ARTICLES
https://doi.org/10.1038/s42256-020-00280-0
nature machine intelligence

Estimation of continuous valence and arousal levels from faces in naturalistic conditions

Antoine Toisoul[1]✉, Jean Kossaifi[1,2], Adrian Bulat[1], Georgios Tzimiropoulos[1] and Maja Pantic[2]

Facial affect analysis aims to create new types of human–computer interactions by enabling computers to better understand a person's emotional state in order to provide ad hoc help and interactions. Since discrete emotional classes (such as anger, happiness, sadness and so on) are not representative of the full spectrum of emotions displayed by humans on a daily basis, psychologists typically rely on dimensional measures, namely valence (how positive the emotional display is) and arousal (how calming or exciting the emotional display looks like). However, while estimating these values from a face is natural for humans, it is extremely difficult for computer-based systems and automatic estimation of valence and arousal in naturalistic conditions is an open problem. Additionally, the subjectivity of these measures makes it hard to obtain good quality data. Here we introduce a novel deep neural network architecture to analyse facial affect in naturalistic conditions with a high level of accuracy. The proposed network integrates face alignment and jointly estimates both categorical and continuous emotions in a single pass, making it suitable for real-time applications. We test our method on three challenging datasets collected in naturalistic conditions and show that our approach outperforms all previous methods. We also discuss caveats regarding the use of this tool, and ethical aspects that must be considered in its application.

Facial affect analysis is an active field of research that aims at automatically estimating the emotions of a person in order to provide new types human–computer interaction. However there has typically been a gap between the state of the art in psychology and what is done in computer vision. In particular, most existing work in computer vision is focused on a simplistic setting, namely that of predicting discrete classes of 'prototypical' expressions of emotions displayed in laboratory conditions. Meanwhile, psychologists have moved away from these coarse classes that do not reflect the range of emotional display shown by humans on a daily basis in naturalistic, everyday situations[1]. Instead, they focus on dimensional measures of affective display, the most notable of which are valence (how negative or positive the emotional display is) and arousal (how calming or exciting the emotional display looks like)[2,3] (Fig. 1). In addition, while we are interested in performing affective display analysis in naturalistic (in-the-wild) conditions, most existing work focuses on controlled, laboratory conditions.

This mismatch between theory and practice can be explained by various factors: (1) the difficulty of collecting large corpora of emotional data in naturalistic conditions and (2) the difficulty of accurately labelling large amounts of data[3,4]. To be able to generalize well to unseen real-life situations, deep learning requires both of these. This state of affairs has recently started evolving with the introduction of datasets collected in the wild and accurately annotated for valence and arousal (for example, AffecNet[5], AFEW-VA[6] and SEWA[7]).

Virtually all existing works approach the task as a series of disjoint steps[4]. First a face detector is run on an image to detect every face present in it. Each face is then cropped to remove the background using the bounding box predicted by the face detector. Facial landmarks (fiducial points) are detected on each face and employed to project the face to a canonical coordinate frame to remove translation, rotation and scaling of the face in the face box. These aligned images are finally used as the input to a machine learning algorithm that can interpret facial information such as emotions. In contrast with existing work, our approach uses a single network to estimate three types information in a single pass: facial landmarks, discrete emotions and continuous emotions. This leads to a great improvement in affect estimation performance since important facial features around fiducial points can be used to build an attention mechanism. It also facilitates the implementation of such a single-step method as a lightweight model that runs in real time. Figure 2 compares the traditional approach to our method.

Specifically, in this paper, we make the following contributions:

- We propose a novel method for continuous valence and arousal estimation from images of facial display recorded in naturalistic conditions, which outperforms state-of-the-art methods by a large margin.
- We do so by proposing a novel deep neural network architecture (Fig. 3) that jointly performs facial alignment and correctly predicts both discrete and continuous emotion labels in a single pass.
- We simplify the pipeline of emotion recognition from facial imagery and provide a series of improvements to achieve a better performance at this task.

Results

We devised a convolutional neural network that takes as input an image of a person's face and estimates the person's affect on that image. This estimation is done in terms of continuous values of valence and arousal levels. We show that our method outperforms all existing methods by a large margin and compare it to the inter-agreement between human annotators.

Baseline comparison. In order to compare to the performance of a traditional deep learning approach to the target problem, we implemented a baseline approach to valence and arousal estimation. Specifically, we trained a ResNet-18 neural network[8]directly on the face images from the datasets[5–7] and on augmented data (that is to

[1]Samsung AI, Cambridge, UK. [2]Imperial College London, Department of Computing, London, UK. ✉e-mail: a.toisoul@samsung.com

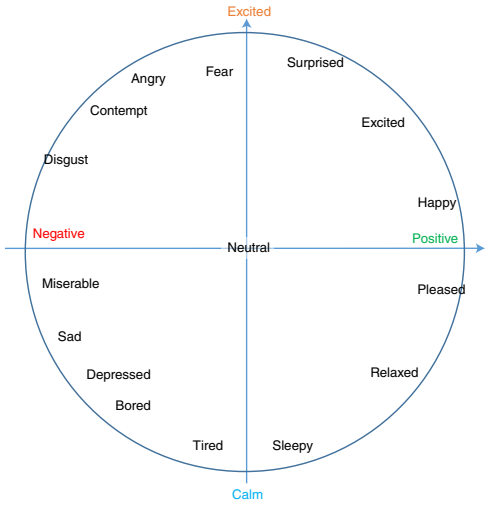

Fig. 1 | Valence and arousal circumplex. Valence and arousal space with the corresponding location of a few discrete emotional classes.

say small rotations, scaling, translations and image flips are applied during training). The network is trained to maximize the concordance correlation coefficient (CCC)[9], a metric that consistently incorporates both elements of relative error and correlation. This metric has been widely adopted for affect estimation[5,7,10–15].

