



# Human Observers and Automated Assessment of Dynamic Emotional Facial Expressions: KDEF-dyn Database Validation

Manuel G. Calvo<sup>1,2</sup>, Andrés Fernández-Martín<sup>3</sup>, Guillermo Recio<sup>4\*</sup> and Daniel Lundqvist<sup>5</sup>

<sup>1</sup> Department of Cognitive Psychology, Universidad de La Laguna, San Cristóbal de La Laguna, Spain, <sup>2</sup> Instituto Universitario de Neurociencia (IUNE), Universidad de La Laguna, Santa Cruz de Tenerife, Spain, <sup>3</sup> Department of Health Sciences, Universidad Internacional de la Rioja, Logroño, Spain, <sup>4</sup> Institute of Psychology, Universität Hamburg, Hamburg, Germany, <sup>5</sup> Department of Clinical Neuroscience, Karolinska Institutet, Stockholm, Sweden

## OPEN ACCESS

### Edited by:

Tjeerd Jellema,  
University of Hull, United Kingdom

### Reviewed by:

Xunbing Shen,  
Jiangxi University of Traditional  
Chinese Medicine, China  
Alessio Miolla,  
Università degli Studi di Padova, Italy

### \*Correspondence:

Guillermo Recio  
guillermo.recio@gmail.com

### Specialty section:

This article was submitted to  
Emotion Science,  
a section of the journal  
Frontiers in Psychology

**Received:** 17 May 2018

**Accepted:** 05 October 2018

**Published:** 26 October 2018

### Citation:

Calvo MG, Fernández-Martín A,  
Recio G and Lundqvist D (2018)  
Human Observers and Automated  
Assessment of Dynamic Emotional  
Facial Expressions: KDEF-dyn  
Database Validation.  
Front. Psychol. 9:2052.  
doi: 10.3389/fpsyg.2018.02052

Most experimental studies of facial expression processing have used static stimuli (photographs), yet facial expressions in daily life are generally dynamic. In its original photographic format, the Karolinska Directed Emotional Faces (KDEF) has been frequently utilized. In the current study, we validate a dynamic version of this database, the KDEF-dyn. To this end, we applied animation between neutral and emotional expressions (happy, sad, angry, fearful, disgusted, and surprised; 1,033-ms unfolding) to 40 KDEF models, with morphing software. Ninety-six human observers categorized the expressions of the resulting 240 video-clip stimuli, and *automated face analysis* assessed the evidence for 6 expressions and 20 facial action units (AUs) at 31 intensities. Low-level image properties (luminance, signal-to-noise ratio, etc.) and other purely perceptual factors (e.g., size, unfolding speed) were controlled. Human recognition performance (accuracy, efficiency, and confusions) patterns were consistent with prior research using static and other dynamic expressions. Automated assessment of expressions and AUs was sensitive to intensity manipulations. Significant correlations emerged between human observers' categorization and automated classification. The KDEF-dyn database aims to provide a balance between experimental control and ecological validity for research on emotional facial expression processing. The stimuli and the validation data are available to the scientific community.

**Keywords:** facial expression, dynamic, action units, KDEF, FACET

## INTRODUCTION

Research on facial expression processing (see reviews in Nelson and Russell, 2013; Calvo and Nummenmaa, 2016) has generally utilized static faces as stimuli, obtained from standardized databases such as the Pictures of Facial Affect (PoFA; Ekman and Friesen, 1976), the Karolinska Directed Emotional Faces (KDEF; Lundqvist et al., 1998), the NimStim Stimulus Set (Tottenham et al., 2002), the Radboud Faces Database (RaFD; Langner et al., 2010), FACES (Ebner et al., 2010) and others (for a review and evaluation, see Cowie et al., 2005; Anitha et al., 2010; Sandbach et al., 2012). Yet, in social encounters and face-to-face communication, facial expressions are generally

dynamic. Further, research has shown that motion benefits affect recognition (see Krumhuber et al., 2013; Calvo et al., 2016; Wingenbach et al., 2016). Accordingly, it is important to use dynamic stimuli for investigating recognition of facial expressions.

A number of *dynamic* expression databases have been developed, generally involving on-line video recordings of facial activity, which represent a valuable advance (e.g., van der Schalk et al., 2011; Banziger et al., 2012; Kaulard et al., 2012; Zhang et al., 2014; O'Reilly et al., 2016; Wingenbach et al., 2016). Krumhuber et al. (2017) have reviewed and discussed the major issues of 22 dynamic expression databases. In the current study, the proposal of a new stimulus set (KDEF-dyn) aims to make a contribution by taking two issues into account. First, the control of possible *perceptual confounds* with non-expressive factors that may affect expression recognition. They involve low-level image properties of the stimuli, such as illumination and light source, size of the face relative to the background, head-face orientation, or changes in facial appearance like hair, make up, eyeglasses, jewelry, etc. They may be difficult to control for in video-recordings of spontaneous expressions. Yet, to unequivocally attribute emotion recognition to facial expression *per se*, all the facial stimuli across types of expressions must be comparable on these non-expressive factors. Further, the control of such factors may be critical for paradigms using neurophysiological (such as event-related potentials, ERPs; see Naples et al., 2015) or eyetracking (e.g., probability of first fixation in a particular face region, or pupillometry; e.g., Calvo and Nummenmaa, 2011) measures, which are particularly sensitive to physical image properties. To this end, all the face stimuli in our KDEF-dyn set are standardized in size, resolution, location, and frontal view, in addition to multiple low-level image properties (luminance, contrast, etc.).

A second issue is concerned with the objective validation of expressions and component facial actions across *multiple intensities*. According to Valstar et al. (2015, 2017), many existing benchmark databases show expressions at fixed intensities (generally, the apex or maximum intensity) or do not support the evaluation of intensity effects. Computational algorithms have been developed to automatically detect Facial Action Coding System (FACS) action units (AUs; Ekman et al., 2002), which are anatomical changes in the facial morphology that can be associated to specific emotions (e.g., AU12 or lip corner puller, to happiness; or AU4, brow lowerer, to anger; etc.). Manual FACS-coding by expert raters (van der Schalk et al., 2011; Banziger et al., 2012), and also automated computation (Lucey et al., 2010; Cosker et al., 2011; Mavadati et al., 2013; Zhang et al., 2014), have been applied to dynamic expression databases only on the apex. The estimation at multiple intensities is, however, required because, in real life, expressions vary in intensity, which is often a critical cue to interpret their meaning. Accordingly, we computed the objective evidence of each of six basic expressions and also the evidence of each of 20 AUs, across 31 intensities from neutral (0% intensity) to emotional (100% intensity) in 3.33% intensity steps. This adds to recent work (Calvo et al., 2016; Wingenbach et al., 2016) regarding the role of intensity on the categorization of dynamic expressions. This approach will be particularly useful for expression discrimination studies, e.g., the lowest intensity

or threshold at which a particular emotion is recognized and differentiated from others and from neutral faces.

With these two issues in mind, in the current study we developed and validated a dynamic version (KDEF-dyn) of the original KDEF database in static format (Lundqvist et al., 1998), to extend research possibilities. The photographic KDEF stimuli have been validated in large norming studies (Calvo and Lundqvist, 2008; Goeleven et al., 2008), and widely used in behavioral (e.g., Calvo et al., 2013; Sanchez et al., 2014; Gupta et al., 2016) and neurophysiological (e.g., Bublitzky et al., 2014; Calvo and Beltrán, 2014; Adamaszek et al., 2015) research. The original KDEF database has been cited in over 1,980 published articles, according to Google Scholar<sup>1</sup> (accessed 18.09.2018). We took advantage of this research on the static KDEF stimuli to produce dynamic expressions of 40 different models, each portraying the six basic emotions.

To *develop* dynamic expressions, we applied morphing animation software (FantaMorph, v. 5.4.2; Abrosoft) to the original KDEF photographs. For each encoder and emotion, we created a 1,033-ms video-clip of 31 frames starting with a neutral face and ending with a full-blown emotional face. Thus, we tried to mimic real-life expressions and approximate the average natural speed of emotional expression development from a neutral face, since apex of facial expression is generally reached within 1 s for basic emotions (Pollick et al., 2003; Hoffmann et al., 2010). Admittedly, dynamic morphing creates linear movement, which can make expressions appear as less natural than on-line video recordings. Nevertheless, although non-linear changes are generally judged as more natural than linear motion, morphing does not necessarily compromise naturalness (Cosker et al., 2010, 2015). In fact, dynamically morphed facial expressions have often been employed in prior research on facial emotion recognition, with behavioral (Hoffmann et al., 2010; Fiorentini and Viviani, 2011; Recio et al., 2013; Calvo et al., 2016) and neurophysiological (Popov et al., 2013; Harris et al., 2014; Recio et al., 2014; Vrticka et al., 2014) measures being sensitive to expression manipulations. The morphing technique involves some advantages, such as fine-grained control and standardization of expressive intensity, unfolding speed, and duration. We chose this approach as a balance between (reduced) ecological validity and (enhanced) experimental control.

