Web Intelligence 17 (2019) 41–52
DOI 10.3233/WEB-190399
IOS Press

A functional data analysis approach for continuous 2-D emotion annotations

Karan Sharma [a,*], Marius Wagner [a], Claudio Castellini [a], Egon L. van den Broek [b], Freek Stulp [a] and Friedhelm Schwenker [c]

[a] *Robotics and Mechatronics Center, DLR – German Aerospace Center, Wessling, Germany*
E-mails: karan.sharma@dlr.de, marius.wagner@dlr.de, claudio.castellini@dlr.de, freek.stulp@dlr.de
[b] *Department of Information and Computing Sciences, Utrecht University, Utrecht, The Netherlands*
E-mail: vandenbroek@acm.org
[c] *Institute of Neural Information Processing, Ulm University, Ulm, Germany*
E-mail: friedhelm.schwenker@uni-ulm.de

Abstract. The standard paradigm in *Affective Computing* involves acquiring one/several markers (e.g., physiological signals) of emotions and training models on these to predict emotions. However, due to the internal nature of emotions, labelling/annotation of emotional experience is done manually by humans using specially developed annotation tools. To effectively exploit the resulting subjective annotations for developing affective systems, their quality needs to be assessed. This entails, (i) evaluating the variations in annotations, across different subjects and emotional stimuli, to detect spurious/unexpected patterns; and (ii) developing strategies to effectively combine these subjective annotations into a ground truth annotation. This article builds on our previous work by presenting a novel *Functional Data Analysis* based approach to assess the quality of annotations. Specifically, the bivariate annotation time-series are transformed into functions, such that each resulting functional annotation then becomes a sample element for analysis like *Multivariate Functional Principal Component Analysis* (MFPCA) that evaluate variation across all annotations. The resulting scores from MFPCA provide interesting insights into annotation patterns and facilitate the use of multivariate statistical techniques to address both (i) and (ii). Given the presented efficacy of these methods, we believe they offer an exciting new approach to assessing the quality of annotations.

Keywords: Functional data analysis, affective computing, continuous annotation, self-reporting

1. Introduction

In the past two decades, the number of gadgets that humans interact with has been on the rise. This trend continues as the next generation of gadgets (e.g., personal robot-companions, autonomous cars) enter our lives. This ever-increasing technologisation brings numerous challenges pertaining to effective and continuously engaging user-interaction. Addressing these challenges effectively is the goal of *Affective Computing* (AC), which aims to develop systems that can recognise and process human emotions [26], such that they continuously adapt to the user's needs [31]. To this end, researchers in this field often investigate how, e.g., physiological signals [32,40], speech [2,32], facial expressions [32,36] and other modalities [31], act as markers for emotional experience. The aim being, that if the relation between these signals and emotions can be robustly modelled, then machines can 'learn' to recognise and adapt to their users' emotional state. Given the far-reaching impact such a technology would have on user-interaction, interest in the field of AC has been steadily growing. Accordingly, in recent years, several interesting research/applications in AC have come to the fore. These include, among others, an affective music player that adapts the music being played to user's emotional state [39], an investigation on the role of AC in monitoring workplace health and

*Corresponding author. E-mail: karan.sharma@dlr.de.

safety [38], and a study on the use of AC methods for continuous pain intensity assessment [21].

In spite of the burgeoning interest and research efforts, several hurdles restrict a more widespread utilisation of AC. Principal among these is the internal nature of human emotions that leads to them not being easily accessible to external entities [40]. To address this, the link between measurable emotion modalities and internal emotions needs to be established. This is still largely an unsolved problem, but a commonly followed approach in laboratory settings involves eliciting emotional response from humans using stimuli like pictures [25], videos [33,40], music [25], etc., while simultaneously acquiring modalities and annotations pertaining to the emotional experience. These annotations are usually provided in form of either discrete emotion categories [31,40] (e.g., *fear*, *joy*, etc.) or in terms of *Valence* and *Arousal (V–A)* values as per the continuous 2-dimensional *Circumplex model of Affect* [30]. Traditionally, these annotations were manually acquired using Likert-scale based questionnaires, where a single emotional-label or V–A pair-value represents the emotional response to the stimulus. However, in recent years, there has been a growing realisation that this approach does not adequately represent the emotional response elicited by dynamic stimuli (e.g., videos) [20,25,36]. Thus, greater emphasis is now placed on continuous annotation of emotional experience using specially developed interfaces like FEELTRACE [7], GTrace [8] and EmuJoy [25]. While these interfaces laid the groundwork for continuous annotation, in recent years, several shortcomings with respect to their setup, annotation-strategy and usability, have been widely reported [3,23,24,43]. To improve upon these, we developed a new *Joystick-based Emotion Reporting Interface* (abbreviated as *DLR–JERI* or *JERI*) [33], that: (i) uses a joystick instead of the less ergonomic (and commonly used) mouse (see Fig. 1), (ii) unlike some current interfaces, allows for simultaneous V–A annotation, and (iii) guides users through more widely interpretable Self-Assessment Manikin (SAM) [4]. This interface has also been formally evaluated through a user-study, where it was rated as having 'excellent' usability [33].

Irrespective of the annotation interface used, a common issue is the quality of the acquired subjective annotations [20,24,41]. This issue can be best summarised in form of the following questions: (a) given an emotional stimulus, do the annotations exhibit some agreement?, (b) do different stimuli with similar intended V–A attributes lead to similar annotations?, (c) how do annotation patterns vary across different stimuli and are they discriminable from each other?, and (d) given multiple annotations, how to best determine the underlying ground truth annotation? Addressing these questions is highly relevant when collecting emotional corpora and/or developing emotion prediction models, as undesired effects, such as diverging annotator behaviour, ill-chosen stimuli, etc., can be detected and mitigated [20,28]. For continuous annotations, addressing these quality issues is analytically more challenging [20]. Existing approaches to the same, including our previous work [1,33,34], fall into the following two categories. In the first, the continuous aspect is ignored. Accordingly, (a) and (b) are addressed using inter-rater reliability measures [24,43], (c) using ANOVA (or MANOVA for multivariate case) models [29,33], and (d) by simply calculating point-by-point arithmetic mean [14,37]. The approaches in the second category account for the continuous nature, but only for the univariate case. Thus, (b) and (c) can be addressed using univariate sequence analysis [33], and (d) using regression-based approaches [33,35]. For Likert-scale or univariate continuous annotations, these approaches are often sufficient. However, for continuous and bivariate annotations (as for JERI), approaches that account for the inherent correlation between valence and arousal need to be developed.

