This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TAFFC.2017.2772882, IEEE Transactions on Affective Computing

Continuous, real-time emotion annotation: A novel joystick-based analysis framework

Karan Sharma, Claudio Castellini, Freek Stulp, Egon L. van den Broek

Abstract—Emotion labels are usually obtained via either manual annotation, which is tedious and time-consuming, or questionnaires, which neglect the time-varying nature of emotions and depend on human's unreliable introspection. To overcome these limitations, we developed a continuous, real-time, joystick-based emotion annotation framework. To assess the same, 30 subjects each watched 8 emotion-inducing videos. They were asked to indicate their instantaneous emotional state in a valence-arousal (V-A) space, using a joystick. Subsequently, five analyses were undertaken: (i) a System Usability Scale (SUS) questionnaire unveiled the framework's excellent usability; (ii) MANOVA analysis of the mean V-A ratings and (iii) trajectory similarity analyses of the annotations confirmed the successful elicitation of emotions; (iv) Change point analysis of the annotations, revealed a direct mapping between emotional events and annotations, thereby enabling automatic detection of emotionally salient points in the videos; and (v) Support Vector Machines (SVM) were trained on classification of 5 second chunks of annotations as well as their change-points. The classification results confirmed that ratings patterns were cohesive across the participants. These analyses confirm the value, validity, and usability of our annotation framework. They also showcase novel tools for gaining greater insights into the emotional experience of the participants.

Index Terms—Affective Computing, Emotion in human-computer interaction, Tools and methods of annotation, Time-series analysis, Change-point analysis, Pattern Recognition

1 Introduction

By robustly estimating emotions in real-time, machines can improve interaction experiences for humans. For example, in Human-Robot Interaction (HRI) scenarios the robot could modify its behaviour to reduce human anxiety [1]. To this end, wearable sensors for measuring physiological descriptors of emotions (e.g., galvanic skin response, heart rate, etc.) are often used [1], [2], [3]. The data from these sensors need to be linked to the internal emotions experienced by the human, and this association step is still a largely unsolved problem. The standard paradigm in a laboratory setting is to provide emotion-inducing stimuli (e.g., videos [2], music [4], and/or photos [4]) to the participants and measure their affective response using biosignals [1], [2], [5], speech signals [6], and/or computer-vision based approaches [5]. The *ground truth* for the subject's emotional experience is then either obtained through Likert-scale based post-stimuli questionnaires or often manually annotated using discrete emotion labels [3]. Both these rating methods are unsuitable when using dynamic stimuli (e.g., videos), as they do not consider the time-varying nature of emotions [4], [5]. Secondly, discrete emotion labels are also insufficient to define the strength of emotional experiences [3], [7].

Thus, researchers have recently started using annotation tools that allow for continuous reporting. Most of these tools are based on the 2-dimensional *circumplex model* of emotion by Russell [8], wherein emotional labels (e.g., *fear*, *joy*, etc.) are represented in continuous dimensions of *valence* and *arousal*. Common 1D and 2D tools include, *CMS* [9], *GTrace* [10], *CARMA* [11], and *FEELtrace* [12], *EMuJoy* [4], *DARMA* [13], respectively. In-spite the abundance of available tools, several challenges remain. First, there are inherent cognitive and physical loads associated with continuous annotation [14], [15], [16]. This problem is especially evident when using mouse-based tools, where the user has to continuously press a button during annotation [16], [17]. Second, although simultaneous annotation of valence and arousal in a 2D space would allow for a more comprehensive reporting of emotional experience [4], [13], [18]; it is often not pursued, either due to the lack of tools or due to concerns regarding cognitive overloading. To address these shortcomings, two approaches have come to fore: (i) researchers have gradually started using joysticks, which have been reported to more intuitive to use than mouse-based tools [4], [13], [15], [17], [19]; and (ii) 2D tools that allow for simultaneous acquisition of valence and arousal ratings have become increasingly prevalent [4], [13], [14], [20]. However, approaches that combine both these desired aspects are still in their nascency; as most current joystick-based implementations acquire only 1D (valence/arousal) from each user [9], [15], [21] and most simultaneous acquisition approaches have used mouse-based tools [4], [14], [18].

To this end, in this work we describe, test and assess a new joystick-based annotation framework aimed at continuous and simultaneous annotation of dynamic stimuli. While the setup and purpose of our framework (first introduced in 2014 [22]) are similar to those of the recently published *DARMA* [13]; our framework is, to the best of our knowledge, one of the first implementations focussed on realising the aforementioned aims. The use of a joystick is central to our framework as most joysticks incorporate a return spring and hence automatically realign to the center under no force. Thus, in comparison to mouse-based setups, the user is consciously aware of the joystick's position in the User-Interface (UI) without continuously looking at it. This property makes joystick-based setups more intuitive to

- *K. Sharma, C.Castellini and F.Stulp are with the Robotics and Mechatronics Center, DLR – German Aerospace Center, Wessling, Germany. E-mail: karan.sharma@dlr.de, claudio.castellini@dlr.de and freek.stulp@dlr.de*
- *Egon L. van den Broek is with the Department of Information and Computing Sciences, Utrecht University, Utrecht, The Netherlands. E-mail: vandenbroek@acm.org*

use, in turn also reducing the cognitive load associated with the annotation procedure. Secondly, the 2D setup of our framework allows users to provide a more comprehensive description of their emotional experience than is possible by annotating each dimension separately [3], [8], [13].

The rest of the paper is organised as follows. Section 2 introduces the setup, the usability questionnaire, and the psychophysical experiment that was undertaken to acquire annotations using this framework. The usability analysis is presented in Section 3.1, where the results on simplicity, intuitiveness and usability of the system are highly positive, thus validating the setup of this framework. Since, annotation frameworks are used both in psychology and affective computing research [13], [15], the analysis presented here aims to address common challenges faced in both these domains. In Sections 3.2 and 3.3, multivariate statistical and sequence analysis are presented that assess the consistency in mean and continuous rating patterns, respectively. A robust regression based approach for combining multiple subjective Continuous Annotations (CA) into a single representative CA is also presented in Section 3.3. Then, in Section 3.4, a novel approach for automatically determining emotionally salient events in the stimuli, by undertaking *Change Point Analysis* (CPA) on the annotations, is presented. Following on, in Section 3.5, a standard classification method is trained on the *change-points* detected in the annotations to investigate the coherence of change-points across different subjects. Through the usability- and the data-analyses presented in this work, we postulate: (i) that our joystick based framework is a viable alternative to commonly used computer-mouse based systems; and (ii) that the analysis methodology presented here makes relevant contributions to the field of CA data analysis.