To *validate* the KDEF-dyn database, we followed two approaches, each with several measures. First, we collected data from *human observers* in an expression categorization task including measures of (a) correct recognition responses, i.e., the probability that they coincided with the intended KDEF expression, (b) reaction times indicating processing efficiency, and (c) the probability of confusions across different expressions, for each of the six basic emotions. Second, with Emotient FACET software (v. 6.1.2667.3; iMotions), we performed *automated facial expression analyses* (Bartlett and Whitehill, 2011; Olderbak et al., 2014; Cohn and De la Torre, 2015; Girard et al., 2015; Dente et al., 2017) of (a) the probability of each expression to be detected, as a function of spatial maps of

<sup>1</sup>[https://scholar.google.com/scholar?cites=93971208802805184&as\\_sdt=2005](https://scholar.google.com/scholar?cites=93971208802805184&as_sdt=2005)

facial features, and also (b) the probability of each of 20 AUs to be activated, i.e., muscle movements, according to FACS (Ekman and Friesen, 1978; Ekman et al., 2002). The automated analyses of expressions and AUs were performed for 31 intensities (including the neutral baseline) of each emotional facial expression (including apex), while the human recognition measures were obtained for the maximum expressive intensity only. These measures indicate to what extent each KDEF stimulus is consistently categorized, the objective evidence for each facial expression configuration, and the specific morphological features.

The current KDEF-dyn database contributes to existing databases of dynamic facial expression stimuli in several respects. First, the combined validation approach (with both 'subjective' human categorization data and 'objective' automated assessment data) provides researchers with empirical and theoretical criteria to select stimuli depending on various dimensions (recognition accuracy and efficiency, susceptibility to specific confusions, and automated classification of expressions and AUs). In a dataset file (see **Supplementary Dataset S1**), each stimulus can be ordered according to each of these measures. Second, due to the standardization of expression unfolding speed and duration for all the stimuli, the present database allows for a fine-grained investigation of emotion recognition as a function of expressive intensity. We provide evidence values from automated analysis of expressions and AUs for each frame of each video-clip. In a dataset file (see **Supplementary Dataset S2**), such values are shown for each of 31 intensity levels of each stimulus, from 0 (neutral) to 100% (full-blown emotion). Third, another novel contribution involves the control of multiple non-expressive perceptual factors (e.g., low-level image properties) that might otherwise confound expression recognition differences. In a dataset file (see **Supplementary Dataset S3**), each stimulus has been quantified in terms of such perceptual factors across each of 31 expressive intensity levels. Potential applications and limitations will be considered in the Section "Discussion."

## MATERIALS AND METHODS

### Participants

Ninety-six university undergraduates (56 females and 40 males; aged 18–30 years;  $M = 21.2$  years) from different courses (Psychology, Medicine, Law, Economics, and Education) participated voluntarily for payment (5 €) or course credit, after signing written informed consent. Four more participants were excluded from the analyses because their mean correct recognition rate was below 50% for three or more expressions. An *a priori* power calculation using *G\*Power* (v. 3.1.9.2; Faul et al., 2007) showed that 46 participants would be sufficient to detect a medium effect size (Cohen's  $d = 0.60$ ) at  $\alpha = 0.05$ , with power of 0.98. As this was a norming study of stimulus materials, a larger participant sample was used to obtain stable and representative average scores for each stimulus. The study was approved by the Ethics Committee of University of La Laguna (protocol CEIBA2017-0227), and was conducted in accordance with the Declaration of Helsinki 2008.

### Stimuli

The color photographs of 40 posers (20 females and 20 males) in frontal view from the KDEF database (Lundqvist et al., 1998) displaying six emotional facial expressions (happiness, sadness, anger, fear, disgust, and surprise) were used. The KDEF identities (see **Supplementary Dataset S1**) were the same as in a previous norming study using photographic stimuli (Calvo and Lundqvist, 2008). For the current study, 240 dynamic video-clip versions (1,033-ms duration) of the original KDEF photographs were constructed. The face stimuli were morphed with FantaMorph (Abrosoft) computer software. For each expression of each poser, we created a 1,033-ms sequence of 31 (33.33-ms) frames smoothly increasing expressive intensity at 30 frames per second (fps), starting with a neutral face as the first frame (frame 0; original KDEF), and ending with an emotional face (happy, sad, etc.) as the final frame (frame 30; original KDEF). Video-clips are shown as supporting information (see **Supplementary Dataset S4**). A very similar or identical procedure and display duration was used previously (Schultz and Pilz, 2009; Johnston et al., 2013; Wingenbach et al., 2016). Each face stimulus subtended a visual angle of  $10.6^\circ$  (height)  $\times$   $8^\circ$  (width) at a 70-cm viewing distance (this approximates the size of a real face, i.e.,  $18.5 \times 13.8$  cm, from a 1-m distance).

### Procedure

The 96 participants were presented with all 240 video-clips (40 posers  $\times$  6 expressions) in six blocks of 40 trials each, and a short break after each block. Block order was counterbalanced, and trial order and type of expression were randomized within each block. The stimuli were displayed on a computer screen (12-in TFT LED LCD with a  $1,366 \times 768$  resolution) by means of E-Prime 2.0 software. Participants were told that short videos of faces with different expressions would be presented, and were asked to indicate which expression was shown on each trial, by pressing a key out of six, as soon and as accurately as possible, with their dominant index finger. Between trials, the index finger was placed at a predetermined location in the middle of the spacebar, equidistant from all six response keys (from 4 to 9). During the instructions, the six basic expressions were identified, as well as the location of the keys to be pressed for each category. Twelve video-clips of two additional, non-KDEF encoders displaying six emotional expressions served as practice trials.

The sequence of events on each trial was as follows. After an initial 500-ms central fixation cross on a screen, a video-clip showed a facial expression that unfolded for 1,033 ms. Following face offset, graphical instructions appeared on the screen for responding: Six small boxes were arranged horizontally, numbered from 4 to 9, with each box/number associated to a verbal label (e.g., 4: happy; 5: sad, etc.). The assignment of expressions to numbers was counterbalanced across participants. For categorizing each expression, participants pressed one key (from 4 to 9) in the upper row of a standard computer keyboard. The selected response and reaction times (RTs; from the video-clip offset) were recorded. There was a 1,500-ms intertrial interval.

## Design and Measures

We used a within-subjects experimental design, with expressive category (happiness, sadness, anger, fear, disgust, and surprise) as a factor. As dependent variables, we measured hits, i.e., the probability that responses coincided with the displayed expression (e.g., responding “happy” when the face stimulus was intended to convey happiness), and RTs. In addition, we identified the type of confusions, i.e., the probability that each target (the actually displayed expression) was categorized as each of the other five, non-target expressions (e.g., if the target was anger on a trial, the five non-targets were happiness, sadness, disgust, fear, and surprise). These measures, along with those involving automated expression analysis (see below), are provided as supplementary data for each KDEF-dyn stimulus (see **Supplementary Dataset S1**).

## Automated Facial Expression Analysis

In addition to the human observers’ performance measures, we subjected the video stimuli to automated face analysis by means of Emotient FACET software, which is assumed to detect facial features (e.g., mouth corners) and feature groups, and then to classify the image as belonging to a particular emotional expression category by comparing the resulting output maps with template images. Recently, FACET has been used in psychological and applied research (see Dente et al., 2017). The automated analysis provides two types of measures (see Gordon et al., 2011; Olderbak et al., 2014): (a) *expression* evidence scores for each category: joy, anger, surprise, fear, disgust, sadness, and contempt, in addition to neutral; and (b) AUs evidence scores (for 20 AUs: 1, 2, 4, 5, 6, 7, 9, 10, 12, 14, 15, 17, 18, 20, 23, 24, 25, 26, 28, and 43), according to FACS (Ekman et al., 2002); see also (Cohn et al., 2007; Cohn and De la Torre, 2015). AUs are anatomically related to the movement of specific face muscles (e.g., AU12 involves the contraction of the zygomaticus major muscle, which draws the angle of the mouth superiorly and posteriorly to allow for smiling).

We obtained expression and AU evidence scores for each of 31 frames across the 1,033-ms unfolding, for each poser and expression (see **Supplementary Dataset S2**). The FACET evidence scores quantify the odds (in decimal logarithmic scale) of each expression or AU to be present in a given face stimulus, and can be transformed into probabilities ( $p$ ) with the formula  $p = 1/(1 + 10^{-\text{evidence score}})$ . An evidence score of zero indicates chance level (0.50/0.50). Positive values indicate greater probabilities that a given expression or AU is present, and negative values indicate greater probabilities that an expression or AU is unlikely to be present in the stimulus. All evidence scores above 1 will approach the probability value of 1, and all evidence scores below  $-1$  will approach a 0 probability. This implies that evidence scores (in odds ratios) are more discriminative than probabilities to detect subtle changes, and the former are more suitable for statistical tests because they tend to be normally distributed. The evidence scores ranged in a continuous scale between  $-12$  and  $12$ . We conducted Kolmogorov–Smirnov and Levene’s tests to exam the assumptions of ANOVA regarding normality and homoscedasticity, respectively. Results revealed

that most residuals of the evidence scores for expressions and AUs were normally distributed and homoscedastic (for multivariate ANOVA with the evidence scores used as dependent variables and expression category as a fixed factor; see **Supplementary Dataset S2**).

