used for judging similarity between facial expressions. The choice probabilities based on physical similarity measures also explained variance of the behavioral choices, with more variance explained by optic flow ($r = 0.74$, 54% variance explained, $p < 0.0001$) than Gabor similarity ($r = 0.60$, 36% variance explained, $p < 0.01$). This finding suggests that motion cues measured by optic flow are closer to cues used for perceiving motion similarity than cues extracted by the biologically motivated V1-based Gabor similarity.

As the animation stimuli were based on time courses of facial action activation, we expected the choice probabilities based on facial action activation to be correlated with choice probabilities based on the physical characteristics of the animations (optic flow and Gabor similarity measures). We found that choice probabilities based on facial action similarity measure significantly correlated with choice probabilities based on both Gabor similarity measure ($r = 0.83$, $p < 0.001$) and on optic flow-based similarity measure ($r = 0.88$; $p < 0.001$). These results suggest that, unsurprisingly, the time courses of facial action activation used to create the animations capture physical properties of the resulting animation well. However, compared to image-based measurements, facial action time courses reflect these physical properties in a semantically meaningful and sparse representation.

3.3. Selecting the best fitting model

We investigated which model based on the three objective measures of similarity could best explain the behavioral choice pattern. In this analysis, the 24 behavioral choice proportions (6

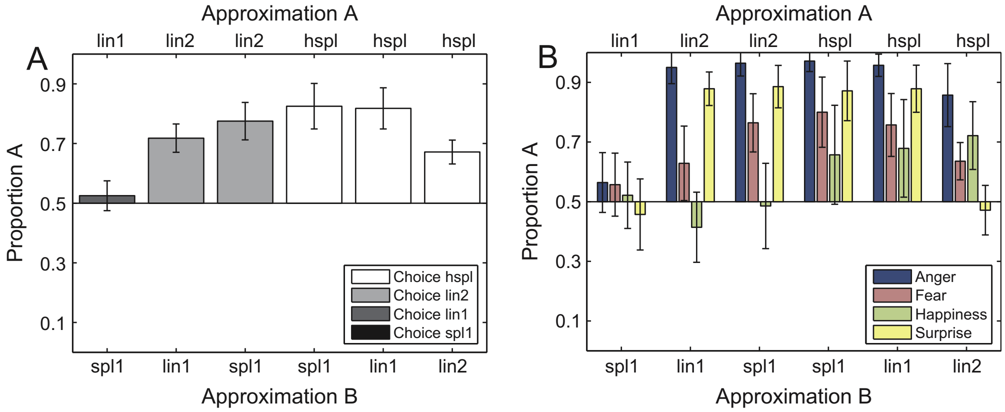

Fig. 5. Behavioral results. Proportion of approximation A choices for each pair of two approximations A and B (A), and as a function of facial expression (B). Error bars indicate 95% confidence interval (CI) in all plots. A choice proportion of 0.5 indicates that both approximations were chosen equally often (50%), a proportion of 1 indicates that approximation A was always chosen in this pair. The approximations can be ranked in decreasing order of observers' choice proportion: hspl > lin2 > lin1 = spl1.

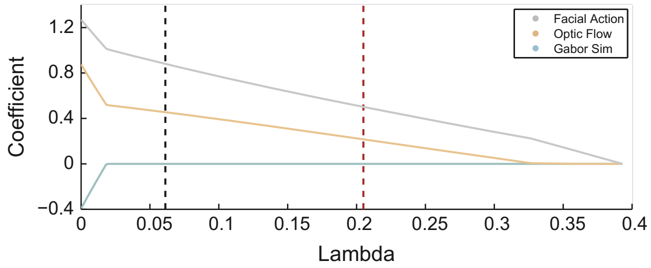

Fig. 6. The result of elastic net fitting using $\alpha = 0.5$ and 10-fold cross validation. Predictor coefficients (y-axis; facial action activation in grey, optic flow in light blue, Gabor similarity in orange) are plotted as a function of the λ regularization parameter (x-axis). The black dashed vertical line represents the λ value with minimal mean squared error (MSE). The λ value with minimal MSE plus one standard deviation is shown by the red dashed line. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