Our approach. In contrast with existing methods in the field, we propose a single network that detects facial landmarks and estimates both categorical and continuous emotions. This method takes advantage of the features learnt by a face-alignment network (FAN)[16] used for facial points detection, which are also relevant for the emotion recognition task. This increases the performance of the complete model. In addition, we also introduce a series of steps that further improve the performance of the model. These steps include: a joint prediction of categorical and continuous emotions to make the network more robust to outliers in the dataset; an attention mechanism[17] that drives the focus on regions of the face that are relevant for affect estimation; a student–teacher training framework, known as knowledge distillation[18,19], that smooths labels learnt by the network; and finally a specifically tailored loss function that leads to the optimization of metrics related to affect recognition. The performance of our network can be seen in Tables 1, 2, 3 and 4. A diagram of the architecture of our network, called EmoFAN, is shown in Fig. 2 and qualitative results are presented in Fig. 4. Additional qualitative results can be found in the Supplementary Information.

Results on AffectNet. Results on the AffectNet dataset[5] are given in Table 1. Our method surpasses all existing methods by a very large margin on all metrics. In particular, we outperform the previous state of the art[14] in terms of the CCC values by 17% on the valence estimation task and by 20% on the arousal estimation task (corresponding to a 0.11 increase in the CCC values for both tasks). This is despite the fact that the approach in ref. [14] was trained on the AffectNet dataset plus an additional 2.5 million proprietary images. Similarly, our approach outperforms the single-shot detection method in ref. [15] by a very large margin. In our experiments, we also found incorrect labelling to be a big issue in the AffectNet dataset. Cleaning the validation and test sets of these incorrect labels is crucial to be able to accurately assess the performance of the models. On manually cleaned test and validation sets, the improvement on CCC over the state of the art is even greater with an increase of 0.2 for valence and 0.21 for arousal. Details on how the manual cleaning was performed are provided in the Supplementary Information.

Results on SEWA. Results on the SEWA dataset[7] are presented in Table 2. Our method outperforms the state of the art[20] in all metrics. In particular, for the CCC values, we obtain a relative improvement of 38% for the valence estimation task and 56% for the arousal estimation task (corresponding to an increase in the CCC values of 0.18 for the valence and 0.22 for the arousal).

Results on AFEW-VA. Results on the AFEW-VA dataset[6] are given in Table 3. Following previous work[6,21], we employed a five-fold person-independent cross-validation strategy to train our networks. The baseline using a ResNet-18 showed a very poor performance due to the very small size of the AFEW-VA dataset (around 30,000 images). To overcome this problem, we trained our EmoFAN network on the AffectNet dataset and then fine tuned it on the AFEW-VA dataset using a five-fold person-independent cross-validation strategy. The resulting network outperforms the state of the art on the AFEW-VA dataset by a very large margin on all metrics. In particular, for the CCC values, we achieve an improvement of 33% for the valence estimation task and 18% for the arousal estimation task (corresponding to an increase in the CCC values of 0.17 for valence and 0.1 for arousal).

Comparison to human inter-agreement. Comparison to human inter-agreement has also been made. As with any computer vision application, supervised learning methods for affect analysis from facial imagery are trained using ground-truth labels generated by human observers who are asked to manually annotate target facial imagery in terms of discrete emotions and affect dimensions (valence and arousal). Whereas in standard computer vision applications (for example, object detection) the ground truth has low inter-observer variability, annotation of facial imagery in terms of displayed emotions is particularly difficult due to the lack of context knowledge needed for accurate annotation (for example, the cultural background of the recorded person, their personality and the current task). This results in higher inter-observer variability, which is usually reported for each facial imagery dataset.

Human inter-agreement is provided for both AffectNet[5] and SEWA[7] datasets (Table 4). Interestingly, for the AffectNet dataset, if our method is treated as another annotator, its average agreement with human annotators is at least as good as that of any other annotator when estimating valence, and it outperforms the average agreement that any annotator has with other annotators when estimating arousal, with CCC values being higher by 0.2 on average. This confirms the research in psychology that found that humans are much better at judging facial affect in terms of valence than at estimating how calm or excited a person is based on their facial expression (arousal)[22,23]. On the SEWA dataset, our method's average agreement with human annotators is much higher than the average agreement human annotator reach among themselves, with CCC values being 0.38 higher for both the valence and arousal.

