Image and Vision Computing 65 (2017) 23–36

Contents lists available at ScienceDirect

Image and Vision Computing

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

AFEW-VA database for valence and arousal estimation in-the-wild ☆

Jean Kossaifi [a,*], Georgios Tzimiropoulos [b], Sinisa Todorovic [c], Maja Pantic [a,d]

[a] Department of Computing, Imperial College London, London SW7 2AZ, UK
[b] School of Computer Science, The University of Nottingham, Nottingham NG8 1BB, UK
[c] School of EECS, Oregon State University, Corvallis, OR 97331, USA
[d] EEMCS, University of Twente, The Netherlands

ARTICLE INFO

Article history:
Received 8 June 2016
Received in revised form 6 February 2017
Accepted 7 February 2017
Available online 12 February 2017

Keywords:
Continuous affect estimation in-the-wild
Dimensional affect recognition in-the-wild
Valence
Arousal
Facial expressions
Dimensional emotion modelling

ABSTRACT

Continuous dimensional models of human affect, such as those based on valence and arousal, have been shown to be more accurate in describing a broad range of spontaneous, everyday emotions than the more traditional models of discrete stereotypical emotion categories (e.g. happiness, surprise). However, most prior work on estimating valence and arousal considered only laboratory settings and acted data. It is unclear whether the findings of these studies also hold when the methodologies proposed in these works are tested on data collected in-the-wild. In this paper we investigate this. We propose a new dataset of highly accurate per-frame annotations of valence and arousal for 600 challenging video clips extracted from feature films (also used in part for the AFEW dataset). For each video clip, we further provide per-frame annotations of 68 facial landmarks. We subsequently evaluate a number of common baseline and state-of-the-art methods on both a commonly used laboratory recording dataset (*Semaine database*) and the newly proposed recording set (*AFEW-VA*). Our results show that geometric features perform well independently of the settings. However, as expected, methods that perform well on constrained data do not necessarily generalise to uncontrolled data and vice-versa.

1. Introduction

A lot of work has been done in the field of Facial Expression Recognition to detect Facial Action Units [1–3] or, directly, a discrete set of basic or non-basic emotions. However, recently, psychologists and researchers in computer vision often focus on continuous dimensional affect analysis such as the analysis of valence and arousal [4,5].

When we speak about "automatic continuous analysis of valence and arousal in-the-wild", we refer to the automatic estimation of:

1. Intensity of valence – how negative or positive the experience is, and
2. Intensity of arousal – how calming or exciting the experience is,

as shown in the target recordings of spontaneous facial behaviour acquired in unconstrained (uncontrolled) conditions.

This problem is challenging for a number of reasons [6]. In general, spontaneous facial expressions are characterized by subtle, minimal facial deformations which are difficult to track, and frequent out-of-plane head movements whose effects are difficult to remove. Additionally, similar intensity levels of valence (and arousal) may share a number of common changes in motion and morphology of facial features. These challenges require a *fine-grained* approach which would be capable of identifying the most relevant facial parts and their subtle movements to attain *continuous* dimensional facial affect estimation.

A literature review suggests that most of these challenges have not been satisfactorily addressed yet. This is mostly due to the fairly limited amount of relevant data. In particular, widely used datasets for evaluating approaches to valence and arousal estimation are all captured in laboratory, controlled settings, with a limited range of face poses and occlusion. Since state-of-the-art methods typically ground their valence and arousal estimation on such data, it remains unclear whether these methods still perform well in videos collected in-the-wild.

Moreover, widely used datasets in the field contain recordings of emotional reactions elicited by a rather limited number of tasks, or in a restrictive setting. This simplifies the problem and again, it remains unclear whether the methods tested on such data could perform well on data recorder in unconstrained conditions. In general, we observe that there is scant work in the field focusing on incorporating robustness mechanisms for addressing uncertainty and noise of real-world

☆ This paper has been recommended for acceptance by Mohammad Soleymani.
* Corresponding author.
E-mail address: jean.kossaifi@gmail.com (J. Kossaifi).

http://dx.doi.org/10.1016/j.imavis.2017.02.001
0262-8856/

settings. Consequently, we believe that current best performers on existing valence and arousal benchmark datasets may exhibit serious limitations on data collected in-the-wild.

To address this lack of data recorded in the wild, we present a new dataset – called *AFEW-VA* – of highly accurate per-frame annotations of valence and arousal for 600 challenging, real-world video clips extracted from feature films (also used in part for the AFEW database [7,8]). Added to these are per-frame annotations of 68 facial landmarks. The dataset has been made publicly available.[1]

Other contributions of the paper are:

- We introduce a novel annotation scheme that addresses the problems exhibited by existing annotation tools, namely delays between the annotation and the video, laps in concentration, inaccuracy of the annotations due to the sensitivity of the joystick or slider and inability to annotate remotely or online.
- We provide baseline results on the newly introduced database and compare the performance of various features.
- We compare baseline and state-of-the-art methods for performance on both a standard controlled database (Semaine [9]) and the new database (*AFEW-VA*).

The results show that methods performing well in controlled environments do not necessarily perform as well in unconstrained conditions. They also demonstrate the descriptive power of geometric features.

2. Related work for continuous valence and arousal estimation

While there is a lot of work on audio-based valence and arousal estimation [4], we focus here on video-based and audio/video-based valence and arousal estimation given that the majority of published methods address this problem. Both valence and arousal are defined as continuous emotional dimensions. Therefore, it seems suitable to study them directly in the continuous domain. Even though much of the early work considered coarse levels of valence and arousal (e.g. positive vs negative), and posed the problem as one of classification [10], more recent work casts the problem in the continuous domain [4,6,11].

Very important progress in the field came with the introduction of the AVEC challenges, in 2011. It started with a subset of the Semaine dataset [40], it originally formulated the problem as one of classification, using binarized values (± 1), before moving to continuous annotations in 2012, still on the Semaine data [41]. The best results obtained that year were an average Pearson Correlation Coefficient (PCC) of 0.456 [15]. The 2013 and 2014 editions of the challenge used the audio-visual depression language corpus [42,43]. The best results obtained on that corpus were lower than in the first year with a PCC of 0.1409 in 2013 [22] but improved to 0.5946 for the best performer in 2014 [23]. Note however that the latter performance was obtained using metadata and a template matching based method (i.e. prototypical valence and arousal episodes) rather than a classical regression based method using audio-video features. From the 2015 edition of the AVEC challenge, the RECOLA dataset was used [44]. RECOLA is a challenging dataset and yet the results obtained on this data were good. For instance, the best performer in the 2015 edition of the challenge obtained an average PCC of 0.685 [33], while the best performer in 2016 obtained an average PCC of 0.731 [34]. Note that the latter results were obtained with a fusion of audio, video and physiological features. Using only video, the average PCC reported is of 0.482. For a detailed summary of methods for valence and arousal estimation, the interested reader is referred to the recent surveys [4,6]. In what follows, we explain the most common approaches to automatic, continuous estimation of valence and arousal, summarised in Table 1.

There are several approaches to valence and arousal estimation, the most obvious of which is static regression, such as linear regression [25,37], partial least squares regression [22], or Support Vector Machine for Regression (SVR) [12,21,38], which is also frequently used as a baseline method [11,41-43]. In [21], SVR is combined with canonical correlation analysis (CCA) to iteratively fuse predictions. SVR is also employed in the work of [23] but its use differs in that template trajectories for each emotional dimension are first built and then matched to a new testing sample using metadata as features. Meta-data is also used in [16] where audio, video and contextual (meta-data) features are combined in a multimodal fuzzy inference system. More powerful kernel based methods can be used such as Nadaraya-Watson kernel regression [15] or Doubly Sparse Relevance Vector Machine [3] that can impose sparsity on both the kernels and the training samples.

When dealing with several features, e.g. geometric and appearance features, or multiple modalities, e.g. audio and video, the question of how to fuse them arises. There are four main types of fusions: feature level (early fusion), decision level (late fusion), model-level and output-associative fusion [10,45]. In the early-fusion case, the features are combined most commonly by simply concatenating them and using the output for estimation (e.g. as is done in [30] for predicting valence). However, in general, this approach tends to create very high dimensional feature vectors and lead to overfit.

Late fusion is the process of first generating separate estimations from each modality before combining – fusing – them into one final estimation. This fusion can be achieved in numerous ways, from simple mapping, e.g. averaging, as in [29], to more complex methods such as linear and multi-linear regression [15,25,26,28,30,37,38], SVR [39], random forests [32] or Kalman filters based [34–36]. This type of late fusion is akin to stacking, a classical ensemble learning technique.

Ensemble methods such as Random Forests or boosting are indeed frequently chosen for valence and arousal estimation because of their robustness. Random Forests are used in [36] to obtain single-cue predictions that are combined with linear regression. [28] combines boosted Regression Trees and Linear Regression with a special regularization to account for temporal correlation. In [19], a random-forest based method is used to jointly recover 3D facial landmarks and the continuous emotional labels. In [29], random forests and gradient boosting are used and their results refined with stochastic gradient descent. Finally, a CCA-based ensemble method is proposed in [24], inspired by the earlier work on Correlated-Spaces Regression of [17].

Model-level fusion consists of letting the model combine the features. An example is the work of [3] where the model jointly selects the most relevant training example and kernels. Correlated Spaces Regression [17] performs model-level fusion by finding a latent correlated subspace between all the features and the labels. A related approach is the work by Panagakis et al. [5] (this is a robust method, tested for interest intensity estimation in the wild, an emotional dimension closely related to arousal). In [12], the different types of fusion are compared using LSTM based models. Model-level fusion is found to largely outperform feature level fusion but underperform compared to output-associative fusion.

Output-associative fusion is a hybrid fusion method that has the potential to leverage both the advantage of early fusion and the self-dependency of the targets. It was introduced for continuous valence and arousal estimation in the seminal work of Nicolaou et al., first with Output-Associative Bidirectional Long Short Term Neural Networks (OA-BLSTMs) [12], then OA-Relevance Vector Machines (OA-RVMs) [13,14]. The same approach was recently used in the AVEC challenge by the work [27], which combines early, output-associative and late fusion.

[1] http://ibug.doc.ic.ac.uk/resources/afew-va-database/.

Table 1
Recently published methods for continuous valence and arousal estimation on existing databases. Modalities are indicated as A: audio, V: video, P: physiological and M: meta-data. Methods are abbreviated as LR: linear regression, RF: random forest, SVR: support vector machines for regression, N-W: Nadaraya-Watson, CSR: Correlated-Spaces Regression, OA: output associative, RVM: relevance vector machines, PLS: Partial Least Squares, CCRF: Continuous Conditional Random Field, RNN: recurrent neural network, (B)LSTM: (bidirectional) long short term memory, DNN: deep neural networks, SGD: stochastic gradient descent. For each work, the best results in term of Pearson's product-moment correlation coefficient (PCC) on the testing set is reported when available. *: only the average result is reported.