approximation pair types × 4 facial expressions) were the predicted measure and the 3 × 24 choice probabilities obtained from the three objective similarity measures were predictors. As the choice probabilities based on image cues were highly correlated with the choice probabilities based on facial action activation, we chose elastic net as regularized regression and variable selection method (Zou & Hastie, 2005). The results of the elastic net fitting using $\alpha = 0.5$ (we chose α to equally weight between lasso and ridge regression) and 10-fold cross validation are shown in Fig. 6. The standardized coefficients for each predictor (facial action activation in grey, optic flow in light blue, Gabor similarity in orange) are plotted as a function of λ. With increasing values of λ, elastic net retains optic flow and facial action activation as nonzero coefficients while the latter is set to zero last. Note that for $\lambda = 0$, the coefficients are equivalent to ordinary least squares regression. The dashed vertical lines represent λ with minimal mean prediction squared error ($\lambda = 0.06$, MSE = 0.03; dashed black line) and the MSE plus one standard deviation ($\lambda = 0.21$; red dashed line) as calculated by cross-validation. The best fitting model with $\lambda = 0.06$ explained 61% of the variance in the behavioral choices ($R = 0.78$; $F(13) = 33.77$; $p < 0.0001$). The fitted coefficients for this model were Beta = 0.88 for facial action activation, Beta = 0.46 for optic flow and Beta = 0 for Gabor similarity. The results suggest that the best model (i.e., the highest prediction accuracy with minimal predictors) to predict the behavioral choices is based on both facial action activation and optic flow.

4. Discussion

When we see a person smile, we see how local face parts such as the mouth and the eyebrows move naturally over time. In this study, we show that observers are highly sensitive to deviations from the natural motions of face parts. Using different approximations of these natural motions, we found that observers could not only discriminate between the different approximations, but that they were sensitive to the amount of information about the natural motion given by those approximations. The more information about the natural original motion (e.g., control points along the natural motion curve) was used to create an approximation, the more similar to the original animation it appeared. These results are consistent with previous findings (e.g., Cosker, Krumhuber, & Hilton, 2010; Curio et al., 2006; Furl et al., 2010; Pollick et al., 2003; Wallraven et al., 2008) and emphasize the importance of the quality of the approximation of facial motion in order to study perception of dynamic faces. Our results extend this previous work by showing a quantitative relationship between, on the one hand, the amount of information about natural motion contained in an approximation, and on the other hand, the perceived similarity between an approximation and a reproduction of natural motion.

It is important to notice that the perceived similarity of an approximation to the original animation varied with the type of facial expressions shown in our study. Facial movements may have different complexities in terms of their dynamics (e.g., linear versus nonlinear) depending on the type of facial expression (e.g., conversational or speech movements). Wallraven et al. (2008), for example, used different techniques to create animations of basic emotional and more subtle conversational expressions. The authors found a recognition advantage for expressions based on natural facial motion compared to linear morphs, with a stronger effect for conversational than for emotional expressions. Cosker, Krumhuber, and Hilton (2010) have also reported a perceptual advantage for natural facial motion dynamics compared to linear motion, and that advantage depended on the type of facial action. Consistent with these findings, in our study linear approximations (e.g., lin1, lin2) were chosen more often for the facial expressions fear, happiness and surprise compared to the facial expression anger which contained speech movements. This suggests that, perhaps unsurprisingly, different kinds of approximations work best for different facial expressions. Testing a wide range of expressions performed by many actors could allow formulation of suggestions about which approximations best reproduce the spatio-temporal dynamics of specific facial expressions. However such an undertaking was beyond the scope of the current study.

Another aim of the experiment was to investigate which characteristic of facial motion observers used to judge the similarity between the original animation and the approximations. This will help understanding which aspect of facial motions we are most sensitive to and therefore needs to be preserved to create adequate approximations. We computed choice probabilities for three similarity measures based on facial action activation, optic flow and Gabor similarity and compared these objective measures to the pattern of perceptual similarity. We found that the similarity measure based on facial action activation best accounted for the variance in the choice behavior. This similarity measure is based on a high-level cue and represents the similarity of face deformations over time (e.g., the way the eyebrows move in two animations). To our knowledge, this is the first demonstration that objective similarity measures can predict perceptual similarity of facial movements.