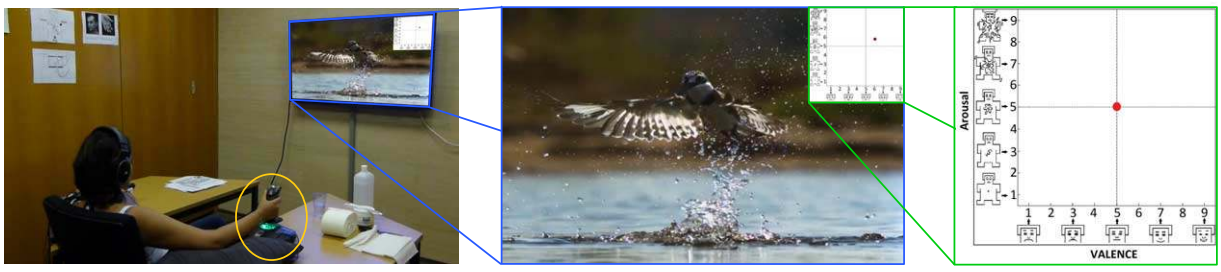

Fig. 1. Typical annotation setup using JERI with the joystick circled (left). A video-stimulus with the embedded annotation user-interface (center), where the Self-Assessment Manikin (SAM) were added to valence-arousal axes (right).

Table 1
The type, label, source, intended valence-arousal attributes and duration of each video used in the experiment

Type	Label	Source (year of release)	Intended attributes		Duration [s]
			Valence	Arousal	
Amusing	am1	Hangover (2009)	mid/high	mid/high	185
	am2	When Harry met Sally (1989)	mid/high	mid/high	173
Boring	bo1	Europe travel advisory (2013)	low	low	119
	bo2	Japanese tea ceremony (2012)	low	low	160
Relaxing	re1	Pristine beach (2011)	mid/high	low	145
	re2	Zambezi river (2011)	mid/high	low	147
Scary	sc1	Shutter (2004)	low	high	197
	sc2	Mama (2008)	low	high	144

To this end, in this work we extend our previous effort [33] by presenting a novel *Functional Data Analysis (FDA)* based approach to address the aforementioned quality issues. The main advantage of FDA techniques is that, unlike the aforementioned discretised approach, they are applied to complete bivariate annotations. The contribution of this work is twofold. First, a foundational framework for converting annotations into functions is presented. Second, several FDA based approaches to evaluate the quality of annotations are developed. To the best of our knowledge, this is the first attempt at using FDA for analysing continuous annotations in AC.

The rest of the paper is organised as follows. In Section 2, the experiment setup, annotation data, processing steps for converting annotations into functions and a theoretical background for *Multivariate Functional Principal Component Analysis (MFPCA)*, are presented. Section 3 presents the results of the analyses undertaken on functional annotations. Namely, (i) MFPCA analysis to investigate the variance, (ii) a distance-based measure to showcase concordance of annotation patterns, (iii) the use of Fisher Discriminant Ratio and (iv) Gaussian Support Vector Machines to discern the discriminability between annotations, and (v) an approach to ascertain ground truth annotations, are presented. A discussion of the results is also presented. Lastly, in Section 4, the conclusions and the outlook from this work are presented.

2. Methods

2.1. Setup

To test and validate annotations from JERI, an experiment involving 30 volunteers (15 males, age 28.6± 4.8 years and 15 females, age 25.7 ± 3.1 years; range of age 22–37 years) was set up. The experiment was approved by the DLR Ethics Committee. The participants of the study watched 8 videos and self-reported their experienced affect (in form of simultaneous V–A annotations) by appropriately positioning the joystick pointer (i.e., the red dot in Fig. 1) in the annotation interface. The videos used for the test were selected so as to elicit the following 4 intrinsic emotions: *amusement*, *boredom*, *relaxation*, and *scaredness*. For every emotion, 2 videos were used, so as to facilitate comparisons. The source of the excerpted videos and their expected intrinsic V–A are listed in Table 1. More details on the setup can be found in [33].

2.2. Annotation data

The raw annotation data acquired from the experiment is in form of 2-D sequence of points within the interval $[0.5, 9.5]$ (see Fig. 2). Each data point has a time-stamp associated with it. Since there were 8 videos and 30 participants in the study, a single data point in an annotation sequence can be expressed as:

$$x_{vs}(t_k) = \left[x_1(t_k), x_{2(}t_k)\right] \in \mathbb{R}^2 \tag{1}$$

where $v = 1, 2, \ldots, 8$ for videos, $s = 1, 2, \ldots, 30$ for subjects, $t_k \in \{t_1, \ldots, t_{n_{vs}}\}$ are timestamps and n_{vs} is the number of points in an annotation sequence for video v by participant s. Through pre-processing n_{vs} is made uniform to n_v across all subjects for a given video v.

2.3. Basis expansion using P-splines

The fundamental notion of FDA is that discrete data points $(x_1, x_2, \ldots, x_n)$, such that $x_i \in \mathbb{R}$, observed at

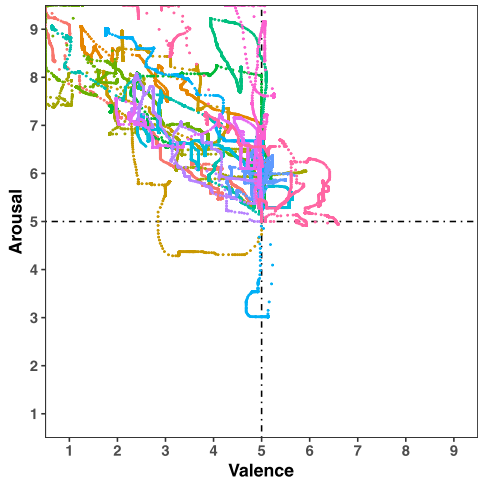

Fig. 2. Annotations from 15 subjects (different colours) for video sc2.

time/location $(t_1, t_2, \ldots, t_n)$ are generated by an underlying continuous process (i.e., smooth function $\tilde{x}$) [42]. Accordingly,

$$x(t) = \tilde{x}(t) + \varepsilon(t) \tag{2}$$

where ε is the noise/error. To ascertain the function $\tilde{x}$ from the observed data x, which is non-linear (see Fig. 2), basis expansion $(B_1(\cdot), B_2(\cdot), \ldots, B_d(\cdot))$ is often used. Thus, the function is expressed as a linear combination of these basis functions [11,17]:

$$\tilde{x}(t) = \sum_{j=1}^{d} \gamma_j B_j(t) \tag{3}$$

where B are the basis functions, γ are the model parameters and d is the number of basis functions used. Amongst the several possible basis expansions, *penalised B-splines* (P-splines), which are a modified form of *basis-splines* (B-splines), were used. To elaborate the workings of P-splines basis expansion, first the innards of B-splines are presented.