2 Methods

An experiment involving thirty volunteers (15 males, age 28.6±4.8 years and 15 females, age 25.7±3.1 years; range of age 22-37 years) who watched a series of 8 videos and simultaneously self-reported their affect state using the joystick-based annotation framework was designed. The 4 target emotions that the selected videos were expected to elicit are: amusement, boredom, relaxation, and scaredness. Thus, for every emotion, 2 videos were used.

To evoke the intended emotional response using videos, first their emotional content/label needs to be determined. To this end, an initial pool of 20 videos was drawn from several commonly used pre-labelled video-sets [23], [24], [25] and other sources. Through an internal review and evaluation, 8 videos were subsequently shortlisted for the experiment (see Table 1). To avoid carry-over effects in the experiment, a unique ordering of these 8 videos for each participant was pseudo-randomly generated, such that no two videos of the same emotion type followed each other in a sequence. The ordered videos were then interleaved by 2'-long blue screens which were not annotated by the participants, thus allowing them to rest during the experiment.

TABLE 1: Details and intended valence/arousal attributes of each video in the experiment.

Label	Title (year of release)	Genre	Intended attributes valence	Intended attributes arousal	Dur. [s]
amusing-1	Hangover (2009)	comedy	mid/high	mid/high	185
amusing-2	When Harry met Sally (1989)	romantic comedy	mid/high	mid/high	173
boring-1	Europe travel advisory (2013)	monologue	low	low	119
boring-2	Japanese tea ceremony (2012)	monologue	low	low	160
relaxed-1	Pristine beach (2011)	nature docu.	mid/high	low	145
relaxed-2	Zambezi river (2011)	nature docu.	mid/high	low	147
scary-1	Shutter (2004)	horror	low	high	197
scary-2	Mama (2008)	horror	low	high	144
-	Blue screen	(transition)	-	-	120

The annotation framework was developed using the data acquisition and graphical programming software *LabVIEW* [26]. Thus, this implementation also enables easy acquisition and synchronizing of sensor data (e.g., ECG, GSR etc.) with the annotations. The open-source *VLC media player* was used for video playback. Since, a large 42" flat-panel TV was used in the experiment, the annotation UI was superimposed in the upper-right corner of the video playback window (see Figure 1); however, other configurations are also possible. The valence (horizontal) and arousal (vertical) axes of the UI also included Self-Assessment-Manikins (SAM) that serve as visual guides to the participant for determining her valence-arousal levels [27]. The instantaneous position of the joystick in the UI is denoted by a red pointer (see Figure 1). The joystick was sampled at 20 Hz, which is in line with recent research on human motor control [28].

The experiment was approved by the DLR Ethics Committee. Before the experiment, each participant was provided a thorough written and oral description of the experiment, and was asked to sign an informed consent form. The participant then sat in a chair, with her hands resting comfortably on tables at either sides, and operated the joystick with her dominant hand. The chair was set at a comfortable viewing position from a TV screen fixed on the wall. The room was silent and darkened for an enhanced viewing experience. High quality headphones were provided to ensure the best sound effects. Before starting the experiment, the participant was instructed to relax, watch the videos, and indicate her perceived affective experience by appropriately positioning the red pointer on the interface. The participant was instructed to rate her "feeling while watching the video, and not the emotional content of the video". Moreover, she was told to move the joystick only when her feeling changed. To help the participant in habituating to the annotation procedure and framework, she practised on 5 videos that elicited distinct emotions. Also, any questions regarding annotation were addressed. After this, the actual experiment, containing the 8 videos and spanning approximately 40 minutes, was started.

At the end of the experiment, the participant was asked to provide feedback on the interface's usability, by answering the System Usability Scale (SUS) questionnaire [29]. As the standard SUS questionnaire can cause confusion among the respondents [29], an all-positive version of the questionnaire was used (see Table 2). The questionnaire also had a field for providing general comments on the system.

3 Results

3.1 Usability

The 'all-positive' SUS is a Likert scale based questionnaire where the responses are in the integer range of 1 — 5 (i.e., from *strongly disagree* to *strongly agree* [29]). The SUS questions, and the mean and standard deviation of the responses from all participants, are shown in Table 2. All responses are more than 'neutral' (i.e. response value of 3), especially for questions on simplicity (q.2) and intuitiveness (q.8) of the system. The SUS score per participant is calculated by first,

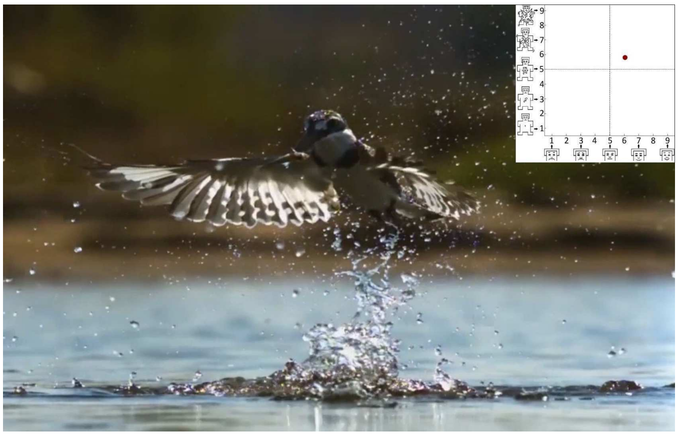

Fig. 1: The annotation UI embedded in a video.

scaling the responses to range 0 — 4, and then multiplying the sum of the scaled responses to all questions by 2.5. The resulting score thus lies in a range of 0 — 100. The mean of all per participant scores gives the average SUS score (i.e., 80.17±9.95) for the annotation system.

TABLE 2: The System Usability Scale (SUS) questions and the average responses.