Applications and ethical considerations

Given that people react to and emote in response to a very wide range of stimuli (including media content, interpersonal conversations, social situations, and casual observations of people and nature around them), there are a large number of potential applications for facial affect analysis methods. The method presented in this article is non-obtrusive, using only facial imagery collected with commercially available cameras (rather than specialized cameras), which increases the number of potential applications. On the purely commercial side, the usage of such technology has proved to be very successful in market analysis, for fast and accurate judgement of whether large cohorts of paid observers like or dislike certain

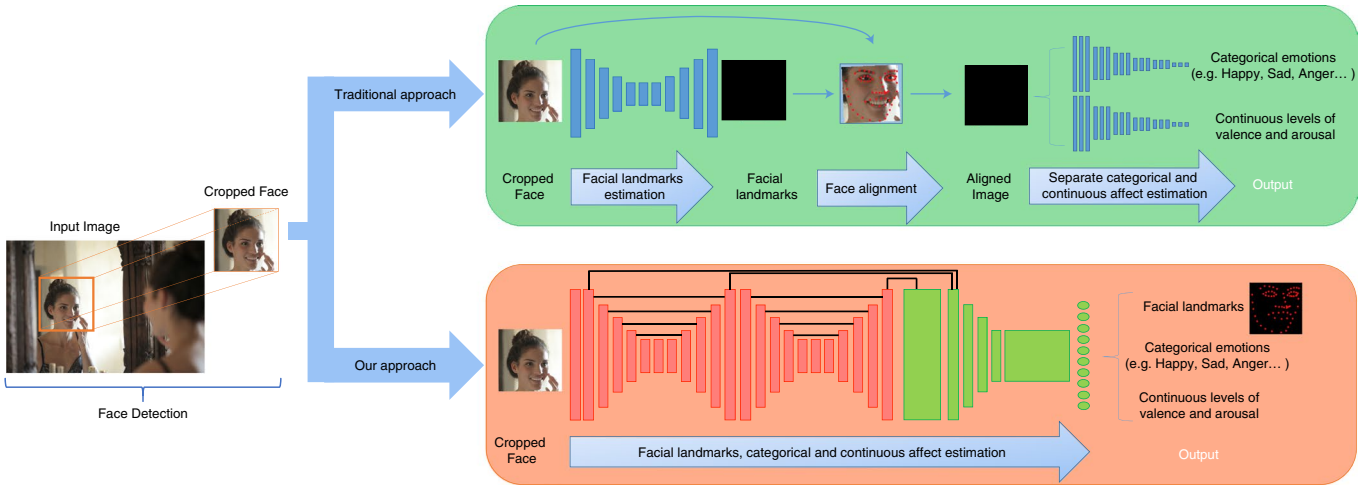

Fig. 2 | Overview of the traditional approach to facial affect recognition contrasted with our methodology. An input image contains one or several faces with different orientations, translations and scaling. A face detector is employed to find the location of each face and extract them from the input image. Top, the traditional approach first finds facial landmarks to align each face and then estimates either categorical or continuous emotions from the aligned faces (multi-step approach). Bottom, our approach directly estimates facial landmarks, discrete and continuous emotions with a single deep neural network allowing real-time predictions (single-step approach).

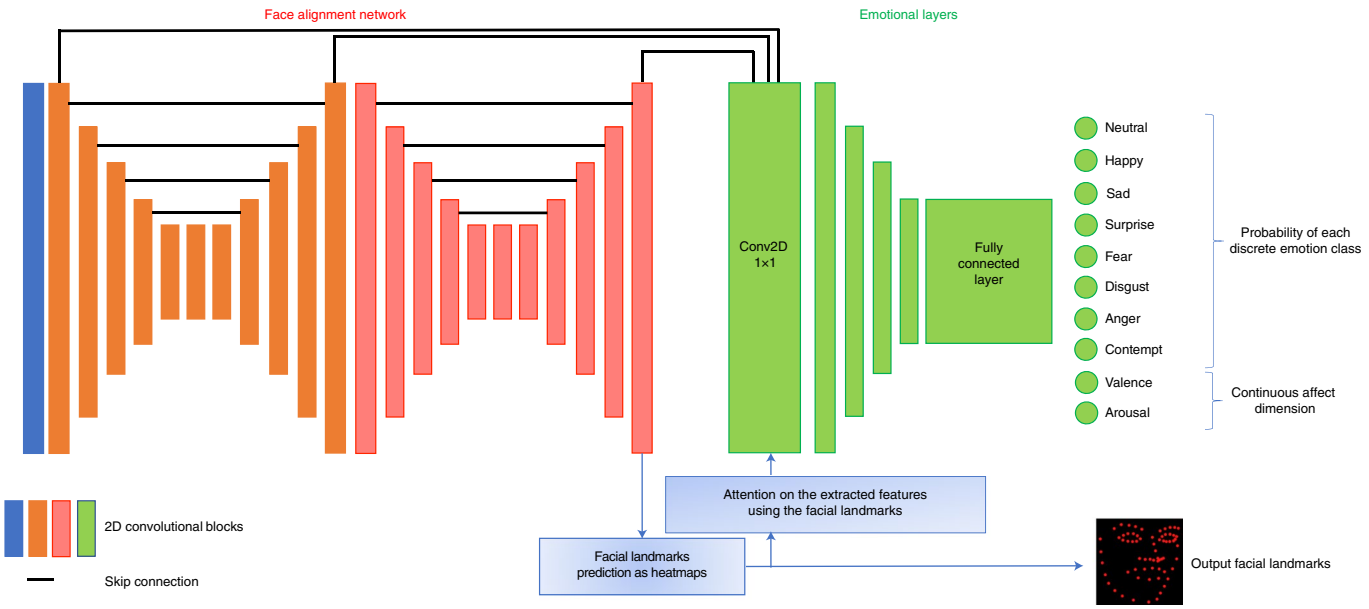

Fig. 3 | Overview of the architecture of our model. EmoFAN builds on top of the face-alignment network[16] to predict jointly discrete emotional classes, continuous affect dimensions and fiducial landmarks on the face. An attention mechanism built using these facial landmarks is employed on the feature maps to attend to the most relevant parts of the face.