B-splines were used for the given dataset because of their efficacy in representing non-linear and non-periodic data [17]. B-splines are a form of *piecewise-polynomials* representation, whereby the input data is divided into contiguous intervals and local polynomials are used to represent the underlying function in that interval [9]. Thus, B-splines representation is contingent to the 'knots' defining the intervals and the degree of the polynomials used. A basis function B_j^0 of degree $= 0$ at some knot κ_j can be expressed as:

$$B_j^0(t) = \begin{cases} 1 & \text{if } \kappa_j \leqslant t < \kappa_{j+1}, \\ 0 & \text{otherwise,} \end{cases}$$
$$j = 1, 2, \ldots, d-1 \tag{4}$$

Higher degree polynomial basis functions (e.g., degree $= l$) are accordingly expressed recursively [11], as:

$$B_j^l(t) = \frac{t - \kappa_{j-l}}{\kappa_j - \kappa_{j-l}} B_{j-1}^{l-1}(t) + \frac{\kappa_{j+1} - t}{\kappa_{j+1} - \kappa_{j+1-l}} B_j^{l-1}(t) \tag{5}$$

The recursive definition leads to an augmented knot sequence, where $2l$ outer knots are added to m initial knots (where, $m =$ #interior $+ 2$ boundary-knots), leading to total length of knots sequence to be $= m + 2l$. The number of basis functions is $d = m + l - 1$.

Given the discrete data and the basis functions, *Least Squares minimisation* [11] is used to compute the estimated model parameters $\hat{\gamma}$, such that:

$$\hat{\gamma} = \arg\min_{\gamma} \sum_{j=1}^{n} \big(x(t_j) - \tilde{x}(t_j)\big)^2 = \arg\min_{\gamma} \sum_{j=1}^{n} \left(x(t_j) - \sum_{k=1}^{d} \gamma_k B_k(t_j)\right)^2 \tag{6}$$

Equation (3) can be stated in a vectorised form as:

$$\tilde{\mathbf{x}} = \mathbf{Z}\boldsymbol{\gamma} \tag{7}$$

Accordingly, through further evaluation Eq. (6) can be expressed as:

$$\hat{\boldsymbol{\gamma}} = \arg\min_{\boldsymbol{\gamma}} \|\mathbf{x} - \mathbf{Z}\boldsymbol{\gamma}\|_2^2 = \mathbf{Z}^{+}\mathbf{x} \tag{8}$$

where $\mathbf{x}$ and $\boldsymbol{\gamma}$ are $n \times 1$ and $d \times 1$ vectors, respectively, and $\mathbf{Z}^{+}$ is the pseudoinverse of $\mathbf{Z}$, which is a $n \times d$ matrix of basis functions, such that:

$$\mathbf{Z} = \begin{bmatrix} B_1^l(t_1) & \cdots & B_d^l(t_1) \\ \vdots & & \vdots \\ B_1^l(t_n) & \cdots & B_d^l(t_n) \end{bmatrix} \tag{9}$$

As is evident from Eqs (8) and (9), $\hat{\boldsymbol{\gamma}}$ is dependent on the number of basis functions d used for representation. Thus, a sub-optimal value of d, i.e., too large or small, leads to overfitting or underfitting, respectively [17]. To address this problem, the *smoothing parameter* λ is used to *regularise* the least squares estimate, leading to the following modified form of Eq. (6):

$$\sum_{j=1}^{n}\big(x(t_j)-\tilde{x}(t_j)\big)^2+\lambda\int\big(\tilde{x}''(t)\big)^2\,dt \tag{10}$$

Here, λ penalises the curvature of the 'fit' function to prevent overfitting, thus resulting in the moniker *Penalised B-splines* (P-splines) [11]. Accordingly, based on Eq. (8), the Penalised Least Squares (PLS) estimate for γ can now be evaluated as:

$$\hat{\boldsymbol{\gamma}}=\big(\mathbf{Z}^T\mathbf{Z}+\lambda\mathbf{K}\big)^{-1}\mathbf{Z}^T\mathbf{x} \tag{11}$$

where the matrix $\mathbf{K}$ is the penalty matrix whose elements are the integrated products of second order derivatives of B-spline basis functions [9].

By substituting the value of $\hat{\boldsymbol{\gamma}}$ from Eq. (11) into Eq. (7), the estimated function $\tilde{\mathbf{x}}$ can be evaluated, as follows:

$$\begin{aligned}\tilde{\mathbf{x}}&=\mathbf{Z}\big(\mathbf{Z}^T\mathbf{Z}+\lambda\mathbf{K}\big)^{-1}\mathbf{Z}^T\mathbf{x}\\&=\mathbf{S}_{\lambda,Z}\mathbf{x}\end{aligned} \tag{12}$$

where $\mathbf{S}$ is known as *smoother matrix* [17]. This equation can be extended to 2-D annotation data, such that the estimated annotation function for video i and subject j, i.e., $\tilde{\mathbf{x}}_{ij}$, can be expressed as:

$$\tilde{\mathbf{x}}_{ij}=(\tilde{\mathbf{x}}_1,\tilde{\mathbf{x}}_2)_{ij}=\big(\mathbf{S}_{\lambda,Z}^{(i,j)}\mathbf{x_1},\mathbf{S}_{\lambda,Z}^{(i,j)}\mathbf{x_2}\big) \tag{13}$$

2.4. Parameters for functional representation

From Eqs (9) and (12) it is evident that the functional representation is contingent to three parameters: (i) the degree l of the basis-splines, (ii) the number d of basis functions, and (iii) the smoothing parameter λ. Optimal values of these parameters are required to obtain a 'good' and denoised approximation of the underlying function $\tilde{x}(t)$. These values can be determined individually for different subjects and videos but here one suitable set of values was determined.

To this end, splines of order 4 (degree $l=3$), i.e., *cubic splines*, were chosen because: (i) they have continuous first and second derivatives at the knots, thus the derivatives of obtained functional representation are also smooth, and (ii) the smoothing parameter λ in Eq. (10) is defined in terms of the squared second derivative. The choice of d is less trivial, as large values result in a richer basis representation, but tend to overfit the data and require considerable computational effort ($\mathcal{O}(nd^2+d^3)$ flops) for the evaluation of least squares minimisation [17]. While the former problem was mitigated through regularisation, the latter requires that an optimal d, such that $d\ll n$ is chosen. To this end, the longest video, i.e., sc1 with $n=3939$, was used to evaluate d that would also generalise to other videos. Accordingly, for different values of d (where d was increased in steps of 200 in the range of $d=220,\ldots,n$), the computation-time and the 'quality of fit' (quantified by Sum of Squared Errors (SSE)) were computed and consequently $d=820$, that represents a balance between quantities, was chosen. Finally, the smoothing parameter λ was determined using *Generalised Cross-Validation* (GCV) [17], where the GCV(λ) criterion was calculated for different videos, as:

$$\mathrm{GCV}(\lambda)=\frac{1}{n}\sum_{i=1}^{n}\left(\frac{x(t_i)-\tilde{x}(t_i)}{1-\mathrm{tr}(\mathbf{S})/n}\right)^2 \tag{14}$$

and the optimal λ was determined by comparing GCV(λ) for different values of λ and selecting the one with minimum GCV(λ) across all videos. Accordingly, $\lambda^*=10^{-4}$ was chosen.