Question	Response
I think that I would like to use this system frequently.	3.53±0.97
I found the system to be simple.	4.47±0.57
I thought the system was easy to use.	4.10±0.80
I think that I could use this system without the support of a technical person.	4.23±0.86
I found the various functions in this system were well integrated.	4.27±0.70
I thought there was a lot of consistency in this system.	4.50±0.73
I would imagine that most people would learn to use this system very quickly.	4.37±0.85
I found the system very intuitive.	4.30±0.75
I felt very confident using the system.	4.00±1.05
I could use the system without having to learn anything new.	4.30±0.84

Discussion. The responses to the individual questions on the questionnaire (see Table 2) show that the participants generally 'agree' to all questions. Also, the average SUS score of the system (80.17) is considered as an 'excellent' score for system usability [29]. These results indicate that the participants had positive experience while annotating and found it to be simple, consistent and intuitive. Several participants informally reported that the joystick, adds an element of "gamification" (i.e., excitement) to the annotation procedure, and that they would prefer it to a computer-mouse due to its ergonomic properties.

3.2 Consistency of Mean Valence-Arousal Ratings

The mean valence-arousal (V-A) ratings, by each participant for each of the 8 videos, are shown as scatter plots in Figure 2 (left). In these plots, a clear trend is visible — the ratings pertaining to 2 videos of the same emotion type tend to cluster in the same regions of the UI. For example, scary-1 and scary-2 videos tend to be in the upper-left quadrant. Also, videos pertaining to different emotion type (e.g., amusing and boring) tend to cluster in different regions of the UI.

To formally test the differences among the rating patterns, the mean ratings were analysed using Multivariate Repeated-Measures ANOVA (aka RM MANOVA) for which the videos were the *within-subjects* factor and the mean V-A ratings were the dependent variables (DVs). The main effect of videos on the ratings was highly significant ($F(14, 406) = 40.9$; $p < .001$; Pillia's trace $= 1.17$; $\eta_p^2 = .58$). Subsequently, two univariate RM-ANOVAs to individually test the effect of the *within-subjects* factor on valence and arousal ratings were performed. Since the data lacked sphericity, Greenhouse-Geisser estimates ($\epsilon = .59$ for valence and $\epsilon = .56$ for arousal, respectively) were used to adjust the degrees of freedom for the univariate RM-ANOVAs (the corrected degrees of freedom are: $df = \epsilon(k-1)$ and $df_{error} = \epsilon(k-1)(n-1)$). The RM ANOVAs using Greenhouse-Geisser estimates for valence ($F(4.13, 119.89) = 34.63$; $p < .001$; $\eta_p^2 = .54$) and arousal ($F(4.09, 118.68) = 88.34$; $p < .001$; $\eta_p^2 = .75$) were also highly significant ($p < .001$). Thus, the MANOVA and the ANOVAs establish that rating patterns differ across 8 videos in the experiment. To precisely determine which ratings were differing/similar from each other, post-hoc Bonferroni pairwise comparisons of the V-A ratings were performed. The results of these comparisons are presented in form of symmetric matrix plots in Figure 2 (center and right).

Discussion. For the video stimuli, the mean ratings should ideally emulate the expected V-A attributes listed in Table 1. To this end, the plots in Figure 2 show that this was generally the case. In particular, they show that the video ratings tend to form clusters together that are distinct from ratings for other video-types, thus exhibiting agreement in the subjective ratings. They also show that in most cases the experimental manipulation of the participants' emotion state was successful in a supposed manner. However, in some cases, as is often reported in literature [2], [3], it was not the case. For example, videos of same emotion type are expected to have same V-A attributes — this is not valid for relaxed videos, as the arousal ratings here are statistically significantly different from each other. Similarly, videos pertaining to different emotions should have different V-A attributes i.e., the valence/arousal can be alike, but both the attributes shouldn't be similar. However, for amusing-2 and relaxed-2, and boring-2 and relaxed-1, this was not the case. The V-A ratings were here not significantly different from each other. We believe, that this unexpected divergence of ratings can be ascribed to misestimation of V-A attributes, rather than to errors in the annotation process.

Though this analysis provides insights into the consistency of ratings, it is however limited, as it doesn't account for the continuous nature of annotations. This topic is further investigated in the next subsection (see 3.3).

3.3 Consistency of Continuous V-A Ratings

In this subsection, sequence dissimilarity (hereafter referred to as 'distance') analysis is used as the continuous analogous to mean ratings analysis presented in the last subsection (see 3.2). To this end, first the high-frequency artifacts in the annotations were removed using a 1-D Savitzky-Golay filter with a time window of $1s$ and a polynomial of degree 3. Then, the 30 subjective annotation trajectories for each video are combined into their respective Characteristic Trajectory (CT). Combining multiple annotations is a common challenge in the field of CA analysis [17]. Approaches to this problem range from the trivial, where the simple point-by-point mean [12]/median [20] is calculated, to the fairly sophisticated, where inter-annotator and reaction delays are also addressed to obtain *ground truth* for audiovisual features [15], [19], [30]. Therefore,for analysing of consistency

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TAFFC.2017.2772882, IEEE Transactions on Affective Computing

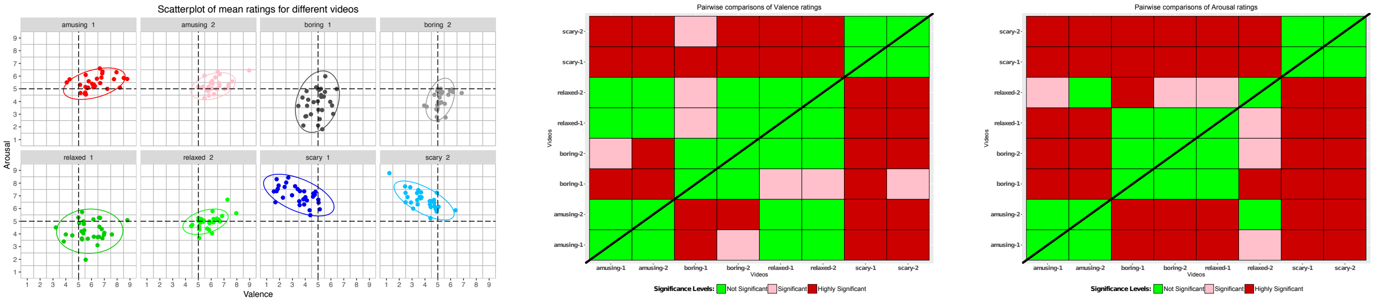

Fig. 2: Scatter plot with data ellipses for each video (left). Pairwise comparisons of the average valence (center) and arousal (right) ratings across videos, with colors depicting different significance levels ($p > .05$, not significant; $.001 \leq p \leq .05$, significant; $p < .001$, highly significant).