Recent methods for continuous valence and arousal estimation				Results (PCC)	
Paper	Model	Modalities	Database	Valence	Arousal
[12]	OA BLSTM-NN	AV	Semaine subset	0.796	0.642
[13]	OA-RVM	V	Semaine subset	–	–
[14]	OA-RVM	AV	Semaine subset	–	–
[15]	N-W kernel regression	AV	Semaine (AVEC'12)	0.341	0.612
[16]	Fuzzy inference system	AVM	Semaine (AVEC'12)	0.42	0.42
[17]	CSR	AV	Semaine subset	0.21	0.46
[18]	SVR and CCRF	V	Semaine (AVEC'12)	0.343	0.341
[19]	RF	V	Semaine (AVEC'12)	0.454	0.564
[3]	Doubly sparse RVM	V	Semaine (AVEC'12)	0.31	0.31
[20]	Time-delay NN	V	Semaine (AVEC'12)	0.308	0.444
[20]	Time-delay NN	V	AVEC'13	0.127	0.155
[21]	SVR	AV	AVEC'13	0.135	0.132
[22]	PLS regression	AV	AVEC'13	0.141*	
[23]	SVR	AVM	AVEC'14	0.587	0.633
[24]	CCA	V	AVEC'14	0.381	0.391
[25]	LR	AV	AVEC'14	0.493	0.620
[26]	Deep belief network	AV	AVEC'14	0.528	0.58
[27]	OA RVM	AVP	AVEC'15	0.588	0.740
[28]	LR + boosted regression trees	AV	AVEC'15	0.501	0644
[29]	RF + gradient boosting + SGD	AVP	AVEC'15	0.490	0.687
[30]	RNN	AVP	AVEC'15	0.590	0.746
[31]	LSTM-RNN	AVP	AVEC'15	0.627	0.781
[32]	DNN	AVP	AVEC'15	–	–
[33]	Deep BLSTM-RNN	AVP	AVEC'15	0.616	0.753
[34]	LSTM + kalman filter	AVP	AVEC'16	0.689	0.774
[35]	SVR + kalman filter	AVP	AVEC'16	–	–
[36]	RF + LR	AVP	AVEC'16	0.634	0.776
[37]	LR	AVP	AVEC'16	–	–
[38]	SVR + LR	AVP	AVEC'16	–	–
[39]	BLSTM-RNNs	AVP	RECOLA	–	–

Alternatively, since valence and arousal are in general slow-varying signals, methods accounting for short-term temporal correlations have been applied. Continuous Conditional Random Fields (CCRF) are used in [18] on top of SVRs trained separately for each modality. Various types of neural networks have also been used, including Time-Delay Neural Networks [20], Recurrent Neural Networks (RNN) [30] and Long-Short Term Memory RNN (LSTM-RNN) in [12,31,39].

Recent successes of deep models have lead to a comeback of neural network based approaches; however even though deep methods are increasingly applied to various vision-based and audio-visual problems, there are scant applications dedicated to continuous valence and arousal estimation. In particular, no end-to-end method has been used to predict valence and arousal directly from videos. In [32], single-cue predictions are obtained with Deep Neural Networks (DNN) but combined using late fusion with a linear regression. Similarly, in [26], deep belief networks are trained on different features and the results are combined using linear regression. Deep bidirectional LSTM recurrent neural networks are used to perform multi-modal prediction using audio, video and physiological modalities [33].

The ground-truth annotations in our dataset, in the range from −10 to 10, are sufficiently fine to allow for continuous-based estimation of valence and arousal levels. Therefore, our dataset can be used for evaluating and ranking the above approaches.

3. Existing datasets for valence and arousal estimation

The problem of automated valence and arousal estimation from video has been introduced relatively recently in computer vision and several databases have been collected for that purpose, summarised in Table 2. The data collected is annotated by human raters either continuously using a FEELTRACE-like tool [46] or discretely using a Self-Assessment Manikin (SAM) [47].

However, these datasets all focus on laboratory or controlled environments. The Humaine project[2] [48] aimed at making data available for affect estimation and the Humaine database includes 23 recordings that were annotated with a FEELTRACE-like software.

The Sustained Emotionally coloured Machine-human Interaction using Nonverbal Expression Dataset (SEMAINE) [9] is a richly annotated corpus containing recordings of interactions between a human and a machine-like agent in a Sensitive Active Listener (SAL) scenario [49]. In particular the Solid SAL scenario, where the operator's role is played by a human instructed to act in the character of a SAL agent, contains 21 sessions (75 character interactions) that were annotated with FEELTRACE for valence and arousal.

In the Belfast naturalistic dataset,[3] 209 clips from TV recordings and 30 from interviews were annotated using FEELTRACE for valence and arousal. Categorical labels were also provided for core emotions. The Belfast induced dataset [50] contains laboratory recordings of responses to various emotion induction tasks. This set contains exclusively subjects from Northern Ireland who self-reported their emotional state using a questionnaire. 37 of the recordings were also annotated with FEELTRACE for emotional Intensity and Valence by multiple annotators (6 to 258).

The Multi-Modal Affective Database for Affect Recognition and Implicit Tagging (MAHNOB-HCI) [51] contains audio-visual and EEG

[2] http://emotion-research.net/toolbox/bydate/toolboxdatabase.2006-09-26.5667892524.
[3] http://sspnet.eu/2010/02/belfast-naturalistic/.

Table 2
Available data available for facial valence and arousal estimation from videos.

Database	Subjects demography	Annotation type for valence and arousal	Amount of data	Elicitation method	Environment	Illumination
SAL [46]	4 UK subjects: 50% females	Continuous, Feeltrace, 4 raters	23 recordings, 11 hours	induced (SAL)	controlled	controlled
Semaine [7]	20 participants	Continuous, Feel-trace, 6 to 8 raters	06:30 hours	induced (SAL)	controlled	controlled
Belfast naturalistic	125 subjects, 31 males, 94 females	Continuous, Feeltrace, 7 raters	298 clips of 10 to 60 s.	TV / interviews	controlled	controlled
Belfast induced [48] (set 1)	44 females + 70 males from Northern Ireland	Continuous, Feeltrace, 6 to 258 raters	37 videos of 5 to 60 s.	laboratory-based test	controlled	controlled
Belfast induced [48] (set 2)	48 females + 42 males from Northern Ireland	Valence only, Continuous, Feeltrace, 1 rater	650 videos, 458mn	laboratory-based test	controlled	controlled
VAM-faces [51]	20 German speakers of the talk-show	5-points SAM, Linkert-like scale (5 points from -1 to 1), 7 to 8 raters	1867 images	talk-show	controlled	controlled
MAHNOB-HCI [49]	27 participants, 11 males, 16 females	Discrete, per video, self-report using SAM, Linkert scale	20 videos of 34.9 to 117 s.	video clips	controlled	controlled
DEAP [50]	32 participants, 50% females, age 19-37	Discrete, per video, self-report using SAM, Linkert scale	40 videos of 1 mn	music (+clip when available)	controlled	controlled
AMFED[53]	140 males, 102 females	Binary self reports	242 videos, 168.359 frames	videos	webcam	Indoor In-the-Wild
Recola [54]	27 French-speaking participants	Continuous, Feeltrace, 7 raters	clips of 5 mn for all 27 subjects	online interactions	controlled	controlled
AVEC'13 [40]	292 subjects, age 18-63	Continuous, Feeltrace, one naive raters per dimension	340 videos, 240 hours	human-computer interaction	webcam	Indoor In-the-Wild
AVEC'14 [41]	84 subjects, age 18-63	Continuous, Feeltrace, 3+ naive raters	300 videos of 6s to 4mn 8s.	human-computer interaction	webcam	Indoor In-the-Wild
Mimicry [55]	48 participants & 12 confederates, 26% female, 95% S.Europe	Continuous, Feel-Trace, ≈ 5 raters	11 hours of recording	naturalistic interactions - discussion	controlled	controlled
AFEW-VA (this work)	age 8-76, 240 subjects, 52% female	Per frame annotations in [-10, 10], 2 expert raters	600 video clips	movie actors	Indoor & Outdoor In-the-Wild	Indoor & Outdoor In-the-Wild

recordings of participants who also self-reported their valence and arousal levels using a SAM self-report Manikin.

Similarly, in the DEAP dataset [52], EEG and video of the participants were recorded while they watched 40 video clips of 1 min each. They also self-assessed for arousal, valence, liking and dominance again with a SAM manikin.

The VAM-Faces dataset [53] is a collection of recordings of the German talk-show "Vera Am Mittag". A total of 1867 images were annotated in a discretized 5-point scale for valence, activation and dominance by 8 to 34 annotators. Sparse annotations are also provided for six basic emotion categories (*happiness, anger, sadness, disgust, fear, surprise*) + neutral.

In the MAHNOB Laughter [54] and the AMFED dataset [55], participants were recorded with a webcam while watching videos. While the MAHNOB Laughter database was annotated for the type of laughter only (acted, spontaneous, speech-laughter), the participants recorded for the AMFED database were asked to provide binary self-assessment for familiarity with, liking of, and desire to watch the stimuli videos again.[4] In the AMFED database, 168,359 frames were also labelled for 10 symmetrical FACS-AU, 4 asymetrical FACS-AU, 2 head movements, smile and expressiveness.

The RECOLA dataset [56] had 46 participants, all from the department of psychology of the Université de Fribourg-Universitat, Switzerland. The participants, all French-speaking (20 French, 5 Italian and 2 German), were asked to perform specific computer interaction tasks under a constant lighting while being recorded with a webcam. Biosignal (EDA and ECG) were also recorded. Self-reported labels for Valence and Arousal were obtained using a SAM Manikin for integer values from 1 to 9. 27 videos of 5 min[5] were released and annotated using a FEELTRACE-like tool[6] by 6 French-speaking raters (3 females and 3 males). To obtain smoothness, the annotations obtained using the FEELTRACE-like tool were post-processed using a piecewise cubic interpolation.

The MAHNOB Mimicry Dataset [57] contains 11 h of recording of 12 confederates in naturalistic interaction - discussion with their 48 counterparts. Most of the videos were annotated by 5+ annotators in terms of valence and arousal using a custom online tool.

Audio-Visual Emotion recognition Challenges (AVEC) were organised to benchmark valence and arousal estimation methods. AVEC'11 [40] uses the Semaine Solid-SAL dataset but used binarized (±1) values while the next competition, AVEC'12 [41] uses the original continuous annotations. AVEC'13 [42] uses a subset of the audio-visual depression language corpus that has been annotated for continuous valence and arousal values by one naive annotator per dimension for each video. AVEC'14 [43] uses a subset of the data used for the previous version of the challenge. All clips were re-annotated by at least 3 (naive) annotators for valence, arousal and dominance. The most recent versions of the challenge – AVEC '15 [44] and AVEC'16 [11] – departed from the previous challenges and used the RECOLA database [56] along with the provided annotations. For a review of older datasets for affect recognition, please refer to [10].

[4] (Did you like it? Have you seen it before? Would you watch it again?).

[5] Videos from 27 participants over the 46 participants (16 females and 19 males).

[6] Annotated using the ANNEMO framework https://diuf.unifr.ch/diva/recola/annemo.html.