2.5. Multivariate Functional Principal Component Analysis (MFPCA)

Functional-PCA (FPCA) is analogous to PCA as the main aim of both these methods is to determine the dominant modes of variation in the data [19]. FPCA is however applied to functions, instead of multi-dimensional vectors as in PCA [27]. Standard FPCA approaches are not suitable for the given dataset, as they – (i) only operate on univariate functions, (ii) require that all functions are of same time-duration. Therefore, for the annotation functions, the state-of-the-art MFPCA [16] was used. MFPCA is an advanced statistical method and its thorough presentation is beyond the scope of this article. The rest of this subsection presents an abridged introduction to MFPCA, such that an intuition about it can be developed. More details on MPFCA can be found in [15,16].

At its core, the used MFPCA method is based on the truncated *Karhunen–Loève* expansion. According

to which, a function can be approximated by a finite sum of its functional principal components [22], so the basis expansion defined in Eq. (3) can be modified as:

$$X^{(j)}(t_j) = \sum_{m=1}^{d} \gamma_m B_m(t_j) = \sum_{m=1}^{K} \theta_m^{(j)} b_m^{(j)}(t_j) \quad (15)$$

where $X^{(j)}$ and t_j are the jth univariate function and its time-duration, respectively. $b_m^{(j)}$ are orthonormal basis functions with coefficients $\theta_m^{(j)}$. The estimation of MFPCA can then be generalised by using univariate functional basis expansion with weighted scalar product, where the weights allow for rescaling such that the univariate principal component expansion in Eq. (15) can be extended to the multivariate case. Thus, given weights $w_1, \ldots, w_p > 0$ for each univariate function and demeaned observations $x_1, \ldots, x_N$ of X with estimated basis function coefficients $\hat{\theta}_{i,m}^{(j)}$ for each element, the eigen-analysis problem for MFPCA can then be expressed as:

$$(N-1)^{-1}\mathbf{B}\mathbf{D}\boldsymbol{\Theta}^T\boldsymbol{\Theta}\mathbf{D}\mathbf{c} = v\mathbf{c} \quad (16)$$

where $\mathbf{B} \in R^{K_+ \times K_+}$ with $K_+ = \sum_{j=1}^{p} K_j$ is a block diagonal matrix of scalar products $\langle b_m^{(j)}, b_n^{(j)} \rangle_2$ of univariate basis functions associated with each element $X^{(j)}$. The matrix $\mathbf{D} = \text{diag}(\mathbf{w}_1^{1/2}, \ldots, \mathbf{w}_p^{1/2})$ accounts for the weights. The matrix $\boldsymbol{\Theta}$ with rows $(\hat{\theta}_{i,1}^{(1)}, \ldots, \hat{\theta}_{i,K_1}^{(1)}, \ldots, \hat{\theta}_{i,1}^{(p)}, \ldots, \hat{\theta}_{i,K_p}^{(p)})$ contains the scores for each observation and $(N-1)^{-1}\boldsymbol{\Theta}^T\boldsymbol{\Theta}$ is an estimate for the covariance matrix $\mathbf{Q}$. The vectors $\mathbf{c}$ and v are the eigenvectors and eigenvalues, respectively. The associated scores $\rho_{i,m}$ for ith functional observation on the mth functional principal component can be then evaluated as:

$$\hat{\rho}_{i,m} = (\hat{v}_m)^{1/2}\left(\hat{c}_m^T\hat{\mathbf{Q}}_w\hat{c}_m\right)^{-1/2}\boldsymbol{\Theta}_{i,\cdot}\mathbf{D}\hat{c}_m \quad (17)$$

3. Results

3.1. MFPCA on all annotations

Based on methods presented in Sections 2.3 and 2.4, 240 (30 subjects × 8 videos) bivariate annotation functions were generated. Similar to PCA, the MFPCA method is contingent to the number of principal components M used for the representation [16]. To this end, the optimal M was determined based on how much of the total variance is explained by the different principal components [19]. For the annotation functions, M was set to 10, as based on scree plot analysis the subsequent increase in explained variance for $M > 10$ was insignificant. Thus, MFPCA (with $M = 10$) applied to 240 annotation functions, results in a 240×10 score matrix where each row contains scores for a function on the 10 functional Principal Components (fPC). The scores on the first and second fPC (i.e., 1.fPC and 2.fPC) for all functions are visualised in Fig. 3. These two fPC account for 88% explained variance, which increases to 95% for the first 4 fPC.

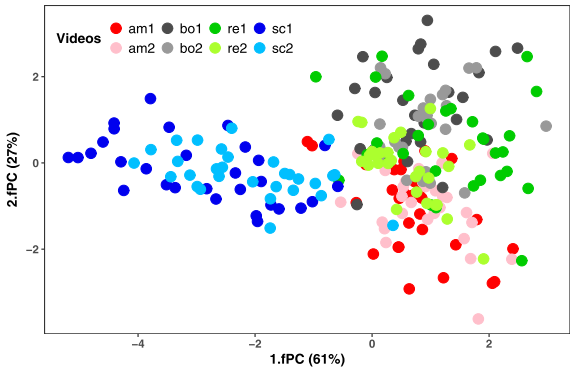

Fig. 3. Scatterplot of the MFPCA scores on the first two fPC for all annotations of the videos (see Table 1) used in the experiment.

Discussion. Fig. 3 shows that the annotations for *scary* videos (i.e., sc1 and sc2) have low scores on 1.fPC and are easily separable from other videos. The *amusing*, *boring* and *relaxed* annotations are not separable based on 1.fPC scores and only marginally separable based on their 2.fPC scores. Table 1 provides an initial reasoning for this result, where besides the scary videos, the V–A attributes of other videos are comparatively similar to each other. Another expected result from Fig. 3 is that annotation scores of 2 videos of the same emotion type tend to cluster together.