Table 1 | Results on the AffectNet dataset

		Valence				Arousal			
Network	**Acc.**	**RMSE**	**SAGR**	**PCC**	**CCC**	**RMSE**	**SAGR**	**PCC**	**CCC**
AffectNet baseline[5]	0.58	0.37	0.74	0.66	0.60	0.41	0.65	0.54	0.34
Face-SSD[15]	-	0.44	0.73	0.58	0.57	0.39	0.71	0.50	0.47
VGG-FACE+2M images[14]	0.60	0.37	0.78	0.66	0.62	0.39	0.75	0.55	0.54
ResNet-18	-	0.39	0.78	0.66	0.66	0.34	0.77	0.60	0.60
Ours (original set)	**0.62**	**0.33**	**0.81**	**0.73**	**0.73**	**0.30**	**0.81**	**0.65**	**0.65**
Ours (clean set)	**0.75**	**0.29**	**0.84**	**0.82**	**0.82**	**0.27**	**0.80**	**0.75**	**0.75**

Our methods performs best on all metrics. In particular, we outperform all previous works, including VGG-FACE+2M images, in which the authors trained their model on AffectNet with an additional 2.5 million synthetic images. Unless specified, results were computed on the original test set.

Table 2 | Results on the SEWA dataset

	Valence				Arousal			
Network	**RMSE**	**SAGR**	**PCC**	**CCC**	**RMSE**	**SAGR**	**PCC**	**CCC**
SEWA baseline[7]	–[a]	–	0.32	0.31	–	–	0.18	0.20
VGG16 CNN+TRL[20]	0.33	–	0.50	0.47	0.39	–	0.44	0.39
ResNet-18	0.37	0.60	0.35	0.35	0.37	0.69	0.35	0.29
Our method[b]	**0.32**	**0.70**	**0.66**	**0.65**	**0.35**	**0.77**	**0.61**	**0.61**

[a]The dashes in the table mean that no results were provided for these metrics by the authors. [b]Our method outperforms the baseline approach as well as the previous methods on all metrics.

Table 3 | Results on the AFEW-VA dataset

	Valence				Arousal			
Network	**RMSE**	**SAGR**	**PCC**	**CCC**	**RMSE**	**SAGR**	**PCC**	**CCC**
AFEW-VA baseline[6]	0.27	–	0.41	–	0.23	–	0.45	–
VGG-FACE+109k images[14]	0.48	–	0.56	–	0.27	–	0.61	–
AffWild Net[21]	–	–	0.51	0.52	–	–	0.58	0.56
ResNet-18	0.43	0.42	0.05	0.03	0.41	0.68	0.06	0.05
Ours	**0.23**	**0.65**	**0.70**	**0.69**	**0.22**	**0.81**	**0.67**	**0.66**

VGG-FACE+109k images means that the authors trained a VGG-FACE network on AFEW-VA with an additional 108,864 synthetic images.

products and adverts. This was the crux of the successful EU SEWA project (https://sewaproject.eu) and is the main business of the two very successful small and medium enterprise companies RealEyes (https://www.realeyesit.com/) and Affectiva (https://www.affectiva.com/). On the purely research-oriented side of the application spectrum, the technology for facial affect analysis has proved indispensable in various psychological and psychiatric studies. For example, the ability to accurately extract emotional information from the face of the person one is communicating with plays a major role in prosociality[24] and this capacity is often found to be altered in numerous psychiatric conditions characterized by impaired social functioning, such as schizophrenia (psychology[25,26]; computer-based study[27]), depression (psychology[28]; computer-based study[29]) and autism spectrum disorder (psychology[30]; computer-based study[31]). Automatic facial affect technology of the kind presented here opens up tremendous potential to measure affective behaviour indicators that heretofore resisted measurement because they were too subtle or fleeting to be measured by the human eye.

It is important to note here that, although our objective is to build robust technology for automatic facial affect analysis to be used in 'humane' applications like those listed above, the technology could be potentially misused. Along with other information, this technology could be employed to more robustly identify and trace subjects and their behavioural patterns over a variety of channels (for example, telephone, webcam) and use this information for targeting in political or other aims. In turn, we argue that every application using automatic facial affect technology of the kind presented here needs to be properly audited and that the risks and merits of this technological solution relative to other solutions need to be clearly communicated to the public and all users of the target application. Specifically, we argue that adequate auditing and redress procedures, in line with the procedures that biometric technologies are now making[32] and in relation to ethical concerns specific for affective computing technology[33], are needed.