These existing datasets all exhibit one or more of the following limitations. (a) The videos are recorded under controlled conditions, e.g. illumination is uniform, background is static, and there is a limited amount of head pose variation and occlusion. (b) Although a range of affective states are displayed and recorded, emotions are elicited by a limited number of tasks, e.g., in [50], all subjects underwent exactly the same tasks. (c) Although there is usually a sufficient number of annotators per video, annotations are obtained using trace-style tools [46] which have been shown to produce low correlation for both intra and inter-rater valence and arousal annotations [43,50,58],[7] or only sparse annotations are provided using SAM.

To address the aforementioned limitations, we compile a new dataset by providing ground-truth valence and arousal annotations for video clips extracted from feature films (also used in part for the AFEW'14 challenge [7,8]). The new dataset – which we call AFEW-VA – contains 600 video clips selected from movies, many of which have been recorded in notably challenging conditions. Although emotions are acted in the movies, the actors live in their roles and realistically portray relevant behaviours and emotions, so the clips can be regarded as showing (nearly) spontaneous human emotions. This is in agreement with the study of believability of portrayal of emotions by professional actors that has shown that lay judges rate these emotional portrayal highly believable [59]. Unlike existing benchmark datasets, the AFEW-VA presented here contains a wide spectrum of facial expressions, elicited in various conditions with natural head pose movements, complex backgrounds, and under severe occlusions. The actors whose video clips are included in AFEW-VA show a large diversity in age and ethnicity.

4. Online tool for per-frame annotation of valence and arousal

Existing tools are based on the FEELTRACE tool [46] and allow users to annotate videos in the valence and arousal quadrant in real time while watching it, using a joystick.

This presents several drawbacks including:

1. Delays between the annotation and the video
2. Lapse in concentration
3. Inaccuracy of the annotation due to the sensitivity of the joystick or slider
4. Very few of these tool allow online remote-annotations.

To address these drawbacks, we developed for the database an online annotation tool (a screenshot of which is shown Fig. 1) that allows several people to annotate video clips per-frame, for valence and arousal, remotely. The first three drawbacks are addressed by the *per-frame* annotation and the ability to go back and forth in the sequence to correct labels, resulting in highly accurate annotations. The second and third drawbacks are further addressed by the ability to mark the annotations in different states (e.g. as to be checked) and the ability to add comments. The last drawback is inherently addressed since the developed tool is online-based. A random clip is presented to the annotator that can then watch the clip frame by frame. When familiar with the clip, the annotator can then go back and forth and annotate it frame by frame either using the sliders or using the keyboard, avoiding inaccuracy. When using the keyboard the application has been programmed to allow first annotation of the whole video for valence then for arousal. Finally the annotations can be saved, marked as to be checked again, in case of doubt or lapse in concentration, or marked to be done again completely. Extensive documentation can be found in the webpage of the project that contains instructions and a wiki.

The annotation tool is made publicly available.[8] It is written in Python as a Flask application backed with a MongoDB database. Easy to deploy and use, it allows any number of users to annotate easily and precisely video frames for valence and arousal intensity. In addition it can be easily extended to handle more annotations (e.g. discrete emotions).

5. AFEW-VA: the new dataset in-the-wild

5.1. Data

AFEW-VA consists of 600 videos extracted from features films (also used in part for the AFEW dataset [7]). The videos range from short (around 10 frames) to longer clips (more than 120 frames), and display various facial expressions. They are captured under challenging indoor and outdoor conditions such as complex cluttered backgrounds, poor illumination, large out-of-plane head rotations, variations in scale, and occlusions.

In total, we annotated more than 30,000 frames with *per frame* levels valence and arousal intensities in the range of −10 to 10. Fig. 2 shows the distribution of the values of valence and arousal present in our dataset. It matches the expected distribution and, as can be seen, there is a wide range of values for both valence and arousal. In some videos, we observe a significant signal change in valence and arousal across the frames. In some other videos, the temporal change of valence and arousal is negligible. Fig. 3 shows the distribution of the annotations in the valence and arousal circle. As can be observed, our data show large variations in valence and arousal values and complement well existing databases. For comparison, the distribution of valence and arousal values of three widely used databases, namely AVEC'14 [43], SEMAINE [9] and RECOLA [56] are presented in Fig. 4. Our database presents a large variation in the values of valence and arousal while extreme values are less frequent in the other databases.

5.2. Annotations

We provide *per frame* annotations of valence and arousal levels for all frames within all videos clips of our dataset. Unlike most other datasets, our annotations are not produced using trace-style tools, but are attained per frame by two expert annotators, a male and a female, FACS AU coding certified. Both annotators have annotated all videos together, therefore discussing all disagreements and coming up with a unique solution. Hence, AFEW-VA annotations are detailed and highly accurate. The range of annotation levels for both valence and arousal is from −10 to 10, resulting in a total of 21 levels. Fig. 5 shows the temporal evolution of the valence and arousal signals for an example video from our dataset, along with some representative frames. To assess rater reliability, 10% of the clips were randomly selected and re-rated by the same two annotators to compute intra-rating consistency. The attained Pearson product-moment correlation coefficient was 0.87.

We provide accurate location for 68 landmarks (see Fig. 6) including both interior and boundary points of the face. The annotation of these facial landmarks was done in a semi-automatic way. First, the face was detected in the first frame of each video sequence using the tree-based deformable part model (DPM) of [60]. Although the tree-based DPM additionally provides the location of facial landmarks, we use this algorithm only as a face detector, since the accuracy of landmark localization of this approach on our challenging videos is not satisfactory. The bounding box of the detected face was subsequently used to initialize the Gauss-Newton generative part-based AAM model of [61] for facial landmarks localization. The landmarks

[7] Inter-rater correlation reported in [43] is as low as 0.4.

[8] http://ibug.doc.ic.ac.uk/resources/valencearousal-online-annotation-tool/.

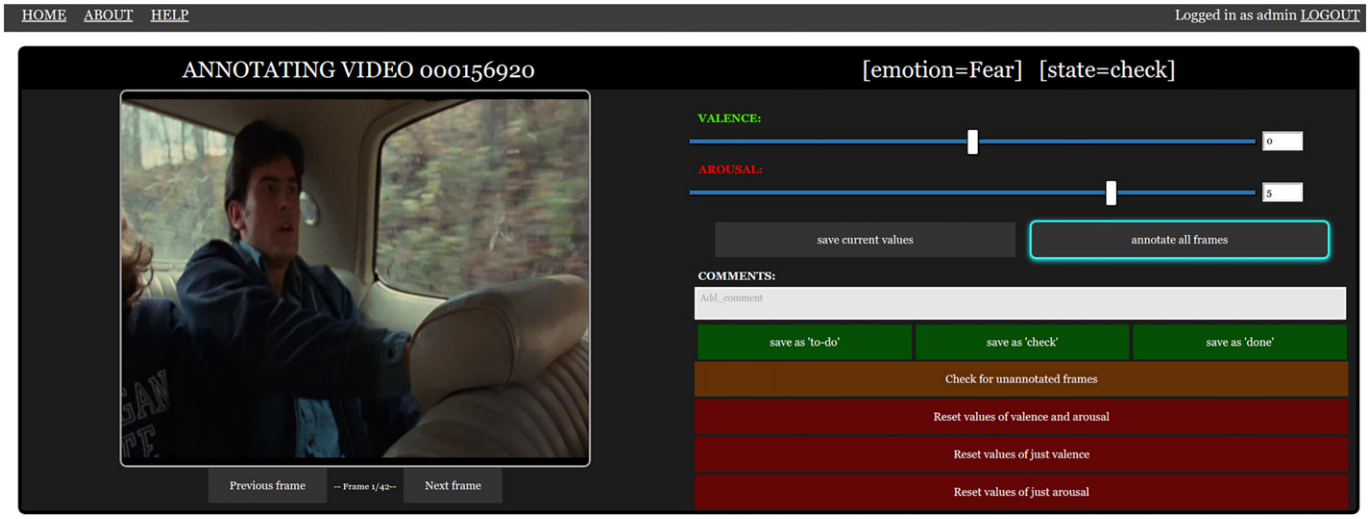

Fig. 1. Screenshot of the annotation tool developed and used to annotate the AFEW-VA dataset.

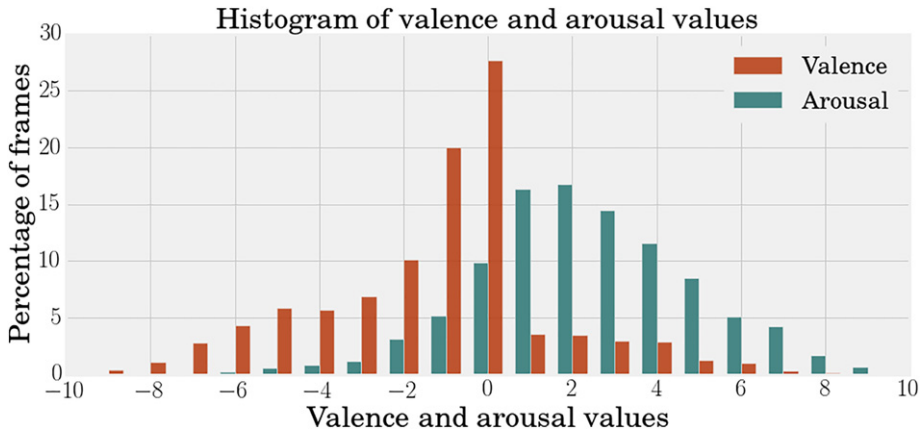

Fig. 2. Distribution of the valence and arousal values in our annotated dataset.

detected in each frame were then used to initialize landmark locations in the next frame. In this way, the landmarks were tracked across the video. Based on the common assumption of relatively smooth and slow varying trajectories of the landmarks, our tracker was capable of automatically estimating moments when the tracking went off. In such cases, it was re-initialized using the tree-based DPM for face detection. The tracking results were visually inspected and the landmarks were manually corrected when minor errors were present. Only satisfactorily tracked clips were kept. The landmark tracking results are publicly released together with the AFEW-VA database.[9]

Examples of tracked faces from AFEW-VA are shown in Fig. 6.

6. Performance measures

Given a ground-truth and a prediction, performance is measured using the root-mean-square error (*RMSE*), the Pearson product-moment correlation coefficient (*CORR*) and the intra-class correlation coefficient (*ICC*).

Let θ be a series of n ground-truth labels, $n \in \mathcal{N}$ and $\hat{\theta}$ a series of n corresponding prediction labels.

The RMSE is then defined as

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(\hat{\theta}(i) - \theta(i)\right)^2} \tag{1}$$

The correlation coefficient is

$$\mathrm{COR}(\hat{\theta}, \theta) = \frac{\mathrm{COV}(\hat{\theta}, \theta)}{\sigma_{\hat{\theta}}\sigma_{\theta}} = \frac{E\left[(\hat{\theta} - \mu_{\hat{\theta}})(\theta - \mu_{\theta})\right]}{\sigma_{\hat{\theta}}\sigma_{\theta}} \tag{2}$$

Both RMSE and CORR are standard measures, widely used for measuring valence and arousal estimation accuracy [4].