3.2. Analysis of annotation patterns

To determine the efficacy of both, the used stimuli and the annotation strategy, in eliciting desired V–A response from the participants, an analysis of concordance in annotations is necessary. To this end, the Euclidean distances between the MFPCA scores of annotations on the first 4 fPC (95% explained variance) were used. Such that, relatively higher distance values signify low concordance between the annotations. For

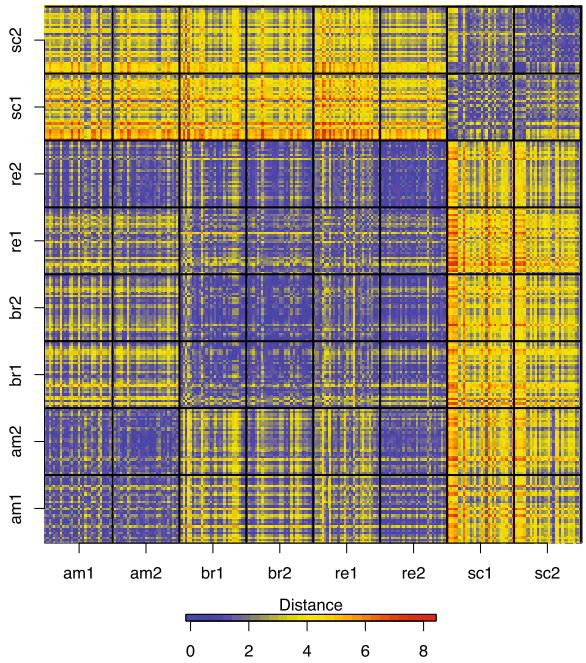

Fig. 4. Distance-heatmap for all annotations, where distances are inversely related to concordance between annotations.

the given 240 bivariate annotation functions, this analysis evaluates the pair-wise distance between annotations resulting in $240 \times (240-1)/2$ unique distance values. Due to the large number of values, this result is best presented in form of a symmetric 240×240 heatmap (see Fig. 4), where the continuous colourmap (ranging from blue–yellow–red) signifies increasing distance between the annotations.

Discussion. A major trend evident from Fig. 4 is that within a given type (i.e., within-type concordance), the annotations for *scary*, *amusing* and *boring* videos, with largely blue regions, are relatively concordant to each other. This result is along expected lines. However, the *relaxed* videos only exhibit marginal within-type concordance. The within-video concordance, that is characterised by the distance amongst annotations for a video, is high for most videos except re1 and sc1. The unexpected results for *relaxed* videos (specifically re2) and sc1, can be attributed to stronger than expected concordance with *amusing* and *boring* videos, and comparatively high between-subject disagreement, respectively.

3.3. Separability of annotations

The analyses in Sections 3.1 and 3.2 present an intuition about the separability of annotations, which can now be formally evaluated using discriminant analysis. To this end, we used *Fisher's Discriminant Ratio* (FDR) to determine the relative cluster separability between the scores of all videos on different fPC [10]. FDR between two videos for each fPC is calculated as:

$$FDR_{i,j} = \frac{(\mu_i - \mu_j)^2}{\sigma_i^2 + \sigma_j^2} \tag{18}$$

where i and j are the videos whose separability is being evaluated. μ and σ^2 are the arithmetic mean and variance of the scores across all subjects for these videos. Since the first 4 fPC account for 95% explained variance, FDR between the videos was calculated on each of these fPC and the results are presented in Fig. 5.

Discussion. In Fig. 5 (leftmost), the comparatively large FDR values of the *scary* videos for first fPC (61.38% explained variance) indicate higher separability from other videos. Similarly, FDR values for the 2.fPC (second to left, 26.89% explained variance) indicate that this fPC is essential for separating the *boring* from *amusing* videos. Also, re1 is marginally separable from *amusing* videos in this fPC, so is re2 from *boring* videos. These results however also indicate that re1 and re2 are not highly discriminable from *boring* and *amusing* videos, respectively. The issue of low separability of re2 from *amusing* videos persists when comparing FDR values in the third (second to right, 3.9% explained variance) and fourth (rightmost, 2.84% explained variance) fPC. Also, the FDR values in 3.fPC and 4.fPC show that these fPC allow for discrimination between sc1 and sc2.

3.4. Classification of annotations

The results presented in Sections 3.1 and 3.3 indicate that two (in turn, four) principal components of the annotations can already be used for a qualitative analysis of the characteristics of each type which are invariant across subjects. We now turn to a quantitative analysis of separability using a standard classification method. To this aim, we use the complete score vector (10-D) for the annotations as the input space to a Support Vector Machine (SVM, [5]) with Radial Basis Function kernel, while the label of each annotation is the type of clip. As a measure of accuracy we employ the balanced classification accuracy, defined as one minus the balanced error rate (average of the classification errors for each type of clip). We employ the standard SVM library *libsvm* [6] and let it perform one-

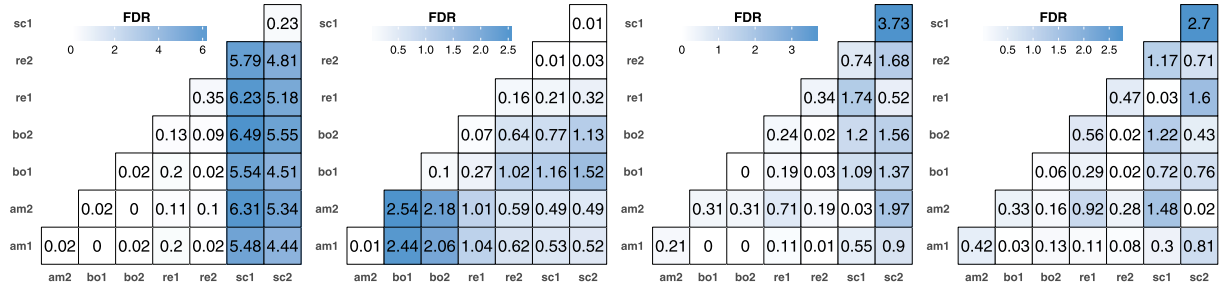

Fig. 5. FDR values from the first four fPC (left to right).

Table 2

Classification balanced accuracy [%] and number of Support Vectors, per video type and overall, for the 25/5 training/testing set [33], and for the leave-one-subject-out (LoSo) schema (mean values ± one standard deviation, all percentages rounded to the nearest integer)

	Classification balanced accuracy [%]					Number of support vectors				
	Amusing	Boring	Relaxed	Scary	Overall	Amusing	Boring	Relaxed	Scary	Overall
25/5	80	70	70	100	80	29	39	45	7	120
LoSo	80 ± 31	68 ± 36	47 ± 39	92 ± 27	72 ± 16	39 ± 4	45 ± 3	56 ± 2	13 ± 3	152 ± 12

versus-one multiclass classification [18]; notice that, since the problem has four classes, a balanced classification accuracy of 25% represents the minimum acceptable result (chance level).

We perform two kinds of analyses, each characterised by a different way of generating the training and testing set. In the first analysis, analogously to what is done with characteristic time sequences in [33], the training set consists of the annotations from the 25 participants chosen in that paper (200 annotations), whereas the testing set consists of the data from the remaining 5 participants (40 annotations). The optimal hyper-parameters of the SVM, C and γ, are found via grid search [5] within ranges of $\gamma = 10^{(\log \frac{1}{10} + [-2,\ldots,2])}$ and $C = 10^{[0,\ldots,3]}$, in steps of 0.1. The resulting optimal model (with $C^* = 200$ and $\gamma^* = 0.04$) has 120 support vectors (29, 39, 45, 7 in turn for amusing, relaxed, boring and scary videos) and obtains a balanced classification accuracy of 80% (80%, 70%, 70%, 100% – round percentages are to be expected since the testing set only contains 10 annotations per type of clip).

