Perceptual integration of kinematic components in the recognition of emotional facial expressions

Enrico Chiovetto*
Section Computational Sensomotorics, Department of Cognitive Neurology, University Clinic, Tübigen, Germany

Cristóbal Curio*
Cognitive Systems Group, Department of Computer Science, Reutlingen University, Germany
Max Planck Institute for Biological Cybernetics, Tübingen, Germany

Dominik Endres†
Theoretical Neuroscience Group, Dept. of Psychology, UMR Marburg, Germany
Section Computational Sensomotorics, Department of Cognitive Neurology, UKT, Germany

Martin Giese†
Section Computational Sensomotorics, Department of Cognitive Neurology, University Clinic Tübigen, Germany

According to a long-standing hypothesis in motor control, complex body motion is organized in terms of movement primitives, reducing massively the dimensionality of the underlying control problems. For body movements, this low-dimensional organization has been convincingly demonstrated by the learning of low-dimensional representations from kinematic and EMG data. In contrast, the effective dimensionality of dynamic facial expressions is unknown, and dominant analysis approaches have been based on heuristically defined facial "action units," which reflect contributions of individual face muscles. We determined the effective dimensionality of dynamic facial expressions by learning of a low-dimensional model from 11 facial expressions. We found an amazingly low dimensionality with only two movement primitives being sufficient to simulate these dynamic expressions with high accuracy. This low dimensionality is confirmed statistically, by Bayesian model comparison of models with different numbers of primitives, and by a psychophysical experiment that demonstrates that expressions, simulated with only two primitives, are indistinguishable from natural ones. In addition, we find statistically optimal integration of the emotion information specified by these primitives in visual perception. Taken together, our results indicate that facial expressions might be controlled by a very small number of independent control units, permitting very low-dimensional parametrization of the associated facial expression.

Introduction

It is a long standing hypothesis in human motor control that the central nervous system (CNS) relies on a low-dimensional organization to produce complex and flexible movements (Bizzi, Cheung, d'Avella, Saltiel, & Tresch, 2008; Flash & Hochner, 2005). According to this hypothesis, a low number of invariant modules (usually referred to as motor primitives or synergies) are linearly combined to generate the desired motor behavior. Although several definitions of primitives have been given in the literature, each one relying on a different mathematical model, they can be divided in two main categories, namely spatial and temporal synergies. Spatial muscle synergies have been, for instance, defined as groups of muscles covarying together in time (Cheung et al., 2009; Ting & Macpherson, 2005; Torres-Oviedo & Ting, 2007), while temporal primitives have instead been described, in the muscle space, as temporal profiles of muscle activations (Chiovetto, Berret, & Pozzo, 2010; Dominici et al., 2011; Ivanenko, Poppele,

Citation: Chiovetto, E., Curio, C., Endres, D., & Giese, M. (2018). Perceptual integration of kinematic components in the recognition of emotional facial expressions. *Journal of Vision*, *18*(4):13, 1–19, https://doi.org/10.1167/18.4.13.

https://doi.org/10.1167/18.4.13 Received October 16, 2017; published April 13, 2018 ISSN 1534-7362

& Lacquaniti, 2004). Similarly, in the kinematic and kinetic space, primitives have been defined as temporal patterns of degrees-of-freedom covariation (Berret, Bonnetblanc, Papaxanthis, & Pozzo, 2009; Chiovetto & Giese, 2013; Kaminski, 2007). The relevance of motor synergies has been widely demonstrated by the analyzes accomplished on the kinematic and electromyographic data associated with a large set of movements including, for instance, hand movements (Santello, Flanders, & Soechting, 1998, 2002), arm reaching movements (d'Avella, Portone, Fernandez, & Lacquaniti, 2006), or more complex whole-body motor behaviors (Chiovetto et al., 2010; Chiovetto & Giese, 2013; Ivanenko et al., 2004). There is, however, still another important class of complex movements that has not been studied in this context: facial expressions. In contrast to many other goal-oriented movements, dynamic facial expressions are interesting because they form a crucial signal for social interaction in primates, e.g., conveying emotional states (Niedenthal, Mermillod, Maringer, & Hess, 2010). The main goal of our study was to investigate the existence of a possible low-dimensional organization underlying the generation of emotional facial expressions and to understand how the primitives at the base of such a synergistic organization contribute to the generation of an emotional percept.

A considerable amount of effort has been spent on finding good spatial synergies for facial movements. Early approaches were based on principal component analysis (PCA; Kalberer & Gool, 2001; Kshirsagar, Molet, & Magnenat-Thalmann, 2001). However, it was soon recognized that facial movements encode relevant information in higher order statistical dependencies, and hence independent component analysis (ICA) was employed for extracting such synergies, both for animation (Mueller, Kalberer, Proesmans, & Gool, 2005) and for recognition (Bartlett, Movellan, & Sejnowski, 2002).

Temporal primitives for facial movements are much less studied, but the relevance of time for facial expression processing has been noted (Bartlett, 2010). Most animation approaches encode the dynamics by storing key-poses from spatial synergies with spline interpolation of different orders between these poses (Dobs et al., 2014; Ezzat, Geiger, & Poggio, 2002; Jack, Garrod, Yu, Caldara, & Schyns, 2012; Kalberer & Gool, 2001; Kshirsagar et al., 2001; Mueller et al., 2005). For recognition, hand-crafted spatio-temporal Gabor filters have been shown to be slightly superior to purely spatial Gabor filters (Bartlett et al., 2006; Littlewort et al., 2011; Wu, Bartlett, & Movellan, 2010). However, with the notable exception of Delis et al. (2016), there is very little work attempting to find temporal synergies in facial movements.

The most well-established approach used in psychology to taxonomize human facial expression is the so-called FACS (Facial Action Coding System) developed by Ekman and Friesen in the late 70s (Ekman & Friesen, 1978). In this framework, every facial expression, including emotional expressions (Ekman et al., 1987; Ekman, Friesen, & Ellsworth, 1972), is decomposed and analyzed in terms 44 elementary movements, referred to as Action Units (AUs), each of which is generated by the action of one or more facial muscles. AUs can therefore be seen as spatial primitives (similar to the ones defined in Chiovetto, Berret, Delis, Panzeri, & Pozzo, 2013; Ting & Macpherson, 2005) that, after being scaled in time, can be superimposed to generate the desired facial expression. Despite of its widespread use, however, we consider FACS an incomplete description of facial movements, because it mainly provides a spatial description of facial expressions, ignoring the organization in the temporal domain. The importance of dynamics in the perception of emotional whole-body movements and temporal expressions has been indeed demonstrated by several recent studies (Roether, Omlor, Christensen, & Giese, 2009; Reinl & Bartels, 2014). Consequently, we expected similar results for facial movements. In addition, the dimensionality provided by FACS might still be substantially higher than the dimensionality of the space of dynamic facial expressions, when such a space is parametrized in a highly efficient manner. The dimensionality of the FACS was originally derived based on anatomical constraints, motivated by the isolated activation of individual facial muscles. In addition, the original FACS aimed to promote a universal and interculturally valid system of coding elements, and not to the minimal parametrization of typical expressions in natural settings. The goal of this study was to find automatically the dimensionality of a group of typically occurring facial expressions by parametrizing them in terms of a minimum number of learned movement primitives. In addition, we wanted to study how such primitives contribute to the perception of emotional styles from facial actions.

We addressed these questions by combining a novel machine learning algorithm for the learning of movement primitives (Chiovetto, d'Avella, & Giese, 2016; Chiovetto & Giese, 2013) and a motion-retargeting system for 3D facial animation (Curio et al., 2006; Curio et al., 2010), which is based on 3D models of FACS. Exploiting this parametrization of the 3D structures of facial expressions, we determined a minimum number of spatio-temporal primitives that capture the major part of the AU variance associated with emotional facial expressions. The estimated primitives were used to generate stimuli for a psychophysical experiment assessing classification rates and emotional expressiveness ratings for stimuli containing combinations of the extracted components. We investigated how the emotional content carried by the

different components extracted from the facial movement is integrated in perception. Specifically, we investigated different cue fusion models, including Bayesian ones inspired by Ernst and Banks (2002) and Körding et al. (2007), trying to account for the obtained experimental results. We found that only two spatio-temporal primitives (in contrast to the 44 AUs) were sufficient for an almost perfect reconstruction of the original expressions. In addition we found that each component contributed significantly to all tested expressions. This small number of necessary primitives is confirmed by psychophysical experiments and by results derived by Bayesian model comparison (Endres, Chiovetto, & Giese, 2013). This finding implies that the efficient dimensionality of the space of dynamic facial expressions, especially for the ones that are frequently used, might be much smaller than what is suggested by the FACS.

In the following, we first describe the face animation system that we used to generate our experimental stimuli. Then we describe the computational methods that we used to estimate a minimal number of spatio-temporal movement primitives to approximate our data. This is followed by a description of the visual stimulus generation, the psychophysical experiments, and statistical methods applied for their analysis. Finally we also describe different cue fusion models that were fitted to our data.

Methods

Facial animation platform

To identify the AU time-varying activation coefficients (weights) associated with several emotional facial expressions and to generate the visual stimuli for the psychophysical experiments, we used a facial animation system that was previously developed (Curio et al., 2006; Curio, Giese, Breidt, Kleiner, & Bülthoff, 2008). The method is based on a morphable face model generated from 3D face scans. Our animation method is characterized by two main steps. The first step consists on approximating the face movements, recorded with a motion capture system, by a linear superposition of static peak frames of 3D facial AUs according to the FACS. For each time step t, a vector of morphing weights $\mathbf{W}^*(t)$ was identified, which specifies how much the individual AUs contribute to the approximation of the present expressions. The second step of the method synthesizes dynamic facial expressions by linearly combining 3D scans of the AU peak frames with the same vector $\mathbf{W}^*(t)$, to create photo-realistic animations. In the following, we explain these two steps briefly (see Curio et al. 2006; Curio et al., 2008; de la Rosa, Giese, Bülthoff, & Curio, 2013) for more details. We did not model any spatial rotation or translation of the head, and all the expressions were displayed with the same orientation of the head in space.