Most recently the intra-class correlation coefficient ICC(3, 1) [62] has been used for facial expression [63] and pain [3] estimation. In this work, all mentions of ICC refer to ICC(3, 1). For two samples (here θ and $\hat{\theta}$), the ICC is defined as:

$$\mathrm{ICC}(\hat{\theta}, \theta) = \frac{2 \times \mathrm{COV}(\hat{\theta}, \theta)}{\sigma_{\hat{\theta}}^2 + \sigma_{\theta}^2} = \frac{2 \times E\left[(\hat{\theta} - \mu_{\hat{\theta}})(\theta - \mu_{\theta})\right]}{\sigma_{\hat{\theta}}^2 + \sigma_{\theta}^2} \tag{3}$$

In all experiments, we report performance in term of all three metrics, RMSE, CORR and ICC.

[9] http://ibug.doc.ic.ac.uk/resources/afew-va-database/

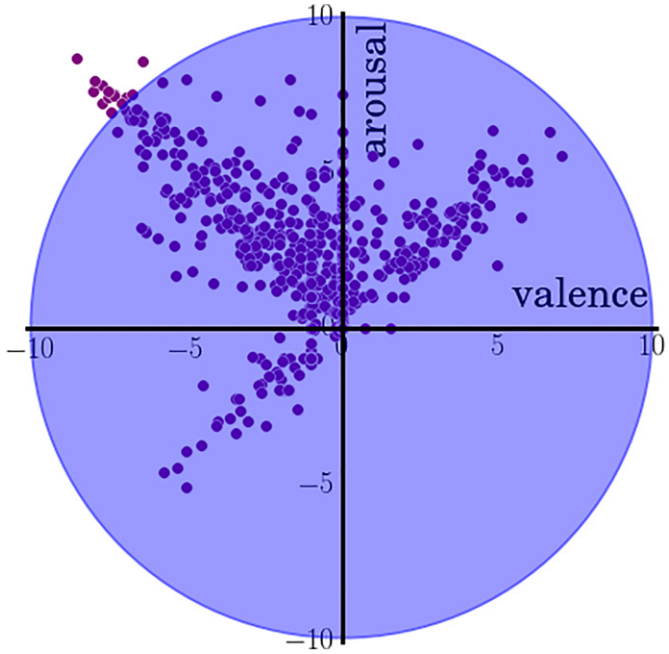

Fig. 3. Distribution of the valence and arousal values in AFEW-VA.

7. Facial features

7.1. Appearance features

We compare several robust appearance features including the same SIFT [64] features used during landmark localisation [61]. We also experiment with local binary patterns [65], which are widely used for facial affect analysis [66–68]. In this work we use uniform local binary patterns with a neighbour set of 8 on a circle of radius 2 $\left(LBP^{u}_{8,2}\right)$, binned in a N-dimensional histogram. This configuration has been shown to give good results for pain analysis [66] and facial Action Unit detection [67]. We also use a 2-dimensional Discrete Cosine Transform [69], where we kept only the N first coefficients following the jpeg zigzag-scheme [70]. It was shown to produce very good results for pain estimation from faces [66].

Hybrid-SIFT Owing to their success for landmarks detection, we use the same SIFT features [64] used during the landmark detection process [71,72]. These are dense features, extracted at the detected landmark locations and extracted in a canonical coordinate frame (Fig. 7c): there the images are normalized according to the similarity transform estimated from the detected landmarks. Using the position of the landmark allows us to incorporate geometric information into the descriptor, hence the name Hybrid. These features are the same as those used for the landmark detection process, therefore allowing for a unified, elegant framework for the entire processing of the videos.

We consider two variants of the LBP features:

Block-LBP The first configuration is coined *Block-LBP*. For a given image, the face is first aligned onto a base shape (the mean shape of the dataset) using a piecewise affine warping from the triangulated mesh of the 68 facial landmarks of the image to the triangulated mesh of the base shape. The resulting aligned image is then divided into a 10 × 10 grid of non-overlapping patches of size 18 × 14 pixels. One 59-D histogram is consequently extracted from each patch. Finally the image is represented by the 5900-D concatenation of all 100 histograms.

Hybrid-LBP encodes some geometric information by using the location of the landmarks. Instead of using a piecewise affine warping to align the faces, a simple translational model is used to scale and translate all the images onto a common coordinate frame. Patches of size 27×27 are then extracted around each landmark. Similarly to the block-LBP, each patch is represented by a 59-D histogram resulting into a global descriptor of size 4012.

We also consider several variants of the DCT features:

Holistic-DCT This descriptor is obtained by aligning the images onto a base shape using piecewise affine warping and extracting DCT features over the whole warped face. The first 500 coefficients are then kept, following the zigzag scheme [70]. This descriptor has been shown to produce good results for facial affect estimation [66].

Block-DCT As in the LBP case, images are normalised using piecewise affine warping and the resulting aligned image is divided into a 10 × 10 grid of non-overlapping patches of size 18 × 14 pixels. The first 59 coefficients are kept following the zigzag scheme, resulting in a global descriptor of size 5900.

Hybrid-DCT Coined *Hybrid-DCT*, this descriptor is similar to *Hybrid-LBP* and incorporates geometric information by extracting the descriptor around the landmarks positions. After scaling and translating all the images onto a common coordinate frame, patches of size 27 × 27 are extracted around each landmark. Each patch is represented by the first 59 coefficient kept in a zigzag order, resulting into a global descriptor of size 4012.

7.2. Geometric features

As geometric features (or shape features), we used normalized vectors composed of the coordinates $[x_k, y_k]$ for $k \in \{1, \ldots, 68\}$ of the 68 detected facial landmarks. After the face has been detected (Fig. 7a), facial landmarks are detected (Fig. 7b) and manually corrected if needed. We then use the standard normalization aimed at removing variations due to translation, scaling and in-plane rotation, as well as pose variations – namely, yaw and pitch. In particular, our shape normalization follows the approach of [71,72] and leverages a linear shape model built from images annotated with $u = 68$ fiducial points. The annotated shapes are first normalized using Procrustes Analysis. This step removes variations due to similarity transformations, i.e., translation, rotation and scaling. Then, PCA is applied on the normalized shapes to obtain the shape model. The shape model is defined by the mean shape $\mathbf{s}_0$, and n shape eigenvectors $\mathbf{s}_i$. The first two of these can be shown to model pose (pitch and yaw). Let us represent them as columns of matrix $\mathbf{P}^{2u\times 2}$. Additionally, to model similarity transforms, we construct 4 additional bases from $\mathbf{s}_0$ compactly represented as columns of $\mathbf{Q}^{2u\times 4}$. Suppose now that $\mathbf{s}_y \in \mathcal{R}^{2u\times 1}$ is the set of landmarks detected at each frame. Then, our similarity normalized features are given by $\mathbf{s}_{\text{sim}} = \mathbf{s}_y - \mathbf{Q}\mathbf{Q}^T(\mathbf{s}_y - \mathbf{s}_0)$. Additionally, our similarity and pose normalized features (Fig. 7d) are given by $\mathbf{s}_{\text{pose}} = \mathbf{s}_y - \mathbf{Q}\mathbf{Q}^T(\mathbf{s}_y - \mathbf{s}_0) - \mathbf{P}\mathbf{P}^T(\mathbf{s}_y - \mathbf{s}_0)$. Fig. 7 illustrates our shape and texture pre-processing.

7.3. Results and discussion

To compare the performance of various features, we sampled regularly an equal number of frames from each sequence to obtain a set of frames representative of the whole dataset, which we then divided in 5 disjoint folds in a subject-independent manner (i.e. a subject does not appear in two different folds). We then performed a 5-fold cross-validation (i.e. we iteratively used one of the set for testing and the other 4 for training) to predict the valence and arousal values for each frame using a Linear Support Vector Machine for Regression, Tables 3 and 4.

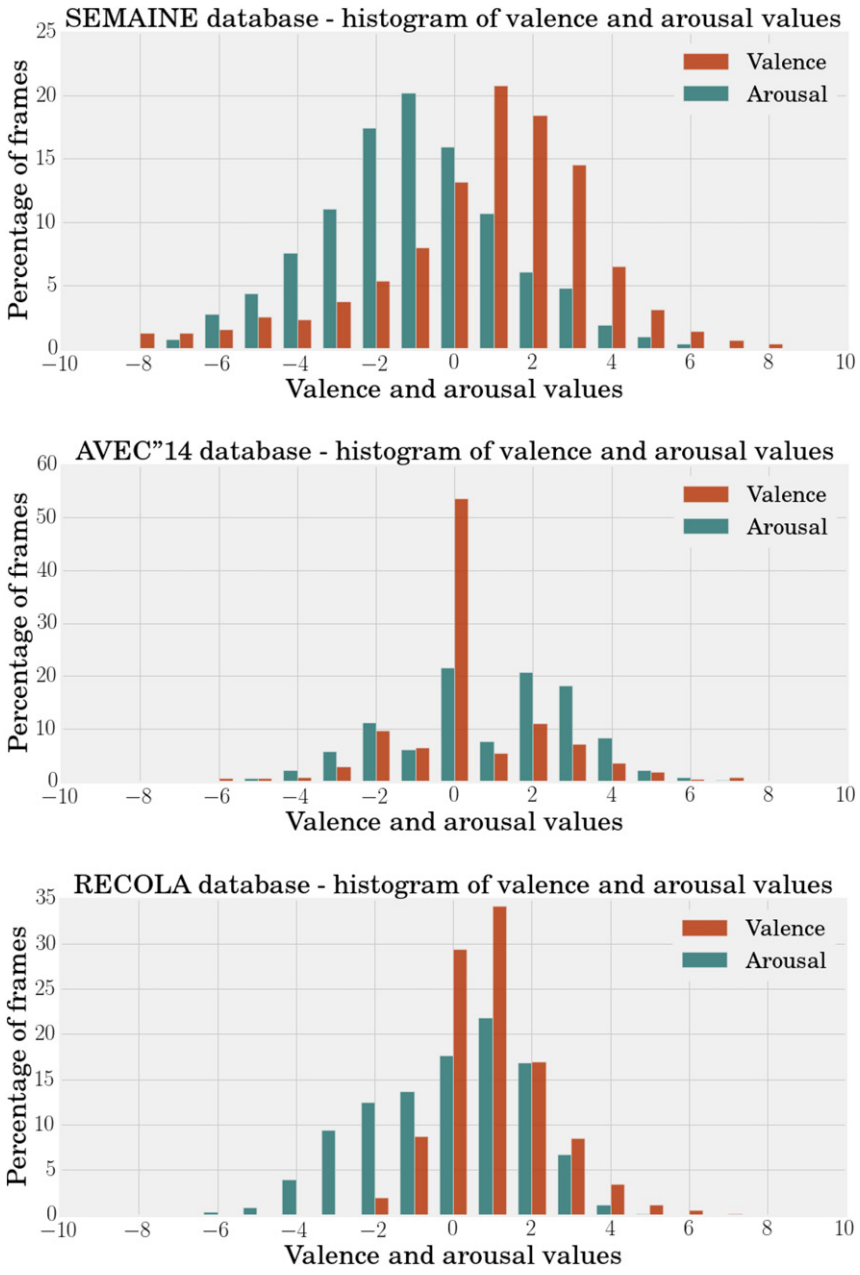

Fig. 4. Distribution of the valence and arousal values in the SEMAINE [9], AVEC'14 [43] and RECOLA [56] datasets.