## Low-Level Stimulus Image Properties

To examine potential physical and perceptual differences among expression categories across the 1,033-ms unfolding display, we computed (with Matlab 7.0, The Mathworks) the following low-level image statistics of each neutral face and the respective emotional faces for each of 31 frames, at consecutive expressive intensity levels, from 0% intensity (i.e., neutral face) to full-blown emotion (i.e., 100% intensity), in 3.33% steps: mean and variance of luminance, RMS or root mean square contrast, skewness, kurtosis, SNR or signal-to-noise ratio, and entropy. Each low-level property was analyzed by means of a (6: Expression Stimulus)  $\times$  31 (Intensity Levels) ANOVA. All the measures were sensitive to the effects of intensity, all  $F_s(30,7020) \geq 38.44$ ,  $p < 0.0001$ ,  $\eta_p^2 \geq 0.14$ ), but, importantly, the main effect of expression was never significant (all  $F_s < 1$ , except for skewness:  $F(5,234) = 1.51$ ,  $p = 0.19$ , *ns*; see **Supplementary Dataset S3**). Accordingly, the face stimuli of the different expressions did not significantly differ in such physical properties. This rules out purely perceptual factors as responsible for the differences observed in categorization performance by human observers or automated facial expression classification (see below).

## RESULTS

We wanted to relate human observers’ performance and automated facial expression analysis, which had to be conducted for each *stimulus*. Further, the study aimed to obtain and provide other researchers with validation measures for each *stimulus* (i.e., KDEF model identity). Accordingly, the statistical analyses were performed on the stimuli as the error term. This means that the recognition performance scores of the 96 participants were averaged for each of the 240 video-clip stimuli, which served as the units of analysis, with an  $N = 40$  for each expression category. All the multiple *post hoc* comparisons in the following analyses involved Bonferroni corrections (with a  $p < 0.05$  threshold).

## Analyses of Recognition Performance and Confusions by Human Observers

For response *accuracy*, a one-way (6: Expression) ANOVA yielded significant effects,  $F(5,234) = 32.07$ ,  $p < 0.0001$ ,  $\eta_p^2 = 0.41$ . *Post hoc* contrasts revealed significantly better recognition of happiness, surprise, and anger, than sadness and disgust, which were recognized better than fear (see **Table 1**). The correct response *reaction times*,  $F(5,234) = 69.91$ ,  $p < 0.0001$ ,  $\eta_p^2 = 0.60$ , were faster for happiness than for any other expression, followed by surprise and anger (which did not differ from each other), and by disgust and sadness (which did not differ from each other), with fear being recognized more slowly than the other categories. Pairwise (Pearson) *correlations* between response accuracy and

**TABLE 1** | Mean proportion (%) of hits and confusions in human observers' responses, and reaction times (for hits only) for each target (stimulus) expression.

Expression stimulus	Expression response					
	Happiness	Surprise	Anger	Sadness	Disgust	Fear
Happiness	<b>98.5<sup>a</sup></b>	1.0 <sup>b</sup>	0.0 <sup>b</sup>	0.0 <sup>b</sup>	0.3 <sup>b</sup>	0.2 <sup>b</sup>
Surprise	2.8 <sup>b</sup>	<b>93.7<sup>a</sup></b>	0.1 <sup>c</sup>	0.0 <sup>c</sup>	0.2 <sup>c</sup>	3.2 <sup>b</sup>
Anger	0.2 <sup>c</sup>	0.8 <sup>bc</sup>	<b>91.7<sup>a</sup></b>	1.3 <sup>bc</sup>	3.6 <sup>b</sup>	2.4 <sup>b</sup>
Sadness	0.5 <sup>c</sup>	0.7 <sup>c</sup>	1.8 <sup>c</sup>	<b>80.7<sup>a</sup></b>	5.7 <sup>b</sup>	10.6 <sup>b</sup>
Disgust	0.1 <sup>d</sup>	0.5 <sup>d</sup>	13.4 <sup>b</sup>	4.7 <sup>bc</sup>	<b>77.8<sup>a</sup></b>	3.5 <sup>c</sup>
Fear	0.8 <sup>d</sup>	18.5 <sup>b</sup>	1.1 <sup>d</sup>	2.5 <sup>d</sup>	8.5 <sup>c</sup>	<b>68.6<sup>a</sup></b>
Hits	98.5 <sup>a</sup>	93.7 <sup>a</sup>	91.7 <sup>a</sup>	80.7 <sup>b</sup>	77.8 <sup>b</sup>	68.6 <sup>c</sup>
Hit RTs	868 <sup>a</sup>	1,061 <sup>b</sup>	1,140 <sup>b</sup>	1,253 <sup>c</sup>	1,229 <sup>c</sup>	1,431 <sup>d</sup>

Within each expression stimulus category (horizontally), scores with different letters across expression response (i.e., on the same line) are significantly different in post hoc multiple contrasts ( $p < 0.05$ , Bonferroni corrected); expressions sharing a letter are equivalent. Boldface for hits in columns.

**TABLE 2** | Mean raw evidence scores (odds ratios) of each expression (response) for each target (stimulus) expression.

Expression stimulus	Expression response						
	Happiness	Surprise	Anger	Sadness	Disgust	Fear	Neutral
Happiness	<b>6.4<sup>a</sup></b>	-8.6 <sup>e</sup>	-7.7 <sup>d</sup>	-9.2 <sup>e</sup>	-5.4 <sup>c</sup>	-3.9 <sup>b</sup>	-11.9 <sup>f</sup>
Surprise	-5.6 <sup>d</sup>	<b>3.5<sup>a</sup></b>	-3.9 <sup>c</sup>	-6.1 <sup>d</sup>	-4.5 <sup>c</sup>	0.8 <sup>b</sup>	-4.8 <sup>cd</sup>
Anger	-6.3 <sup>e</sup>	-4.8 <sup>d</sup>	<b>1.7<sup>a</sup></b>	-2.8 <sup>c</sup>	-0.4 <sup>b</sup>	-2.8 <sup>c</sup>	-2.6 <sup>bc</sup>
Sadness	-4.6 <sup>e</sup>	-4.1 <sup>e</sup>	-2.3 <sup>d</sup>	<b>1.7<sup>a</sup></b>	-1.3 <sup>c</sup>	-0.2 <sup>b</sup>	-2.2 <sup>cd</sup>
Disgust	-5.4 <sup>d</sup>	-8.2 <sup>e</sup>	-1.6 <sup>b</sup>	-5.7 <sup>d</sup>	<b>4.2<sup>a</sup></b>	-3.8 <sup>c</sup>	-8.2 <sup>e</sup>
Fear	-4.0 <sup>d</sup>	-1.3 <sup>b</sup>	-2.9 <sup>cd</sup>	-3.1 <sup>cd</sup>	-2.1 <sup>bc</sup>	<b>1.6<sup>a</sup></b>	-4.4 <sup>de</sup>
Target	6.4 <sup>a</sup>	3.5 <sup>b</sup>	1.7 <sup>c</sup>	1.7 <sup>c</sup>	4.2 <sup>b</sup>	1.6 <sup>c</sup>	

Automated analysis computed by Emotient FACET software. Within each expression stimulus category (horizontally), scores with different letters across expression response (i.e., on the same line) are significantly different in post hoc multiple contrasts ( $p < 0.05$ , Bonferroni corrected); expressions sharing a letter are equivalent. Boldface for correct responses to target (stimulus) expressions. Target: correct classification of each stimulus.

reaction times for all the expressions showed that reaction times decreased as accuracy increased (Happiness:  $r = -0.67$ ; Surprise:  $r = -0.72$ ; Anger:  $r = -0.78$ ; Sadness:  $r = -0.64$ ; Disgust:  $r = -0.81$ ; Fear:  $r = -0.71$ ; all  $ps < 0.0001$ ;  $N = 40$ ).

For the analysis of *confusions*, a 6 (Expression Stimulus)  $\times$  6 (Expression Response) ANOVA was conducted. Interactive effects,  $F(25,1170) = 836.53$ ,  $p < 0.0001$ ,  $\eta_p^2 = 0.95$ , were decomposed by means of separate one-way (6: Expression Response) ANOVAs for each expression stimulus. See the mean scores and multiple contrasts in **Table 1**. Facial *happiness*,  $F(5,195) = 11922.15$ ,  $p < 0.0001$ ,  $\eta_p^2 = 1$ , was very unlikely to be confused. *Surprise*,  $F(5,195) = 2952.68$ ,  $p < 0.0001$ ,  $\eta_p^2 = 0.99$ , was slightly confused with fear and happiness. *Anger*,  $F(5,195) = 1625.02$ ,  $p < 0.0001$ ,  $\eta_p^2 = 0.98$ , was slightly confused with disgust and fear. *Sadness*,  $F(5,195) = 427.46$ ,  $p < 0.0001$ ,  $\eta_p^2 = 0.92$ , was confused with fear and disgust more than with other expressions. *Disgust*,  $F(5,195) = 228.31$ ,  $p < 0.0001$ ,  $\eta_p^2 = 0.85$ , was confused with anger and sadness, followed by fear. Finally, *fear*,  $F(5,195) = 315.88$ ,  $p < 0.0001$ ,  $\eta_p^2 = 0.89$ , was confused with surprise, followed by disgust.