For static faces, Gabor jets which model simple and complex cells in early stages of visual processing have successfully predicted perceived similarity between facial identities (Yue et al., 2012) and facial expressions (e.g., Lyons et al., 1998; Susskind et al., 2007). Here, we adapted this image-based measure to motion stimuli. We found that Gabor similarity explained the least variance in the behavioral choice pattern among the three similarity measures, with much lower predictive power than reported for static faces (see Yue et al., 2012). Furthermore, the best fitting model did not include Gabor similarity as a predictor. There are two possible explanations for this finding. On the one hand, compared to real faces, our stimuli lack high-spatial frequency contents both in the spatial and temporal domains (e.g., freckles or skin wrinkling during expressions).This could potentially reduce the efficacy of the Gabor similarity which is based on the available frequency content of the image sequences as measured by Gabor filters. However, Gabor filters based on known properties of neurons found in early visual cortex (e.g., Jones & Palmer, 1987) remove these high spatial frequencies. In line with this, psychophysical studies found that human observers mainly use mid spatial frequencies to recognize faces (e.g., Costen, Parker, & Craw, 1996; Näsänen, 1999) and neuroimaging studies have shown that low spatial frequency information play an important role in the brain responses to fearful stimuli (Vuilleumier et al 2003; Vlamings et al, 2009). It is beyond the scope of our current study to determine the extent to which high spatial frequencies may contribute to the similarity judgments for dynamic expressions. On the other hand, the low predictive performance in our experiment could be due to differences in static versus dynamic faces. In line with the latter possibility, neurophysiological and psychophysical studies (e.g., Duffy & Wurtz, 1991; Morrone, Burr, & Vaina, 1995; Wurtz et al., 1990) reported that motion stimuli are processed at later stages in the visual hierarchy that are more responsive to optic flow features than to Gabor measures.

Given the importance of optic flow for the processing of natural motion stimuli (e.g., Bartels, Zeki, & Logothetis, 2008), we hypothesized that objective similarity between stimuli measured by optic flow might explain the behavioral similarity judgments we observed. We indeed found that optic flow was an important predictor of the perceptual choices. However, the contribution to the behavioral variance by optic flow was smaller than for facial action activation. This finding suggests that the overall similarity in low-level motion might capture subtle differences in face motion stimuli which are less relevant for observers' decisions than the spatio-temporal dynamics of local face parts. In the future, it would be interesting to investigate how responses of higher-level models of biological motion which combine Gabor filters and optic flow measures (e.g., Giese & Poggio, 2003) relates to human responses to facial motion.

We used animations of synthetic faces as stimuli in our study. While videos of faces capture much of the visual experience of perceiving real-life faces, it is difficult to precisely quantify the spatio-temporal information in videos, let alone to systematically manipulate this information to address the questions raised in this study. Still the question arises whether the reported results can be generalized to faces in real life. Evidence from psychophysical and neuroimaging studies investigating static (e.g., Dyck et al., 2008; Ishai, Schmidt, & Boesiger, 2005; Wilson, Loffler, & Wilkinson, 2002) and dynamic face processing (e.g., Mar et al., 2007; McDonnell, Breidt, & Bülthoff, 2012; Moser et al., 2007) indicates that synthetic faces are processed by similar mechanisms as natural faces. However, contradictory results have also been reported (e.g., Han et al., 2005; Moser et al., 2007). These differences in results may be due to differences in the naturalness of the synthetic face stimuli across these studies, highlighting the need to capture natural facial motions with a high degree of fidelity. As we have strived to generate avatars with motion as natural as possible, we believe that our results would generalize to real-life faces if the same tests could be run under controlled conditions. With the techniques available today, we do not expect any of our similarity measures to account better for the perceived similarity between videos than perceived similarity between our animations, since videos contain much more irrelevant information that would need to be discounted for a sensitive analysis (e.g., head movements, different backgrounds, hair).