We observe that normalised shape performs well overall and gives the best results in term of correlation and ICC for Arousal. This is not surprising as Arousal is typically characterised by large vertical changes, directly reflected in the position of the facial landmarks. This also shows that accurate landmark localisation is crucial. Valence, on the other hand can be much more subtle and low valence does not necessarily translate in large shape variations (e.g. sadness), hence the need of appearance features. Amongst these, Hybrid descriptors, unsurprisingly, give the best results, as they encode not only appearance information but also, inherently, some geometric information. In particular, Hybrid-DCT and Hybrid-SIFT perform very well. For the remaining experiments, we decide to use Norm-Shape and Hybrid-DCT – owing to its good performance and low dimensionality compared to SIFT.

Perhaps surprisingly, we notice that, overall, valence is better predicted than arousal on Semaine, Table 4, while the opposite is true for our AFEW-VA dataset. This might be due to the distribution of the valence and arousal labels, see Fig. 4. The second observation is that the performance of all features is more homogeneous in the case of the Semaine database. In particular, LBP features perform better on the controlled dataset than on the In-The-Wild one. So do features extracted from the warped appearance in the canonical coordinate frame. This is most likely due to the near frontal pose of all faces and to the controlled conditions. The difference in pose and environment is well demonstrated when comparing a warped appearance of exemplary faces from AFEW-VA and those from Semaine (Fig. 8). As can be seen, the piecewise-affine warping does not deform much the faces that are in a nearly frontal view as is the case in controlled environments but it does deform them largely in unconstrained conditions where large changes in head pose and occlusion are typical. This result illustrates the need of in-the-wild datasets – features found to work well on controlled datasets do not necessarily perform well in more realistic settings. As can be seen from Tables 3 and 4, our geometric features perform very well in both cases –controlled

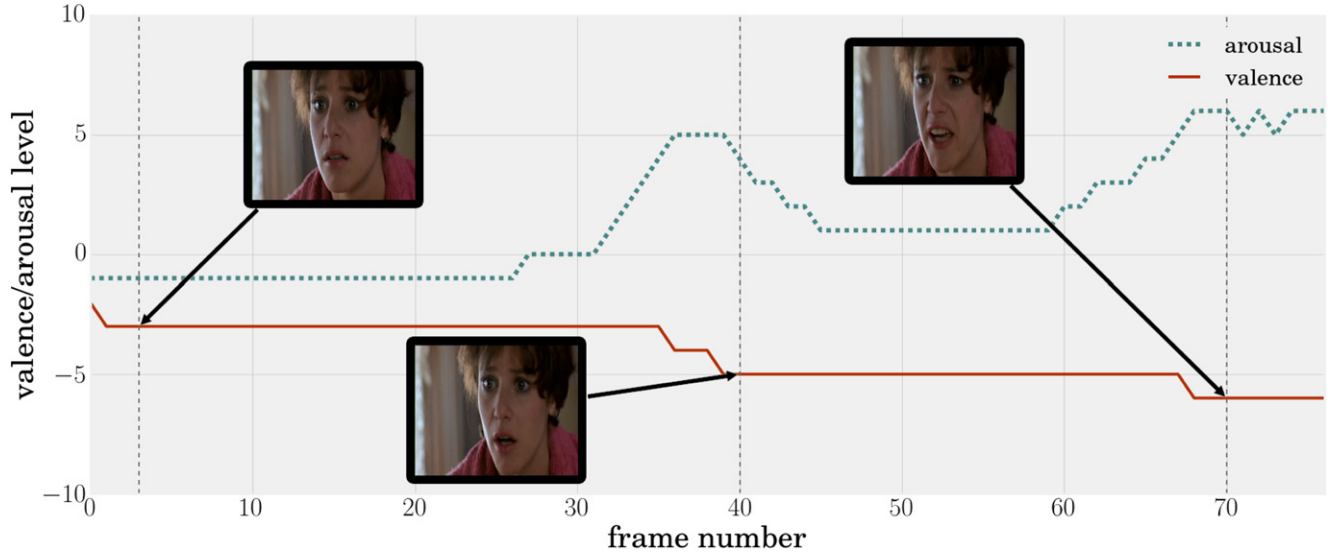

Fig. 5. Example of annotated valence and arousal levels for a sample video from our dataset along with some representative frames.

and unconstrained settings. This highlights the descriptive power of these features when it comes to facial expression tracking and analysis. This has been shown by previous research too [73,74] and re-confirmed here again.

8. Valence and arousal estimation

In this section, we evaluate a number of common baselines and state-of-the-art methods. In all cases, results were obtained as previously by performing a person-independent (i.e. the same participant did not appear simultaneously in two different folds) 5-folds cross-validation to predict per-frame values for valence and arousal, by successively training on 4 of the folds and testing on the remaining one. In all cases we optimised the respective parameters of each method by grid-search cross-validation on the training set.

8.1. Methods

We report the performance of several baselines and state-of-the-art methods, covering most widely used Machine-Learning methods (Statistical, Bayesian, Tree-based and Deep Learning):

Support Vector Machine for Regression (SVR) We used a linear Support Vector Machine for regression as our first baseline. SVR is a common choice for continuous affect recognition and is usually applied to a high-dimensional descriptor representing a concatenation of different features (e.g., LBP-TOP extracted on a dense grid of aligned faces) [11,41,43]. In [18] SVR is combined with Continuous Conditional Random Field (CCRF) or Correlation Aware CCRF. In [14], SVR is extended to Output-Associative RVM, and used with shape features to predict valence and arousal levels; but even this approach can be regarded as SVR-based. In this paper, we used the Scikit-Learn implementation [75] and validated the regularization parameter in the set $\{10^{-5}, ..., 10^{3}\}$ and the tolerance in the set $\{10^{-6}, \cdots, 10^{-2}\}$.

Bag of Words (BoW) is our second baseline. This approach is widely used in action recognition and is also common in facial expression recognition [76]. Our Bag-of-visual words baseline uses our Hybrid-DCT features and provides *per-frame* predictions, unlike in some prior work, to allow for a fair comparison. We used a vocabulary of 100 words and a Linear SVR as the back-end regression method.

Multiple Kernel Learning (MKL) We also provide results for a Multiple Kernel Learning approach which has been shown to produce state-of-the-art results for the task of Facial Action Unit detection [77,78] as well as for continuous facial behaviour estimation [3]; we adopt it here for the task of valence and arousal estimation. We implemented this approach using an RBF kernel for the shape features (*Norm-Shape*) and another one for the appearance features (*Hybrid-DCT*), using the Simple MKL library ([79]).

Conditional Random Field Conditional Random Field (*CRF*) has previously been used to estimate discretized levels of valence and arousal [80]. Here, we apply the method on sequences of 10 frames. The pystruct implementation was used [81] and the regularization was validated via grid-search.

Tree-based Random Forest (*RF*) have been shown to perform best on a large range of supervised learning problems [82]. They are widely used and recently demonstrated state-of-the-art results for continuous emotion recognition [19]. The scikit-learn [75] implementation of random forests was used with a number of trees in the forest was validated in the range {50, 100, 500, 1000}.

Ordinal Regression Ordinal Regression has been showed to produce state-of-the-art results for Facial Expression [83] and Facial Action Unit Estimation [84]. We evaluate Ordinal Regression (*OR*) and, like in the original paper [83] which implementation we used, we applied it to shape features, after having applied PCA and kept only the 20 most informative dimensions. Additionally, the labels were compressed into 7 labels instead of the 21 original ones. The regularizer was validated in the set $\{10^{-5}, \cdots, 10^{3}\}$.

Deep learning We trained two Deep CNN models for valence and arousal classification, using as input cropped faces re-sized to the dimension defined in the input layer of each network (*RGB-Images*). In both networks the output layer is modified to contain output units equal to the 21 levels of arousal and valence in our problem. Two different DCNN architectures were evaluated. First, a modified DCNN structure defined in [85] (*DCNN* in the result table). The model is trained from scratch using the randomly sampled frames from the videos. Second, a fine tuned, modified, AlexNet [86] coined *FT-DCNN*. We fine tuned the model learned on ImageNet [87].

Fig. 6. Examples of tracked landmarks from our newly annotated dataset.

8.2. Results and discussion

Table 5 shows the results obtained on AFEW-VA as explained with all methods, using geometric (*Norm-shape*) and appearance (*Hybrid DCT*) features, while Table 6 shows the results obtained on Semaine.

By construction, the ICC tends to be lower than the correlation. SVR performs very well, performing best in term of ICC. In comparison, the bag-of-words approach does not perform as well probably because the vocabulary learnt is not as informative as the original shape/DCT features. Random Forests, on the other hand, perform

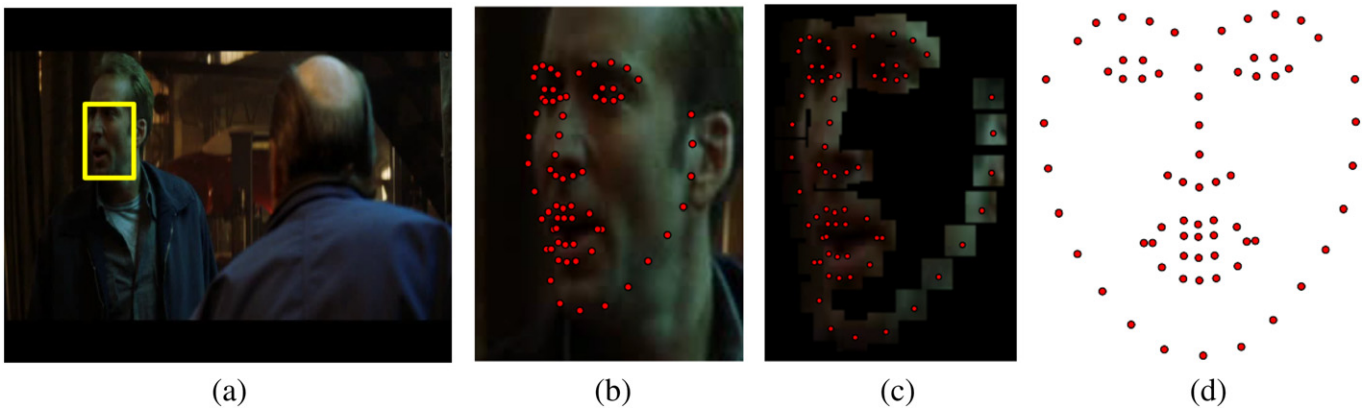

Fig. 7. Preprocessing steps for shape normalization. First the face is detected using the tree-based deformable part model of [60]. The resulting bounding box is used to initialize landmark detection using a gradient based method of [61]. Around each landmark, robust descriptors such (SIFT, LBP, DCT) are extracted as appearance features. We also normalize the shape formed by the detected landmarks, and use it as an additional shape descriptor for valence and arousal estimation.