in annotations acquired from our framework, an intermediate approach we used. This approach, specifically Robust Local Polynomial Regression (RLPR), does not address delays, but is robust against highly diverging annotations that unduly influence the shape of resulting CT [31]. In RLPR, first a linear regression model is fit to the data to calculate its residuals $\hat{\epsilon_i}$ [31]. The 'robust' weights r_i for RLPR are then calculated applying Tukey's bi-weight function $\mathcal{B}$ to these residuals:

$$r_i = \mathcal{B}\left(\frac{\hat{\epsilon_i}}{6\hat{q_{0.5}}}\right), \tag{1}$$

where $\hat{q_{0.5}}$ is the median of $|\hat{\epsilon_i}|$. This approach makes RPLR more robust to outliers than weighted least squares regression [31]. For calculating CTs, RLPR with a 2nd degree polynomial and a span of 3 seconds was used. The resulting CTs are shown in Figure 3, where, as expected, they span the same quadrants as their mean ratings.

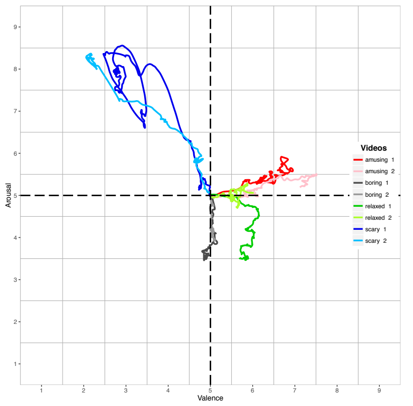

Fig. 3: Characteristic trajectories for each video.

Grid-based discretisation, with a grid resolution based on the UI design (9x9 tiles), was then used to convert the 2D CTs to 1D discrete sequences. The motivation for the same was that sequence analysis methods can detect similar, but not necessarily concurrent [32], patterns in the sequences. Thus, common patterns between CTs, irrespective of their position in a given CT, can be identified and used to determine the distances between the videos. To quantify these distances, we defined two metrics: the Unique States (d_{US}) and Longest Common Sub-Sequence (d_{LCSS}) distance [33].

A unique sequence $\tilde{S}$ is determined by extracting only the distinct elements of any given sequence S. Thus, $\tilde{S}$ contains all the unique tiles (see Figure 3) that a trajectory traversed. The d_{US} between any two unique sequences $(\tilde{S_1}, \tilde{S_2})$ is then calculated as:

$$d_{US}(\tilde{S_1}, \tilde{S_2}) = 1 - \frac{C(\tilde{S_1}, \tilde{S_2})}{\sqrt{|\tilde{S_1}||\tilde{S_2}|}}, \tag{2}$$

where $|\tilde{S_1}|$ is the length (i.e., total no. of elements) of the unique-states sequence $\tilde{S_1}$ and $C(\tilde{S_1}, \tilde{S_2})$ is number of common elements in the unique sequences $\tilde{S_1}$ and $\tilde{S_2}$.

Unlike d_{US}, d_{LCSS} is calculated over the complete length of the sequences. The LCSS for any two sequences (S_1, S_2) is a sequence (S_{1-2}^{LCSS}) containing all elements that occur in the same order, but are not necessarily contiguous [33]. The d_{LCSS} between S_1, S_2 is calculated as:

$$d_{LCSS}(S_1, S_2) = 1 - \frac{L(S_1, S_2)}{\sqrt{|S_1||S_2|}}, \tag{3}$$

where $L(S_1, S_2)$, $|S_1|$ and $|S_2|$ are the lengths of S_{1-2}^{LCSS}, S_1 and S_2 , respectively. The resulting pairwise distances d_{US}, d_{LCSS} for all sequences are shown in form of symmetric matrix plots in Figures 4 and 5, respectively.

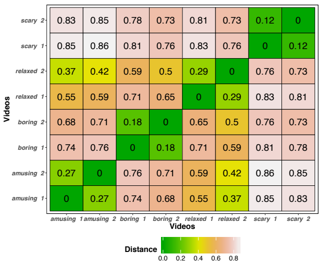

Fig. 4: Distance Matrix plot for d_{US}.

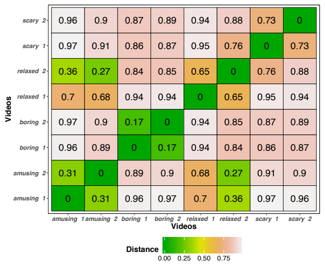

Fig. 5: Distance Matrix plot for d_{LCSS}.

Discussion. As previously mentioned in the discussion of Section 3.2, the videos' ratings are deemed to be consistent if they emulate the expected V-A attributes listed in Table 1. Thus, ratings of videos from the same/different emotion label should be accordingly close/aside from each other. As can be observed in Figures 4 and 5, the aforementioned conditions are generally valid in most cases, barring a few exceptions.

An interesting d_{US} result is that besides being similar to each other, the amusing sequences are also relatively close to the relaxed-2 sequence ($d_{US}(\text{amusing-1}, \text{relaxed-2}) = 0.37$, $d_{US}(\text{amusing-2}, \text{relaxed-2}) = 0.42$). Similarly, based on their d_{LCSS} distance, amusing-2 and relaxed-2 ($d_{LCSS} = 0.27$) are most similar to each other. A similar result was observed in Section 3.2, where V-A ratings for amusing-2 and relaxed-2 were not significantly different. The d_{LCSS} between any two given sequences is generally greater than their corresponding d_{US} distance. This can be attributed to the fact that unlike d_{US}, d_{LCSS} distances also account for the temporal ordering of the elements/states in the sequences. Thus, for example, $d_{LCSS} = 0.73$ and $d_{US} = 0.12$ for the scary videos. The difference in temporal ordering of events in scary videos is also evident from Figure 3, where the ”swirls” in CT of scary-1 show that the tension builds and drops several times, and for scary-2 video the tension gradually build-ups throughout the video. By accounting for the time-varying nature of emotion annotations, this analysis augments the ratings' consistency analysis presented in Section 3.2.