AU motion capture model fitting

To compute the temporal evolutions of the morphing weights of the AUs for a specific dynamic expression, a standard least-square optimization problem was solved. In brief, for each time instant t, the error between the kinematic vector of the facial expression and the kinematic vector resulting by the superposition of the vector representing the AUs (that were previously recorded) was minimized over each weighting coefficient $W_i(t)$. See Figure 1, left side. In the following, we indicate by $M_E(t)$ the kinematic data associated with the facial expression and by $\tilde{M}_E(t)$ the approximated facial kinematics. Let M_N indicate the vector of the static neutral reference face and $M_{AU,i}(t)$ the kinematic vector associated with the *ith* Action Unit. The weights were estimated by solving the nonnegative least-square problem

$$\begin{aligned} &\underset{W_i(t)}{\text{minimize}} \left\| M_E(t) - \tilde{M}_E(t) \right\|^2 \\ &= \left\| M_E(t) - \left(M_N + \sum_{i=1}^{N} W_i(t) M_{AU,i}(t) \right) \right\|^2 \\ &\text{subject to } W_i(t) \geq 0,\ i = 1, \ldots, N. \end{aligned} \quad (1)$$

for every time step. N indicates the total number of AU (in our case $N = 17$). The computed optimal weight vectors are denoted by $\mathbf{W}^*(t)$. The nonnegativity constraint is justified by the fact that AUs, as equivalent of muscle activations, should always be nonnegative. The time course of the AU activation coefficients were obtained solving this optimization problem separately for all time steps t instead of solving the optimization over all time steps simultaneously. Both optimization approaches provide, however, equivalent solutions given the linearity of the problem. Figure 2 shows examples of these time courses for two different emotional expressions, respectively disgust and happy.

Synthesis of facial expression animation

The optimal AU vectors $\mathbf{W}^*(t)$ were used to generate photorealistic dynamic facial expressions. This was achieved by superimposing, at each time instant t, the 3D scans of the shapes of the single AUs modulated by the morphing weights in $W_i^*(t)$. See Figure 1, right. The shapes of the single action units were recorded from one single actor that executed the action units individually. As a consequence, all the facial expres-

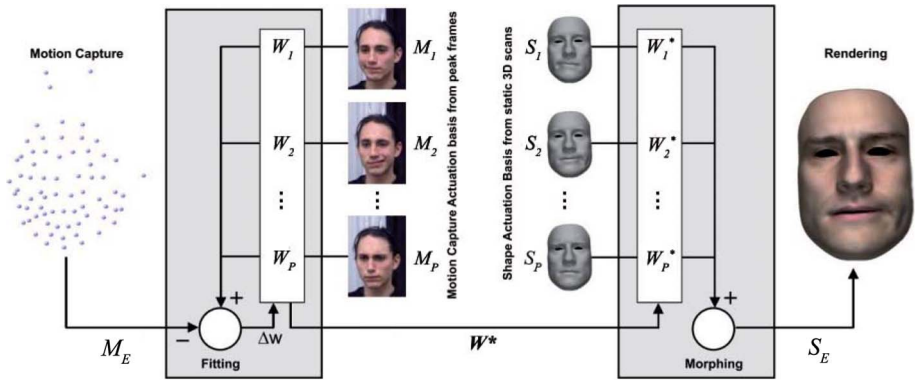

Figure 1. Overview of our facial animation system: Motion capture data of the expression M_E is approximated by a superposition of action units M_i with the mixing weights W_i. The estimated linear weights **W*** define the linear weights of the 3D shapes S_i of individual AUs, which are linearly combined together to obtain the animated face shape S. The figure was adapted from (Curio et al., 2006), according to the ACM (Association for Computing Machinery) copyright policy and rules.

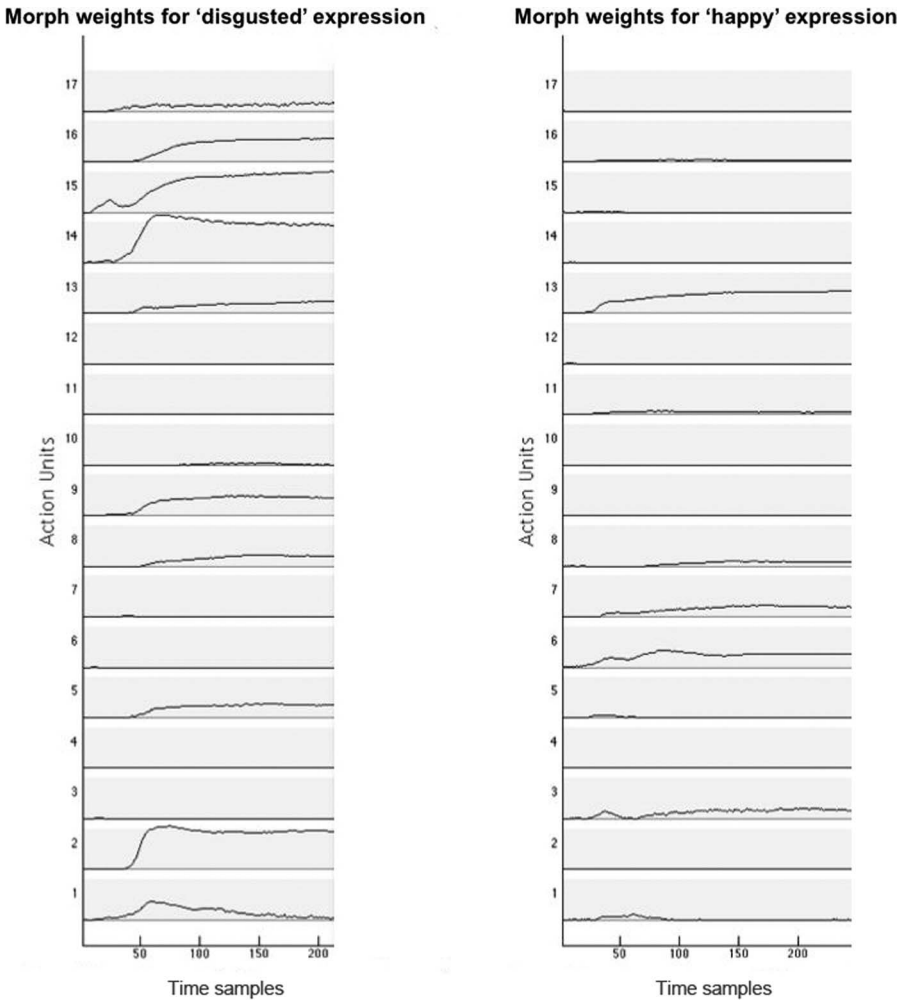

Figure 2. Estimated morph weight time courses $W_i^*(t)$ of expression "disgusted" (left) and "happy" (right) for each AU. The figure was adapted from de la Rosa, Giese, Bülthoff, and Curio (2013), according to the ARVO (Association for Research in Vision and Ophthalmology) copyright policy and rules.

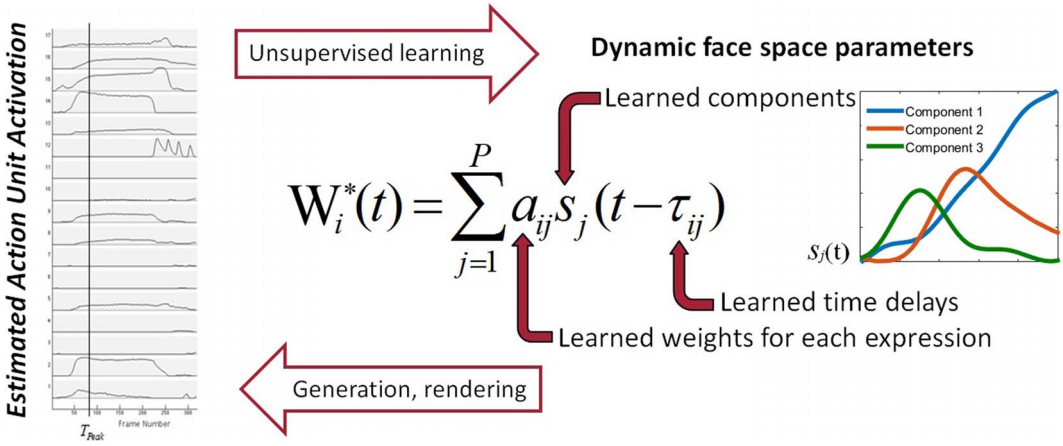

Figure 3. Figure summarizing the procedure underlying the generation of the artificial AU temporal profiles based on Equation 3. The parameters of the anechoic mixing model are learned from a set of AU profiles associated with several emotional facial expressions via unsupervised learning. The learned parameters are then modified to generate synthetic AU coefficients to use for rendering of new facial expressions to use as visual stimuli. The right panel in the figure illustrates the first three identified components.

sions generated for this study were always associated with the same identity. Indicating by $S_E(t)$ the 3D shape of the reconstructed face (parametrized by 3D polygons), by S_N the shape associated with a neutral reference face and by $S_{AU,i}$ the 3D scan of the shape corresponding to the peak of the *ith* AU, expressions were computed using the equation

$$S_E(t) = S_N + \sum_{i=1}^{N} W_i^*(t) S_{AU,i}(t) \quad (2)$$

In order to further reduce the dimensionality of the functions $W_i^*(t)$ over time, we applied unsupervised learning techniques.

Dimensionality reduction

To investigate the spatio-temporal structure underlying emotional dynamic facial expressions, we used the approach that is illustrated in Figure 3. We used unsupervised dimensionality reduction techniques to obtain a low-dimensional description of the spatio-temporal structure of the AU coefficients associated with different emotional expressions. In order to generate training data, all identified AU profiles were first segmented in time. Only the temporal interval starting at the initiation of the facial movement until the time of peak expression was considered. The AU temporal profile within this time interval was then standardized to a temporal window with $T = 100$ time steps. The dimensionality reduction techniques were applied to the data matrix $\mathbf{X}$, whose rows consist of the identified AU temporal profiles $\mathbf{W}^*(t)$ associated with all facial expressions that were produced by a professional actor. $\mathbf{W}^*(t)$ was sampled at T points in time, resulting in a size of $\mathbf{X}$ of 561 (17 AU $\times$ 11 expressions $\times$ 3 repetitions) by 100 time samples. Facial expressions considered were facial scrunch, mouth opening, agreement, confusion, disagreement, disgust, fear, happiness, and surprise, thinking for problem solving, and thinking to remember. The first method that we applied for dimensionality reduction was Nonnegative Matrix Factorization (NMF), as developed by D. D. Lee and Seung (1999). The second method is called FADA (Fourier-based Anechoic Demixing Algorithm), which we developed (Chiovetto & Giese, 2013) inspired by the previous work by Omlor and Giese (2007a, 2007b). Differently from NMF, which is based on an instantaneous mixing model, the FADA algorithm is based on an anechoic mixture equation, Equation 3, as used in acoustics for the modeling of acoustic mixtures in reverberation-free rooms (Bofill, 2003; Emile & Common, 1998; Yilmaz & Rickard, 2004). This model assumes that N_s observable acoustic signals x_i ($i = 1, 2,...N_s$) are caused by the superposition of P acoustic source functions (signals) $s_j(t)$, where time-shifted versions of these source functions are linearly superposed with the mixing weights a_{ij}. The time shifts are specified by the time delays τ_{ij}, and in the acoustical model are determined by the traveling times of the signals. The model has the following mathematical form:

$$\mathrm{W}_i^*(t) = \sum_{j=1}^{P} a_{ij} s_j(t - \tau_{ij}) \quad (3)$$

Note that for the special case that $\tau_{ij} = 0$ for all pairs (i, j), this Equation 3 coincides with the classical instantaneous mixing model underlying NMF, except for the positivity constraints. As the anechoic algorithm

used by Omlor and Giese (2007a, 2007b), FADA is based on Equation 3 but it includes additional smoothness priors for the source functions. The introduction of such priors is justified by the observations that biological data usually have limited bandwidth and by the fact that such priors improve substantially the robustness of the estimation method. A detailed description of the FADA algorithm can be found in the Supplementary material.