Our results suggest that observers extract spatio-temporal characteristics of facial motion stimuli and make their judgments based on a sparse but semantically meaningful representation rather than on low-level physical properties of the stimuli. Given the social importance of facial motion, it is likely that despite the fact that observers were asked to perform a simple similarity judgment task, they automatically extracted and analyzed the semantic content of facial motion. To further test this hypothesis in a future experiment, one could use nonsense facial motion stimuli (e.g., by scrambling the frames) or inverted face motion stimuli to investigate whether image-based measures better explain the perceived similarity of such stimuli.

5. Conclusions

We draw three main conclusions from our study. First, our results demonstrate how exquisitely sensitive the human perceptual system is to degradations of the spatio-temporal properties of natural facial motion: observers discriminated the subtle differences between the different approximations and preferred animations containing more information about the natural facial motion dynamics. Second, the perceived similarity of an approximation depended on the type of facial expression, which shows that the use of simple approximations, such as linear interpolations, is not appropriate to reproduce all types of facial expressions. Third, our approach allowed a quantitative explanation of observers' perceptual choices revealing the importance of high-level cues in the processing of facial motion. These findings suggest that to understand facial motion processing, we need more advanced analyses than for static images, going beyond the analysis of image-based properties. These conclusions validate attempts to capture and render semantically meaningful information in facial motion. Using better approximations will open the door to in-depth studies of how humans judge and perceive natural facial motion, what information in facial motion they rely on when performing different tasks involving facial motion, and what neural mechanisms underlie the processing of these different kinds of information. We believe that such methods are essential for a systematic, quantitative analysis of the incredible amount of information that can be conveyed by facial motion and have important implications for theories and models of facial motion perception.

Acknowledgments

The project was supported by Perceptual Graphics Project PAK 38 CU 149/1-1/2 funded by the Deutsche Forschungsgemeinschaft (DFG) and EU Project EC FP7-ICT-249858 TANGO. Further, we would like to thank Mario Kleiner for his valuable input and support by implementing the motion analysis system to acquire the stimuli for this study; Prof. Martin Giese for advice on the experimental design; Dr. Stephan de la Rosa for advice on statistics; and Prof. Heinrich H. Bülthoff for support. There was no conflict of interest.

Appendix A

A.1. Facial animation procedure

A.1.1. Acquiring and post-processing facial motion capture data

We captured facial movements of a non-professional female actor using a seven-cameras optical motion capture system (NaturalPoint Optitrack) running at 100 Hz, and OptiTrack Expression software (version 1.8.0, NaturalPoint, Inc., Corvallis, OR, USA). The positions of 41 reflective markers (37 markers on the actor's face and 4 markers on a headband, see left image of Fig. 2A) were tracked by six infra-red cameras, while an additional synchronized scene camera recorded a gray-scale video of the actor performing the facial movements (see Fig. 2A).

At the beginning of the motion capture session, 30 facial actions were captured from the actor. These actions are listed in Table A1. Although the selected facial actions were mainly based on FACS, the actor and the instructor were not certified FACS experts. The actor received verbal instructions for each facial action and was instructed to perform the movement as intensely and as clearly as possible, with as little co-activation as possible in other facial regions corresponding to other facial actions. From these recordings, we manually selected the frame displaying the maximum intensity for each of the 30 facial actions (e.g., eyebrow raising: when the eyebrows were maximally raised).

The actor then performed four emotional facial expressions (anger, fear, happiness and surprise) that involved a wide range of facial motion distributed across different regions of the face (e.g., mouth, eyebrows). To induce the expressions as naturally as

Table A1
The 30 recorded facial actions and their semantic meaning specifying which part of the face moves and in which way.