Secondly, we perform a deeper analysis, namely a Leave-one-Subject-out (LoSo) classification: for each of the 30 subjects in turn, her annotations are used as the testing set, and the annotations of the remaining 29 subjects constitute the training set. The ranges for the grid search of the optimal parameters are as above. The balanced classification accuracy is 71.67% ± 15.72% (average value plus/minus one standard deviation). Table 2 summarises the results (percentages rounded to the nearest integer).

Discussion. The classification accuracy in both problems is high if compared to that found in our previous work [33]: in that case the overall accuracy was 52% or 60% (as opposed to 80% and 72% ± 16% in the present case) but it must be remarked that the two datasets are hardly comparable: in the former case the input space was represented by *points in time* whereas here we deal with samples, each one of which represents a whole trajectory in the V–A space. Overall, it seems that MFPCA is hereby to some extent capturing essential information about a trajectory, and this is reflected in the results, much higher than chance levels. As expected, the accuracy values per-type-of-video confirm the impressions of the previous sections, namely, that the *scary* videos are clearly different from all others, that the *amusing* ones are too, albeit to a lesser extent, and that *boring* and *relaxed* videos are harder to discriminate from other ones. The analysis of the number of support vectors confirms the previous claims: very few SVs are necessary to form a support for the scary videos (13 ± 3), more of them for the amusing ones (39 ± 4) and even more for boring and relaxed videos.

3.5. Characteristic annotation

Given multiple subjective annotations pertaining to an emotional stimulus, a major challenge in AC is to determine a *ground truth* (also known as *characteris-*

Table 3
Calculation of the characteristic annotation for videos used in the experiment: The video label, the number of fPC required to account for 95% explained variance, the variance explained (in %) by the first two fPC and the number of annotations detected to be outlying based on robust mahalanobis distance

Video label	No. of fPC for 95% explained variance	Explained variance (in %) by 1.fPC and 2.fPC	No. of outlying annotations
am1	6	75.3	5
am2	6	71.1	7
bo1	3	88.7	0
bo2	6	77.3	9
re1	4	84.4	3
re2	6	74.1	3
sc1	6	81	0
sc2	6	75.4	7

tic) annotation for that stimulus. To this end, the aforementioned MFPCA approach can be used. This entails, first, undertaking the MFPCA analysis separately for each video stimulus. The M used for this analysis was determined to be 10, which is the same as used for the MFPCA analysis for all annotations (see Section 3.1). However, the number of fPC that account for 95% explained variance varies across videos (see Table 3). Second, in the resulting fPC space, robust *Mahalanobis Distance* (*MD*), which is a co-variance weighted distance measure between each annotation and the cluster center [12], was calculated as follows:

$$MD = \left(\hat{\boldsymbol{\rho}}^T \mathbf{C}^{-1} \hat{\boldsymbol{\rho}}\right)^{-1/2} \tag{19}$$

where $\mathbf{C}$ is the co-variance matrix and $\hat{\boldsymbol{\rho}}$ are the estimated fPC scores. For evaluating *MD*, only the scores on the first two fPC (>70% explained variance) were used (see Table 3) as an initial analysis revealed that scores on higher fPC (i.e, >2), which comparatively explain less variance than the first two fPC, skew the distance measure and thereby lead to false positives. Third, the 'outlying' annotations for each video were determined by enforcing a standard threshold (i.e., α) level of 0.95 [13] on the resulting distribution of distances (see Table 3). Lastly, the characteristic annotations were then evaluated by removing the diverging annotation functions and recalculating the mean functional annotation for each video. The results of the aforementioned steps for am1 are shown in Fig. 6.

Discussion. Table 3 shows the number of 'outlying' annotations detected for each video. An interesting result here is for bo1 and sc1, as no 'outlying' annotations were detected. This can be attributed to the variance-based nature of this approach. Specifically, if the annotations for a video have high within-video variance, then this approach may fail to detect any annotation as diverging. Nevertheless, this approach provided expected results for most videos in the dataset. The left plot in Fig. 6 shows that by removing the 'outlying' annotations for am1 there is evident, albeit small, change in the mean score. This behaviour varies across videos, where for some, the change was large and for others, insignificant. Also, for am1 video the overall expected V–A attributes were mid to high valence and arousal and the center plot in Fig. 6 shows that this approach was quite successful in detecting annotations that diverge from these attributes. However, this result also varied across videos, such that annotations that appeared to be diverging were not always detected. The resulting characteristic annotation for the

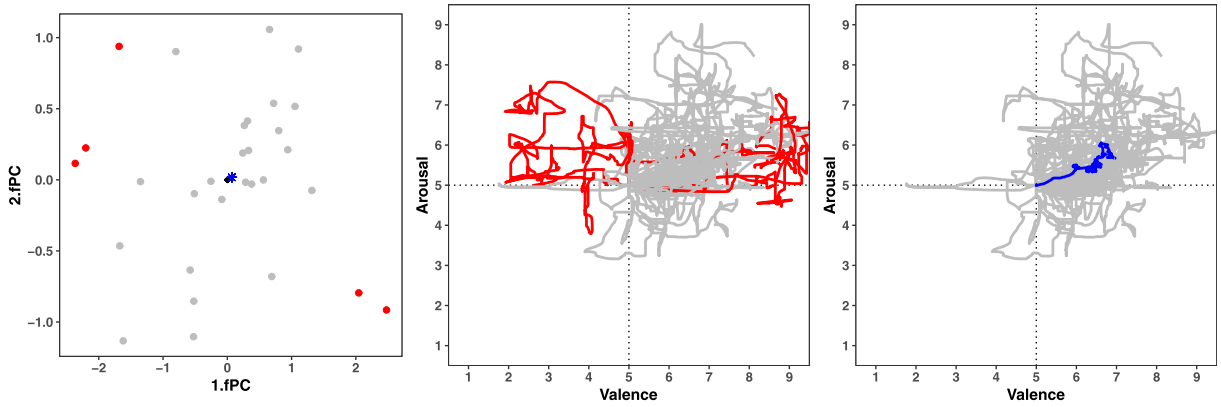

Fig. 6. Calculation of characteristic annotation for am1. (Left) scores in the first 2 fPC: Outliers (red), original-mean (black) and adjusted-mean (blue star). (Center) the outlying (red) and non-diverging (gray) annotations in the original valence-arousal space. the resulting characteristic annotation (blue) in the valence-arousal space.

am1 also conforms with the expected overall V–A attributes for this video. This result was along expected lines for most videos, except the *relaxed* videos, owing to the same reasons as presented in Sections 3.3 and 3.2, namely, that the resulting annotations for these videos do not partially conform with their expected V–A attributes.

4. Conclusions

A major goal of AC is to develop systems that can detect, and react to, human emotions. Yet, as emotions are personal and intimate, it is hard to define *what such systems are supposed to detect*, that is, what a sensible *annotation* of emotions can possibly be. The problem is even harder when these annotations are acquired continuously. In this work, a FDA approach that addresses these challenges was presented.