By construction, the values of the AU activation profiles are nonnegative. The FADA algorithm does not include a nonnegativity constraint for the identified source functions. To overcome this discrepancy, the FADA algorithm was not applied directly to the matrix X but to the matrix $d\mathrm{X}/dt$, the rows of which are the derivatives of the AU profiles in the matrix X. Once the components were identified, the original data were approximated by combining the integrals of the identified components. The constant values to add to the components after their integration were identified using an optimization procedure minimizing the error between actual and reconstructed data. An additional constraint was also imposed on the constant parameters in order to assure the nonnegativity after summation of the corresponding integrated source functions.

Visual stimulus generation and psychophysical experiments

In this study we carried out two experiments. First we carried out a "Turing test" in order to investigate to what extent the approximation of facial expressions generated using our model could be perceived as natural and human-like. A Turing test is a method, commonly used in the field of artificial intelligence, for determining how well a computer (or artificial agent) is capable of imitating human-like performance or behavior. In addition, we also carried out a classification and rating experiment, aiming to study how different spatio-temporal primitives of facial expressions are integrated in perception. We recruited 12 participants (age ranged from 21 to 43 years, mean = 28.4). The participants had normal vision or corrected-to-normal using contact lenses or glasses. Participants gave their written, informed consent form prior to the experiment. The study was conducted in line with Max Planck Society policy and was previously approved by ethics committee of the University of Tübingen.

For the generation of the visual stimuli for the psychophysical experiments, we exploited the graphical animation platform described above (Figure 1). More in detail, facial expressions were generated using the morphing weight trajectories $\mathbf{W}^*(t)$ obtained either from the motion capture data through the optimization procedure (described by Equation 1), or reconstructed mixing the temporal sources s_j identified by the FADA algorithm based on the anechoic mixture, Equation 3. For the psychophysical experiments, only 3 of the 11 initial expressions were rendered, namely fear, disgust, and surprise. To test the invariance of the shapes of the identified sources s_j across expressions, we used a leave-one-out approach. More specifically we compared the sources identified from the data associated with each single emotion, with the ones identified from all the other expressions. To compute the similarity between two sets of sources, we followed an iterative procedure. For each pair of sources in the two groups we first computed, for each temporal delay between the sources, the similarity index S, quantified as the dot product between two components, normalized with respect to their norms. The index S represents the cosine of the angles between the vectors identified by the two components. When the index is equal to 1, the components are proportional to each other, while $S = 0$ implies that they are orthogonal. We then removed from the data the pair of sources with the highest S value and repeated this procedure until only one pair of sources was left. We found that, on average, the similarity between the groups of sources was $S = 0.96 \pm 0.02$, indicating a very high level of invariance of the shapes of the sources across expressions.

Turing test

In order to determine the minimum number P of source functions that is required for the generation of photorealistic emotional expressions based on Equation 3, we designed a Turing test. Participants sat in front of a screen and were presented a series of visual stimuli. Each stimulus consisted of two rendered facial animations appearing side-by-side on the screen and showing one of three emotional expressions (disgust, fear, and pleasant surprise). One was rendered using the original kinematic data collected during the motion capture recordings. The other one was generated based on Equation 3 using either $P = 1$, $P = 2$ or $P = 3$ source functions. A series of examples of the visual stimuli used for the Turing test can be found in the Supplementary material. At any presentation of the stimuli, the order of the positions (left/right) of the two animations was chosen randomly, and the two animations ran simultaneously for three times. After the presentation of the stimulus, participants were asked to choose which one of the two animations was most natural. The aim of this experiment was to identify the minimum number P from which original and synthetic expressions became indistinguishable. Twelve subjects participated in the experiment. Each participant was presented a total of 108 stimuli (3

emotions × 3 levels of model complexity × 12 repetitions). Once all the data were collected, we computed, for each model complexity P and for each emotion, the probability that participants could discriminate between original and synthesized expression as the ratio between the number of correct answers over the total number of stimulus presentations. A discrimination probability equal to 0.5 (chance level) indicates that participants could not discriminate correctly which of the two stimuli was the original emotional expression.

In order to test whether the visual stimuli that we generated were actually distinguishable from "natural" stimuli animated with motion capture data, we classified the stimuli using an ideal observer model. This model parametrized the face pictured in individual frames by PCA and applied a SVM classifier for the distinction between natural and model-based animations. Opposed to the participants, this classifier could distinguish between natural and approximated animations with high reliability (on average, the probability of correct classification was 0.99 ± 0.04).

Classification and rating experiment

The goal of the second experiment was to determine the contribution of the individual source functions to the class of facial expressions. Once the model complexity P was determined, we generated another set of visual stimuli by varying the contribution of each of the P source function s_j for the synthesis of $\mathbf{W}^*(t)$. To this end we introduced P additional morphing parameters $\gamma_j \in [0, 0.33, 0.66, 1]$ in Equation 3 so that

$$\mathbf{W}^*(t) = \sum_{j=1}^{P} \gamma_j a_{ij} s_j (t - \tau_{ij}) \quad (4)$$

Since we found that $P = 2$ is sufficient (see Results), we generated 48 visual stimuli in total (4^2 morphing levels by 3 emotions). Each participant was presented with one of these stimuli, one at the time, and was asked to indicate, pressing one out of three buttons, which emotion such stimulus corresponded to and to rate the intensity of the stimulus by choosing a integer R between 0 and 6. Each stimulus was presented 15 times. The aim of this experiment was to test the extent to which each source function s_j contributed to the perception of emotional content. The probability with which participant could discriminate correctly the emotion associated with the presented facial expression was computed, for each morphing level, as the ratio between the correct number of answers over the total number of times the stimulus was presented. To avoid biases in the data due to subjective differences among the participants in the judgment of the emotional content, for each participant the ratings were normalized to [0, 1], applying the following formula:

$$R_{st} = \frac{R - R_{min}}{R_{max} - R_{min}}. \quad (5)$$

where R_{min} and R_{max} indicate respectively the minimal and maximal ratings given by the participant.

Statistics

A 3 × 3 repeated measure analysis of variance (ANOVA) was carried out in order to investigate any influence of the emotion type (disgust, fear, and pleasant surprise) and model complexity (one, two, or three sources) on the capability of the participants to discriminate between actual and recorded expressions and the ones rendered with the AU reconstruction provided by Equation 3. For all ANOVAs in this study, Greenhouse–Geisser adjustment was used when the data violated the sphericity assumptions. A one-sample Wilcoxon signed-rank test was then used to assess, for each expression and for each model complexity, whether participants could discriminate between the actual animated expression and the one based on the model. Separately for each expression, 4 × 4 repeated measure ANOVAs were used to investigate statistically significant effects of the morphing parameters γ_1 and γ_2 (values: 0, 0.33, 0.66, and 1) on the classification and rating performance of the participants. All statistical tests were implemented using SPSS (V.22; SPSS, Inc., Chicago, IL). For all tests the significance level was set to 5%.

Model type and order selection

To determine whether human facial emotion perception is (approximately) Bayes-optimal, we compared human perception results from psychophysical experiments to Bayesian model selection. More specifically, we applied an approximate Bayesian model type and order (i.e., number of sources) selection criterion, which we called LAP, based on a Laplace approximation to the model evidence (Laplace, 1774; Bishop, 2007). In brief, the criterion is based on the optimization of the following log-probability function:

$$\mathrm{LAP} = \underbrace{\log(p(\mathbf{X}|\Theta^\star_M, M))}_{\log-\text{likelihood}} + \underbrace{\log(p(\Theta^\star_M|M))}_{\log-\text{prior}} + \underbrace{\frac{\dim(\Theta)}{2}\log(2\pi) - \frac{1}{2}\log(|\mathbf{H}|)}_{\log-\text{posterior}-\text{volume}} \quad (6)$$

where $\mathbf{X}$ is the observable data, Θ_M is a vector of model parameters for a model indexed by M, M is a tuple (model type, model order) in which the model type is either a smooth anechoic mixture determined with the

FADA algorithm (Chiovetto & Giese, 2013) or a synchronous (i.e., undelayed) mixture computed with NMF, $\Theta^{\star}_{M}$ is a tuple of optimized parameters, and $\mathbf{H}$ is the Hessian matrix of second derivatives with respect to $\Theta^{\star}_{M}$. Once we have evaluated Equation 6 for all M, we can select that M (i.e., the model type and the model order, which maximizes the model evidence, since we have no a-priori preference for any M.

In the analysis, we compared the LAP criterion to two other standard model complexity estimators: the Bayesian information criterion (BIC; Schwarz, 1978) and the Akaike information criterion (AIC; Akaike, 1974). A detailed description of the derivation of the LAP, AIC, and BIC criteria can be found in the Supplementary material and in Endres et al. (2013).