3.4 Change-Point Analysis (CPA)

As the emotional content in a dynamic stimulus (e.g., video) varies, we anticipate that the CA pertaining to that stimulus, also change in tandem. Thus, major change-points in CA would signify emotionally relevant moments/points that can be used to identify and isolate emotionally salient segments in the stimuli [17]. Based on this premise, CPA (i.e., a method that detects distributional shifts in time series [34]) was used to determine the change-points in the valence and arousal time-series from the CT of the stimuli.

To undertake CPA, the non-parametric *E-Divisive with Medians (EDM)* [35] algorithm was used. The EDM identifies change-points by detecting divergence in the mean of a time-series, but can also detect changes in distribution [35]. Given a time series, $Z_1, Z_2, \ldots, Z_n$, a change-point τ is the point that splits the time series into two segments: $X_\tau = \{Z_1, Z_2, \ldots, Z_\tau\}$ and $Y_\tau(\kappa) = \{Z_{\tau+1}, Z_{\tau+2}, \ldots, Z_\kappa\}$. Subject to the conditions $1 < \delta \leq \tau$ and $\tau + \delta \leq \kappa \leq n$, the estimated change-point location, $\hat{\tau}$, is calculated as:

$$(\hat{\tau}, \hat{\kappa}) = \underset{\tau,\kappa}{\arg\max}\, \tilde{\mathcal{Q}}(X_\tau, Y_\tau(\kappa); \alpha, \delta), \tag{4}$$

where α is the indexing parameter used to scale the distance between the distributions and δ is the reduced number of observations taken from head and tail of distributions for X and Y, respectively [35]. Also, $\tilde{\mathcal{Q}}(X_\tau, Y_\tau(\kappa); \alpha, \delta)$ is the scaled sample divergence measure that serves as the test statistic for permutation tests (presented later in this Section). The statistic $\tilde{\mathcal{Q}}(X_\tau, Y_\tau(\kappa); \alpha, \delta)$ is based on robust sample divergence measure, $\tilde{\mathcal{E}}(X_\tau, Y_\tau(\kappa); \alpha, \delta)$, and is calculated as:

$$\tilde{\mathcal{Q}}(X_\tau, Y_\tau(\kappa); \alpha, \delta) = \frac{\tau\kappa}{\tau + \kappa} \tilde{\mathcal{E}}(X_\tau, Y_\tau(\kappa); \alpha, \delta). \tag{5}$$

The measure $\tilde{\mathcal{E}}(X_\tau, Y_\tau(\kappa); \alpha, \delta)$ is robust against anomalies as it is calculated using medians instead of sample means:

$$\tilde{\mathcal{E}}(X_n, Y_m; \alpha, \delta) = 2m^{\alpha,\delta}_{XY} - m^{\alpha,\delta}_{XX} - m^{\alpha,\delta}_{YY}, \tag{6}$$

where $m^{\alpha,\delta}_{XX}$ and $m^{\alpha,\delta}_{YY}$, and $m^{\alpha,\delta}_{XY}$ are the within-sample and between-sample distances, respectively. These are calculated as follows:

$$\begin{aligned} m^{\alpha,\delta}_{XX} &= \text{median}\{|x_i - x_j|^\alpha : 1 \leq i < j \leq \delta \quad \text{or} \quad i + 1 = j\}, \\ m^{\alpha,\delta}_{XY} &= \text{median}\{|x_i - y_j|^\alpha : n - \delta + 1 \leq i \leq n, 1 \leq j \leq \delta\}. \end{aligned} \tag{7}$$

While large values of the test statistic $\tilde{\mathcal{Q}}(X_\tau, Y_\tau(\kappa); \alpha, \delta)$ correspond to a significant change in distribution [35], we wish to precisely determine the statistical significance of the estimated change-point $\hat{\tau}$. For this, knowledge of the underlying distribution is required, but is not available in this scenario. Therefore, the significance of $\hat{\tau}$ is determined through permutation tests. Under the null hypothesis that a change-point does not exist, the permutation test is performed as follows: First, the observations in the time-series are permuted to obtain a new time-series. Then, a new change-point is estimated from the permuted time-series by applying the aforementioned estimation procedure. After R random permutations, the corresponding test statistic $\tilde{\mathcal{Q}}^{(r)}$ is used to calculate the approximate p-value as follows: $\hat{p} = \#\{r : \tilde{\mathcal{Q}}^{(r)} \geq \tilde{\mathcal{Q}}\}/(R + 1)$. For the EDM algorithm, $R = 199$ and the significance level of 0.05 is used to determine the significance of the change-point.

The procedure above determines the location estimate and statistical significance of a single change-point. This procedure can be iteratively applied to determine multiple change-points (refer [34]), as was done in our case.

TABLE 3: Change-points for each video.

	Valence		Arousal	
Video	CPs	range	CPs	range
amusing-1	3	1.96	4	0.92
amusing-2	1	2.47	2	0.56
boring-1	5	0.36	2	1.55
boring-2	11	0.12	2	1.02
relaxed-1	3	1.16	1	1.56
relaxed-2	4	1.00	4	0.58
scary-1	3	2.54	4	3.58
scary-2	1	2.92	2	3.30

Table 3 lists the number of resulting change-points and the range, for the characteristic V-A time-series (deduced from CTs). As is evident, these values vary across videos. A qualitative analysis of the detected change-points reveals that they can be almost exactly mapped to emotional events in the videos. For example, Figure 6 shows the change-points for scary-1 (numbered and marked with a green vertical line in the Figure) and the segments created by these change-points (shaded pink/blue in the Figure). A manual evaluation of correspondence between the change-points (CP) and the emotional content of the scary-1 video reveals:

- *arousal CP#1 at 47secs*: gradual rise in arousal; the protagonist starts to realise that something is wrong.
- *valence CP#1 at 57s*: drop in valence; the ghost appears for the first time.
- *valence CP#2, arousal CP#2 at 89s*: drop in arousal, rise in valence; the ghost disappears from the frame.
- *valence CP#3, arousal CP#3 at 116s*: rise in arousal, drop in valence; the ghost unexpectedly re-appears.
- *arousal CP#4 at 122s*: gradual drop in arousal; the protagonist has successfully avoided the ghost.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TAFFC.2017.2772882, IEEE Transactions on Affective Computing

Segments A and B in the scary-1 time-series highlight the emotionally salient sectors for this video. Segment A is characterised by a drop in valence and an increase in arousal, while segment B is characterised by an opposite change in V-A levels.