We would also like to stress that the facial affect analysis approach presented here is neither capable of recognizing 'innermost emotions' of people nor is it sensitive to cultural differences existing in display of various emotions. Specifically, the facial affect technology proposed here is able to analyse only what is portrayed on someone's face, nothing else. As a result, if someone chooses to smile in order to mask his or her feelings, the technology would fail to correctly recognize the actual affective state of that person and will recognize just the smile (that is 'happy' state and positive valence). Furthermore, neither emotional displays nor perception of emotional displays are universal; facial expressions are displayed and interpreted differently depending on the cultural background of subjects and annotators[34,35]. As a result, the data used to train facial affect analysis technologies may be biased towards facial displays present in the training data (for example, if the majority of subjects are European Americans, the method would be biased towards facial expressions typically displayed by European Americans[36]). Similarly, the data used to train facial affect analysis technologies may be biased towards how facial displays are interpreted in terms of emotions in the culture of the annotators. One version of the method presented in this paper has been trained on the AffectNet database. This database contains facial imagery of mainly European American people, which was also annotated mainly by European American annotators. Hence, the version of our method trained on the AffectNet dataset is not expected to perform with the same high accuracy presented in this paper on data coming from non-European American cultural backgrounds or on the data being annotated by non-European American annotators. This is the reason why the recent SEWA dataset[7] employs culture-dependent annotations, meaning that annotators are from the same culture as the person whose facial affect is being annotated. However, even for the version of our method trained on the SEWA dataset, it will be sensitive to cultural differences in facial affect displays only for the six cultures for which we have the data in the SEWA database (that is, German, Hungarian, Serbian, British, Greek and Chinese). For other cultures, such as African, South Asian and South American, the method will be insensitive to culture-specific differences in facial affect displays and interpretation. Of course, if such data becomes available, the method could be retrained.

Finally, we would like to emphasize that the above-mentioned auditing and redress procedures need to include auditing of the training data used for building automatic facial affect technology.

Table 4 | Comparison to human inter-agreement

			Valence				Arousal			
Dataset	**Network**	**Acc.**	**RMSE**	**SAGR**	**PCC**	**CCC**	**RMSE**	**SAGR**	**PCC[a]**	**CCC**
AffectNet[5]	Human IA	0.66	0.34	0.82	0.82	0.82	0.36	0.67	0.57	0.55
	Ours	**0.75**	**0.29**	**0.84**	**0.82**	**0.82**	**0.27**	**0.80**	**0.75**	**0.75**
SEWA[7]	Human IA	-	0.24	0.64	0.38	0.27	0.24	0.62	0.33	0.23
	Ours	-	0.32	**0.70**	**0.66**	**0.65**	0.35	**0.77**	**0.61**	**0.61**

[a]For the correlation coefficients PCC and CCC that are of interest in continuous affect estimation, our approach reaches a performance superior to the agreement between expert human annotators on both AffectNet and SEWA databases.

The issue here is that this technology could also be used for seemingly benign applications that could nonetheless end up being severely discriminatory if the technology is not trained properly. Examples include AI-empowered remote-interviewee analysis and/or rental-car driver analysis where behavioural patterns indicative of stress or intoxication could be screened. The applied automatic facial affect technology needs to be trained on appropriate demographically diverse data as to avoid situations in which certain portion of the users would be discriminated against because their age or culturally specific behavioural patterns are under-represented in the used training data. Hence, once again, we argue here that adequate auditing and redress procedures, fully in the line with the procedures for biometric technologies that are now in making[32], which address all issues including privacy considerations, training data, dual use/misuse, and ethical considerations of the choice to use affective technology in the target application, are essential.

Emotion estimation from facial imagery

Our method outperforms state-of-the-art algorithms by a large margin at the emotion recognition task. In what follows we explain the details of our approach and the changes that we have made in the architecture of our approach that led to this highly improved performance.

Datasets. In our experiments, we employ three datasets of videos and images, collected in naturalistic conditions and annotated by human expert annotators in terms of valence and arousal levels.

AffectNet. A large-scale facial imagery dataset[5] annotated in terms of discrete and continuous emotion labels (valence and arousal). It contains more than a million images downloaded from the Internet along with the annotation of 66 facial landmarks. Among these, 450, 000 images were manually annotated by twelve human annotators. The dataset contains a very large demographic variety of subjects.

AFEW-VA. A dataset of 600 video clips[6], spanning 30,000 frames with high-quality annotations of valence and arousal levels, and 68 facial landmarks, all highly accurately annotated per frame.

SEWA. A large-scale multimodal dataset containing over 2,000 minutes of audio and video data, and richly annotated in terms of 59 facial landmarks and continuous valence and arousal levels. It contains 398 different subjects from six different cultures, almost uniformly spanning an age range of 20–80.

Performance metrics. The performance metrics used in our evaluation studies are: the RMSE, which must be minimized, and the SAGR, PCC and CCC, which all must be maximized. If Y is the predicted label, $\hat{Y}$ is the ground-truth label and μ_Y and σ_Y correspond to the mean and the standard deviation of Y, respectively, then these metrics are defined as follows.