## Automated Assessment of Expressions With FACET

The evidence scores for each expression were subjected to a 6 (Expression Stimulus)  $\times$  7 (Expression Response,

i.e., the six basic emotions plus neutral) ANOVA. Main effects of expression stimulus,  $F(5,234) = 73.25$ ,  $p < 0.0001$ ,  $\eta_p^2 = 0.61$ , and response,  $F(6,1404) = 142.17$ ,  $p < 0.0001$ ,  $\eta_p^2 = 0.38$ , and an interaction,  $F(30,1404) = 152.43$ ,  $p < 0.0001$ ,  $\eta_p^2 = 0.77$ , emerged. To decompose the interaction, separate one-way (7: Expression Response) ANOVAs were conducted for each expression stimulus. All the expressions were correctly classified (e.g., facial happiness was classified as joy), with target responses being significantly higher (after Bonferroni corrections) than alternative responses (e.g., happiness classified as surprise, etc.), which were assigned negative scores: Facial *happiness*,  $F(6,234) = 636.60$ ,  $p < 0.0001$ ,  $\eta_p^2 = 0.94$ ; *surprise*,  $F(6,234) = 150.16$ ,  $p < 0.0001$ ,  $\eta_p^2 = 0.79$ ; *anger*,  $F(6,234) = 66.31$ ,  $p < 0.0001$ ,  $\eta_p^2 = 0.63$ ; *sadness*,  $F(6,234) = 61.98$ ,  $p < 0.0001$ ,  $\eta_p^2 = 0.61$ ; *disgust*,  $F(6,234) = 196.70$ ,  $p < 0.0001$ ,  $\eta_p^2 = 0.86$ ; and *fear*,  $F(6,234) = 31.44$ ,  $p < 0.0001$ ,  $\eta_p^2 = 0.45$ . The interaction reflected the fact that the correct response scores were higher for happy expressions, followed by disgust and surprise (which did not differ from each other), followed by anger, sadness, and fear (which did not differ from one another), as indicated by a one-way (6: Expression Stimulus) ANOVA,  $F(5,234) = 64.34$ ,  $p < 0.0001$ ,  $\eta_p^2 = 0.58$ , and multiple *post hoc* comparisons. See the mean scores and contrasts in **Table 2**.

## Automated Assessment of Expressive Intensity With FACET

To examine expression classification by FACET as a function of expressive intensity, we conducted a 6 (Stimulus Expression)  $\times$  31 (Intensity Levels: 0% or neutral, 3.3%, 6.7%, etc., and 100% or full-blown emotion) ANOVA on the evidence scores. Effects of expression,  $F(5,7254) = 420.79$ ,  $p < 0.0001$ ,  $\eta_p^2 = 0.23$ , intensity,  $F(30,7254) = 593.43$ ,  $p < 0.0001$ ,  $\eta_p^2 = 0.71$ , and an interaction,  $F(150,7254) = 23.66$ ,  $p < 0.0001$ ,  $\eta_p^2 = 0.33$ , emerged. Separate one-way (Intensity: 31) ANOVAs were performed for each expression to determine the intensity threshold, i.e., when significant evidence of each emotion started relative to the neutral face baseline. Facial happiness,  $F(30,1209) = 232.76$ ,  $p < 0.0001$ ,  $\eta_p^2 = 0.85$ , started to be correctly classified as such at 13.3% intensity ( $p = 0.003$ , after Bonferroni corrections); disgust,  $F(30,1209) = 146.76$ ,  $p < 0.0001$ ,  $\eta_p^2 = 0.78$ , at 20.0% intensity ( $p = 0.002$ ); surprise,  $F(30,1209) = 109.37$ ,  $p < 0.0001$ ,  $\eta_p^2 = 0.73$ , at 23.3% ( $p = 0.012$ ); anger,  $F(30,1209) = 43.38$ ,  $p < 0.0001$ ,  $\eta_p^2 = 0.52$ , at 26.7% ( $p = 0.02$ ); fear,  $F(30,1209) = 52.47$ ,  $p < 0.0001$ ,  $\eta_p^2 = 0.57$ , at 26.7% ( $p = 0.039$ ); and sadness,  $F(30,1209) = 44.45$ ,  $p < 0.0001$ ,  $\eta_p^2 = 0.53$ , at 36.7% intensity ( $p = 0.007$ ). **Figure 1** shows the pattern of automated expression classification as a function of expressive intensity.

## Automated Assessment of Action Units (AUs) With FACET

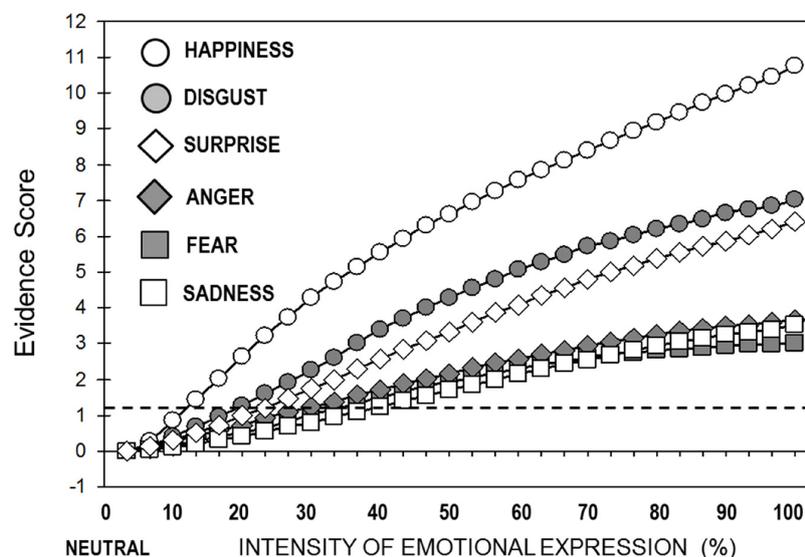
The evidence scores (at 100% intensity of expression) of AUs were subjected to a 6 (Expression Stimulus)  $\times$  20 (AUs) ANOVA. Effects of expression,  $F(5,234) = 30.69$ ,  $p < 0.0001$ ,  $\eta_p^2 = 0.40$ , AUs,  $F(19,4446) = 433.60$ ,  $p < 0.0001$ ,  $\eta_p^2 = 0.65$ , and an interaction,  $F(95,4446) = 100.63$ ,  $p < 0.0001$ ,  $\eta_p^2 = 0.68$ , emerged. For all

the AUs, there were significant differences across expressions, all  $F_s(5,234) \geq 23.64$ ,  $p < 0.0001$ ,  $\eta_p^2 \geq 0.34$ . **Table 3** shows the 100% intensity AU scores.

To interpret the interaction and determine the association of specific AUs to particular expressions, we used two complementary approaches. First, we examined whether, for each AU and emotional expression, the scores were positive and above 0 (thus revealing that an AU was in fact present), by means of *t*-tests for dependent samples. Significant differences appeared for all the AUs in boldface in **Table 3**, all  $t_s(39) \geq 5.53$ ,  $p < 0.0001$ ,  $d \geq 0.87$ . Second, for each AU, we examined whether scores were higher for each emotional expression (at any intensity level from 3.33 to 100%) relative to those for the neutral face, in one-way (31: Intensity level) ANOVAs, followed by Bonferroni ( $p < 0.05$ ) corrections. Significant differences appeared for all the AUs in boldface in **Table 3**,  $F_s(30,1170) \geq 59.62$ ,  $p < 0.0001$ ,  $\eta_p^2 = 0.61$ . **Figure 2** shows the variations in the selected AUs (those that fulfilled both criteria, i.e., significantly above 0 and above neutral faces) across expressive intensities. In sum, facial happiness or joy was significantly characterized by AUs 6, 12, and 25; surprise, by AUs 1, 2, 5, 25, and 26; anger, by AUs 4 and 7; sadness, by AUs 1, 4, and 15; disgust, by AUs 4, 6, 7, 9, and 10; and fear, by AUs 1, 5, and 25.

## Relationships Between Human Observers' Performance (Responses and RTs) and Automated Assessment With FACET (Evidence Scores of Expressions and AUs)

Intra-class correlation (*ICC*, 2) analyses revealed high classification consistency between the automated evidence



**FIGURE 1** | Automated assessment of expressive intensity. Mean automated (FACET) difference (emotional minus neutral) evidence scores of each type of expression (response) across intensity levels for each expression stimulus. Above the dotted line: significantly different from the 0% (neutral) baseline (happy: 13.3% of intensity; disgust: 20.0%; surprise: 23.3%; anger and fear: 26.7%; sadness: 36.7%).

**TABLE 3** | Mean raw evidence scores (odds ratios) of action units (AUs) for each expression (100% expressive intensity).

Action Units		Expression response					
		Happiness	Surprise	Anger	Sadness	Disgust	Fear
AU1	Inner brow raiser	-1.11	<b>1.51</b>	-1.76	<b>1.38</b>	-2.23	<b>1.58</b>
AU2	Outer brow raiser	-0.71	<b>1.93</b>	-1.75	-0.15	-1.60	0.71
AU4	Brow lowerer	-1.65	-1.08	<b>1.86</b>	<b>1.55</b>	<b>1.70</b>	0.92
AU5	Upper lid raiser	-1.41	<b>1.99</b>	0.31	-0.19	-0.97	<b>1.26</b>
AU6	Cheek raiser	<b>2.88</b>	-2.32	-0.24	-0.28	<b>1.02</b>	-0.72
AU7	Lid tightener	0.43	-1.10	<b>0.89</b>	0.09	<b>1.28</b>	-0.15
AU9	Nose wrinkle	-2.49	-5.22	0.19	-2.45	<b>3.48</b>	-3.25
AU10	Upper lip raiser	-0.40	-1.79	0.43	-0.17	<b>3.55</b>	-0.34
AU12	Lip corner puller	<b>4.06</b>	-1.60	-1.80	-1.04	-1.29	-0.76
AU14	Dimpler	-1.73	-2.62	-1.94	-1.37	-3.58	-1.88
AU15	Lip corner depressor	-1.98	-1.87	-0.99	<b>0.98</b>	0.16	-1.22
AU17	Chin raiser	-1.79	-2.53	-0.31	0.25	0.49	-2.02
AU18	Lip puckerer	-9.79	-3.04	-1.69	-2.08	-4.97	-3.83
AU20	Lip stretcher	-0.23	-0.74	-1.48	-0.08	-0.07	0.37
AU23	Lip tightener	-1.58	-1.13	-0.23	-0.80	-0.89	-1.09
AU24	Lip pressor	-2.89	-3.38	-1.07	-0.97	-2.54	-2.77
AU25	Lips part	<b>2.07</b>	<b>2.58</b>	-1.41	-1.31	0.94	<b>1.48</b>
AU26	Jaw drop	-0.05	<b>2.27</b>	-2.44	-1.76	-1.82	0.02
AU28	Lip suck	-3.42	-4.84	-3.80	-2.99	-6.07	-3.54
AU43	Eyes closed	-3.45	-0.96	-1.20	-1.47	-0.99	-1.33