Facial action number	Semantic	Facial action number	Semantic
1	Neutral	16	Lips open
2	Eyebrows lowered	17	Mouth open
3	Eyebrows raised	18	Mouth wide open
4	Eyes wide open	19	Lower lip down
5	Eyes squint	20	Mouth stretched
6	Eyes closed	21	Dimpler
7	Nose wrinkled	22	Smile, mouth closed
8	Upper lip up	23	Right lips up
9	Upper lip up, teeth showed	24	Left lips up
10	Right mouth corner up	25	Smile, mouth open
11	Left mouth corner up	26	Lip corners up
12	Chin up	27	Pucker
13	Lip corners down	28	Lips funnel
14	Right lip corner down	29	Lips tight
15	Left lip corner down	30	Lips pressed

possible, we used a "method-acting protocol" in which the actor is verbally given a particular background scenario designed to elicit the desired facial expression (see Kaulard et al., 2012). Three of the recorded expressions (fear, happiness, and surprise) started from a neutral expression and proceeded to the target expression. For the facial expression anger, we chose a background scenario leading up to an anger expression that contained visual speech (i.e., "speak angrily to someone"). This facial expression increased the range of spatio-temporal profiles of facial motion tested in our study.

Using OptiTrack Arena Expression software, the facial motion data was post-processed as follows. First, the markers were labeled according to their position on the face. Triangulation errors were manually removed from the marker position time courses and rarely occurring gaps in time courses were filled in by cubic spline interpolation. Second, rigid head motion was removed from the motion capture data by aligning the four recorded head markers to their positions at the start of the motion capture. Third, the remaining non-rigid component of the motion data was loaded into Matlab (version R2010b, The MathWorks, Inc., Natick, MA, USA) using the MoCap Toolbox (Burger & Toiviainen, 2013), and filtered with a low pass filter (digital Butterworth filter, cut-off frequency = 10 Hz, order = 2) to reduce jitter in the marker time courses.

A.1.2. Analyzing facial motion capture data

The post-processed motion capture data for each expression were decomposed into time courses of the constituent facial actions (see Table A1). These time courses were obtained by linearly combining the marker positions of the set of static facial actions to the marker positions at each time point of the recorded expression (see Curio et al., 2006 for further details). The activation for each facial action at each time point ranged from 0 (no activation) to 1 (maximum intensity). We first identified the peak of the facial expression by summing all 30 facial action activations at each time point and selecting the frame that had the largest sum. For each facial action, we then selected sequences of 1s duration (100 frames) that contained this peak at the end of the sequence.

A.1.3. Facial motion retargeting

The time courses were transferred onto a female 3D head model designed in Poser 8 (SmithMicro, Inc., Watsonville, CA, USA). We manually altered the model using Poser's in-built animation parameters to create 30 facial action "shapes" corresponding to the 30 facial actions performed by the actor (at their maximum intensity). Fig. 2B shows the facial actions "neutral" (facial action 1), "mouth open" (facial action 17) and "smile" (facial action 22) as motion capture data (middle) and the corresponding facial action shape (right). Each of the facial action shapes was exported in OBJ format from Poser into the animation software 3ds Max 2012 (Autodesk, Inc., San Rafael, CA, USA). The 3D coordinates of all the facial action shapes were in correspondence (e.g., the tip of the nose is represented by the same vertex across all shapes). This correspondence allowed us to take a weighted linear combination (i.e., morph) of the neutral action shape with all the other action shapes (sometimes referred to as weighted morphing). For example, increasing the weight of any particular action shape (e.g., mouth open) adds increasing amounts of that action shape to the neutral action shape. To synthesize a complex facial expression, the facial action shapes were weighted by their activation at each frame (time step) and combined with the neutral action shape. This combination was done in 3ds Max.

References

Ambadar, Z., Schooler, J. W., & Cohn, J. F. (2005). Deciphering the enigmatic face. The importance of facial dynamics in interpreting subtle facial expressions. *Psychological Science, 16*(5), 403–410.

Bartels, A., Zeki, S., & Logothetis, N. K. (2008). Natural vision reveals regional specialization to local motion and to contrast-invariant, global flow in the human brain. *Cerebral Cortex, 18*(3), 705–717. http://dx.doi.org/10.1093/cercor/bhm107.