Cue fusion models

The FADA algorithm can discover modular movement primitives, i.e., simple component movements that are linearly combined to form complex natural movements (see Equation 3). Since we are interested in the perception of facial emotion expressions, we wondered in which form the individual primitives contribute to the perception of emotions, and how the contributions of different primitives interact. To this end, we investigated the hypothesis that the primitives can be viewed as "cues" that are fused by the observer to generate a unified percept. In the domain of multimodal perception, it has been demonstrated that human performance can come close to a Bayesian ideal observer in haptic-visual or audio-visual cue integration tasks (Ernst & Banks, 2002; Körding et al., 2007), or proprioceptive-visual integration informed by an internal forward model (Beck, Endres, Lindner, & Giese, 2014). It can be shown (see Supplementary material) that in our case the observer is described by the following linear model:

$$R(C_1, C_2) = \beta_1 R(C_1, 0) + \beta_2 R(0, C_2) + \beta_0 \quad (7)$$

where the cues C_1 and C_2 are the amplitude-scaled versions of the movement primitives, which can be achieved by multiplying the emotion-specific weights $\mathbf{W}_{j,i}$ for primitive i with a constant (positive) factor, the β_i are scalar coefficients and R is the perceived emotion strength associated with a specific facial expression. More specifically, it can be shown that $\beta_1 = \beta_2 = 1$ and therefore that

$$R(C_1, C_2) = R(C_1, 0) + R(0, C_2) - \beta_0 \quad (8)$$

We hypothesized therefore that the expected emotional strength rating for C_1, $C_2 > 0$ can be computed by summing up the strength ratings measured when one of the cues is zero, minus a bias term β_0. Given the existence of some evident nonlinear saturation effects for high morphing levels (γ_1, γ_2; see Figure 4), we also tested a cue fusion model with an output nonlinearity of the form

$$\begin{aligned}\tilde{R}(C_1, C_2) &= f(R(C_1, C_2)) \\ &:= z_0(1 - \exp(-z_1 R(C_1, C_2))) \quad (9)\end{aligned}$$

where $\tilde{R}(C_1, C_2)$ is the rating reported by the participants, and $R(C_1, C_2)$ is the rating predicted by the linear Bayesian cue fusion model, Equation 8. Combining therefore Equation 9 with Equation 8, we obtain the following equation for the predicted rating with output nonlinearity:

$$\begin{aligned}&\tilde{R}(C_1, C_2) = \\ &\quad f\big(f^{-1}(R(C_1, 0)) + f^{-1}(R(0, C_2)) - \beta_0\big). \quad (10)\end{aligned}$$

Results

Turing test, classification and rating experiments

Using the NMF and FADA algorithm, we performed dimensionality reduction on the temporal profiles associated with the activation coefficients of 17 AUs characterizing 11 different dynamic facial expressions. The identification of the AU activation coefficients was carried out by optimizing the kinematic matching between actual and synthesized expressions (Curio et al., 2006). For each dimensionality reduction algorithm separately, we used either one, two or three source functions to approximate the AU activation coefficients associated with three facial expressions (disgust, fear, and pleasant surprise) and use them to generate synthesized facial expression to use as stimuli of the Turing test to determine the minimum number of components needed to generate facial expressions that were indistinguishable from the actual ones. For the Turing test based on the source functions identified using NMF, we test 13 participants (results are shown in Figure 4A). Data were analyzed using a 3×3 repeated measure ANOVA with emotion (disgust, fear, and pleasant surprise) and model complexity (1, 2, or 3 sources) as factors and it revealed a significant main effect of the emotion displayed on the percentage of correct classification, $F(1.31, 1.44) = 10.88$, $p = 0.003$, $\eta_p = 0.48$. A one sample Wilcoxon signed-rank test (see the results in Table 1) indicated that the participants could never discriminate above chance level between the actual recorded expressions and the ones obtained using the data predicted by the model when the displayed emotions were disgust and fear. In the case of surprise, differently, they could discriminate above

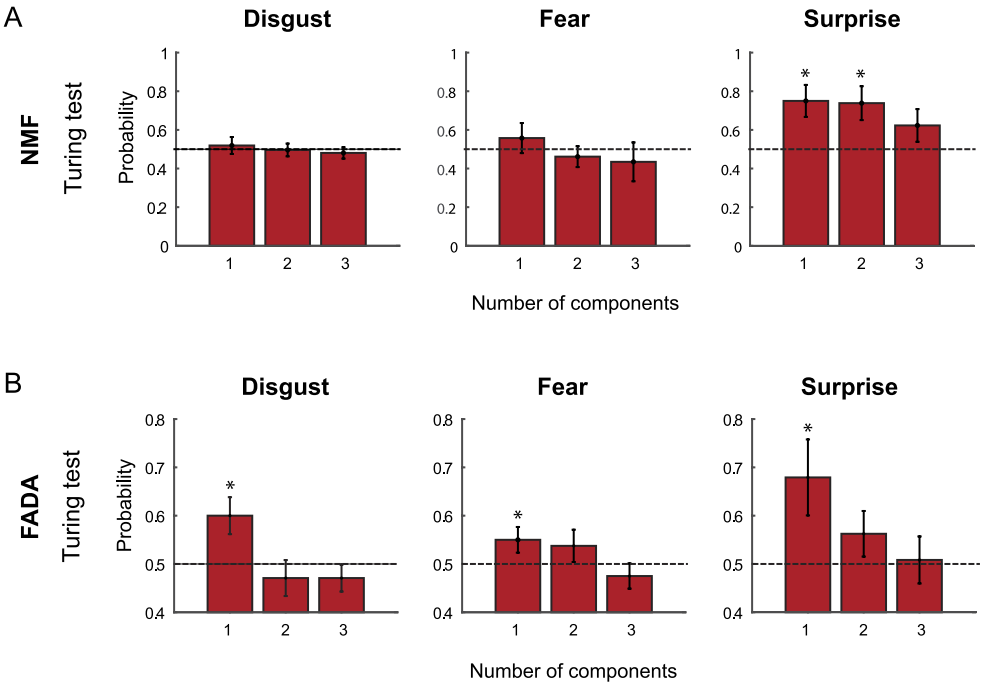

Figure 4. Results of the Turing tests. (A) For each tested expression, the probability with which participants were able to discriminate between the actual recorded expression from the one rendered by using the estimated temporal AU profiles based on the instantaneous mixture of the sources identified using NMF as function of the number *P* of sources is shown. Horizontal dashed lined indicate the chance level (probability = 0.5), at which original and reconstructed expressions are indistinguishable. Asterisks indicate significant differences from chance level. (B) Probabilities with which participants were able to discriminate between the actual and rendered expressions by using the estimated temporal AU profiles based on Equation 3 as function of the number *P* of sources.

chance when either one or two sources were used in the model. The Turing test based on the sources identified using the FADA algorithm was run on 12 other participants (results are shown in Figure 4B). Even in this case, data were analyzed using a 3×3 repeated measure ANOVA with emotion and model complexity as factors, and it revealed a significant main effect of the model complexity on the percentage of correct classification, $F(2, 22) = 12.64$, $p < 0.001$, $\eta_p = 0.54$. A one-sample Wilcoxon signed-rank test (see Table 1) indicated that the participants could discriminate between the recorded expressions and the ones obtained using the data predicted by the model only when one single source was used to render the facial expressions (Figure 4A). In all these cases, the observed probability was always statistically significantly higher than chance level.

In order to access the contribution of each component to the expressiveness of the synthesized facial movements, we designed a classification and rating experiment. The two groups of participants that took part in the Turing tests were presented also with synthesized facial expressions obtained by combining two morphed temporal components according to either the instantaneous linear mixture model (and identified using the NMF algorithm) or the anechoic model (and identified using the FADA algorithm). They were then asked to report which expression they had been presented and to rate the level of expressiveness of the facial animation. Regarding the classification performance, participants were able to recognize with very

	Disgust			Fear			Surprise		
	1 source	2 sources	3 sources	1 source	2 sources	3 sources	1 source	2 sources	3 sources
NMF	$Z = 38$, $p = 0.654$	$Z = 37$, $p = 0.872$	$Z = 29$, $p = 0.417$	$Z = 57$, $p = 0.416$	$Z = 32$, $p = 0.343$	$Z = 35.5$, $p = 0.483$	$Z = 72.5$, $p = 0.008$*	$Z = 80.5$, $p = 0.014$*	$Z = 58.5$, $p = 0.125$
FADA	$Z = 52$, $p = 0.011$*	$Z = 19.5$, $p = 0.41$	$Z = 16$, $p = 0.233$	$Z = 32$, $p = 0.048$*	$Z = 53$, $p = 0.266$	$Z = 16.5$, $p = 0.253$	$Z = 48.5$, $p = 0.031$*	$Z = 23$, $p = 0.127$	$Z = 30.5$, $p = 0.759$

Table 1. Results of the Wilcoxon signed-rank tests. *Note*: *, signifies the cases in which the discrimination probability was significantly different from chance level.

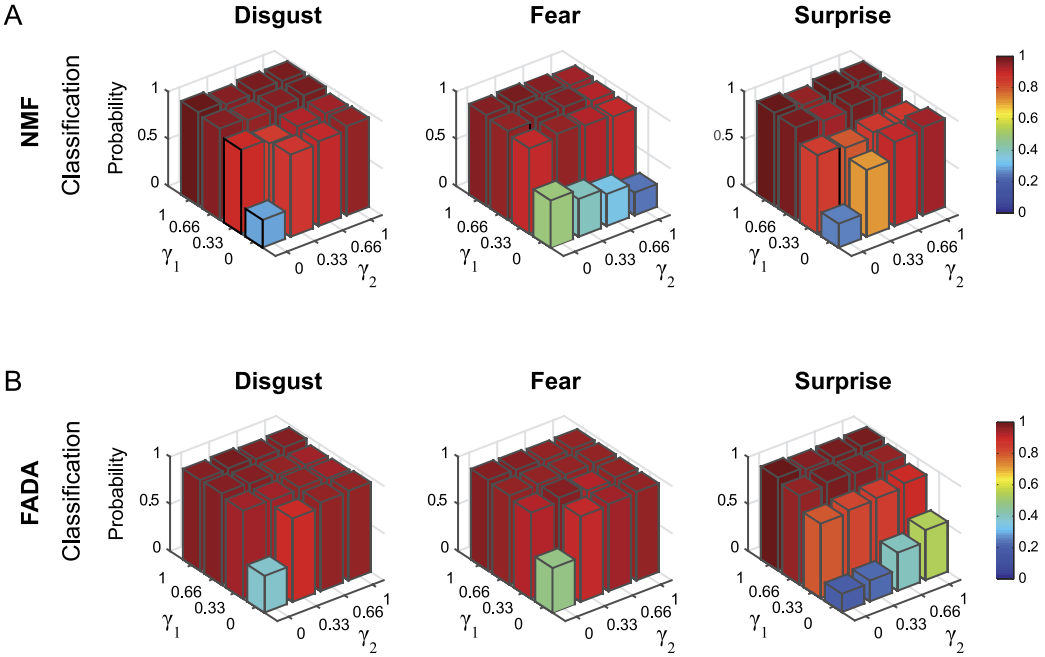

Figure 5. Results of expression classification. (A) The graphs show, for each expression and morphing level determined by the pair (γ_1, γ_2), the average probability with which participants could recognize the actual emotion from the expression generated using the temporal functions identified using the NMF algorithm. (B) Results of expression classification relative to the expressions that were based on the anechoic mixture of the source functions identified using the FADA algorithm.

high precision the presented emotion as long as one of the morphing parameters was different from 0 (i.e., as long as $\gamma_1 > 0$ or $\gamma_2 > 0$). Results from the expressions generated using the instantaneous mixture of the NMF sources are shown in Figure 5A. For each expression, a 4×4 repeated measure ANOVA with γ_1 and γ_2 (taken the values 0, 0.33, 0.66, and 1) as factors revealed a significant main effect on the percentage of correct classification of both factors. In addition, these main effects were always qualified by an interaction between the factors (see Table 2). Similarly to the instantaneous case, participants were able to recognize with very high precision the presented emotion as long as one of the morphing parameters was different from 0 and also when the rendered expressions were based on AU coefficients resulting from the anechoic mixture of the sources functions identified using the FADA algorithm (Figure 5B). Even in this case, a 4×4 repeated measure ANOVA with γ_1 and γ_2 as factors revealed, for each expression, a significant main effect on the percentage of correct classification of both factors. Also the interaction effect was always significant (see Table 2).