Automated analysis computed by Emotient FACET Software. Boldface: AU evidence scores for emotional faces significantly higher than those for neutral faces and above 0. They represent the AUs specifically associated with each expression.

scores and hits from human raters, separately for each emotional category ( $N = 40$ ; Happiness:  $ICC = 0.93$ ; Surprise:  $ICC = 0.94$ ; Anger:  $ICC = 0.89$ ; Sadness:  $ICC = 0.95$ ; Disgust:  $ICC = 0.76$ ; Fear:  $ICC = 0.65$ ; all  $ps < 0.001$ ; 95% CI).  $ICCs$  were calculated as consistency between the proportion of hits for each KDEF model (averaged across all 96 human observers) and the evidence scores recalculated into probabilities as  $p = 1/(1 + 10^{-\text{evidence score}})$ . Also, RTs for observers' hits were negatively related to automated evidence of expressions (Happiness:  $r = -0.45$ ; Surprise:  $r = -0.51$ ; Anger:  $r = -0.40$ ; Sadness:  $r = -0.41$ ; Disgust:  $r = -0.58$ ; Fear:  $r = -0.47$ ; all  $ps \leq 0.01$ ;  $N = 40$ ).

In addition, there were positive correlations between specific AUs and the probability of human categorization responses. Most of the significantly related (all  $ps < 0.0001$ ;  $N = 240$ ) AUs were those that typically characterize each expression: The probability that observers categorized expressions (a) as *happy* was related to AU6 ( $r = 0.67$ ) and AU12 ( $r = 0.90$ ); (b) as *surprised*, to AU1 ( $r = 0.45$ ), AU2 ( $r = 0.73$ ), AU5 ( $r = 0.68$ ), AU25 ( $r = 0.45$ ), and AU26 ( $r = 0.77$ ); (c) as *angry*, to AU4 ( $r = 0.41$ ), AU7 ( $r = 0.37$ ), and AU23 ( $r = 0.48$ ); (d) as *sad*, to AU1 ( $r = 0.36$ ), AU4 ( $r = 0.34$ ), AU15 ( $r = 0.63$ ), and AU24 ( $r = 0.44$ ); (e) as *disgusted*, to AU4 ( $r = 0.36$ ), AU7 ( $r = 0.50$ ), AU9 ( $r = 0.73$ ), and AU10 ( $r = 0.77$ ); and (f) as *fearful*, to AU1 ( $r = 0.42$ ) and AU5 ( $r = 0.34$ ).

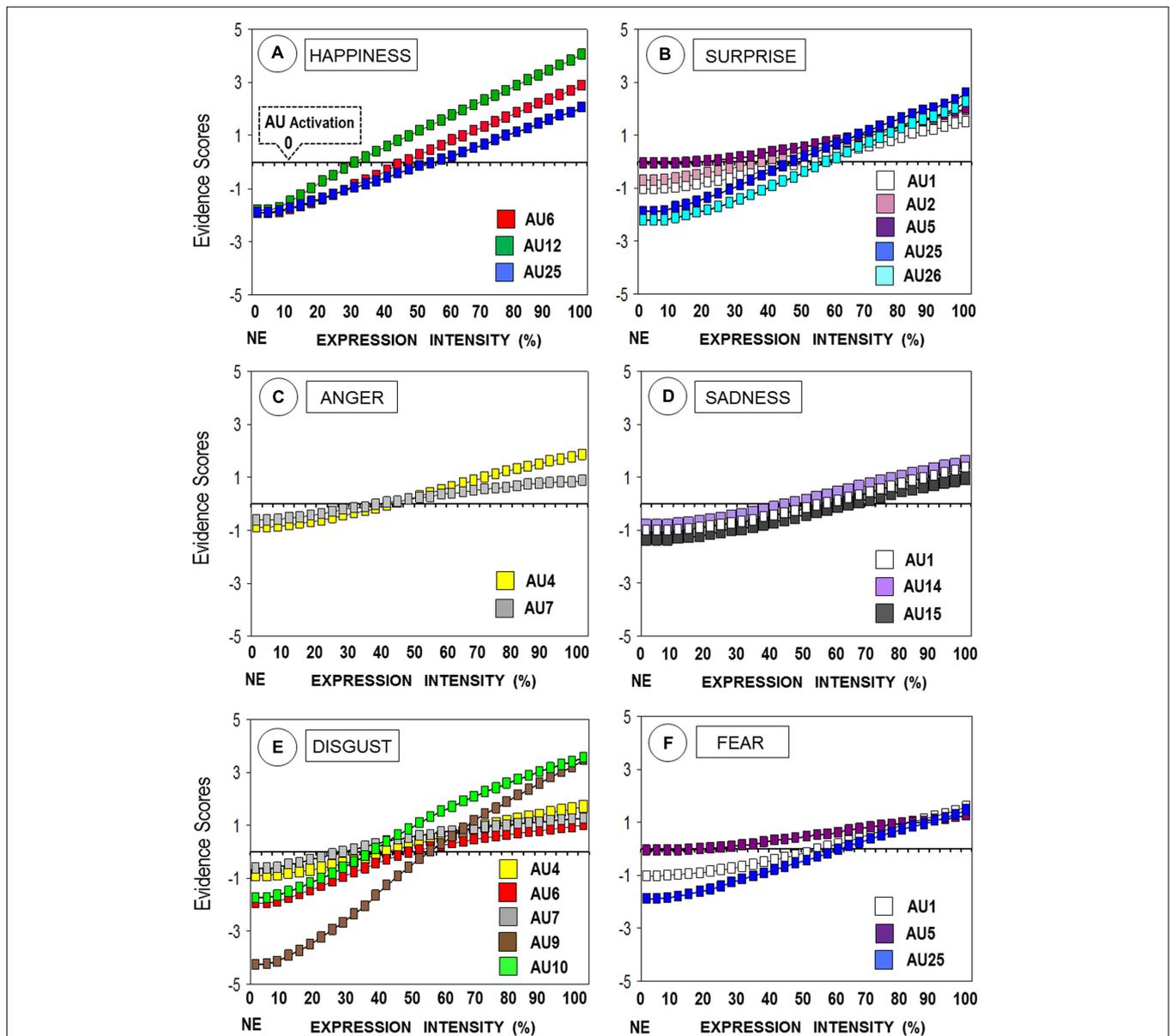
## DISCUSSION

We aimed to provide researchers of emotional facial expression processing with a set of useful and valid dynamic stimuli. To

this end, with agreed time parameters (i.e., unfolding speed to expressive apex within 1 s; Schultz and Pilz, 2009; Hoffmann et al., 2010; Johnston et al., 2013; Wingenbach et al., 2016), we animated static face stimuli of the KDEF database (Lundqvist et al., 1998). The current study examined the resulting KDEF-dyn video-clip stimuli from two complementary approaches: human observer judgments and automated assessment of facial expression. A variety of measures (recognition accuracy, efficiency, and confusions, as well as automated classification of expressions and detection of AUs as a function of intensity, in addition to low-level image properties) were obtained, and are shown on a stimulus level as supplementary data. They will supply researchers with an instrument to select the stimuli as a function of multiple criteria.

## Recognition Patterns of Static and Dynamic Expressions

Human observers correctly recognized all the expressions (as they were intended) well-above chance level ( $M = 85.2\%$ ). Happy faces were recognized better and faster—and fearful faces, less accurately and more slowly—than others, with confusions of fear as surprise, disgust as anger, and sadness as fear. The patterns of recognition accuracy, processing efficiency, and confusions across dynamic expressions converge with those found in prior research for *static expressions*, using different stimulus databases. Regarding *recognition accuracy*, Nelson and Russell reviewed 38 sets of data from 17 studies (Nelson and Russell, 2013): Scores were highest for facial happiness (89%), followed by surprise



**FIGURE 2 |** Distribution of AU evidence scores across levels of expressive intensity. For each expression, AUs were selected if scores were higher than for the neutral face (NE, or 0% expression intensity) and were positive and above the 0 AU activation baseline. AU1: inner brow raiser; AU2: outer brow raiser; AU4: brow lowerer; AU5: upper lid raiser; AU6: cheek raiser; AU7: lid tightener; AU9: nose wrinkle; AU10: upper lip raiser; AU12: lip corner puller; AU15: lip corner depressor; AU25: lips part; AU26: jaw drop. (A) Happiness; (B) Surprise; (C) Anger; (D) Sadness; (E) Disgust; (F) Fear.