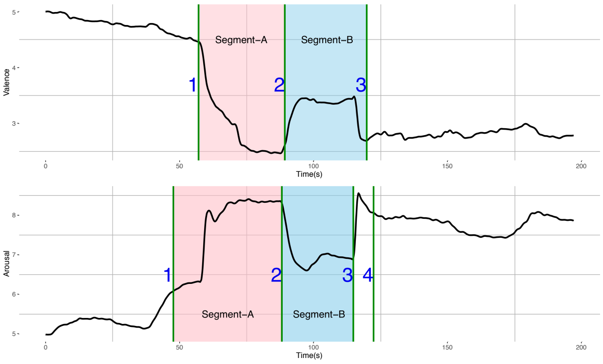

Fig. 6: Time-series (top: valence, bottom: arousal) with change-points and segments for Scary-1.

Discussion. From Table 3 it is evident that the number of change-points for any given video is usually not the same for valence and arousal time-series. This is expected, as the variance in V-A time-series is not the same for most videos. Also, the number of detected change-points is not directly proportional to the range of the time-series.For example, the range in valence for boring-2 is relatively small in comparison to scary-2, yet the number of change-points detected is significantly higher than for scary-2. With 11 change-points in valence, the results for boring-2 do not directly correspond to emotional events in the video, as slight changes in the time-series are also accounted as change-points. However, for videos with relatively large range in time-series, the change-points generally correspond to emotional events (e.g., amusing-1, scary-1&2, etc.). This pattern is evident in the qualitative analysis of scary-1 video, where 3 change-points (1, 2, 3 for both valence and arousal time-series) temporally correspond to each other and also map to events in the video.

3.5 Classification of Ratings

Lastly, in order to assess the usability of CA as ground truth given the visual stimuli (i.e., the videos), we tried to classify features extracted from CA itself, according to the emotional content of each video, irrespective of the subjects. This approach is based on the premise, that reasonably low classification error rates indicate coherence and replicability in the participants' annotation behaviour. For this analysis, first all CA were segmented into non-overlapping *chunks* of 5 seconds each and the mean V-A values for each chunk were extracted. Then, these chunks were labelled according to the emotional label for the given video. For example, amusing-1&2 were labelled as *amusing*. Lastly, these chunks were split into *training* and *test* sets. The training set consisted of labelled data from 25 randomly chosen participants (i.e., 6275 chunks) and the test set consisted of the data from the remaining 5 participants (i.e., 2874 chunks).

An initial visual analysis of the distribution of the chunks in the V-A plane revealed that the problem was highly non-linearly-separable. Hence, a Support Vector Machine Classifier (SVC), with a Radial Basis Function (RBF) kernel, was used for classification [36]. SVCs are often used in the machine learning community because of their efficacy in addressing non-linearly-separable problems. The optimal hyper parameters of the SVC ($\mathcal{C}$ and γ) were found via grid search [36], within ranges of $\gamma = 10^{-2,\dots,2}$ and $\mathcal{C} = 10^{0,\dots,2}$, and in steps of 0.25. The resulting optimal model (with $\mathcal{C} = 3.16$ and $\gamma = 0.56$) had 4473 support vectors, evenly distributed among the four classes (from 956 for scary to 1264 for relaxed videos). Most of the support vectors were concentrated around the center of the V-A plane i.e., the return point induced by the spring-based joystick. The overall classification rate obtained was 51.76% (see Table 4).

As a further, more refined investigation, we then identified chunks which corresponded to emotionally salient change-points (see 3.4). First, we used CPA [34] to identify such change-point chunks, which resulted in 871 and 167 chunks for the training and test sets, respectively. Then, another SVC was trained and tested. The optimal model (with $\mathcal{C} = 56.23$ and $\gamma = 1.78$) had 718 support vectors and the classification rate obtained by this model was 60.22% (see Table 4).

TABLE 4: Classification rates [%] per emotional video type, for all-chunks and change-point chunks.

	Amusing	Boring	Relaxed	Scary	Overall
All-chunks	74.65	32.36	13.45	86.57	51.76
CP chunks	80.77	38.46	30.00	91.67	60.22

Discussion. Table 4 shows that even though the overall classification rates are above the 25% chance level, they are still far from the optimal value. The high concentration of support vectors around the center of the UI makes the classification problem particularly hard. The single-class classification rates vary, with correct classification of chunks for amusing and scary videos being easier than that for chunks of boring and relaxed videos. This result can be attributed to the similarity in both, the V-A attributes (see Table 1) and the participants' annotations (see Sections 3.2 and 3.3), of boring and relaxed states. The table also shows that classification rates for change-point chunks are always better than their corresponding all-chunks rates. The reason for the same being that unlike the all-chunks scenario, most change-point classification support vectors are not located around the center of the UI. Even though the overall classification rates range from 50%–60%, the high classification rates for amusing and scary videos indicate the existence of cohesive ratting patterns among the participants.

4 Conclusion

Emotional memory is short-termed [2]; thus, it is highly desirable that the human participants report/annotate their affect state as soon as possible, ideally continuously and *while* the emotional stimuli are presented. However, this annotation exercise must be as unobtrusive as possible, such that it doesn't influence the emotions of the participant annotating her emotional experience.