Table 3
Comparison of the performance of the different features on AFEW-VA.

	Arousal			Valence		
Features	RMSE	CORR	ICC	RMSE	CORR	ICC
Block-LBP	2.503	0.253	0.211	2.881	0.220	0.164
Hybrid-LBP	2.402	0.279	0.193	2.807	0.238	0.134
Holistic-DCT	3.742	0.287	0.286	3.658	0.131	0.130
Block-DCT	4.188	0.118	0.106	4.088	0.254	0.252
Hybrid-DCT	**2.318**	0.381	0.318	**2.670**	**0.374**	0.290
Hybrid-SIFT	2.492	0.323	0.295	2.754	**0.374**	**0.341**
Norm-shape	2.446	**0.426**	**0.356**	2.829	0.293	0.21

Table 4
Comparison of the performance of the different features on SEMAINE [9].

	Arousal			Valence		
Features	RMSE	CORR	ICC	RMSE	CORR	ICC
Block-LBP	**2.275**	0.143	0.103	2.342	**0.283**	**0.159**
Hybrid-LBP	2.386	0.081	0.061	2.576	**0.262**	0.138
Holistic-DCT	2.303	**0.281**	**0.271**	3.566	0.06	0.057
Block-DCT	3.146	0.146	0.144	2.823	0.207	**0.203**
Hybrid-DCT	**2.278**	0.18	0.128	**2.287**	0.18	**0.162**
Hybrid-SIFT	2.507	0.1	0.061	2.981	0.167	0.147
Norm-shape	**2.249**	**0.272**	**0.226**	**2.323**	**0.35**	**0.331**

very well, with a higher correlation than all other methods when using Hybrid-DCT. CRF, however, does not give good results, probably because of the short temporal dependencies in the video and due to the difficulty of predicting 21 classes. Ordinal regression also under-performs, most likely because the features are not ordinal with respect to the labels, i.e. variations in valence or arousal are not always directly correlated with the landmark shifts. Similarly, learning a DCNN from scratch does not provide good result. The main reason might be that there is not enough samples to train the model. More interestingly the modified AlexNet does not provide significant improvement. We believe there are two reasons for this: a) The re-sized inputs have an insufficient resolution. b) The face images are very different from the images in ImageNet, i.e they come from a completely different distribution. As a result fine-tuning does not provide informative feedback signals to the lower layers of the network. Finally, the Multiple Kernel Learning approach successfully combines shape and appearance information, producing very good results, and the best RMSE for both valence and arousal.

These results are comparable to the baselines provided with existing databases. In the AVEC 2013 challenge, the Pearson's correlation coefficient on the testing set using video features was reported as 0.076 for valence and 0.134 for arousal [42]. In the 2014 standing this was of 0.188 for valence and 0.206 for arousal [43]. RMSE was not reported for the 2013 and 2014 standings. However, in the 2015 challenge, CORR was 0.354 for arousal and 0.490 for valence, while the RMSE was between 0.164 and 0.113 [44]. Note that the values of valence and arousal ranged between −1 and 1, making these RMSE values comparable with the ones reported here. Although our labels range from −10 to 10 these can easily be normalised to range between −1 and 1.

However, interestingly, and as hypothesised, the same methods that perform best on AFEW-VA do not necessarily perform best on Semaine, and the results are generally much more homogeneous on the latter. In particular, on the Semaine data, the Random Forest approach does not perform as well and is outperformed by CRF applied on geometric features. The modified AlexNet does not do well in term of correlation for arousal but obtains the lowest RMSE and performs comparatively better for valence.

Surprisingly, overall, CORR and ICC are, lower on the Semaine data than on the AFEW-VA data. This might be because the Semaine data is not very versatile in the sense that it contains too few strong intensity examples. This causes the features to end up being too similar, with too much data and too little facial deformation (See Fig. 8). As seen in Fig. 4, in the Semaine database, for a large majority of the frames, the arousal is negative. This is in contrast to AFEW-VA where positive arousal forms a majority. Another assumption is therefore that methods could learn better on the AFEW-VA data what is typical for positive vs negatve.

The representational power of geometric features is again confirmed, and, combined with these, SVR, CRF and MKL yield the best overall results in term of CORR and ICC. These results on the Semaine data are comparable with those recently reported on the same data [3,14,41].

9. Conclusion

We motivated the problem of estimating valence and arousal on data collected in-the-wild. A new dataset called *AFEW-VA* was presented that provides *per frame* highly accurate annotations of valence and arousal along with facial landmarks for 600 video clips extracted from feature films, and carefully selected for evaluation under challenging, unconstrained conditions. Different state-of-the-art robust features were compared to select the most appropriate ones for the

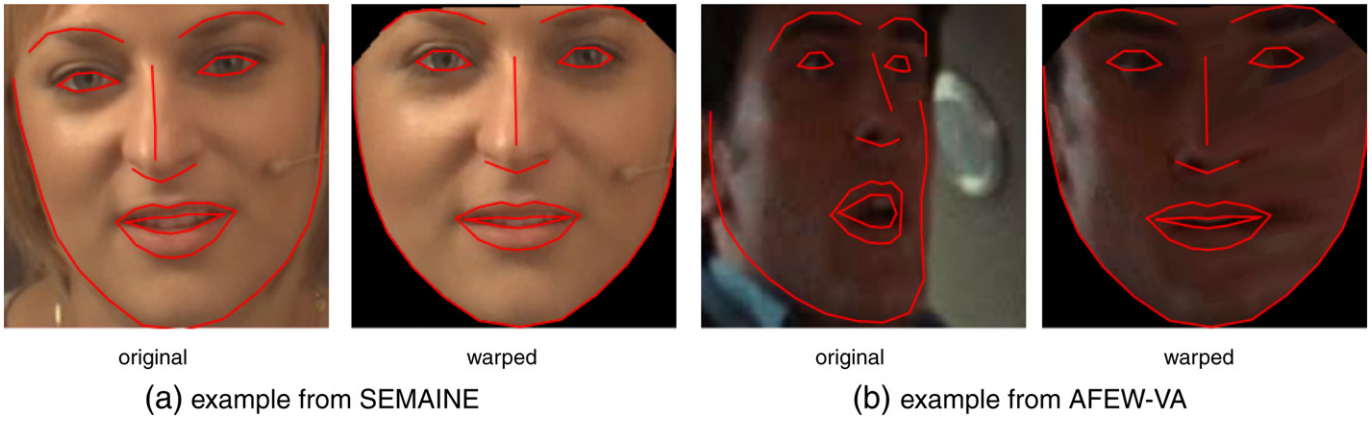

Fig. 8. Examples of original and corresponding warped images from SEMAINE and AFEW-VA. In controlled environment with limited illumination and pose variation (a), the warped image is very similar to the original image while in real in-the-wild conditions (b), the warped image presents large deformations.

Table 5
Comparison of the performance of different methods on AFEW-VA.

		Arousal			Valence		
Method	Features	RMSE	CORR	ICC	RMSE	CORR	ICC
SVR	Norm-shape	2.446	0.426	**0.356**	2.829	0.293	0.21
SVR	Hybrid-DCT	2.318	0.381	0.318	2.670	0.374	**0.290**
RF	Norm-shape	2.256	0.411	0.303	2.672	0.365	0.267
RF	Hybrid DCT	2.275	**0.45**	0.2	2.687	**0.407**	0.154
CRF	Norm-shape	2.815	0.330	0.326	3.289	0.244	0.242
CRF	Hybrid DCT	2.912	0.214	0.21	3.421	0.137	0.137
DCNN	RGB-Images	4.6	0.25	–	4.1	0.17	–
FT-DCNN	RGB-Images	3.9	0.31	–	3.7	0.26	–
BoW	Hybrid-DCT	2.467	0.25	0.194	2.907	0.124	0.071
OR	Norm-shape	–	0.28	0.23	–	0.25	0.20
MKL	Shape+DCT	**2.229**	**0.445**	0.340	**2.639**	0.401	0.274

Table 6
Comparison of the performance of different methods on SEMAINE. The valence and arousal intensity values, originally ranging from −1 to 1, have been scaled by a factor of 10 to range in [−10, 10].

		Arousal			Valence		
Method	Features	RMSE	CORR	ICC	RMSE	CORR	ICC
SVR	Norm-shape	2.249	**0.272**	**0.226**	2.323	**0.35**	**0.331**
SVR	Hybrid-DCT	2.278	0.18	0.128	2.287	0.17	0.131
RF	Norm-shape	2.50	0.123	0.117	2.221	0.23	0.152
RF	Hybrid DCT	2.674	0.08	0.077	**2.087**	0.150	0.134
CRF	Norm-shape	2.466	**0.266**	**0.245**	3.05	**0.275**	**0.230**
CRF	Hybrid DCT	2.475	0.10	0.094	2.814	0.173	0.151
FT-DCNN	RGB-Images	**1.608**	0.109	–	2.173	**0.268**	–
BoW	Hybrid-DCT	2.920	0.199	0.18	2.222	0.166	0.158
OR	Norm-shape	–	0.10	0.09	–	0.18	0.14
MKL	**Shape+DCT**	2.366	0.23	0.15	2.575	**0.296**	**0.198**

task of valence and arousal prediction in-the-wild. A number of baselines as well as state-of-the-art methods were also evaluated on this new dataset. In addition, the results of the same features and methods were investigated on a classical, widely used, controlled dataset.

Our results illustrate our hypothesis that features and methods that work well in controlled environments do not necessarily perform as well in unconstrained conditions. They also confirmed the descriptive power of geometric features and highlighted the importance of accurately tracked facial landmarks. Finally, they demonstrate our dataset to be challenging, both complementary to and extending existing benchmark datasets. It is useful for static, in-the-wild, valence and arousal estimation but less so for dynamic models due to the short duration of some of the clips and to the low inter-variability of the expressions displayed in between the frames.

Acknowledgements

The work of Jean Kossaifi and Maja Pantic has been funded by the European Community Horizon 2020 [H2020/2014-2020] under grant agreement no. 645094 (SEWA). For this work, Sinisa Todorovic has been supported in part by grant NSF RI 1302700.

The authors would also like to thank Behrooz Mahasseni for running the DCNN experiments.