Bernstein, L. E., Demorest, M. E., & Tucker, P. E. (2000). Speech perception without hearing. *Perception & Psychophysics, 62*(2), 233–252.

Bruce, V., & Young, A. W. (1986). Understanding face recognition. *British Journal of Psychology, 77*(3), 305–327.

Burger, B., & Toiviainen, P. (2013). MoCap Toolbox-A Matlab toolbox for computational analysis of movement data. In R. Bresin (Ed.), *Proceedings of the sound and music computing conference 2013* (pp. 172–178). Stockholm, Sweden: KTH Royal Institute of Technology.

Calder, A. J., & Young, A. W. (2005). Understanding the recognition of facial identity and facial expression. *Nature Reviews Neuroscience, 6*(8), 641–651. http://dx.doi.org/10.1038/nrn1724.

Cosker, D., Krumhuber, E., & Hilton, A. (2010). Perception of linear and nonlinear motion properties using a FACS validated 3D facial model. In *Proceedings of the 7th symposium on applied perception in graphics and visualization* (pp. 101–108). New York, NY: ACM Press.

Costen, N., Parker, D., & Craw, I. (1996). Effects of high-pass and low-pass spatial filtering on face identification. *Perception & Psychophysics, 58*(4), 602–612.

Cunningham, D. W., & Wallraven, C. (2009). The interaction between motion and form in expression recognition. In *Proceedings of the 6th symposium on applied perception in graphics and visualization* (pp. 41–44). New York, NY: ACM Press.

Curio, C., Breidt, M., Kleiner, M., Vuong, Q. C., Giese, M. A., & Bülthoff, H. H. (2006). Semantic 3D motion retargeting for facial animation. In *Proceedings of the 3rd symposium on applied perception in graphics and visualization* (pp. 77–84). New York, NY: ACM Press.

Duffy, C., & Wurtz, R. (1991). Sensitivity of MST neurons to optic flow stimuli. I. A continuum of response selectivity to large-field stimuli. *Journal of Neurophysiology, 65*(6), 1329–1345.

Dyck, M., Winbeck, M., Leiberg, S., Chen, Y., Gur, R. C., & Mathiak, K. (2008). Recognition profile of emotions in natural and virtual faces. *PLoS One, 3*(11), e3628. http://dx.doi.org/10.1371/journal.pone.0003628.

Ekman, P., & Friesen, W. V. (1978). *Facial action coding system: A technique for the measurement of facial movement*. Palo Alto, CA: Consulting Psychologists Press.

Ekman, P., Friesen, W. V., & Hager, J. C. (2002). *The facial action coding system*. Salt Lake City, UT: A Human Face.

Furl, N., van Rijsbergen, N. J., Kiebel, S. J., Friston, K. J., Treves, A., & Dolan, R. J. (2010). Modulation of perception and brain activity by predictable trajectories of facial expressions. *Cerebral Cortex, 20*(3), 694–703. http://dx.doi.org/10.1093/cercor/bhp140.

Giese, M. A., & Poggio, T. (2003). Neural mechanisms for the recognition of biological movements. *Nature Reviews Neuroscience, 4*(3), 179–192. http://dx.doi.org/10.1038/nrn1057.

Han, S., Jiang, Y., Humphreys, G. W., Zhou, T., & Cai, P. (2005). Distinct neural substrates for the perception of real and virtual visual worlds. *NeuroImage, 24*(3), 928–935. http://dx.doi.org/10.1016/j.neuroimage.2004.09.046.

Haxby, J. V., Hoffman, E. A., & Gobbini, M. I. (2002). Human neural systems for face recognition and social communication. *Biological Psychiatry, 51*(1), 59–67.

Hoerl, A. E., & Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics, 12*(1), 55–67.

Horn, B. K. P., & Schunck, B. G. (1981). Determining optical flow. *Artificial Intelligence, 17*(1), 185–203. http://dx.doi.org/10.1016/0004-3702(81)90024-2.

Ishai, A., Schmidt, C. F., & Boesiger, P. (2005). Face perception is mediated by a distributed cortical network. *Brain Research Bulletin, 67*(1), 87–93. http://dx.doi.org/10.1016/j.brainresbull.2005.05.027.