The joystick-based annotation framework presented in this work goes in this direction. The use of a joystick is especially advantageous as: (i) it allows for continuous and simultaneous acquisition of V-A annotations, (ii) it helps mitigate the cognitive load of the annotation procedure by providing *proprioceptive feedback* to the annotator; and (iii) joysticks are generally more ergonomic than computer mice. Also, unlike other commonly used computer-mouse based frameworks, the annotator does not need to continuously press any buttons in our setup. Thus, it seems reasonable to claim that joystick-based CA framework offers clear advantages over mouse-based approaches. But, to explicitly address this question, a comparative study between joystick- and mouse-based approaches needs to be undertaken. Secondly, a comparative study between consequent (V-A individually) and simultaneous (V-A together) annotation strategies should also be undertaken. Nevertheless, given that the framework was generally well received (see Section 3.1) and that we received no negative comments on simultaneous V-A annotation, we can conclude that our framework is a viable alternative to other approaches mentioned in Section 1.

But, are these emotional experiences consistent with the intended V-A attributes of the stimuli? And if so, can the CA be used to detect the differences in emotional experiences evoked by these stimuli? The results presented in Sections 3.2 and 3.3 provide an affirmative answer to these questions. Particularly, the analyses of mean (see Figure 2) and continuous (see Figures 4 & 5) V-A ratings shows that the annotations are mostly consistent with the intended V-A attributes of the stimuli. Section 3.3 also presents a robust regression based approach for determining CTs from multiple subjective annotation trajectories. While this approach does not addresses the problem of inter-annotator and reaction delay, it is nonetheless useful in preventing highly diverging trajectories from unduly influencing the shape of CT.

Similarly, given their continuous nature, can CA be used to discern *when* important events occur in the stimuli? To address this question, we undertook CPA (see Section 3.4) on the CTs. Through this analysis, we showed that in certain cases, the change-points detected in CTs directly map onto the most salient moments in the video. For example, the change-points in scary-1's CT correspond to the most scariest moments in that video. Thus, through CPA on CA, emotionally salient events in the stimuli can be automatically identified and extracted for further analyses.

Lastly, to determine cohesiveness and replicability in participants' rating patterns, we used a SVC to classify the annotation and change-point chunks (see Section 3.5). Where, correct classification for scary and amusing stimuli was easier than for boring and relaxing stimuli. Thus, it was easier to discern cohesiveness and replicability in emotional experiences, where the participants reacted in line with intended V-A attributes of the stimuli. Nevertheless, the results here indicate that for a given stimuli the participants generally react in the same manner.

The presented approach indeed has some drawbacks. For example, the — albeit small — cognitive load imposed on the participants while annotating. Moreover, it has emerged during this study that the video stimuli must be carefully selected in order to elicit the desired affect state. Particularly, we noticed that the annotations for amusing-2 video from several participants had unexpectedly large sections in the negative valence quadrants i.e., it was rated as not so pleasant. Given its ironical sexual content, this we speculate, is the reason for the unexpected result. The future work should also investigate inter-annotator and reaction delays. CPA is a promising method for detecting salient events from CA and could also possibly be used for addressing delays in annotations.

Nonetheless, given the simplicity and high acceptance of the proposed framework, and the statistically coherent results obtained while using it, we plan to use it for several experimental/practical applications. For example, the annotations acquired from this framework can be used for the development and calibration of emotion detection systems that use, for example, physiological signals to predict a human's affect state. These systems can then be used in a multitude of scenarios (e.g., in cooperative robotics environments). We also plan to further develop the annotation framework and make it publicly available.

Acknowledgements

The authors would thank Jossin Antony for helping with the data acquisition during the experiment. We also want to thank Bernhard Weber for his help with the MANOVA analysis. Lastly, we thank all participants in our experiment.

References

[1] P. Rani, N. Sarkar, C. A. Smith, and L. D. Kirby, "Anxiety detecting robotic system - towards implicit human-robot collaboration," *Robotica*, vol. 22, pp. 85–95, 1 2004.

[2] 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*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2010, pp. 21–47.

[3] E. L. van den Broek, "Affective Signal Processing (ASP): Unraveling the mystery of emotions," Ph.D. dissertation, Human Media Interaction (HMI), Faculty of Electrical Engineering, Mathematics, and Computer Science, University of Twente, Enschede, The Netherlands, 2011.

[4] F. Nagel, R. Kopiez, O. Grewe, and E. Altenmueller, "Emujoy: Software for continuous measurement of perceived emotions in music," *Behavior Research Methods*, vol. 39, no. 2, pp. 283–290, 2007.

[5] M. Soleymani, S. Asghari Esfeden, Y. Fu, and M. Pantic, "Analysis of eeg signals and facial expressions for continuous emotion detection," *IEEE Transactions on Affective Computing*, 2015.

[6] M. E. Ayadi, M. S. Kamel, and F. Karray, "Survey on speech emotion recognition: Features, classification schemes, and databases," *Pattern Recognition*, vol. 44, no. 3, pp. 572 – 587, 2011.

[7] M. Soleymani, S. Koelstra, I. Patras, and T. Pun, "Continuous emotion detection in response to music videos," in *Face and Gesture 2011*, March 2011, pp. 803–808.

[8] J. A. Russell, "Core affect and the psychological construction of emotion," *Psychological review*, vol. 110, 2003.

[9] D. S. Messinger, T. D. Cassel, S. I. Acosta, Z. Ambadar, and J. F. Cohn, "Infant smiling dynamics and perceived positive emotion," *Journal of Nonverbal Behavior*, vol. 32, no. 3, p. 133, 2008.

[10] R. Cowie and M. Sawey, "Gtrace-general trace program from queens, belfast," 2011.

[11] J. Girard, "Carma: Software for continuous affect rating and media annotation," *Journal of Open Research Software*, vol. 2, no. 1, 2014.

[12] 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.

[13] J. M. Girard and A. G. C. Wright, "Darma: Software for dual axis rating and media annotation," *Behavior Research Methods*, Jun 2017.

[14] 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)*, May 2011, pp. 2376–2379.

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

[16] G. N. Yannakakis and H. P. Martnez, "Grounding truth via ordinal annotation," in *Affective Computing and Intelligent Interaction (ACII), 2015 International Conference on*, Sept 2015, pp. 574–580.

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

[18] H. Egermann, F. Nagel, E. Altenmüller, and R. Kopiez, "Continuous measurement of musically-induced emotion: A web experiment," *International Journal of Internet Science*, vol. 4, no. 1, pp. 4–20, 2009.