References

[1] P. Ekman, W.V. Frisen, L. Erika, Facial Action Coding System (FACS), Weidenfeld & Nicolson, Salt Lake City; London, 2002.
[2] O. Rudovic, V. Pavlovic, M. Pantic, Context-sensitive dynamic ordinal regression for intensity estimation of facial action units, IEEE Trans. Pattern Anal. Mach. Intell. 37 (2015) 944–958.
[3] S. Kaltwang, S. Todorovic, M. Pantic, Doubly sparse relevance vector machine for continuous facial behavior estimation, IEEE Trans. Pattern Anal. Mach. Intell. 38 (99) (2016) 1748–1761.
[4] H. Gunes, B. Schuller, Categorical and dimensional affect analysis in continuous input: current trends and future directions, Image Vis. Comput. 31 (2) (2013) 120–136.
[5] Y. Panagakis, M.A. Nicolaou, S. Zafeiriou, M. Pantic, Robust correlated and individual component analysis, IEEE Trans. Pattern Anal. Mach. Intell. 38 (8) (2016) 1665–1678.
[6] E. Sariyanidi, H. Gunes, A. Cavallaro, Automatic analysis of facial affect: a survey of registration, representation, and recognition, IEEE Trans. Pattern Anal. Mach. Intell. 37 (6) (2015) 1113–1133.
[7] A. Dhall, R. Goecke, S. Lucey, T. Gedeon, Collecting large, richly annotated facial-expression databases from movies, IEEE MultiMedia (3) (2012) 34–41.
[8] A. Dhall, R. Goecke, J. Joshi, K. Sikka, T. Gedeon, Emotion Recognition in the Wild Challenge 2014: Baseline, Data and Protocol, 16Th ACM International Conference on Multimodal Interaction, ACM. 2014.
[9] G. McKeown, M. Valstar, R. Cowie, M. Pantic, M. Schröder, The SEMAINE database: annotated multimodal records of emotionally colored conversations between a person and a limited agent, IEEE Trans. Affect. Comput. 3 (1) (2012) 5–17.
[10] Z. Zeng, M. Pantic, G.I. Roisman, T.S. Huang, A survey of affect recognition methods: audio, visual, and spontaneous expressions, IEEE Trans. Pattern Anal. Mach. Intell. 31 (1) (2009) 39–58.
[11] M. Valstar, J. Gratch, B. Schuller, F. Ringeval, D. Lalanne, M. Torres Torres, S. Scherer, G. Stratou, R. Cowie, M. Pantic, AVEC 2016: Depression, Mood, and Emotion Recognition Workshop and Challenge, Proceedings of the 6Th International Workshop on Audio/Visual Emotion Challenge, AVEC '16, ACM, New York, NY, USA, 2016, pp. 3–10.
[12] M.A. Nicolaou, H. Gunes, M. Pantic, Continuous prediction of spontaneous affect from multiple cues and modalities in valence-arousal space, IEEE Trans. Affect. Comput. 2 (2) (2011) 92–105.
[13] M.A. Nicolaou, H. Gunes, M. Pantic, Output-Associative RVM Regression for Dimensional and Continuous Emotion Prediction, Proceedings of IEEE International Conference on Automatic Face and Gesture Recognition (FG'11),Santa Barbara, CA, USA, 2011. pp. 16–23.
[14] M.A. Nicolaou, H. Gunes, M. Pantic, Output-associative RVM regression for dimensional and continuous emotion prediction, Image Vis. Comput. 30 (3) (2012) 186–196.
[15] J. Nicolle, V. Rapp, K. Bailly, L. Prevost, M. Chetouani, Robust Continuous Prediction of Human Emotions Using Multiscale Dynamic Cues, Proceedings of the 14Th ACM International Conference on Multimodal Interaction, ICMI '12, ACM, New York, NY, USA, 2012, pp. 501–508.
[16] C. Soladié, H. Salam, C. Pelachaud, N. Stoiber, R. Séguier, A Multimodal Fuzzy Inference System Using a Continuous Facial Expression Representation for Emotion Detection, Proceedings of the 14Th ACM International Conference on Multimodal Interaction, ICMI '12, ACM, New York, NY, USA, 2012, pp. 493–500.
[17] M.A. Nicolaou, S. Zafeiriou, M. Pantic, Correlated-Spaces Regression for Learning Continuous Emotion Dimensions, Proceedings of the 21St ACM International Conference on Multimedia, MM '13, ACM, New York, NY, USA, 2013, pp. 773–776.
[18] T. Baltrusaitis, N. Banda, P. Robinson, Dimensional Affect Recognition Using Continuous Conditional Random Fields, Automatic Face and Gesture Recognition (FG), 2013 10Th IEEE International Conference and Workshops On, IEEE. 2013, pp. 1–8.
[19] H. Chen, J. Li, F. Zhang, Y. Li, H. Wang, 3D Model-Based Continuous Emotion Recognition, 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015. pp. 1836–1845.
[20] H. Meng, N. Bianchi-Berthouze, Y. Deng, J. Cheng, J. Cosmas, Time-delay neural network for continuous emotional dimension prediction from facial expression sequences, IEEE Trans. Cybern. 46 (2015) 916–929.
[21] E. Sánchez-Lozano, P. Lopez-Otero, L. Docio-Fernandez, E. Argones-Rúa, J.L. Alba-Castro, Audiovisual Three-Level Fusion for Continuous Estimation of Russell's Emotion Circumplex, Proceedings of the 3Rd ACM International Workshop on Audio/Visual Emotion Challenge, AVEC '13, ACM, New York, NY, USA, 2013, pp. 31–40.
[22] H. Meng, D. Huang, H. Wang, H. Yang, M. AI-Shuraifi, Y. Wang, Depression Recognition Based on Dynamic Facial And Vocal Expression Features Using Partial Least Square Regression, Proceedings of the 3Rd ACM International Workshop On Audio/Visual Emotion Challenge, AVEC '13, ACM, New York, NY, USA, 2013, pp. 21–30.
[23] M. Kächele, M. Schels, F. Schwenker, Inferring Depression and Affect from Application Dependent Meta Knowledge, Proceedings of the 4Th International Workshop On Audio/Visual Emotion Challenge, AVEC '14, 2014.
[24] H. Kaya, F. Çilli, A.A. Salah, Ensemble CCA for Continuous Emotion Prediction, Proceedings of the 4Th International Workshop On Audio/Visual Emotion Challenge, AVEC '14, 2014.
[25] R. Gupta, N. Malandrakis, B. Xiao, T. Guha, M. Van Segbroeck, M. Black, A. Potamianos, S. Narayanan, Multimodal Prediction of Affective Dimensions And Depression in Human-Computer Interactions, Proceedings of the 4Th International Workshop On Audio/Visual Emotion Challenge, AVEC '14, 2014.
[26] L. Chao, J. Tao, M. Yang, Y. Li, Z. Wen, Multi-Scale Temporal Modeling for Dimensional Emotion Recognition in Video, Proceedings of the 4Th International Workshop On Audio/Visual Emotion Challenge, AVEC '14, 2014.
[27] Z. Huang, T. Dang, N. Cummins, B. Stasak, P. Le, V. Sethu, J. Epps, An Investigation of Annotation Delay Compensation and Output-Associative Fusion for Multimodal Continuous Emotion Prediction, Proceedings of the 5Th International Workshop On Audio/Visual Emotion Challenge, AVEC '15, ACM, New York, NY, USA, 2015, pp. 41–48.

[28] A. Milchevski, A. Rozza, D. Taskovski, Multimodal Affective Analysis Combining Regularized Linear Regression and Boosted Regression Trees, Proceedings of the 5Th International Workshop On Audio/Visual Emotion Challenge, AVEC '15, ACM, New York, NY, USA, 2015, pp. 33–39.

[29] M. Kächele, P. Thiam, G. Palm, F. Schwenker, M. Schels, Ensemble Methods for Continuous Affect Recognition: Multi-Modality, Temporality, and Challenges, Proceedings of the 5Th International Workshop On Audio/Visual Emotion Challenge, AVEC '15, ACM, New York, NY, USA, 2015, pp. 9–16.

[30] S. Chen, Q. Jin, Multi-Modal Dimensional Emotion Recognition Using Recurrent Neural Networks, Proceedings of the 5Th International Workshop On Audio/Visual Emotion Challenge, AVEC '15, ACM, New York, NY, USA, 2015, pp. 49–56.

[31] L. Chao, J. Tao, M. Yang, Y. Li, Z. Wen, Long Short Term Memory Recurrent Neural Network Based Multimodal Dimensional Emotion Recognition, Proceedings of the 5Th International Workshop On Audio/Visual Emotion Challenge, AVEC '15, ACM, New York, NY, USA, 2015, pp. 65–72.

[32] P. Cardinal, N. Dehak, A.L. Koerich, J. Alam, P. Boucher, ETS System for AV+EC 2015 Challenge, Proceedings of the 5Th International Workshop On Audio/Visual Emotion Challenge, AVEC '15, ACM, New York, NY, USA, 2015, pp. 17–23.

[33] L. He, D. Jiang, L. Yang, E. Pei, P. Wu, H. Sahli, Multimodal Affective Dimension Prediction Using Deep Bidirectional Long Short-Term Memory Recurrent Neural Networks, Proceedings of the 5Th International Workshop On Audio/Visual Emotion Challenge, AVEC '15, ACM, New York, NY, USA, 2015, pp. 73–80.

[34] K. Brady, Y. Gwon, P. Khorrami, E. Godoy, W. Campbell, C. Dagli, T.S. Huang, Multi-Modal Audio, Video and Physiological Sensor Learning for Continuous Emotion Prediction, Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge, AVEC '16, ACM, New York, NY, USA, 2016, pp. 97–104.

[35] K. Somandepalli, R. Gupta, M. Nasir, B.M. Booth, S. Lee, S.S. Narayanan, Online Affect Tracking with Multimodal Kalman Filters, Proceedings of the 6Th International Workshop on Audio/Visual Emotion Challenge, AVEC '16, ACM, New York, NY, USA, 2016, pp. 59–66.

[36] M. Amirian, M. Kächele, P. Thiam, V. Kessler, F. Schwenker, Continuous Multimodal Human Affect Estimation Using Echo State Networks, Proceedings of the 6Th International Workshop on Audio/Visual Emotion Challenge, AVEC '16, ACM, New York, NY, USA, 2016, pp. 67–74.

[37] F. Povolny, P. Matejka, M. Hradis, A. Popková, L. Otrusina, P. Smrz, I. Wood, C. Robin, L. Lamel, Multimodal Emotion Recognition for AVEC 2016 Challenge, Proceedings of the 6Th International Workshop on Audio/Visual Emotion Challenge, AVEC '16, ACM, New York, NY, USA, 2016, pp. 75–82.

[38] B. Sun, S. Cao, L. Li, J. He, L. Yu, Exploring Multimodal Visual Features for Continuous Affect Recognition, Proceedings of the 6Th International Workshop on Audio/Visual Emotion Challenge, AVEC '16, ACM, New York, NY, USA, 2016, pp. 83–88.

[39] F. Ringeval, F. Eyben, E. Kroupi, A. Yuce, J.-P. Thiran, T. Ebrahimi, D. Lalanne, B. Schuller, Prediction of asynchronous dimensional emotion ratings from audiovisual and physiological data, Pattern Recogn. Lett. 66 (2015) 22–30. pattern Recognition in Human Computer Interaction.

[40] B. Schuller, M. Valstar, F. Eyben, G. McKeown, R. Cowie, M. Pantic, AVEC 2011 - The first international audio/visual emotion challenge, Lect. Notes Comput. Sci 6975 LNCS (PART 2) (2011) 415–424. (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics).

[41] B. Schuller, M. Valster, F. Eyben, R. Cowie, M. Pantic, AVEC 2012: The continuous audio/visual emotion challenge, Proc. 14th Int'l Conf. Multimodal Interaction Workshops (2012) 449–456.