(83%), which were higher than for sadness and anger (71 and 68%, respectively), followed by disgust and fear (65 and 59%, respectively). This coincides with our own relative differences (see also Tottenham et al., 2009; Recio et al., 2014; Calvo et al., 2016). Such a consistency extends also to *processing efficiency*, as happy faces are typically recognized faster, followed by surprise, while fear is recognized most slowly (Calder et al., 2000; Elfenbein and Ambady, 2003; Palermo and Coltheart, 2004; Calvo and Nummenmaa, 2009). The pattern of *confusions* is also consistent, as they have been found to occur systematically between disgust and anger, and between surprise and fear, and to a lesser extent

between sadness and fear (Palermo and Coltheart, 2004; Calvo and Lundqvist, 2008; Tottenham et al., 2009; Recio et al., 2013).

Further validation comes from prior research using *dynamic expression* stimuli. First, three studies included all six basic expressions in dynamic *morphing format* from three different databases. Calvo et al. (2016) presented real faces (24 models of the KDEF-dyn database) for 1 s. Recio et al. (2014) presented real faces (from the RaFD; Langner et al., 2010) for 600 ms. Recio et al. (2013) displayed computer-generated faces (FACS Gen 2.0; Krumhuber et al., 2012) for 900 ms. The pattern of recognition accuracy across expressions was similar in all three studies, with

happy faces being identified most accurately (also including higher  $A'$  sensitivity; Calvo et al., 2016), and disgusted and fearful faces, least accurately (and lower  $A'$  sensitivity; Calvo et al., 2016). In addition, in all three studies, fear was likely to be confused with surprise, disgust with anger, and there was some confusion between sadness and fear. Second, regarding the dynamic stimulus sets based on on-line *video recordings* (e.g., van der Schalk et al., 2011; Banziger et al., 2012; Kaulard et al., 2012; Zhang et al., 2014; O'Reilly et al., 2016; Wingenbach et al., 2016; see the 22 databases reviewed by Krumhuber et al., 2017), it is difficult to make comparisons because some studies did not measure recognition performance (accuracy, RTs, or confusions), and due to considerable variations in number of expressive categories and display times (among many other methodological differences). The study conducted by Wingenbach et al. (2016) was methodologically more similar to our own. Their relative recognition accuracies and the pattern of RTs across the six basic expressions were comparable to those in the current study. Altogether, this empirical consistency validates the current database.

## Automated Assessment vs. Human Observers

Another major source of validation for the current database involves the use of automated facial expression analysis. First, the automated classification of expressions showed discrimination specificity, with the evidence of each expression being significantly greater for the corresponding stimulus category than for the others. Nevertheless, some expressions, especially, happiness, and also disgust and surprise, were classified better than sadness, anger, and fear (see **Table 2**), which is in total agreement with results obtained with other automated computation algorithms (Lucey et al., 2010). Second, AUs generally discriminated between expressive categories, and this was in accordance with FACS proposals (Ekman et al., 2002; Olderbak et al., 2014). Some AUs characterized expressions more specifically or strongly than others (see **Table 3**), e.g., AU12 for happiness, AU25 for surprise, AUs 9 and 10 for disgust, AU1 for fear, and AU4 for anger and sadness (the AU4 combination with other AUs allowed for a clear discrimination between these two expressions; see **Table 2**). A related pattern has been obtained with different automated AU detection systems (Lucey et al., 2010; Mavadati et al., 2013; Zhang et al., 2014). Third, automated expression classification and also AU evidence scores increased significantly across 3.33% expressive intensity steps between a neutral and an emotional face (see **Figures 1, 2**). The steepness of such a progressive increase as a function of intensity varied for different expressions and AUs. This approach and results regarding intensity represent a novel contribution and further validate the current video-clip stimuli.

Fourth, importantly, significant correlations emerged between human observers' performance and automated evidence of expressions (large effect sizes: Cohen's  $d_s \geq 1.71$ ) and AUs (medium to large effects:  $d_s \geq 0.72$ ). This has implications for expression recognition theories concerning the type of information that is processed and the cognitive processes

involved. Computational models such as EMPATH (Dailey et al., 2002, 2010) and support vector machine (SVM) based techniques (Susskind et al., 2007)—and, presumably, FACET—simulate face processing and expression recognition in humans. In these models, facial expressions are computed by “emotionless machines” on purely perceptual grounds, i.e., physical image properties (the morphological structure of facial configurations and the visual saliency of distinctive facial cues), in the absence of affective processing. Accordingly, the fact that the automated classifications of expressions converged with human observers' judgments in the current study suggests that human expression recognition also relies to a significant extent on the perceptual (devoid of affect) analysis of facial features. Nevertheless, first, while this may be true for photographs or videos of faces, the role of human affective processing is probably greater in actual face-to-face social encounters, when emotional significance becomes relevant for adaptive purposes. Second, it is likely that the morphological facial features of expressions have become associated (through practice) with their affective significance, and thus both would be processed in tandem, therefore explaining the observed correlations.

## Applications and Limitations

The KDEF-dyn database aims to extend the research possibilities of dynamic facial expression stimuli. First, regarding *experimental control*, all the stimuli are equated in multiple image properties that are non-specific of expression—but can act as confounds—(luminance, signal-to-noise ratio, size, orientation, etc.), in addition to standardization of dynamic properties (unfolding speed and duration). Such controls will be particularly useful for neurophysiological and eyetracking research, where the dependent measures are especially sensitive to physical stimulus factors; and also useful for paradigms in which the stimuli must be presented briefly, where display duration needs to be strictly comparable for the different stimuli. A second benefit is related to the role of *expressive intensity*. Instead of considering only the apex, we have established the assessment of expressions and AUs at fine-grained intensities. This is important, as intensity is often critical to interpret the meaning of expressions. By knowing the evidence for each expression and AU at each intensity level, and the time-intensity correspondence in the video-clips (as shown in **Supplementary Dataset S3**), researchers can easily manipulate the display time of the stimuli to investigate the desired intensity (e.g., by cutting, masking, or stopping each video-clip at the respective time point). This approach will be useful for the investigation of visual processing, particularly for studies of expression discrimination thresholds. A third promising application is concerned with the use of these stimuli in the investigation of *cognitive biases* (attentional and interpretative) in *psychopathology*. For example, it has been shown that individuals with clinical levels of social anxiety are especially prone to detect negatively valenced dynamic expressions at low intensities (Gutiérrez-García and Calvo, 2016, 2017; Gutiérrez-García et al., 2018). A reason for the usefulness of this application to psychopathology research is that dynamic information improves identification of facial affect, particularly for lower intensity and subtle stimuli

(Krumhuber et al., 2013), which would increase sensitivity for individuals that are hypervigilant to threat and incongruities in facial expressions.

Researchers should, nonetheless, be aware of potential limitations. First, although standardization of unfolding *speed* is beneficial for experimental control, it can reduce the natural speed variance across expressions. For example, we averaged the 1-s unfolding speed from neutral baseline to emotional apex for all the expressions (see Schultz and Pilz, 2009; Johnston et al., 2013; Wingenbach et al., 2016). However, facial surprise is considered as most natural when it unfolds at a fast pace while sadness is judged as more realistic when the facial expression changes slowly (Sato and Yoshikawa, 2004; Adamaszek et al., 2015). To remedy this potential limitation, it is possible to slow down or speed up the video-clips, by means of video-editing software. Second, we used *posed* instead of spontaneous expressions. The majority of extant dynamic stimulus sets, in fact, include posed expressions, either in response to instructions to perform facial actions or as the enactment of emotional scenarios (van der Schalk et al., 2011; Banziger et al., 2012; Kaulard et al., 2012; O'Reilly et al., 2016; Wingenbach et al., 2016), although some have included spontaneous expressions (Mavadati et al., 2013; Zhang et al., 2014). Posed expressions may lose naturalness and their recognition rates may be inflated, although the former avoid the ambiguity of spontaneous expressions. Third, we used *morphed* expressions. Morphing creates linear movement where all the facial components change at the same time and speed, whereas natural expressions appear to change in a non-linear manner. However, some studies indicate that natural expressions look smooth, uniform, and ballistic (Weiss et al., 1987; Hess et al., 1989), thus actually sharing properties with morphed dynamic expressions. Further, in the current study, automated assessment revealed specificity and sensitivity to expressions and also to AUs in accordance with FACS proposals. This suggests that the possible reduction of naturalness was not critical (see Cosker et al., 2010, 2015).

## CONCLUSION

We present a set of dynamic facial expressions (KDEF-dyn) based on a widely used database of static expressions (KDEF). The new stimuli have been validated by means of several measures from two approaches: expression categorization by human observers and automated analysis of facial expressions and AUs with computer software. Results show good convergence with prior research using static and dynamic expression stimuli. Although not devoid of limitations, this convergence reinforces the validation of the current database, while offering additional

## REFERENCES

Adamaszek, M., Kirkby, K. C., D'Agata, F., Olbrich, S., Langner, S., Steele, C., et al. (2015). Neural correlates of impaired emotional face recognition in cerebellar lesions. *Brain Res.* 1613, 1–12. doi: 10.1016/j.brainres.2015.01.027

advantages: (a) the use of automated facial expression and AU analysis, with significant correlations between human and automated performance; (b) the control of perceptual properties (e.g., size and multiple low-level image statistics) and stimulus dynamic properties (e.g., duration and unfolding speed); and (c) the systematic and fine-grained gradation of expressive intensities of an otherwise relatively large sample of encoders. This will be useful for behavioral, computational, and neurophysiological studies investigating facial expression processing.