[19] M. A. Nicolaou, Y. Panagakis, S. Zafeiriou, and M. Pantic, "Robust canonical correlation analysis: Audio-visual fusion for learning continuous interest," in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2014, pp. 1522–1526.

[20] 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*, vol. 7, no. 4, p. 774, 2007.

[21] E. Dellandréa, L. Chen, Y. Baveye, M. V. Sjöberg, C. Chamaret *et al.*, "The mediaeval 2016 emotional impact of movies task," in *MediaEval 2016 Multimedia Benchmark Workshop Working Notes Proceedings of the MediaEval 2016 Workshop*, 2016.

[22] 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.

[23] J. J. Gross and R. W. Levenson, "Emotion elicitation using films," *Cognition & Emotion*, vol. 9, no. 1, pp. 87–108, 1995.

[24] J. Hewig, D. Hagemann, J. Seifert, M. Gollwitzer, E. Naumann, and D. Bartussek, "A revised film set for the induction of basic emotions," *Cognition and Emotion*, vol. 19, no. 7, pp. 1095–1109, 2005.

[25] E. E. Bartolini, "Eliciting emotion with film: Development of a stimulus set," Master's thesis, Wesleyan University, 2011.

[26] C. Elliott, V. Vijayakumar, W. Zink, and R. Hansen, "National instruments labview: A programming environment for laboratory automation and measurement," *JALA: Journal of the Association for Laboratory Automation*, vol. 12, no. 1, pp. 17–24, 2007.

[27] M. M. Bradley and P. J. Lang, "Measuring emotion: the self-assessment manikin and the semantic differential," *Journal of behavior therapy and experimental psychiatry*, vol. 25, no. 1, pp. 49–59, 1994.

[28] I. D. Loram, H. Gollee, M. Lakie, and P. J. Gawthrop, "Human control of an inverted pendulum: is continuous control necessary? is intermittent control effective? is intermittent control physiological?" *The Journal of physiology*, vol. 589, no. 2, pp. 307–324, 2011.

[29] J. Sauro and J. R. Lewis, "When designing usability questionnaires, does it hurt to be positive?" in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, ser. CHI '11. New York, NY, USA: ACM, 2011, pp. 2215–2224.

[30] S. Mariooryad and C. Busso, "Correcting time-continuous emotional labels by modeling the reaction lag of evaluators," *IEEE Transactions on Affective Computing*, vol. 6, no. 2, pp. 97–108, 2015.

[31] W. L. Martinez and A. R. Martinez, *Computational statistics handbook with MATLAB*, 3rd ed. CRC press, 2007, vol. 22.

[32] C. C. Aggarwal, *Data mining: the textbook*. Springer, 2015.

[33] A. Gabadinho, G. Ritschard, N. S. Mueller, and M. Studer, "Analyzing and visualizing state sequences in r with traminer," *Journal of Statistical Software*, vol. 40, no. 4, pp. 1–37, 2011, iD: unige:16809.

[34] D. S. Matteson and N. A. James, "A nonparametric approach for multiple change point analysis of multivariate data," *Journal of the American Statistical Association*, vol. 109, no. 505, pp. 334–345, 2014.

[35] N. A. James, A. Kejariwal, and D. S. Matteson, "Leveraging cloud data to mitigate user experience from breaking bad," *arXiv preprint arXiv:1411.7955*, 2014.

[36] C. J. C. Burges, "A tutorial on Support Vector Machines for pattern recognition," *Knowledge Discovery and Data Mining*, vol. 2, no. 2, 1998.

Karan Sharma received his M.Sc. in Mechatronics in 2009 from the University of Applied Sciences, Ravensburg-Weingarten, Germany. Since 2009, he is a researcher at the Institute of Robotics and Mechatronics, German Aerospace Center (DLR). His research at the DLR primarily focuses on human safety analysis in the DLR-Robotic Motion Simulator (DLR-RMS) platform. Currently, he is investigating the application of affective computing to simulation as well as robot-human cooperation scenarios.

Claudio Castellini, Ph.D. received a Laurea in Biomedical Engineerings in 1998 from the University of Genova, Italy and a Ph.D. in Artificial Intelligence in 2005 from the University of Edinburgh, Scotland. Since 2009 he is a researcher at the DLR (German Aerospace Center) in Oberpfaffenhofen, concentrating on human-machine interfaces for the disabled and assistive robotics. He is currently (co)author of some 75 papers appeared in international journals, books and peer-reviewed conferences.

Freek Stulp received his doctorate degree in Computer Science from the Technische Universität München in 2007. He was then awarded post-doctoral research fellowships to pursue his research at the Advanced Telecommunications Research Institute International (Kyoto) and the University of Southern California (Los Angeles). After being an assistant professor at the École Nationale Supérieure de Techniques Avancées (ENSTA-ParisTech) in Paris, he is now the head of the department of Cognitive Robotics at the Institute of Robotics and Mechatronics with the German Aerospace Center (DLR), Wessling, Germany. His research interests include autonomous robotics, motion primitives, continual robotic learning and semantic planning, with a focus on applications in robust manipulation and future manufacturing.

Egon L. van den Broek received a MSc in Artificial Intelligence (AI) (2001), a PhD in Social Sciences (2005), and a second PhD in Electrical Engineering, Mathematics, and Computer Science (2011). Currently, he is assistant professor and research director of the Center for Research on data-driven User eXperience (CRUX) at the Utrecht University, founding partner at Information eXperience (IX) BV, and consultant (e.g., for TNO, Philips, and the United Nations). His interests are on pattern recognition, interaction technology, and affective computing. Egon is Editor-in-Chief of Open Computer Science, Area Editor of Pattern Recognition Letters, Section Editor of Journal of Theoretical and Applied Computer Science (JTACS), and Associate Editor of Behaviour & Information Technology. Further, Egon serves as external expert for various agencies (e.g., European Commission), in conference program committees, on boards of advice, and on several journal editorial boards. He frequently serves as invited/keynote speaker, conference chair, and has received several awards, most recently the Journal of the Association for Information Science and Technology (JASIST) 2015 best paper award. Egon has published 160+ scientific articles, guided 80+ students, and has several patent applications.