The results of the rating experiments associated with the NMF algorithm are shown in Figure 6. For each expression, a 4×4 repeated measure ANOVA with γ_1 and γ_1 as factors revealed a significant main effect on the ratings of both γ_1 and γ_2. The interaction effect was significant only for disgust and surprise (see Table 3). Concerning the expressions based on the anechoic mixture of the sources identified with the FADA algorithm, the results of the rating experiments (Figure 6B) showed that participants increased their ratings approximately linearly with the values of the morphing parameters. For each expression, a 4×4 repeated measure ANOVA with γ_1 and γ_2 as factors revealed a significant main effect on the ratings of both factors. The analysis also revealed that these main effects were in most of the cases qualified by an interaction between the factors (see Table 3).

Model selection

To discriminate whether an anechoic model based on the combination of source functions that can be shifted in time could account for the data more efficiently than a classic instantaneous mixture model, we applied a generalized version of a Bayesian model selection criterion based on Laplace-approximation (LAP criterion) that we previously developed (Endres et al., 2013).

We used the LAP criterion to select the best model type and order on a per trial basis. A trial of our experiment consisted of $J = 17$ time-courses of AU activities, and we analyzed a total of 21 trials. We tested both a FADA and a synchronous model with $I = 1, 2, \ldots, 5$ sources. The FADA model had a cutoff-frequency of $0.15\, f_N$, where f_N is the Nyquist frequency of the data; we took the same f_0 for the wave kernel of the LAP. The synchronous model was not regularized for smooth sources ($f_0 \to \infty$). We aimed for 90% explained variance

	Factor 1 (γ_1)	Factor 2 (γ_2)	Interaction ($\gamma_1 \times \gamma_2$)
NMF			
Disgust	$F(1.65, 19.81) = 36.10$, $p < 0.001$, $\eta_p = 0.75$*	$F(3, 36) = 35.35$, $p < 0.001$, $\eta_p = 0.75$*	$F(2.48, 29.72) = 45.32$, $p < 0.001$, $\eta_p = 0.79$*
Fear	$F(1.20, 14.37) = 158.56$, $p < 0.001$, $\eta_p = 0.93$*	$F(3, 36) = 5.79$, $p = 0.002$, $\eta_p = 0.33$*	$F(1.92, 23.05) = 2.60$, $p = 0.009$, $\eta_p = 0.18$*
Surprise	$F(2.09, 25.12) = 21.96$, $p < 0.001$, $\eta_p = 0.65$*	$F(1.46, 17.54) = 26.49$, $p < 0.001$, $\eta_p = 0.69$*	$F(2.47, 29.64) = 35.50$, $p < 0.001$, $\eta_p = 0.75$*
FADA			
Disgust	$F(1.67, 18.37) = 30.50$, $p < 0.001$, $\eta_p = 0.74$*	$F(1.44, 15.84) = 36.40$, $p < 0.001$, $\eta_p = 0.77$*	$F(1.55, 17.01) = 31.33$, $p < 0.001$, $\eta_p = 0.74$*
Fear	$F(1.80, 19.76) = 21.92$, $p < 0.001$, $\eta_p = 0.67$*	$F(1.28, 14.03) = 14.84$, $p = 0.001$, $\eta_p = 0.57$*	$F(1.49, 16.40) = 16.48$, $p < 0.001$, $\eta_p = 0.60$*
Surprise	$F(1.57, 17.27) = 47.6$, $p < 0.001$, $\eta_p = 0.81$*	$F(1.40, 14.84) = 16.53$, $p < 0.001$, $\eta_p = 0.60$*	$F(3.57, 39.24) = 8.21$, $p < 0.001$, $\eta_p = 0.43$*

Table 2. Results of the 4 × 4 ANOVAs on the probabilities of correct classification, for each expression and algorithm. *Note*: *, signifies the cases in which a significant effect was found.

	Factor 1 (γ_1)	Factor 2 (γ_2)	Interaction ($\gamma_1 \times \gamma_2$)
NMF			
Disgust	$F(1.40, 16.67) = 499.66$, $p < 0.001$, $\eta_p = 0.97$*	$F(3, 36) = 92.51$, $p < 0.001$, $\eta_p = 0.89$*	$F(3.83, 45.93) = 22.31$, $p < 0.001$, $\eta_p = 0.65$*
Fear	$F(1.30, 15.63) = 186.27$, $p < 0.001$, $\eta_p = 0.94$*	$F(3, 36) = 3.37$, $p = 0.029$, $\eta_p = 0.22$*	$F(9, 108) = 1.49$, $p = 0.16$, $\eta_p = 0.11$
Surprise	$F(1.45, 17.37) = 206.39$, $p < 0.001$, $\eta_p = 0.95$*	$F(3, 36) = 69.22$, $p = 0.002$, $\eta_p = 0.33$*	$F(9, 108) = 6.12$, $p < 0.001$, $\eta_p = 0.34$*
FADA			
Disgust	$F(1.42, 15.45) = 181.09$, $p < 0.001$, $\eta_p = 0.94$*	$F(1.69, 18.36) = 86.80$, $p < 0.001$, $\eta_p = 0.89$*	$F(9, 99) = 29.76$, $p < 0.001$, $\eta_p = 0.73$*
Fear	$F(1.27, 13.50) = 66.80$, $p < 0.001$, $\eta_p = 0.86$*	$F(1.60, 17.67) = 32.10$, $p < 0.001$, $\eta_p = 0.75$*	$F(9.99) = 14.56$, $p < 0.001$, $\eta_p = 0.57$*
Surprise	$F(1.31, 14.37) = 185.02$, $p < 0.001$, $\eta_p = 0.94$*	$F(1.62, 17.85) = 69.22$, $p < 0.001$, $\eta_p = 0.86$*	$F(3.51, 38.55) = 3.17$, $p = 0.029$, $\eta_p = 0.22$*

Table 3. Results of the 4 × 4 ANOVAs on the emotion ratings, for each expression and algorithm. *Note*: *, signifies the cases in which a significant effect was found.

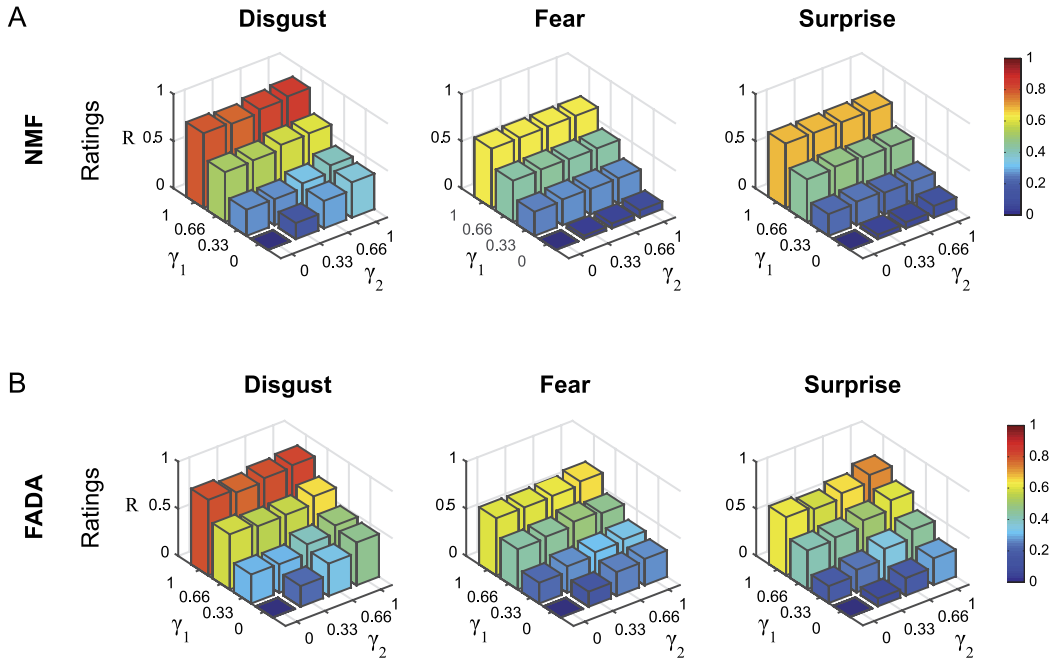

Figure 6. Results of expression rating. Each graph shows the average normalized rating values for each expression in function of each morphing level when the generation of the rendered expressions was based either on the temporal source functions identified using the NMF (A) or FADA algorithm (B).

all models. The parameters v, S of the gamma prior on the weights and the λ of the exponential distribution on the delays are estimated from the data after source extraction. See Supplementary material for more detailed information. To put our results into the context of well-known model selection schemes, we repeated our analysis with the Akaike Information Criterion (Akaike, 1974) and the Bayesian Information Criterion (Schwarz, 1978). All three criteria prefer the anechoic model in every trial, as shown in Figure 7, middle. However, as depicted in Figure 7, right, BIC and AIC would pick models with a larger number of sources than LAP, which yields an average $I = 1.95 \pm 0.50$ sources ($SEM = 0.11$). Therefore, the only criterion which is consistent with the perceptual Turing-test results described above is LAP: Human observers reach chance discrimination level at two sources. A scree plot (Figure 7, left) shows a similar result for FADA algorithm (two sources are