Jack, R. E., Garrod, O. G. B., Yu, H., Caldara, R., & Schyns, P. G. (2012). Facial expressions of emotion are not culturally universal. *Proceedings of the National Academy of Sciences, 109*(19), 7241–7244. http://dx.doi.org/10.1073/pnas.1200155109.

Jones, J. P., & Palmer, L. A. (1987). An evaluation of the two-dimensional Gabor filter model of simple receptive fields in cat striate cortex. *Journal of Neurophysiology, 58*(6), 1233–1258.

Kamachi, M., Bruce, V., Mukaida, S., Gyoba, J., Yoshikawa, S., & Akamatsu, S. (2001). Dynamic properties influence the perception of facial expressions. *Perception, 30*, 875–887. http://dx.doi.org/10.1068/p3131.

Kaulard, K., Cunningham, D. W., Bülthoff, H. H., & Wallraven, C. (2012). The MPI Facial Expression Database—A validated database of emotional and conversational facial expressions. *PLoS One, 7*(3), e32321. http://dx.doi.org/10.1371/journal.pone.0032321.

Kleiner, M. (2010). Visual stimulus timing precision in Psychtoolbox-3: Tests, pitfalls and solutions. *Perception, 39*(ECVP Abstract Supplement), 189.

Krumhuber, E. G., Kappas, A., & Manstead, A. S. R. (2013). Effects of dynamic aspects of facial expressions: A review. *Emotion Review, 5*(1), 41–46. http://dx.doi.org/10.1177/1754073912451349.

Ku, J., Jang, H., Kim, K., Kim, J., Park, S., Lee, J., et al. (2005). Experimental results of affective valence and arousal to avatar's facial expressions. *CyberPsychology & Behavior, 8*(5), 493–503.

LaBar, K. S., Crupain, M. J., Voyvodic, J. T., & McCarthy, G. (2003). Dynamic perception of facial affect and identity in the human brain. *Cerebral Cortex, 13*(10), 1023–1033.

Lades, M., Vorbrüggen, J. C., Buhmann, J., Lange, J., von der Malsburg, C., Würtz, R. P., et al. (1993). Distortion invariant object recognition in the dynamic link architecture. *IEEE Transactions on Computers, 42*(3), 300–311.

Luce, R. D. (1959). *Individual choice behavior: A theoretical analysis*. New York, NY: John Wiley and Sons.

Lyons, M., Akamatsu, S., Kamachi, M., & Gyoba, J. (1998). Coding facial expressions with Gabor wavelets. In *Proceedings of the 3rd international conference on automatic face and gesture recognition* (pp. 200–205). Washington, DC: IEEE Computer Society.

Mar, R. A., Kelley, W. M., Heatherton, T. F., & Macrae, C. N. (2007). Detecting agency from the biological motion of veridical vs animated agents. *Social Cognitive and Affective Neuroscience, 2*(3), 199–205. http://dx.doi.org/10.1093/scan/nsm011.

McDonnell, R., Breidt, M., & Bülthoff, H. H. (2012). Render me real? Investigating the effect of render style on the perception of animated virtual humans. *ACM Transactions on Graphics, 31*(4), 91.

Morrone, M. C., Burr, D. C., & Vaina, L. M. (1995). Two stages of visual processing for radial and circular motion. *Nature, 376*, 507–509.

Moser, E., Derntl, B., Robinson, S., Fink, B., Gur, R. C., & Grammer, K. (2007). Amygdala activation at 3T in response to human and avatar facial expressions of emotions. *Journal of Neuroscience Methods, 161*(1), 126–133. http://dx.doi.org/10.1016/j.jneumeth.2006.10.016.

Näsänen, R. (1999). Spatial frequency bandwidth used in the recognition of facial images. *Vision Research, 39*(23). 3824–33.

Pollick, F. E., Hill, H., Calder, A. J., & Paterson, H. (2003). Recognising facial expression from spatially and temporally modified movements. *Perception, 32*(7), 813–826. http://dx.doi.org/10.1068/p3319.