[42] M. Valstar, B. Schuller, K. Smith, F. Eyben, B. Jiang, S. Bilakhia, S. Schnieder, R. Cowie, M. Pantic, AVEC 2013: The Continuous Audio/Visual Emotion and Depression Recognition Challenge, Proceedings of the 3rd ACM International Workshop on Audio/Visual Emotion Challenge (AVEC '13), 2013. pp. 3–10.

[43] M. Valstar, B. Schuller, K. Smith, T. Almaev, F. Eyben, J. Krajewski, R. Cowie, M. Pantic, AVEC 2014: 3D Dimensional Affect and Depression Recognition Challenge, Proceedings of the 4th ACM International Workshop on Audio/Visual Emotion Challenge (AVEC '14), 2014. pp. 3–10.

[44] F. Ringeval, M. Valstar, E. Marchi, D. Lalanne, R. Cowie, The AV + EC 2015 multimodal affect recognition challenge: bridging across audio, video, and physiological data categories and subject descriptors, Proc. ACM Multimedia Workshops (CCC) (2015) 2–5.

[45] M. Pantic, L. Rothkrantz, Towards an affect-sensitive multimodal human–computer interaction, Invited Paper, Proc. IEEE, Spec. Issue Multimodal Hum. Comput. Interact. (HCI) 91 (9) (2003) 1370–1390.

[46] R. Cowie, E. Douglas-cowie, S. Savvidou, E. Mcmahon, M. Sawey, M. Schröder, Feeltrace: an instrument for recording perceived emotion in real time., ISCA Workshop on Speech &, Emotion (2000) 19–24.

[47] M. Grimm, K. Kroschel, Evaluation of Natural Emotions Using Self Assessment Manikins, IEEE Workshop on Automatic Speech Recognition And Understanding, 2005. pp. 381–385.

[48] E. Douglas-Cowie, R. Cowie, I. Sneddon, C. Cox, O. Lowry, M. McRorie, J.-C. Martin, L. Devillers, S. Abrilian, A. Batliner, N. Amir, K. Karpouzis, The HUMAINE database: addressing the collection and annotation of naturalistic and induced emotional data, Affect. Comput. Intell. Interact. 4738 (2007) 488–500.

[49] M. Schröder, E. Bevacqua, F. Eyben, H. Gunes, D. Heylen, M. ter Maat, S. Pammi1, M. Pantic, C. Pelachaud, B. Schuller, E. de Sevin, M. Valstar, M. Wöllmer, A Demonstration of Audiovisual Sensitive Artificial Listeners, 3rd International Conference on Affective Computing and Intelligent Interaction and Workshops, 2009. pp. 1–2.

[50] I. Sneddon, M. McRorie, The Belfast induced natural emotion database, IEEE Trans. Affect. Comput. 3 (1) (2012) 32–41.

[51] M. Soleymani, J. Lichtenauer, T. Pun, M. Pantic, A multimodal database for affect recognition and implicit tagging, IEEE Trans. Affect. Comput. 3 (1) (2012) 42–55.

[52] S. Koelstra, C. Muhl, M. Soleymani, J.S. Lee, A. Yazdani, T. Ebrahimi, T. Pun, A. Nijholt, I. Patras, DEAP: a database for emotion analysis ;using physiological signals, IEEE Trans. Affect. Comput. 3 (1) (2012) 18–31.

[53] M. Grimm, K. Kroschel, S. Narayanan, The vera am mittag German Audio-visual Emotional Speech Database, 2008 IEEE International Conference on Multimedia and Expo, ICME 2008 - Proceedings, 2008. pp. 865–868.

[54] S. Petridis, B. Martinez, M. Pantic, The MAHNOB laughter database, Image Vis. Comput. J. 31 (2) (2013) 186–202.

[55] D. McDuff, R. El Kaliouby, T. Senechal, M. Amr, J.F. Cohn, R. Picard, Affectiva-mit Facial Expression Dataset (AM-FED): Naturalistic and Spontaneous Facial Expressions Collected 'in-the-wild', IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, 2013. pp. 881–888.

[56] F. Ringeval, A. Sonderegger, J. Sauer, D. Lalanne, Introducing the RECOLA Multimodal Corpus of Remote Collaborative and Affective Interactions, Automatic Face and Gesture Recognition (FG), 2013 10th IEEE International Conference and Workshops on, 2013. pp. 1–8.

[57] S. Bilakhia, S. Petridis, A. Nijholt, M. Pantic, The MAHNOB mimicry database - a database of naturalistic human interactions, Pattern Recogn. Lett. 66 (2015) 52–61. in press.

[58] M.A. Nicolaou, V. Pavlovic, M. Pantic, Dynamic probabilistic CCA for analysis of affective behaviour and fusion of continuous annotations, IEEE Trans. Pattern Anal. Mach. Intell. 36 (7) (2014) 1299–1311.

[59] K.R. Scherer, T. Banziger, E. Roesch, A Blueprint for Affective Computing: a Sourcebook And Manual, first ed., Oxford University Press, Inc., New York, NY, USA, 2010.

[60] X. Zhu, D. Ramanan, Face Detection, Pose Estimation, and Landmark Estimation in the Wild., IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2012. pp. 2879–2886.

[61] J. Kossaifi, G. Tzimiropoulos, M. Pantic, Fast and exact Newton and bidirectional fitting of active appearance models, IEEE Trans. Image Process. 26 (2) (2017) 1040–1053.

[62] P.E. Shrout, J.L. Fleiss, Intraclass correlations: uses in assessing rater reliability, Psychol. Bull. 86 (2) (1979) 420–428.

[63] S.M. Mavadati, M.H. Mahoor, K. Bartlett, P. Trinh, J.F. Cohn, Disfa: a spontaneous facial action intensity database, IEEE Trans. Affect. Comput. 4 (2) (2013) 151–160.

[64] D.G. Lowe, Distinctive image features from scale-invariant keypoints, Int. J. Comput. Vis. (IJCV) 60 (2) (2004) 91–110.

[65] T. Ojala, M. Pietikainen, T. Maenpaa, Multiresolution gray-scale and rotation invariant texture classification with local binary patterns, IEEE Trans. Pattern Anal. Mach. Intell. 24 (7) (2002) 971–987.

[66] S. Kaltwang, O. Rudovic, M. Pantic, Continuous Pain Intensity Estimation from Facial Expressions, Advances in Visual Computing, Lecture Notes in Computer Science 7432, Springer, Heidelberg, 2012, pp. 368–377.

[67] M.F. Valstar, B. Jiang, M. Mehu, M. Pantic, K. Scherer, The First Facial Expression Recognition and Analysis Challenge, Proceedings of IEEE International Conference On Automatic Face and Gesture Recognition (FG'11), FERA 2011 Workshop on Facial Expression Recognition And Analysis Challenge, Santa Barbara, CA, USA, 2011. pp. 921–926.

[68] C. Shan, S. Gong, P.W. McOwan, Facial expression recognition based on local binary patterns: a comprehensive study, Image Vis. Comput. 27 (6) (2009) 803–816.

[69] N. Ahmed, T. Natarajan, K.R. Rao, Discrete cosine transform, IEEE Trans. Comput. C-23 (1) (1974) 90–93.

[70] G.K. Wallace, The JPEG still picture compression standard, Commun. ACM 34 (1991) 31–44.

[71] J. Kossaifi, G. Tzimiropoulos, M. Pantic, Fast Newton Active Appearance Models, Proceedings of the IEEE Intl Conf. on Image Processing (ICIP14), Paris, France, 2014. pp. 1420–1424.

[72] J. Kossaifi, G. Tzimiropoulos, M. Pantic, Fast and Exact Bi-Directional Fitting of Active Appearance Models, Proceedings of the IEEE Intl Conf. on Image Processing (ICIP15), Quebec City, QC, Canada, 2015. pp. 1135–1139.

[73] M. Pantic, Machine analysis of facial behaviour: naturalistic and dynamic behaviour, Philos. Trans. R. Soc. B 364 (2009) 3505–3513.

[74] M.F. Valstar, M. Mehu, B. Jiang, M. Pantic, K. Scherer, Meta-analyis of the first facial expression recognition challenge, IEEE Trans. Syst. Man Cybern. B 42 (4) (2012) 966–979.

[75] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, E. Duchesnay, Scikit-learn: machine learning in Python, J. Mach. Learn. Res. 12 (2011) 2825–2830.

[76] M. Liu, S. Shan, R. Wang, X. Chen, Learning Expressionlets on Spatio-Temporal Manifold for Dynamic Facial Expression Recognition, 2014 IEEE Conference on Computer Vision and Pattern Recognition, 2014. pp. 1749–1756.

[77] T. Senechal, V. Rapp, H. Salam, R. Séguier, K. Bailly, L. Prevost, Facial action recognition combining heterogeneous features via multikernel learning, IEEE Trans. Syst. Man Cybern. B 42 (4) (2012) 993–1005.

[78] S. Kaltwang, S. Todorovic, M. Pantic, Latent Trees for Estimating Intensity of Facial Action Units, 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015. pp. 296–304.

[79] A. Rakotomamonjy, F.R. Bach, S. Canu, Y. Grandvalet, SimpleMKL, J. Mach. Learn. Res. 9 (2008) 2491–2521.
[80] M. Wöllmer, F. Eyben, S. Reiter, B. Schuller, C. Cox, E. Douglas-Cowie, R. Cowie, Abandoning emotion classes-towards continuous emotion recognition with modelling of long-range dependencies, INTERSPEECH 2008 (2008) 597–600.
[81] A.C. Müller, S. Behnke, Pystruct - learning structured prediction in python, J. Mach. Learn. Res. 15 (2014) 2055–2060.
[82] R. Caruana, N. Karampatziakis, A. Yessenalina, An Empirical Evaluation of Supervised Learning In High Dimensions, Proceedings of the 25Th International Conference On Machine Learning, ACM, New York, NY, USA, 2008, pp. 96–103.
[83] O. Rudovic, V. Pavlovic, M. Pantic, Multi-Output Laplacian Dynamic Ordinal Regression For Facial Expression Recognition and Intensity Estimation, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2012), Providence, USA, 2012. pp. 2634–2641.
[84] R. Walecki, O. Rudovic, V. Pavlovic, M. Pantic, Copula Ordinal Regression for Joint Estimation Of Facial Action Unit Intensity, Proceedings of IEEE Intl Conf. on Computer Vision and Pattern Recognition (CVPR 2016), Las Vegas, Nevada, 2016. pp. 1–8.
[85] Y. Sun, X. Wang, X. Tang, Hybrid Deep Learning for Face Verification, IEEE International Conference on Computer Vision (ICCV), IEEE. 2013, pp. 1489–1496.
[86] A. Krizhevsky, I. Sutskever, G.E. Hinton, Imagenet Classification with Deep Convolutional Neural Networks, Neural Information Processing Systems (NIPS), 2012. pp. 1097–1105.
[87] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A.C. Berg, L. Fei-Fei, ImageNet Large Scale Visual Recognition Challenge, Int. J. Comput. Vis. (IJCV) (2015) 1–42.