Root mean square error (RMSE) evaluates how close predicted values are from the target values:

$$\mathrm{RMSE}(Y,\hat{Y}) = \sqrt{\mathbb{E}((Y-\hat{Y})^2)} \tag{1}$$

Sign agreement (SAGR) evaluates whether the sign of the predicted value matches the sign of the target value:

$$\mathrm{SAGR}(Y,\hat{Y}) = \frac{1}{n}\sum_{i=1}^{n}\delta(\mathrm{sign}(y_i),\ \mathrm{sign}(\hat{y}_i)) \tag{2}$$

Pearson correlation coefficient (PCC) measures how correlated predictions and target values are:

$$\mathrm{PCC}(Y,\hat{Y}) = \frac{\mathbb{E}(Y-\mu_Y)(\hat{Y}-\mu_{\hat{Y}})}{\sigma_Y\sigma_{\hat{Y}}} \tag{3}$$

Concordance correlation coefficient (CCC) also incorporates the PCC value but penalizes correlated signals with different means. In other words, if the predicted signal has a trend similar to the target signal but with values that are far from the target values (high error), it then gets penalized with a low CCC (although the PCC is high).

$$\mathrm{CCC}(Y,\hat{Y}) = \frac{2\sigma_Y\sigma_{\hat{Y}}\,\mathrm{PCC}(Y,\hat{Y})}{\sigma_Y{}^2+\sigma_{\hat{Y}}{}^2+(\mu_Y-\mu_{\hat{Y}})^2} \tag{4}$$

Combining geometric and appearance information. The traditional approach of affect analysis from facial imagery is composed of a pipeline that first pre-processes images using geometric information (in the form of facial landmarks), then extracts appearance features to perform a classification or a regression in order to attain affect recognition. This pre-processing removes translation and scaling by cropping the input image around the face, previously aligned based on a set of fiducial points (or landmarks). Therefore, accurate facial landmark detection and face alignment are essential and the accuracy of the detected landmarks has a significant impact on the performance of the algorithm. As the landmarks provided in affect datasets such as AffectNet and SEWA were originally computed using older, less accurate methods, we recomputed them using the state-of-the-art facial landmarks detector of Bulat and Tzimiropoulos[16]. All reported networks were trained using these improved landmarks and all showed an increase in their performance on the two datasets. The improvement obtained by using these more accurate landmarks can be seen in the second row of Table 5 (ResNet-18 with modality F+R).

Based on the above observation that the detection of accurate facial landmarks is essential, we go one step further and propose to merge the landmarks detection and the emotion recognition into a single novel method that incorporates both. In particular, we build on a network trained for the detection of landmarks[16], and

Fig. 4 | Qualitative results of our approach. For each image, the detected face bounding box, the facial landmarks and the corresponding predicted emotion on the valence and arousal circumplex are displayed.

Table 5 | Ablation study on the AffectNet dataset

			Valence				Arousal			
Network	**Modality**	**Acc.**	**RMSE**	**SAGR**	**PCC**	**CCC**	**RMSE**	**SAGR**	**PCC**	**CCC**
ResNet-18	O+R	-	0.39	0.78	0.66	0.66	0.34	0.77	0.60	0.60
ResNet-18	F+R	-	0.37	0.79	0.69	0.69	0.34	0.78	0.61	0.6
ResNet-18	F+R+C	0.59	0.37	0.79	0.70	0.70	0.33	0.79	0.62	0.62
EmoFAN	F+R	-	0.38	0.79	0.70	0.70	0.34	0.78	0.62	0.62
EmoFAN	F+R+C	0.60	0.36	0.80	0.71	0.71	0.33	0.80	0.64	0.64
EmoFAN	F+R+C+A	0.60	0.35	0.81	0.72	0.72	0.33	0.80	0.64	0.64
EmoFAN	F+R+C+A+D	0.62	0.34	0.80	0.72	0.72	0.31	0.80	0.64	0.64
EmoFAN	F+R+C+A+S	0.60	0.35	0.80	0.72	0.72	0.31	0.81	0.64	0.64
EmoFAN	F+R+C+A+S+D	0.62	0.33	0.81	0.73	0.73	0.30	0.81	0.65	0.65
EmoFAN (clean set)	F+R+C+A+S+D	**0.75**	**0.29**	**0.84**	**0.82**	**0.82**	**0.27**	**0.80**	**0.75**	**0.75**

Modalities: O, original landmarks; F, FAN landmarks; R, valence–arousal regression; C, discrete emotion classification; A, attention mechanism; S, shake–shake regularization between MSE, PCC and CCC losses; D, distillation.

first combine features from several layers of the FAN, as well as the heatmaps it predicted, before passing them on to a series of convolutional blocks to predict emotional labels. These features intrinsically encode both low-level facial features (such as edges located at the boundary of the facial parts) and high-level morphological features that contain the location of certain facial regions (that is, eyes, lips). Such geometric features are known to have a strong correlation with facial expressions of emotions[37], which has been empirically observed in practice for both constraint and naturalistic environments[6]. This allows us to get a single network that jointly finds facial landmarks and estimates discrete and continuous emotions. Since the FAN network is pretrained on large datasets of faces containing extreme head poses and various facial expressions, the features it extracts from images are very relevant to affect analysis as well and act as weak supervision for emotion prediction. This novel approach leads to large improvements in performance, reported in the fourth row of Table 5 (EmoFAN with modality F+R).