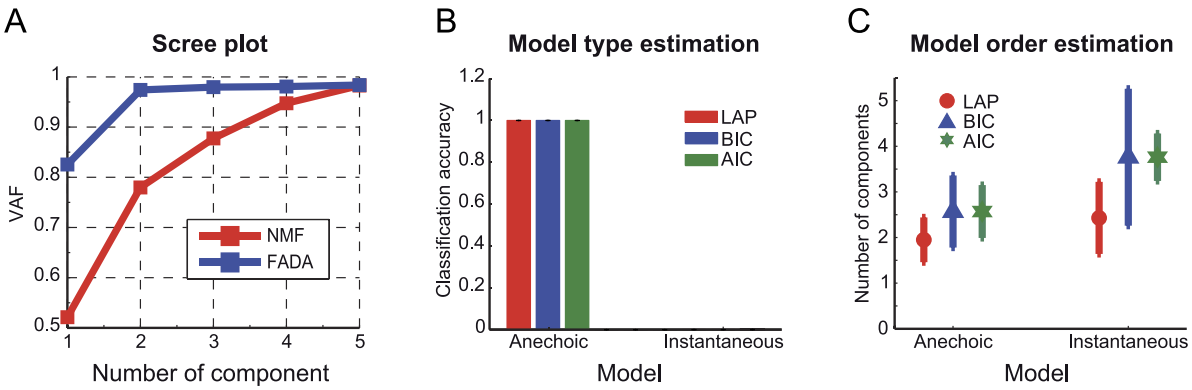

Figure 7. Model comparison results on facial expression data prefer the anechoic mixture with two sources. We compared models with our Laplace approximation to the model evidence (LAP) and two standard plug-in estimators, the Bayesian information criterion (BIC), and the Akaike information criterion (AIC). We analyzed the data trial-by-trial, where one trial is comprised of 17 time-courses of AU activities and the complete data set consisted of 21 trials. (A) Scree plot of the VAF as a function of the number of sources (primitives, components) for synchronous (NFM) and delayed (FADA) models. (B) Model type selection. All three criteria agree on every single trial that the anechoic model is a better explanation of the data than the synchronous one. (C) Model order (number of sources) selection. AIC and BIC typically prefer models with a larger number of sources. On average, LAP yield an average model order across participants of $I = 1.95 \pm 0.50$, which is consistent with human perception.

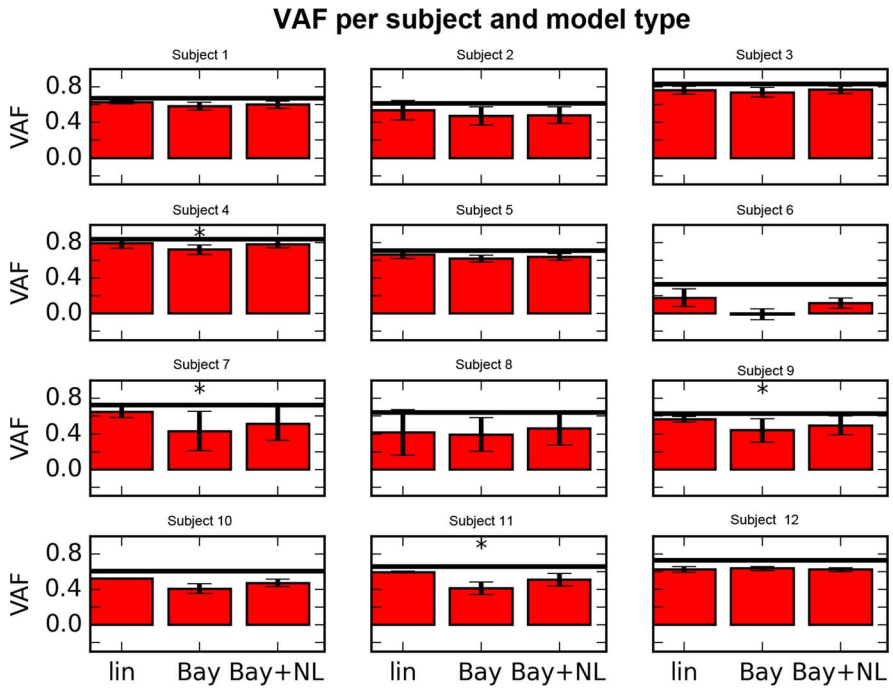

Figure 8. Cross-validation results. Shown is the VAF in the emotion ratings, averaged across emotions. One plot per subject. Error bars indicates standard deviations across emotions. The horizontal line in each graph represents the maximal average VAF, computed assuming that the mean ratings of each morphing level (γ_1, γ_2) are known, i.e., an upper bound on the performance of any model that predicts mean ratings. The abbreviation "lin" represents linear, but inconsistent model, Equation 8. The abbreviation "Bay" represents the linear Bayesian model, Equation 8. "Bay + NL" represents the Bayesian model with saturating output nonlinearity, Equation 10. Stars indicate significant differences ($p < 0.01$) between linear and Bayesian models, evaluated with a Wilcoxon signed rank test. For details, see text.

best), but provides no clear answer for the synchronous NMF model.

Based on these results, for the rest of the analysis we factorized data according to an anechoic model with one to three components (see Figure 3 for the shapes the first three components).

Cue fusion

We compare the three cue fusion models described above: the general, but inconsistent, linear model described by (Equation 7, "lin" in Figure 8), the restricted linear model which follows from the Bayesian cue fusion model with normality assumptions (Equation 8, "Bay"), and the Bayesian model with output nonlinearity (Equation 10, "Bay + NL"). Models are compared by cross-validation with variance accounted for (VAF) as the score on held-out data, which are all data points of one morphing level (γ_1, γ_2), averaged across all levels. For each (γ_1, γ_2), we trained the models on the data from all other morphing levels. Furthermore, to establish an upper bound on the performance of any model that predicts mean ratings, we computed the mean rating for each subject and emotion. Results are shown in Figure 8. Each plot depicts the average predicted VAFs for one participant, averaged across all emotions. Error bars indicates standard deviations. The horizontal line in each plot indicates the average upper bound. All models are on average below the bound; the negative VAF in one participant is a consequence of the cross-validation procedure: The data used for learning the model parameters are disjoint from the data used for evaluation. The linear Bayesian model tends to perform worse than the general linear model. The performance difference is significant ($p < 0.01$, Wilcoxon signed rank test) in four participants, indicated by a star in the figure. The nonlinear model shows no significant difference to the general linear model in any of the participants. Furthermore, its VAF predictions are close to the maximally expectable ones (horizontal lines) in most participants. We can therefore conclude that a Bayesian cue fusion model with output nonlinearity is a good description of the computational process with integrates facial movement primitives for emotion recognition.

Discussion

In this study we aimed to identify the low-dimensional organization of the spatio-temporal kinematic variability of facial expressions. To this end, we applied we applied Nonnegative Matrix Factorization (NMF; D. D. Lee & Seung, 1999) and a blind source separation algorithm (FADA; Chiovetto et al., 2016; Chiovetto & Giese, 2013) to approximate the temporal profiles of the AU activation coefficients by superposition of a very small number of source components, defining movement primitives. Using these techniques and a new Bayesian model selection criterion (Endres et al., 2013), we then investigated which generative model was more suitable to describe the data and the minimum number of invariant components required for a good approximation of a variety of facial expressions. In addition, we studied how information from the different primitives about emotion is integrated in visual perception. We found that only two primitives of the anechoic mixing model were sufficient for an accurate approximation of more than 10 facial expressions. The information contributions of the different primitives were integrated by a mechanism that can be modeled as a linear cue integration model, or by a Bayesian fusion model with appropriate nonlinearity.

Dimensionality reduction and model selection

Inspired by other studies in motor control (Bizzi et al., 2008; Flash & Hochner, 2005) suggesting that the human central nervous system relies on a low-dimensional organization to simplify motor control and learning, we applied the NMF and FADA algorithm to the temporal profiles of the AU coefficients underlying a set of eleven emotional facial expressions. Unexpectedly we found that a model with only two time-shiftable movement primitives could account for the majority of the variance (more than 90% in total, see Figure 7) suitable for the generation of photorealistic facial animations that were, as demonstrated by the Turning test, indistinguishable from the facial expressions generated by using the recorded kinematics. This result was consistent with an estimation of the complexity of the underlying model using a model comparison approach (Endres et al., 2013). Complementing the large number of studies in the motor control literature, we showed in this study not only that emotional dynamic facial expression can be approximated by a low-dimensional kinematic organization, but also that the modules at the base of such organization play a role in perception.

Our analysis relied on the taxonomization of emotional facial expressions according to the FACS system developed by Ekman and Friesen (1978). We used the activation coefficients of the AUs as input for the dimensionality reduction algorithms. The low-dimensional organizations that we identified, therefore, do not dismiss and cannot replace the FACS as a modular framework to characterize emotional facial expressions. Rather, they complement the FACS system, revealing how the AUs (which provide mainly a spatial description of facial expressions) are synergistically recruited over time according to well-defined temporal activation profiles in order to convey a specific emotional content.

In the last three decades, a lot of effort has been spent in computer graphics to develop advanced computational models to use for the generation of realistic, 3D facial animations (Parke, 1972; Y. Lee, Terzopoulos, & Waters, 1995; Parke & Waters, 2008). Multiple physically based methods have been for instance proposed that try to simulate facial muscles and skin to animate face models (Kähler, Haber, Yamauchi, & Seidel, 2002; Sifakis, Selle, Robinson-Mosher, & Fedkiw, 2006; Terzopoulos & Waters, 1990). These physically based approaches are, however, associated with high computational complexity. Our approach is mathematically much simpler, and can be extended easily to reactive real-time animations (Mukovskiy, Land, Schack, & Giese, 2015). The combination of the AUs of FACS system with the spatio-temporal primitives provided by the FADA algorithm provides a very compact parametrization of the emotional space associated with facial movement, which might be also interesting for analysis applications.

It is widely accepted that emotional and affective stimuli, including emotional facial expressions, can be coded in a low-dimensional space spanned by two primary dimensions, valence and arousal (Lang, Greenwald, Bradley, & Hamm, 1993; Larsen & Diener, 1992; Russell, 1980). Valence defines whether the stimulus is perceived as emotionally "positive" or "negative." In opposition, arousal defines the level of physical response that the stimulus elicits and can vary from calming or soothing to exciting or agitating. In a very interesting study, Boukricha and colleagues (Boukricha, Wachsmuth, Hofstätter, & Grammer, 2009) have studied how the AU space maps onto the one spanned by valence and arousal. A very interesting extension could be to extend this approach to movement primitives and their weights as studied here.

In addition to emotional content, dynamic facial movements have been shown to convey also important information about identity (Girges, Spencer, & O'Brien, 2015; Hill & Johnston, 2001; Knappmeyer, Thornton, & Bülthoff, 2003; Thornton & Kourtzi,

2002). Currently, there are two main hypotheses regarding how facial motion aids identification processes (O'Toole, Roark, & Abdi, 2002): One, the "supplemental information hypothesis," according to which idiosyncratic facial movements facilitate identification. According to this hypothesis, different people may express the same emotion with slightly different facial expressions, and these slight differences may be learned and used to discriminate between different identities. By contrast, the "representation enhancement hypothesis" states that facial motion contributes to develop a more accurate 3D face representation. Indeed, facial motion might provide additional information about 3D structure. Our low-dimensional parametrization of facial expressions might be interesting to study the question of identity perception from the viewpoint of spatial versus temporal variation, which are encoded separately by the mixing weights and the time delays of the model.