## AVAILABILITY OF DATA

The KDEF-dyn stimuli and datasets are freely available for scientific purposes, and can be downloaded from <http://kdef.se/versions.html> (KDEF-dyn I).

## AUTHOR CONTRIBUTIONS

MC and DL conceived and designed the experiments. AF-M prepared the materials, performed the experiments, and conducted the statistical analyses. MC wrote the first draft of the manuscript. MC, AF-M, GR, and DL wrote sections and revised the whole manuscript.

## FUNDING

This research was supported by Grant PSI2014-54720-P to MC from the Spanish Ministerio de Economía y Competitividad and Grant RE 3721/2-1 to GR from the Deutsche Forschungsgemeinschaft.

## SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/fpsyg.2018.02052/full#supplementary-material>

**DATASET S1** | Human and automated expression and action unit categorization. Measures of recognition performance by human observers and automated analysis for each KDEF-dyn stimulus.

**DATASET S2** | FACET assessment of intensities. Evidence values for each expression and action units, at each of 31 intensity levels (in 3.33% steps).

**DATASET S3** | Low-level image statistics of intensities. Image values at each of 31 intensity levels (in 3.33% steps).

**DATASET S4** | Stimuli. Video-clip stimuli\_MP4. Two hundred and forty video-clips, separated for each of six emotional expression categories (40 video-clips each).

Anitha, B., Venkatesha, M. K., and Adiga, B. S. (2010). A survey on facial expression databases. *Int. J. Eng. Sci. Technol.* 2, 5158–5174.

Banziger, T., Mortillaro, M., and Scherer, K. R. (2012). Introducing the Geneva Multimodal expression corpus for experimental research on emotion perception. *Emotion* 12, 1161–1179. doi: 10.1037/a0025827

- Bartlett, M., and Whitehill, J. (2011). "Automated facial expression measurement: recent applications to basic research in human behavior, learning, and education," in *Handbook of Face Perception*, eds A. Calder, G. Rhodes, M. Johnson, and J. Haxby (Oxford: Oxford University Press), 489–513.
- Bublatzky, F., Gerdes, A. B. M., White, A. J., Riemer, M., and Alpers, G. W. (2014). Social and emotional relevance in face processing: happy faces of future interaction partners enhance the late positive potential. *Front. Hum. Neurosci.* 8:493. doi: 10.3389/fnhum.2014.00493
- Calder, A. J., Young, A. W., Keane, J., and Dean, M. (2000). Configural information in facial expression perception. *J. Exp. Psychol. Hum. Percept. Perform.* 26, 527–551. doi: 10.1037/0096-1523.26.2.527
- Calvo, M. G., Avero, P., Fernandez-Martin, A., and Recio, G. (2016). Recognition thresholds for static and dynamic emotional faces. *Emotion* 16, 1186–1200. doi: 10.1037/emo0000192
- Calvo, M. G., and Beltrán, D. (2014). Brain lateralization of holistic versus analytic processing of emotional facial expressions. *Neuroimage* 92, 237–247. doi: 10.1016/j.neuroimage.2014.01.048
- Calvo, M. G., Gutiérrez-García, A., Avero, P., and Lundqvist, D. (2013). Attentional mechanisms in judging genuine and fake smiles: eye-movement patterns. *Emotion* 13, 792–802. doi: 10.1037/a0032317
- Calvo, M. G., and Lundqvist, D. (2008). Facial expressions of emotion (KDEF): identification under different display-duration conditions. *Behav. Res. Methods* 40, 109–115. doi: 10.3758/BRM.40.1.109
- Calvo, M. G., and Nummenmaa, L. (2009). Eye-movement assessment of the time course in facial expression recognition: neurophysiological implications. *Cogn. Affect. Behav. Neurosci.* 9, 398–411. doi: 10.3758/CABN.9.4.398
- Calvo, M. G., and Nummenmaa, L. (2011). Time course of discrimination between emotional facial expressions: the role of visual saliency. *Vis. Res.* 51, 1751–1759. doi: 10.1016/j.visres.2011.06.001
- Calvo, M. G., and Nummenmaa, L. (2016). Perceptual and affective mechanisms in facial expression recognition: an integrative review. *Cogn. Emot.* 30, 1081–1106. doi: 10.1080/02699931.2015.1049124
- Cohn, J. F., Ambadar, Z., and Ekman, P. (2007). "Observer-based measurement of facial expression with the Facial Action Coding System," in *The Handbook of Emotion Elicitation and Assessment*, eds J. A. Coan and J. J. B. Allen (New York, NY: Oxford University Press), 203–221.
- Cohn, J. F., and De la Torre, F. (2015). "Automated face analysis for affective computing," in *The Oxford Handbook of Affective Computing*, eds R. A. Calvo, S. DiMello, J. Gratch, and A. Kappas (New York, NY: Oxford University Press), 131–150.
- Cosker, D., Krumhuber, E., and Hilton, A. (2010). "Perception of linear and nonlinear motion properties using a FACS validated 3D facial model," in *Proceedings of the Symposium on Applied Perception in Graphics and Visualization (APGV)*, New York, NY. doi: 10.1145/1836248.1836268
- Cosker, D., Krumhuber, E., and Hilton, A. (2011). "A FACS valid 3D dynamic action unit database with applications to 3D dynamic morphable facial modeling," in *Proceedings of the 13th IEEE International Conference on Computer Vision (ICCV)*, Barcelona. doi: 10.1109/ICCV.2011.6126510
- Cosker, D., Krumhuber, E., and Hilton, A. (2015). "Perceived emotionality of linear and nonlinear AUs synthesised using a 3D dynamic morphable facial model," in *Proceedings of the Conference on Facial Analysis and Animation (FAA)*, New York, NY. doi: 10.1145/2813852.2813859
- Cowie, R., Douglas-Cowie, E., and Cox, C. (2005). Beyond emotion archetypes: databases for emotion modelling using neural networks. *Neural Netw.* 18, 371–388. doi: 10.1016/j.neunet.2005.03.002
- Dailey, M. N., Cottrell, G. W., Padgett, C., and Adolphs, R. (2002). EMPATH: a neural network that categorizes facial expressions. *J. Cogn. Neurosci.* 14, 1158–1173. doi: 10.1162/089892902760807177
- Dailey, M. N., Joyce, C., Lyons, M. J., Kamachi, M., Ishi, H., Gyoba, J., et al. (2010). Evidence and a computational explanation of cultural differences in facial expression recognition. *Emotion* 10, 874–893. doi: 10.1037/a0020019
- Dente, P., Küster, D., Skora, L., and Krumhuber, E. (2017). Measures and metrics for automatic emotion classification via FACET. *Paper Presented at the Artificial Intelligence and Simulation of Behaviour Annual Convention*, Bath.
- Ebner, N. C., Riediger, M., and Lindenberger, U. (2010). FACES—a database of facial expressions in young, middle-aged, and older women and men: development and validation. *Behav. Res. Methods* 42, 351–362. doi: 10.3758/BRM.42.1.351
- Ekman, P., and Friesen, W. V. (1976). *Pictures of Facial Affect*. Palo Alto, CA: Consulting Psychologists Press.
- Ekman, P., and 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., and Hager, J. C. (2002). *Facial Action Coding System*. Salt Lake City, UT: A Human Face.
- Elfenbein, H. A., and Ambady, N. (2003). When familiarity breeds accuracy: cultural exposure and facial emotion recognition. *J. Pers. Soc. Psychol.* 85, 276–290. doi: 10.1037/0022-3514.85.2.276
- Faul, F., Erdfelder, E., Lang, A. G., and Buchner, A. (2007). G\*Power 3: a flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behav. Res. Methods* 39, 175–191. doi: 10.3758/BF03193146
- Fiorentini, C., and Viviani, P. (2011). Is there a dynamic advantage for facial expressions? *J. Vis.* 11, 1–15. doi: 10.1167/11.3.17
- Girard, J. M., Cohn, J. F., Jeni, L. A., Sayette, M. A., and De la Torre, F. (2015). Spontaneous facial expression in unscripted social interactions can be measured automatically. *Behav. Res. Methods* 47, 1136–1147. doi: 10.3758/s13428-014-0536-1
- Goeleven, E., De Raedt, R., Leyman, L., and Verschuere, B. (2008). The Karolinska Directed Emotional Faces: a validation study. *Cogn. Emot.* 22, 1094–1118. doi: 10.1080/02699930701626582
- Gordon, I., Tanaka, J. W., Pierce, M., and Bartlett, M. (2011). Facial expression production and training. *J. Vis.* 11, 565–565. doi: 10.1167/11.11.565
- Gupta, R., Hur, Y. J., and Lavie, N. (2016). Distracted by pleasure: effects of positive versus negative valence on emotional capture under load. *Emotion* 16, 328–337. doi: 10.1037/emo0000112
- Gutiérrez-García, A., and Calvo, M. G. (2016). Social anxiety and trustworthiness judgments of dynamic facial expressions of emotion. *J. Behav. Ther. Exp. Psychiatry* 52, 119–127. doi: 10.1016/j.jbtep.2016.04.003
- Gutiérrez-García, A., and Calvo, M. G. (2017). Social anxiety and threat-related interpretation of dynamic facial expressions: sensitivity and response bias. *Pers. Individ. Differ.* 107, 10–16. doi: 10.1016/j.paid.2016.11.025
- Gutiérrez-García, A., Calvo, M. G., and Eysenck, M. W. (2018). Social anxiety and detection of facial untrustworthiness: spatio-temporal oculomotor profiles. *Psychiatry Res.* 262, 55–62. doi: 10.1016/j.psychres.2018.01.031
- Harris, R. J., Young, A. W., and Andrews, T. J. (2014). Dynamic stimuli demonstrate a categorical representation of facial expression in the amygdala. *Neuropsychologia* 56, 47–52. doi: 10.1016/j.neuropsychologia.2014.01.005
- Hess, U., Kappas, A., McHugo, G. J., Kleck, R. E., and Lanzetta, J. T. (1989). An analysis of the encoding and decoding of spontaneous and posed smiles: the use of facial electromyography. *J. Nonverb. Behav.* 13, 121–137. doi: 10.1007/BF00990794
- Hoffmann, H., Traue, H. C., Bachmayr, F., and Kessler, H. (2010). Perceived realism of dynamic facial expressions of emotion: optimal durations for the presentation of emotional onsets and offsets. *Cogn. Emot.* 24, 1369–1376. doi: 10.1080/02699930903417855
- Johnston, P., Mayes, A., Hughes, M., and Young, A. W. (2013). Brain networks subserving the evaluation of static and dynamic facial expressions. *Cortex* 49, 2462–2472. doi: 10.1016/j.cortex.2013.01.002
- Kaulard, K., Cunningham, D. W., Bulthoff, H. H., and Wallraven, C. (2012). The MPI facial expression database — A validated database of emotional and conversational facial expressions. *PLoS One* 7:e32321. doi: 10.1371/journal.pone.0032321
- Krumhuber, E. G., Kappas, A., and Manstead, A. S. R. (2013). Effects of dynamic aspects of facial expressions: a review. *Emot. Rev.* 5, 41–46. doi: 10.1177/1754073912451349
- Krumhuber, E. G., Skora, L., Küster, D., and Fou, L. (2017). A review of dynamic datasets for facial expression research. *Emot. Rev.* 9, 280–292. doi: 10.1177/1754073916670022
- Krumhuber, E. G., Tamarit, L., Roesch, E. B., and Scherer, K. R. (2012). FACSGen 2.0 animation software: generating three-dimensional FACS-valid facial expressions for emotion research. *Emotion* 12, 351–363. doi: 10.1037/a0026632
- Langner, O., Dotsch, R., Bijlstra, G., Wigboldus, D. H. J., Hawk, S. T., and van Knippenberg, A. (2010). Presentation and Validation of the Radboud Faces Database. *Cogn. Emot.* 24, 1377–1388. doi: 10.1080/02699930903485076