Joint prediction of continuous and discrete affect. Most existing work focuses on predicting either discrete emotions or affect dimensions ('continuous emotions')[5,14,21]. By contrast, we posit that the model is less likely to mislabel the data when having to predict both discrete and continuous emotions in the same facial image. As a result, unlike previous works[5,14,21], we train the network to estimate both discrete and continuous emotions jointly when annotations are available. Our network therefore becomes more robust against outliers in the dataset, that is against images that have been incorrectly annotated either in terms of discrete or continuous emotion labels. To incorporate this, we introduce a novel loss function being a sum of a cross entropy for the categorical loss (discrete emotion classes ; equation (5)) and a CCC loss (equation (6)) for the continuous affect dimension estimation.

$$\mathcal{L}_{\text{categories}}(Y,\hat{Y}) = \text{Cross entropy}\,(Y,\hat{Y}) = -\sum_{i=1}^{n}\hat{y}_i\log(y_i) \tag{5}$$

$$\mathcal{L}_{\text{CCC}}(Y,\hat{Y}) = 1 - \frac{\text{CCC}_{\text{valence}}(Y,\hat{Y}) + \text{CCC}_{\text{arousal}}(Y,\hat{Y})}{2} \tag{6}$$

This results in a further improvement of all metrics as shown in Table 5 (third and fifth rows with modality F+R+C).

Attention mechanism. Not all the information in the image is relevant for emotion classification. In particular, there is a strong evidence that areas around facial landmarks are particularly relevant while the areas on the face boundary are less relevant[6]. This is a consequence of human foveal vision, which focuses our attention on a 'central' part of the object we look at, rather than its periphery. Consequently, humans learnt through evolution to communicate visual signals by contracting central parts of their faces[38]. This information, which is not taken into account by traditional neural networks, can be incorporated via an attention mechanism[17]. The attention mechanism in our network is implemented as a multiplication of the features extracted at different levels in the FAN with the predicted facial landmarks (heatmap). Additional information on this step can be found in the Supplementary Information. This allows the network to better focus on areas of the face that are likely to be important for emotion estimation and reduces the importance of regions that are less useful. The fact that each heatmap represents the probability of the location of each landmark facilitates this. As a result it leads to an improvement in the metrics as shown in the sixth row of Table 5 (EmoFAN with modality F+R+C+A).

Knowledge distillation. Knowledge distillation[19] is a technique to improve network predictions. It works in two steps. First a teacher network is trained on a specific dataset. Then a second network, called the student network, is trained on the same dataset but using the labels predicted by the teacher network instead of the labels given in the dataset. The idea behind this process is that the teacher network has already learnt how to smooth incorrect labels in the dataset. Hence, providing these to the student network gives much cleaner data to learn from. Mathematically, as the output of each network is a distribution over the labels, we minimize the distance between the distribution predicted by the student p_s and the teacher p_t, typically using a KL divergence[18]. The KL divergence is defined, for two probability distributions p_t and p_s corresponding respectively to the predictions from the teacher and the student, as $\text{KL}\,(\text{p}_\text{t}||\text{p}_\text{s}) = \sum_{\text{i}=1}^{\text{n}}\text{p}_\text{t}(\text{i})\log\left(\frac{\text{p}_\text{t}(\text{i})}{\text{p}_\text{s}(\text{i})}\right)$. This results in the following loss term added to the loss function:

$$\mathcal{L}_{\text{distillation}}(p_s,p_t) = \text{KL}\,(\text{p}_\text{t}||\text{p}_\text{s}) = \sum_{\text{i}=1}^{\text{n}}\text{p}_\text{t}(\text{i})\log\left(\frac{\text{p}_\text{t}(\text{i})}{\text{p}_\text{s}(\text{i})}\right) \tag{7}$$

Table 6 | Investigation of the impact of each term of the proposed loss function

		Valence				Arousal			
Variations in loss function	Acc.	RMSE	SAGR	PCC	CCC	RMSE	SAGR	PCC	CCC
EmoFAN with CCC-based loss	0.60	0.36	0.80	0.71	0.71	0.33	0.80	0.64	0.64
EmoFAN with MSE loss	0.59	0.34	0.80	0.70	0.68	0.30	0.82	0.62	0.59
EmoFAN with shake-shake	0.61	0.34	0.80	0.72	0.72	0.31	0.81	0.64	0.64

The results for distillation are given in the seventh row of Table 5 (EmoFAN with modalities F+R+C+A+D), which again show an improvement.

Improving the evaluation. At the core of machine learning is the idea of minimizing the empirical risk as a good approximation to the actual risk; this circumvents the need for evaluating the joint probability distribution over the labels and the data. As a result, a crucial assumption is that the distribution of the data and their labels is the same for the training, test and validation sets. However, for the AffectNet dataset, we found that, on the validation and test sets, a large proportion of labels were incorrect. As a result, the probability distribution of the training, validation and test sets can be drastically different. This phenomenon, known as label shift, severely affects performance, since minimizing the empirical risk on the validation set no longer results in good performance on the test set. This is the reason why, in general, great care should be taken to consider such issues when interpreting the predictions. We therefore cleaned the test and validation sets by manually removing all incorrect labels. This gave us a better way to validate the hyper-parameters of our model and evaluate its performance, resulting in a large improvement in performance as shown in the last row of Table 5 (EmoFAN (clean set)). For details on how the test and validation sets were cleaned, please refer to the Supplementary Information.