Rhodes, G. (1988). Looking at faces: First-order and second-order features as determinants of facial appearance. *Perception, 17*, 43–63.

Roesch, E. B., Tamarit, L., Reveret, L., Grandjean, D., Sander, D., & Scherer, K. R. (2011). FACSGen: A tool to synthesize emotional facial expressions through systematic manipulation of facial action units. *Journal of Nonverbal Behavior, 35*(1), 1–16. http://dx.doi.org/10.1007/s10919-010-0095-9.

Rosenblum, L. D., Yakel, D. A., Baseer, N., Panchal, A., Nodarse, B. C., & Niehus, R. P. (2002). Visual speech information for face recognition. *Perception & Psychophysics, 64*(2), 220–229.

Sarkheil, P., Goebel, R., Schneider, F., & Mathiak, K. (2012). Emotion unfolded by motion: A role for parietal lobe in decoding dynamic facial expressions. *Social Cognitive and Affective Neuroscience*. http://dx.doi.org/10.1093/scan/nss092.

Sato, W., & Yoshikawa, S. (2007). Enhanced experience of emotional arousal in response to dynamic facial expressions. *Journal of Nonverbal Behavior, 31*(2), 119–135. http://dx.doi.org/10.1007/s10919-007-0025-7.

Steyvers, M., & Busey, T. (2000). Predicting similarity ratings to faces using physical descriptions. In M. Wenger & J. Townsend (Eds.), *Computational geometric and process perspectives on facial cognition: Contexts and challenges* (pp. 115–146). Lawrence Erlbaum.

Susskind, J. M., Littlewort, G., Bartlett, M. S., Movellan, J., & Anderson, A. K. (2007). Human and computer recognition of facial expressions of emotion. *Neuropsychologia, 45*(1), 152–162. http://dx.doi.org/10.1016/j.neuropsychologia.2006.05.001.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological), 58*(1), 267–288.

Vlamings, P. H. J. M., Goffaux, V., & Kemner, C. (2009). Is the early modulation of brain activity by fearful facial expressions primarily mediated by coarse low spatial frequency information? *Journal of Vision, 9*(5), 1–13.

Vuilleumier, P., Armony, J. L., Driver, J., & Dolan, R. J. (2003). Distinct spatial frequency sensitivities for processing faces and emotional expressions. *Nature Neuroscience, 6*(6), 624–631.

Wallraven, C., Breidt, M., Cunningham, D. W., & Bülthoff, H. H. (2008). Evaluating the perceptual realism of animated facial expressions. *ACM Transactions on Applied Perception, 4*(4), 1–20. http://dx.doi.org/10.1145/1278760.1278764.

Wilson, H. R., Loffler, G., & Wilkinson, F. (2002). Synthetic faces, face cubes, and the geometry of face space. *Vision Research, 42*(27), 2909–2923.

Wurtz, R., Yamasaki, D., Duffy, C. J., & Roy, J.-P. (1990). Functional specialization for visual motion processing in primate cerebral cortex. *Cold Spring Harbor Symposia on Quantitative Biology, 55*, 717–727.

Xu, X., & Biederman, I. (2010). Loci of the release from fMRI adaptation for changes in facial expression, identity, and viewpoint. *Journal of Vision, 10*, 1–13. http://dx.doi.org/10.1167/10.14.36.Introduction.

Yu, H., Garrod, O. G. B., & Schyns, P. G. (2012). Perception-driven facial expression synthesis. *Computers & Graphics, 36*(3), 152–162. http://dx.doi.org/10.1016/j.cag.2011.12.002.

Yue, X., Biederman, I., Mangini, M. C., von der Malsburg, C., & Amir, O. (2012). Predicting the psychophysical similarity of faces and non-face complex shapes by image-based measures. *Vision Research, 55*, 41–46. http://dx.doi.org/10.1016/j.visres.2011.12.012.

Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology), 67*(2), 301–320. http://dx.doi.org/10.1111/j.1467-9868.2005.00503.x.