- Lucey, P., Cohn, J. F., Kanade, T., Saragih, J., Ambadar, Z., and Matthews, I. (2010). "The Extended Cohn-Kanade Dataset (CK + ): a complete dataset for action unit and emotion-specified expression," in *Proceedings of the Third International Workshop on CVPR for Human Communicative Behavior Analysis (CVPR4HB 2010)*, San Francisco, CA.
- Lundqvist, D., Flykt, A., and Öhman, A. (1998). *The Karolinska Directed Emotional Faces - KDEF [CD-ROM]*. Stockholm: Department of Clinical Neuroscience, Psychology section, Karolinska Institutet.
- Mavadati, S. M., Mahoor, M. H., Bartlett, K., Trinh, P., and Cohn, J. F. (2013). DISFA: a spontaneous facial action intensity database. *IEEE Trans. Affect. Comput.* 4, 151–160. doi: 10.1109/T-AFFC.2013.4
- Naples, A., Nguyen-Phuc, A., Coffman, M., Kresse, A., Faja, S., Bernier, R., et al. (2015). A computer-generated animated face stimulus set for psychophysiological research. *Behav. Res. Methods* 47, 562–570. doi: 10.3758/s13428-014-0491-x
- Nelson, N. L., and Russell, J. A. (2013). Universality revisited. *Emot. Rev.* 5, 8–15. doi: 10.1177/1754073912457227
- Olderbak, S., Hildebrandt, A., Pinkpank, T., Sommer, W., and Wilhelm, O. (2014). Psychometric challenges and proposed solutions when scoring facial emotion expression codes. *Behav. Res. Methods* 46, 992–1006. doi: 10.3758/s13428-013-0421-3
- O'Reilly, H., Pigat, D., Fridenson, S., Berggren, S., Tal, S., Golan, O., et al. (2016). The EU-Emotion Stimulus Set: a validation study. *Behav. Res. Methods* 48, 567–576. doi: 10.3758/s13428-015-0601-4
- Palermo, R., and Coltheart, M. (2004). Photographs of facial expression: accuracy, response times, and ratings of intensity. *Behav. Res. Methods Instrum. Comput.* 36, 634–638. doi: 10.3758/BF03206544
- Pollick, F. E., Hill, H., Calder, A., and Paterson, H. (2003). Recognizing facial expression from spatially and temporally modified movements. *Perception* 32, 813–826. doi: 10.1068/p3319
- Popov, T., Miller, G. A., Rockstroh, B., and Weisz, N. (2013). Modulation of alpha power and functional connectivity during facial affect recognition. *J. Neurosci.* 33, 6018–6026. doi: 10.1523/JNEUROSCI.2763-12.2013
- Recio, G., Schacht, A., and Sommer, W. (2013). Classification of dynamic facial expressions of emotion presented briefly. *Cogn. Emot.* 27, 1486–1494. doi: 10.1080/02699931.2013.794128
- Recio, G., Schacht, A., and Sommer, W. (2014). Recognizing dynamic facial expressions of emotion: specificity and intensity effects in event-related brain potentials. *Biol. Psychol.* 96, 111–125. doi: 10.1016/j.biopsycho.2013.12.003
- Sanchez, A., Vazquez, C., Gomez, D., and Joermann, J. (2014). Gaze-fixation to happy faces predicts mood repair after a negative mood induction. *Emotion* 14, 85–94. doi: 10.1037/a0034500
- Sandbach, G., Zafeiriou, S., Pantic, M., and Yin, L. (2012). Static and dynamic 3D facial expression recognition: a comprehensive survey. *Image Vis. Comput.* 30, 683–697. doi: 10.1016/j.imavis.2012.06.005
- Sato, W., and Yoshikawa, S. (2004). The dynamic aspects of emotional facial expressions. *Cogn. Emot.* 18, 701–710. doi: 10.1080/02699930341000176
- Schultz, J., and Pilz, K. S. (2009). Natural facial motion enhances cortical responses to faces. *Exp. Brain Res.* 194, 465–475. doi: 10.1007/s00221-009-1721-9
- Susskind, J. M., Littlewort, G., Bartlett, M. S., Movellan, J., and Anderson, A. K. (2007). Human and computer recognition of facial expressions of emotion. *Neuropsychologia* 45, 152–162. doi: 10.1016/j.neuropsychologia.2006.05.001
- Tottenham, N., Borscheid, A., Ellertsen, K., Marcus, D., and Nelson, C. A. (2002). *The Nimstim Face Set1*. Available at: <http://www.macbrain.org/faces/index.htm>
- Tottenham, N., Tanaka, J. W., Leon, A. C., McCarry, T., Nurse, M., Hare, T. A., et al. (2009). The NimStim set of facial expressions: judgments from untrained research participants. *Psychiatry Res.* 168, 242–249.
- Valstar, M. F., Almaev, T., Girard, J. M., McKeown, G., Mehu, M., Yin, L., et al. (2015). "FERA 2015 – Second facial expression recognition and analysis challenge," in *Proceedings of the Eleventh IEEE International Conference on Face and Gesture Recognition*, Ljubljana. doi: 10.1109/FG.2015.7284874
- Valstar, M. F., Sanchez-Lozano, T., Cohn, J. F., Jeni, L. A., Girard, J. M., Zhang, Z., et al. (2017). "FERA 2017 - Addressing head pose in the third facial expression recognition and analysis challenge," in *Proceedings of the Twelfth IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017)*, Washington, DC. doi: 10.1109/FG.2017.107
- van der Schalk, J., Hawk, S. T., Fischer, A. H., and Doosje, B. (2011). Moving faces, looking places: validation of the Amsterdam Dynamic Facial Expression Set (ADFES). *Emotion* 11, 907–920. doi: 10.1037/a0023853
- Vrticka, P., Lordier, L., Bediou, B., and Sander, D. (2014). Human amygdala response to dynamic facial expressions of positive and negative surprise. *Emotion* 14, 161–169. doi: 10.1037/a0034619
- Weiss, F., Blum, G. S., and Gleberman, L. (1987). Anatomically based measurement of facial expressions in simulated versus hypnotically induced affect. *Motiv. Emot.* 11, 67–81. doi: 10.1007/BF00992214
- Wingenbach, T. S., Ashwin, C., and Brosnan, M. (2016). Correction: validation of the Amsterdam Dynamic Facial Expression Set – Bath Intensity variations (ADFES-BIV): a set of videos expressing low, intermediate, and high intensity emotions. *PLoS One* 11:e0168891. doi: 10.1371/journal.pone.0168891
- Zhang, X., Yin, L., Cohn, J. F., Canavan, S., Reale, M., Horowitz, A., et al. (2014). BP4D-spontaneous: a high-resolution spontaneous 3D dynamic facial expression database. *Image Vis. Comput.* 32, 692–706. doi: 10.1016/j.imavis.2014.06.002

**Conflict of Interest Statement:** The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2018 Calvo, Fernández-Martín, Recio and Lundqvist. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.