The main result of our study is that, according to the behavioral data as well as with respect to the statistical analysis of the model accuracy, the anechoic model (described by Equation 3) was more suitable for approximating the data than the classic instantaneous mixture model that underlies NMF. This is probably a consequence of an underlying low-dimensional structure in the control of independent groups of muscles. Our experiments showed that, when the visual stimuli were generated by using an instantaneous mixture model using classic NMF (D. D. Lee & Seung, 1999) instead of the anechoic one, only one single component was sufficient to create photorealistic emotional stimuli for "disgust" and "fear," while three components had to be used for "surprise." Other qualitative differences were found for the results concerning the classification and rating experiments. Classification accuracy was reduced, and the linear relationship between morphing parameters and emotional ratings was less marked. Also the average similarity between the components extracted by the two algorithms was quite high ($S = 0.90 \pm 0.03$). The observed differences between the two model types must thus be due to the presence of time delays in the anechoic mixture model, i.e., the type of invariance that is assumed by the model architecture.

In a recent study, Delis and colleagues (Delis et al., 2016) used a tri-factor decomposition algorithm to identify the low-dimensional spatio-temporal organization underlying the categorization of dynamic facial expressions defining the six classic emotions (happy, surprise, fear, disgust, anger, sad). In their analysis, the authors first reduced the dimensionality of the data to code each emotion in terms of a set of invariant AU groups and temporal synergies (the spatial and temporal components) and then used linear-discriminant analysis to find the boundaries that discriminate the six emotions. They also demonstrated the dependence of emotion-recognition accuracy on the stimulus temporal dynamics. Although some of the results reported by Delis and colleagues are similar to those described here (such as, for instance, the existence of a low-dimensional manifold underlying emotional facial expressions), the two studies differ in important aspects, thus making them complementary. First of all the two studies differ on the methods that were used to identify the AU profiles associated with different facial expressions. In the present study the AU activation profiles were identified using an optimization procedure minimizing the error between the original expressions and the ones obtained superimposing the kinematics associated to each single AU. Differently, in Delis et al. (2016) the AU profiles were first randomly generated and combined, and the corresponding facial animations were then classified and rated by human participant during a psychophysical experiment. The second difference between the two studies concerns the generative model used to decompose the data. While the FADA algorithm finds invariance only across time, the space-by-time decomposition sets invariance constraints also in the AU space. It would be interesting in the future to investigate more in detail what are the advantages and disadvantages associated with the use of one model with respect to the other.

Cue fusion

Besides the identification of the low-dimensional architecture underlying the kinematics of emotional facial expressions and its psychological validation, the other remarkable results of our study is the findings regarding the way how the kinematic components identified with FADA are integrated in perception. We tested the hypothesis that the primitives can be seen as "cues" that are fused by an observer to generate a unified percept. Over the past 15 years, a large body of experimental evidence has shown that multistory integration in humans often occurs in a Bayesian-optimal fashion (Clark & Yuille, 2013; Ernst & Banks, 2002). That is, unimodal estimates are combined into a weighted average, where each weight of each unimodal estimate depends on the relative precision of the unimodal information. In this way, the theoretically highest reliability is achieved, higher than the one obtainable from each individual sensory modality.

We found that the emotional strength ratings are consistent with a Bayesian cue fusion model, where the weights of the individual FADA primitives act as cues or features supporting emotion perception. A saturating output nonlinearity, which maps percepts onto ratings, increased the cross-validation performance of the Bayesian cue fusion model close to its maximally possible value. This result is interesting, as

it suggests that kinematic sources that we identified might have a perceptual representation that might be interpreted as complex features or "channels" specifying emotion-specific information in the recognition of dynamic faces. A cue-fusion approach similar to the one we used here was already shown to be successful in the study of emotion conveyed by different styles of walking.

Limitations of the study

The results of our study are limited by a number of technological constraints. First, the identification of the single AUs was based on the recording of the 3D kinematics of a limited number of reflective markers, applied onto a series of key points spread over the facial surface. For animation, the motion of the face surface within these points was interpolated, and this interpolation might have led to the loss of movement variability that may be present in real faces. In addition, the morphable graphical model that we used for this study comprised only 17 out of 44 total AUs. One may argue that these constraints may have affected the dimensionality reduction analysis that we carried out. However, the high level of approximation quality of the motion of the spatial locations within the key markers was already proven previously (Curio et al., 2006). In addition, the unsupervised learning methods were not applied directly to the 3D spatial trajectories of the markers but to the AU activation coefficients, which implicitly encode an average (rather than point-specific) behavior. Moreover, only a subset of AUs is emotion-specific (Friesen & Ekman, 1983) and such AUs were all included in our model. Based on these considerations, we can therefore conclude that the very compact organization that we identified is not a mere consequence of the technological constraints of our experimental setup.

An obvious limitation of this study is the small number of tested emotions. The estimation of the primitives was based on 11 emotional expressions; but cue fusion, due to the limitation set by the duration of the experiments, was constrained to only three emotions. We used, however, a within subject design, where each participant performed indeed both tasks (classification and rating) on all three expressions. This procedure maximized the sensitivity of out studies, on cue integration and of the Turing test. In addition, the expressions were captured under highly controlled laboratory conditions, maximizing expressiveness.

Keywords: dynamic facial expressions, emotions, synergies, motor primitives, emotion perception

Acknowledgments

The research leading to these results has received funding from, Koroibot FP7-ICT-2013-10/611909; EC FP7-ICT-248311 AMARSi; DFG GI 305/4-1, DFG GZ: KA 1258/15-1; CogIMon H2020 ICT-23-2014/644727, HFSP RGP0036/2016. Cristóbal Curio was also supported by Perceptual Graphics Project PAK38 CU 149/1-1/2 funded by the Deutsche Forschungsgemeinschaft (DFG) and by BMBF KollRo 4.0. Dominik Endres acknowledges support from the DFG under IRTG 1901 "The Brain in Action" and the SFB-TRR 135. Enrico Chiovetto has been supported also by the Institutional Strategy of the University of Tübingen (Deutsche Forschungsgemeinschaft, ZUK 63). We acknowledge support by Deutsche Forschungsgemeinschaft and Open Access Publishing Fund of University of Tübingen. Further, we would like to thank Mario Kleiner for his valuable input and technical support.

*EC and CC contributed equally to the work
†DE and MG contributed equally to the work.
Commercial relationships: none.
Corresponding author: Enrico Chiovetto.
Email: enrico.chiovetto@uni-tuebingen.de.
Address: Section Computational Sensomotorics, Department Cognitive Neurology, University Clinic Tübigen, Germany.

References

Akaike, H. (1974). A new look at the statistical model identification. *Automatic Control, IEEE Transactions on*, *19*(6), 716–723, https://doi.org/10.1109/TAC.1974.1100705.

Bartlett, M. S. (2010). Emotion simulation and expression understanding: A case for time. *Behavioral and Brain Sciences*, *33*(6), 434–435.

Bartlett, M. S., Littlewort, G. C., Frank, M. G., Lainscsek, C., Fasel, I. R., & Movellan, J. R. (2006). Automatic recognition of facial actions in spontaneous expressions. *Journal of Multimedia*, *1*(6), 22–35.

Bartlett, M. S., Movellan, J. R., & Sejnowski, T. J. (2002, Nov). Face recognition by independent component analysis. *IEEE Transactions on Neural Networks*, *13*(6), 1450–1464, https://doi.org/10.1109/TNN.2002.804287.

Beck, T., Endres, D., Lindner, A., & Giese, M. A. (2014). Active sampling supported comparison of causal inference models for agency attribution in

goal-directed actions. *Journal of Vision*, *14*(10): 838, https://doi.org/10.1167/14.10.838. [Abstract]

Berret, B., Bonnetblanc, F., Papaxanthis, C., & Pozzo, T. (2009). Modular control of pointing beyond arm's length. *The Journal of Neuroscience*, *29*(1), 191–205.

Bishop, C. M. (2007). *Pattern recognition and machine learning*. Secaucus, NJ: Springer-Verlag New York, Inc.

Bizzi, E., Cheung, V., d'Avella, A., Saltiel, P., & Tresch, M. (2008). Combining modules for movement. *Brain Research Reviews*, *57*(1), 125–133.

Bofill, P. (2003). Underdetermined blind separation of delayed sound sources in the frequency domain. *Neurocomputing*, *55*(34), 627–641. Available from http://www.sciencedirect.com/science/article/pii/S0925231202006318 (Evolving Solution with Neural Networks), http://doi.org/10.1016/S0925-2312(02)00631-8.

Boukricha, H., Wachsmuth, I., Hofstätter, A., & Grammer, K. (2009). Pleasure-arousal-dominance driven facial expression simulation. *3rd International Conference on Affective Computing and Intelligent Interaction and Workshops (ACII)*, (pp. 1–7).

Cheung, V. C., Piron, L., Agostini, M., Silvoni, S., Turolla, A., & Bizzi, E. (2009). Stability of muscle synergies for voluntary actions after cortical stroke in humans. *Proceedings of the National Academy of Sciences, USA*, *106*(46), 19563–19568.

Chiovetto, E., Berret, B., Delis, I., Panzeri, S., & Pozzo, T. (2013). Investigating reduction of dimensionality during single-joint elbow movements: A case study on muscle synergies. *Frontiers in Computational Neuroscience*, *7*, 11, https://doi.org/10.3389/fncom.2013.00011.

Chiovetto, E., Berret, B., & Pozzo, T. (2010). Tridimensional and triphasic muscle organization of whole-body pointing movements. *Neuroscience*, *170*(4), 1223–1238. Available from http://www.sciencedirect.com/science/article/pii/S030645221000984X (reviewed), https://doi.org/10.1016/j.neuroscience.2010.07.006.

Chiovetto, E., d'Avella, A., & Giese, M. (2016). A unifying framework for the identification of motor primitives. *arXiv preprint arXiv:1603.06879*.

Chiovetto, E., & Giese, M. A. (2013). Kinematics of the coordination of pointing during locomotion. *PLoS One*, *8*(11), https://doi.org/10.1371/journal.pone.0079555.

Clark, J. J., & Yuille, A. L. (2013). *Data fusion for sensory information processing systems* (Vol. 105). Berlin: Springer Science & Business Media.

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). Boston, MA.

Curio, C., Giese, M. A., Breidt, M., Kleiner, M., & Bülthoff, H. H. (2008). Exploring human dynamic facial expression recognition with animation. In *International conference on cognitive systems* (pp. 1–6). New York, NY: ACM.