The fundamental challenge of how to best acquire comprehensive, continuous annotations of the emotional experience is addressed by our *Joystick-based Emotion Reporting Interface* (JERI). This approach, involving continuous and simultaneous annotation by the subject, can lead to increased cognitive load; but we have shown in our previous work that the use of joystick mitigates this issue to a large extent. Secondly, given that valence and arousal are inherently related, our approach of simultaneous annotation is advantageous over existing commonly-used approaches. The next challenge pertains to the continuous (and in our case, bivariate) nature of the annotations. The approach presented in this work focusses on converting these annotations into functions for further processing. This adds an extra data processing stage (see Sections 2.3 and 2.4) that is computationally intensive. However, we believe the benefits here outweigh the costs. As, firstly, this approach also 'smoothens' the data by removing perturbations, hence it comprises of the pre-processing stage that needs to be undertaken on this data. Secondly, this conversion into functions retains the time component of annotations, which is often ignored by commonly used analytical methods that ignore auto-correlation conspicuous in this data.

The intended overall emotional (Valence–Arousal) attributes of different video types and labels are shown in Table 1. Accordingly, it is desired that the subjective annotations pertaining to these videos exhibit these attributes. Such that, the annotations for different video types are distinct, and of same video label are similar to each other. These desired properties encompass the next challenge associated with subjective annotation of emotional experience, which was addressed using MFPCA. Accordingly, Section 3.1 presents an exploratory analysis of annotation patterns which was then formalised through concordance analysis presented in Section 3.2. In fact, this analysis also facilitates a comparison of annotations across all videos. The separability of annotations was then initially investigated using FDR in Section 3.3 and later formalised through SVM-based classification analysis in Section 3.4. Most results of these analyses are along expected lines and they demonstrate how MFPCA scores can be successfully used to address the aforementioned challenge.

The unexpected results, specifically for video re2, can be attributed to improper selection of this media as it fails to evoke the desired V–A attributes. Nevertheless, these unexpected results are a testimony to the benefit of the presented analyses, as they demonstrate how undesired effects can be detected and removed. They also demonstrate the benefit and efficacy of MFPCA technique in transforming a complex bivariate function into a 10-D score vector, which in turn facilitates the application of commonly-used statistical techniques on this data.

The MFPCA approach can indeed be used to determine the ground truth (or characteristic) for the different videos, which is another major challenge in AC. To this end, diverging (or 'outlying') annotations for a video were determined using robust MD and removed from the data. Subsequently, characteristic annotation for that video was determined by evaluating the mean functional representation for the remaining annotations. While this approach works as expected for most videos, for some (like br1 and sc1) the MD method doesn't find any outlying annotations. This unexpected result can be attributed to high within-video variances and demonstrates a limitation of this approach. Nonetheless, these results might improve by choosing a different distance measure or changing the criterion used to determine outliers.

At the onset of this paper, we introduced several challenges and current approaches to continuous annotation in AC. The use of our annotation interface (JERI) and the subsequent FDA-based analyses presented here address most, albeit not all, of these challenges. For example, the issue of inter-annotator delay is another major issue faced in continuous annotations. Also, as presented above, our approach has some shortcomings. Nevertheless, given that this is one of the initial, if not the first, attempts at using FDA methods for solving common problems in AC, we plan to

further develop on these methods to improve their performance. Also, other FDA methods, such as landmark registration, can be extended to this type of subjective data to address inter-annotator delay in annotations.

References

[1] J. Antony, K. Sharma, E.L. van den Broek, C. Castellini and C. Borst, Continuous affect state annotation using a joystick-based user interface, in: *Proceedings of Measuring Behavior 2014: 9th International Conference on Methods and Techniques in Behavioral Research*, 2014, pp. 268–271. doi:10.13140/2.1.2507.3929.

[2] M.E. Ayadi, M.S. Kamel and F. Karray, Survey on speech emotion recognition: Features, classification schemes, and databases, *Pattern Recognition* **44**(3) (2011), 572–587. doi:10.1016/j.patcog.2010.09.020.

[3] Y. Baveye, E. Dellandréa, C. Chamaret and L. Chen, Deep learning vs. kernel methods: Performance for emotion prediction in videos, in: *2015 International Conference on Affective Computing and Intelligent Interaction (ACII)*, IEEE, 2015, pp. 77–83. doi:10.1109/ACII.2015.7344554.

[4] M.M. Bradley and P.J. Lang, Measuring emotion: The self-assessment manikin and the semantic differential, *Journal of behavior therapy and experimental psychiatry* **25**(1) (1994), 49–59. doi:10.1016/0005-7916(94)90063-9.

[5] C.J.C. Burges, A tutorial on Support Vector Machines for pattern recognition, *Knowledge Discovery and Data Mining* **2**(2) (1998). doi:10.1023/A:1009715923555.

[6] C.-C. Chang and C.-J. Lin, LIBSVM: A library for support vector machines, *ACM Transactions on Intelligent Systems and Technology* **2** (2011), 27:1–27:27.

[7] R. Cowie, E. Douglas-Cowie, S. Savvidou, E. McMahon, M. Sawey and M. Schröder, 'FEELTRACE': An instrument for recording perceived emotion in real time, in: *ISCA Tutorial and Research Workshop (ITRW) on Speech and Emotion*, 2000, pp. 19–24.

[8] R. Cowie, M. Sawey, C. Doherty, J. Jaimovich, C. Fyans and P. Stapleton, GTrace: General trace program compatible with EmotionML, in: *2013 Humaine Association Conference on Affective Computing and Intelligent Interaction*, IEEE, 2013, pp. 709–710. doi:10.1109/ACII.2013.126.

[9] C. De Boor, *A Practical Guide to Splines*, Applied Mathematical Sciences, Springer, Berlin, 2001.

[10] L. Fahrmeir, A. Hamerle and G. Tutz, *Multivariate Statistische Verfahren*, Walter de Gruyter GmbH & Co KG, 1996.

[11] L. Fahrmeir, T. Kneib, S. Lang and B. Marx, *Regression: Models, Methods and Applications*, Springer Science & Business Media, 2013.

[12] P. Filzmoser, R.G. Garrett and C. Reimann, Multivariate outlier detection in exploration geochemistry, *Computers & Geosciences* **31**(5) (2005), 579–587. doi:10.1016/j.cageo.2004.11.013.

[13] P. Filzmoser and M. Gschwandtner, mvoutlier: Multivariate outlier detection based on robust methods, R package version 2.0.6, 2015.

[14] O. Grewe, F. Nagel, R. Kopiez and E. Altenmüller, Emotions over time: Synchronicity and development of subjective, physiological, and facial affective reactions to music, *Emotion* **7**(4) (2007), 774. doi:10.1037/1528-3542.7.4.774.

[15] C. Happ, funData: An S4 class for functional data, R package version 1.0, 2016.

[16] C. Happ and S. Greven, Multivariate functional principal component analysis for data observed on different (dimensional) domains, *Journal of the American Statistical Association* **113**(522) (2018), 649–659.

[17] T. Hastie, R. Tibshirani and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference and Prediction*, 2nd edn, Springer, 2008.

[18] C.-W. Hsu and C.-J. Lin, A comparison of methods for multiclass support vector machines, *IEEE Transactions on Neural Networks* **13**(2) (2002), 415–425. doi:10.1109/72.991427.

[19] I.T. Jolliffe and J. Cadima, Principal component analysis: A review and recent developments, *Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences* **374**(2065) (2016). doi:10.1098/rsta.2015.0202.

[20] M. Kächele, M. Schels and F. Schwenker, The influence of annotation, corpus design, and evaluation on the outcome of automatic classification of human emotions, *Front. ICT* **3** (2016), 27.

[21] M. Kächele, P. Thiam, M. Amirian, F. Schwenker and G. Palm, Methods for person-centered continuous pain intensity assessment from bio-physiological channels, *IEEE Journal of Selected Topics in Signal Processing* **10**(5) (2016), 854–864. doi:10.1109/JSTSP.2016.2535962.

[22] K. Karhunen, Über lineare Methoden in der Wahrscheinlichkeitsrechnung, *Annales Academiae Scientarum Fennicae* **37** (1947), 3–79.

[23] N. Malandrakis, A. Potamianos, G. Evangelopoulos and A. Zlatintsi, A supervised approach to movie emotion tracking, in: *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2011, pp. 2376–2379. doi:10.1109/ICASSP.2011.5946961.

[24] A. Metallinou and S. Narayanan, Annotation and processing of continuous emotional attributes: Challenges and opportunities, in: *2013 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, IEEE, 2013, pp. 1–8.

[25] F. Nagel, R. Kopiez, O. Grewe and E. Altenmueller, Emujoy: Software for continuous measurement of perceived emotions in music, *Behavior Research Methods* **39**(2) (2007), 283–290. doi:10.3758/BF03193159.

[26] R.W. Picard, *Affective Computing*, MIT Press, Cambridge, MA, USA, 1997.

[27] J. Ramsay and B. Silverman, *Functional Data Analysis*, Springer Science & Business Media, 2005.

[28] V.C. Raykar, S. Yu, L.H. Zhao, G.H. Valadez, C. Florin, L. Bogoni and L. Moy, Learning from crowds, *Journal of Machine Learning Research* **11** (2010), 1297–1322.

[29] A.M. Ruef and R.W. Levenson, Continuous measurement of emotion, in: *Handbook of Emotion Elicitation and Assessment*, 2007, pp. 286–297.

[30] J.A. Russell, Core affect and the psychological construction of emotion, *Psychological Review* **110**(1) (2003), 145. doi:10.1037/0033-295X.110.1.145.

[31] F. Schwenker, R. Böck, M. Schels, S. Meudt, I. Siegert, M. Glodek, M. Kächele, M. Schmidt-Wack, P. Thiam, A. Wendemuth and G. Krell, Multimodal affect recognition in the context of human–computer interaction for companion-systems, in: *Companion Technology – A Paradigm Shift in Human–Technology Interaction*, 2017, pp. 387–408.

[32] F. Schwenker and S. Scherer (eds), *Multimodal Pattern Recognition of Social Signals in Human–Computer-Interaction – 4th IAPR TC 9 Workshop, MPRSS 2016*, Cancun, Mexico, December 4, 2016, Revised Selected Papers, Lecture Notes in Computer Science, Vol. 10183, Springer, 2017. doi:10.1007/978-3-319-59259-6.

[33] K. Sharma, C. Castellini, F. Stulp and E.L.V. den Broek, Continuous, real-time emotion annotation: A novel joystick-based analysis framework, *IEEE Transactions on Affective Computing* (2018), 1–1. doi:10.1109/TAFFC.2017.2772882.

[34] K. Sharma, C. Castellini and E.L. van den Broek, Continuous affect state annotation using a joystick-based user interface: Exploratory data analysis, in: *Measuring Behavior 2016: 10th International Conference on Methods and Techniques in Behavioral Research*, 2016. doi:10.13140/RG.2.1.2285.7841.

[35] I. Sneddon, G. McKeown, M. McRorie and T. Vukicevic, Cross-cultural patterns in dynamic ratings of positive and negative natural emotional behaviour, *PloS One* **6**(2) (2011), e14679. doi:10.1371/journal.pone.0014679.

[36] M. Soleymani, S. Koelstra, I. Patras and T. Pun, Continuous emotion detection in response to music videos, in: *Face and Gesture 2011*, 2011, pp. 803–808. doi:10.1109/FG.2011.5771352.

[37] M. Valstar, B. Schuller, K. Smith, T. Almaev, F. Eyben, J. Krajewski, R. Cowie and M. Pantic, Avec 2014: 3d dimensional affect and depression recognition challenge, in: *Proceedings of the 4th International Workshop on Audio/Visual Emotion Challenge, AVEC '14*, ACM, New York, NY, USA, 2014, pp. 3–10.

[38] E.L. van den Broek, Monitoring Technology: The 21st Century's Pursuit of Well-being? Eu-osha Discussion Paper, European Agency for Safety and Health at Work (EUOSHA), Bilbao, Spain, 6 July 2017.

[39] E.L. van den Broek, J.H. Janssen and J.H.D.M. Westerink, Autonomous closed-loop biofeedback: An introduction and a melodious application, in: *The Oxford Handbook of Affective Computing*, Chapter 35 (Section 5: Applications of Affective Computing), Oxford Library of Psychology, Oxford University Press, Inc., New York, NY, USA, 2015, pp. 472–482.

[40] E.L. van den Broek, V. Lisý, J.H. Janssen, J.H.D.M. Westerink, M.H. Schut and K. Tuinenbreijer, *Affective Man–Machine Interface: Unveiling Human Emotions Through Biosignals*, Springer, Berlin, Heidelberg, 2010, pp. 21–47.

[41] C. Vondrick, D. Patterson and D. Ramanan, Efficiently scaling up crowdsourced video annotation, *International Journal of Computer Vision* **101**(1) (2013), 184–204. doi:10.1007/s11263-012-0564-1.

[42] J.-L. Wang, J.-M. Chiou and H.-G. Müller, Functional data analysis, *Annual Review of Statistics and Its Application* **3** (2016), 257–295. doi:10.1146/annurev-statistics-041715-033624.

[43] G.N. Yannakakis and H.P. Martinez, Grounding truth via ordinal annotation, in: *2015 International Conference on Affective Computing and Intelligent Interaction (ACII)*, 2015, pp. 574–580. doi:10.1109/ACII.2015.7344627.