Curio, C., Giese, M. A., Breidt, M., Kleiner, M., Bülthoff, H. (2010). Recognition of dynamic facial action probed by visual adaptation. In *Dynamic faces: Insights from experiments and computation* (pp. 47–65). Cambridge, MA: MIT Press.

d'Avella, A., Portone, A., Fernandez, L., & Lacquaniti, F. (2006). Control of fast-reaching movements by muscle synergy combinations. *The Journal of Neuroscience*, *26*(30), 7791–7810.

de la Rosa, S., Giese, M., Bülthoff, H. H., & Curio, C. (2013). The contribution of different cues of facial movement to the emotional facial expression adaptation aftereffect. *Journal of Vision*, *13*(1):23, 1–15, https://doi.org/10.1167/13.1.23. [PubMed] [Article]

Delis, I., Chen, C., Jack, R. E., Garrod, O. G., Panzeri, S., & Schyns, P. G. (2016). Space-by-time manifold representation of dynamic facial expressions for emotion categorization. *Journal of Vision*, *16*(8):14, 1–20, https://doi.org/10.1167/16.8.14. [PubMed] [Article]

Dobs, K., Bülthoff, I., Breidt, M., Vuong, Q. C., Curio, C., & Schultz, J. (2014). Quantifying human sensitivity to spatio-temporal information in dynamic faces. *Vision Research*, *100*, 78–87.

Dominici, N., Ivanenko, Y. P., Cappellini, G., dAvella, A., Mondì, V., Cicchese, M., … Lacquaniti F. (2011, November 18). Locomotor primitives in newborn babies and their development. *Science*, *334*(6058), 997–999.

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

Ekman, P., Friesen, W. V., & Ellsworth, P. (1972). *Emotion in the human face: Guidelines for research and an integration of findings*. New York, NY: Pergamon Press.

Ekman, P., Friesen, W. V., O'Sullivan, M., Chan, A., Diacoyanni-Tarlatzis, I., Heider, K., … Ricci-Bitti P. E., et al. (1987). Universals and cultural differences in the judgments of facial expressions of

emotion. *Journal of Personality and Social Psychology*, *53*(4), 712.

Emile, B., & Common, P. (1998). Estimation of time delays between unknown colored signals. *Signal Processing*, *68*(1), 93–100.

Endres, D., Chiovetto, E., & Giese, M. (2013). Model selection for the extraction of movement primitives. *Frontiers in Computational Neuroscience*, *7*, 185. Available from http://www.frontiersin.org/computational_neuroscience/10.3389/fncom.2013.00185/abstract, https://doi.org/10.3389/fncom.2013.00185.

Ernst, M. O., & Banks, M. S. (2002, January 24). Humans integrate visual and haptic information in a statistically optimal fashion. *Nature*, *415*(6870), 429–433.

Ezzat, T., Geiger, G., & Poggio, T. (2002). Trainable videorealistic speech animation. In *Proceedings of the 29th annual conference on computer graphics and interactive techniques*(pp. 388–398). New York, NY: ACM. Available from http://doi.acm.org/10.1145/566570.566594

Flash, T., & Hochner, B. (2005). Motor primitives in vertebrates and invertebrates. *Current Opinion in Neurobiology*, *15*(6), 660–666.

Friesen, W. V., & Ekman, P. (1983). *Emfacs-7: Emotional facial action coding system*. Unpublished manuscript, University of California at San Francisco, San Francisco, CA, *2*(36), 1.

Girges, C., Spencer, J., & O'Brien, J. (2015). Categorizing identity from facial motion. *The Quarterly Journal of Experimental Psychology*, *68*(9), 1832–1843.

Hill, H., & Johnston, A. (2001). Categorizing sex and identity from the biological motion of faces. *Current Biology*, *11*(11), 880–885.

Ivanenko, Y. P., Poppele, R. E., & Lacquaniti, F. (2004). Five basic muscle activation patterns account for muscle activity during human locomotion. *The Journal of Physiology*, *556*(1), 267–282.

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, USA*, *109*(19), 7241–7244, https://doi.org/10.1073/pnas.1200155109.

Kähler, K., Haber, J., Yamauchi, H., & Seidel, H.-P. (2002). Head shop: Generating animated head models with anatomical structure. In *Proceedings of the 2002 ACM siggraph/eurographics symposium on computer animation* (pp. 55–63).

Kalberer, G. A., & Gool, L. V. (2001). Lip animation based on observed 3d speech dynamics. In S. El-Hakim & A. Gruen (Eds.), *Proceedings of SPIE* (Vol. 4309, pp. 16–25).

Kaminski, T. (2007). The coupling between upper and lower extremity synergies during whole body reaching. *Gait & Posture*, *26*(2), 256–262.

Knappmeyer, B., Thornton, I. M., & Bülthoff, H. H. (2003). The use of facial motion and facial form during the processing of identity. *Vision Research*, *43*(18), 1921–1936.

Körding, K. K., Beierholm, U., Ma, W. J., Quartz, S., Tenenbaum, J. B., & Shams, L. (2007). Causal inference in multisensory perception. *PLoS One*, *2*(9), e943.

Kshirsagar, S., Molet, T., & Magnenat-Thalmann, N. (2001). Principal components of expressive speech animation. In *Computer graphics international 2001* (pp. 38–46). Washington, DC, USA: IEEE Computer Society. Available from http://dl.acm.org/citation.cfm?id=647781.735231.

Lang, P. J., Greenwald, M. K., Bradley, M. M., & Hamm, A. O. (1993). Looking at pictures: Affective, facial, visceral, and behavioral reactions. *Psychophysiology*, *30*(3), 261–273.

Laplace, P. S. (1774). Mémoire sur la probabilité des causes par les événemens [Memory on the probability of the causes of the events]. *Savants Étranges*, *6*, 621–656.

Larsen, R. J., & Diener, E. (1992). Promises and problems with the circumplex model of emotion. In M. S. Clark (Ed.), *Review of personality and social psychology: Emotion* (Vol. 13. pp. 25–59). Newbury Park, CA: Sage.

Lee, D. D., & Seung, H. S. (1999, October 21). Learning the parts of objects by non-negative matrix factorization. *Nature*, *401*(6755), 788–791.

Lee, Y., Terzopoulos, D., & Waters, K. (1995). Realistic modeling for facial animation. In *Proceedings of the 22nd annual conference on computer graphics and interactive techniques* (pp. 55–62).

Littlewort, G., Whitehill, J., Wu, T.-F., Butko, N., Ruvolo, P., Movellan, J., & Bartlett, M. (2011). The motion in emotion—A cert based approach to the fera emotion challenge. *Proceedings of the IEEE International Conference on Automatic Face & Gesture Recognition and Workshops*. (pp. 897–902).

Mueller, P., Kalberer, G. A., Proesmans, M., & Gool, L. V. (2005, August). Realistic speech animation based on observed 3d face dynamics. *IEEE Proceedings on Vision, Image & Signal Processing*, *152*, 491–500.

Mukovskiy, A., Land, W. M., Schack, T., & Giese, M. A. (2015). Modeling of predictive human move-

ment coordination patterns for applications in computer graphics. *Journal of WSCG*, *23*(2), 139–146.

Niedenthal, P. M., Mermillod, M., Maringer, M., & Hess, U. (2010. The simulation of smiles (sims) model: Embodied simulation and the meaning of facial expression. *Behavioral and Brain Sciences*, *33*, 417–433, https://doi.org/10.1017/S0140525X10000865.

Omlor, L., & Giese, M. (2007a). Blind source separation for over-determined delayed mixtures. In B. Schölkopf, J. Platt, & T. Hoffman (Eds.), *Advances in neural information processing systems 19* (pp. 1049–1056). Cambridge, MA: MIT Press.

Omlor, L., & Giese, M. A. (2007b). Extraction of spatio-temporal primitives of emotional body expressions. *Neurocomputing*, *70*(10–12), 1938–1942.

O'Toole, A. J., Roark, D. A., & Abdi, H. (2002). Recognizing moving faces: A psychological and neural synthesis. *Trends in Cognitive Sciences*, *6*(6), 261–266.

Parke, F. I. (1972). Computer generated animation of faces. In *Proceedings of the ACM annual conference* (Vol. 1, pp. 451–457).

Parke, F. I., & Waters, K. (2008). *Computer facial animation*. Wellesley, MA: CRC Press.

Reinl, M., & Bartels, A. (2014). Face processing regions are sensitive to distinct aspects of temporal sequence in facial dynamics. *NeuroImage*, *102*, 407–415.

Roether, C. L., Omlor, L., Christensen, A., & Giese, M. A. (2009). Critical features for the perception of emotion from gait. *Journal of Vision*, *9*(6):15, 1–32, https://doi.org/10.1167/9.6.15. [PubMed] [Article]

Russell, J. (1980). A circumplex model of affect. *Journal of Personality and Social Psychology*, *39*(6), 1161–1178.

Santello, M., Flanders, M., & Soechting, J. F. (1998). Postural hand synergies for tool use. *The Journal of Neuroscience*, *18*(23), 10105–10115.

Santello, M., Flanders, M., & Soechting, J. F. (2002). Patterns of hand motion during grasping and the influence of sensory guidance. *The Journal of Neuroscience*, *22*(4), 1426–1435.

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics*, *6*(2), 461–464.

Sifakis, E., Selle, A., Robinson-Mosher, A., & Fedkiw, R. (2006). Simulating speech with a physics-based facial muscle model. In *Proceedings of the 2006 ACM siggraph/eurographics symposium on computer animation* (pp. 261–270).

Terzopoulos, D., & Waters, K. (1990). Physically-based facial modelling, analysis, and animation. *Computer Animation and Virtual Worlds*, *1*(2), 73–80.

Thornton, I. M., & Kourtzi, Z. (2002). A matching advantage for dynamic human faces. *Perception*, *31*(1), 113–132.

Ting, L. H., & Macpherson, J. M. (2005). A limited set of muscle synergies for force control during a postural task. *Journal of Neurophysiology*, *93*(1), 609–613.

Torres-Oviedo, G., & Ting, L. H. (2007). Muscle synergies characterizing human postural responses. *Journal of Neurophysiology*, *98*(4), 2144–2156.

Wu, T., Bartlett, M. S., & Movellan, J. R. (2010). Facial expression recognition using Gabor motion energy filters. In *2010 IEEE computer society conference on computer vision and pattern recognition-workshops* (pp. 42–47).

Yilmaz, Ö., & Rickard, S. (2004). Blind separation of speech mixtures via time-frequency masking. *IEEE Transactions on Signal Processing*, *52*, 1830–1847.