Loss function. For continuous affect prediction, we are mainly interested in maximizing the correlation coefficients between the prediction and the ground-truth annotation, namely PCC and CCC. However each metric encodes important information about the target task (for example, a lower RMSE usually leads to a higher SAGR as the prediction error is lower). Therefore an optimal predictor should be able to maximize all of them while minimizing the RMSE. We encode this information by changing the loss function to a sum of four terms: a categorical loss for discrete emotions, a loss related to minimizing the RMSE, a loss to maximize the PCC and a loss to maximize the CCC. Furthermore, the regression loss is further regularized with shake–shake regularization coefficients[39] α, β and γ chosen randomly and uniformly in the range $[0;1]$ at each iteration of the training process. This ensures that the network does not only focus on the minimization of one of the three regression losses. The full loss minimized by the network is given by:

$$\mathcal{L}(Y,\hat{Y}) = \mathcal{L}_{\text{categories}}(Y,\hat{Y}) + \frac{\alpha}{\alpha+\beta+\gamma}\mathcal{L}_{\text{MSE}}(Y,\hat{Y}) + \frac{\beta}{\alpha+\beta+\gamma}\mathcal{L}_{\text{PCC}}(Y,\hat{Y}) + \frac{\gamma}{\alpha+\beta+\gamma}\mathcal{L}_{\text{CCC}}(Y,\hat{Y}) \quad (8)$$

with:

$$\mathcal{L}_{\text{MSE}}(Y,\hat{Y}) = \text{MSE}_{\text{valence}}(Y,\hat{Y}) + \text{MSE}_{\text{arousal}}(Y,\hat{Y}) \quad (9)$$

$$\mathcal{L}_{\text{PCC}}(Y,\hat{Y}) = 1 - \frac{\text{PCC}_{\text{valence}}(Y,\hat{Y}) + \text{PCC}_{\text{arousal}}(Y,\hat{Y})}{2}. \quad (10)$$

The use of this loss function leads to an overall improvement of the network performance. This can be seen in Table 6, where we study the improvement that each part of the above-defined loss function has led to. In particular, when compared to a baseline trained with a regular CCC loss (first row), the same model trained with our loss function has a lower RMSE and better SAGR without degrading the CCC values (EmoFAN with shake–shake).

Implementation. The implementation was done using open-source software, specifically PyTorch[40] for the deep learning part. We trained the networks using Adam optimizer[41] with a decrease of the learning rate by 10 every 15 epochs. All the hyper-parameters were validated using a randomized grid search. In particular, we validated the weight decay in the range $[0.0, 0.01]$, the learning rate in the range $[0.0001; 0.01]$ and the optimizer's parameters beta1 and beta2 in the range $[0.0; 0.999]$. Additional details and specifications are provided in the Supplementary Information.

Conclusions

In this paper, we investigated a deep learning approach to facial affect analysis in naturalistic conditions with an unprecedented level of accuracy. We confirmed the importance of facial geometric information for this task, information typically encoded by the location of fiducial landmarks on the face. We then highlighted the importance of the attention mechanism to focus on the most relevant part of the image for the target emotion estimation task. We identified a significant issue with the annotation of the data, which can be mitigated by further cleaning the validation and test sets as well as training the deep neural network to estimate both categorical and continuous affect labels simultaneously. The annotation issues on the training set can then be smoothed-out using model distillation. Our method incorporates all the above into a single, end-to-end trainable model which outperforms all existing work on automatic facial affect estimation by a large margin.

Reporting Summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

The datasets analysed during the current study are available from the original authors in the AFEW-VA (https://ibug.doc.ic.ac.uk/resources/afew-va-database/), AffectNet (http://mohammadmahoor.com/affectnet/) and SEWA (https://db.sewaproject.eu/) repositories. The list of cleaned images for the validation and test sets of AffectNet employed in this paper are available on the authors' Github repository (https://github.com/face-analysis/emonet).

Code availability

The pretrained network, testing code and the annotations of the cleaned AffectNet test and validation sets are available at https://github.com/face-analysis/emonet under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International Licence (CC BY-NC-ND).

Received: 13 December 2019; Accepted: 3 December 2020;
Published online: 11 January 2021

References

1. Posner, J., Russell, J. A. & Peterson, B. S. The circumplex model of affect: an integrative approach to affective neuroscience, cognitive development, and psychopathology. *Dev. Psychopathol.* **17**, 715–734 (2005).
2. Russell, J. A circumplex model of affect. *J. Pers. Soc. Psychol.* **39**, 1161–1178 (1980).
3. Gunes, H. & Schuller, B. Categorical and dimensional affect analysis in continuous input: current trends and future directions. *Image Vision Comput.* **31**, 120–136 (2013).
4. Sariyanidi, E., Gunes, H. & Cavallaro, A. Automatic analysis of facial affect: a survey of registration, representation, and recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **37**, 1113–1133 (2015).
5. Mollahosseini, A., Hasani, B. & Mahoor, M. H. AffectNet: a database for facial expression, valence, and arousal computing in the wild. *IEEE Trans. Affect. Comput.* **10**, 18–31 (2019).
6. Kossaifi, J., Tzimiropoulos, G., Todorovic, S. & Pantic, M. AFEW-VA database for valence and arousal estimation in-the-wild. *Image Vision Comput.* **65**, 23–36 (2017).
7. Kossaifi, J. et al. SEWA DB: a rich database for audio-visual emotion and sentiment research in the wild. *IEEE Trans. Pattern Anal. Mach. Intell.* https://doi.org/10.1109/TPAMI.2019.2944808 (2019).
8. He, K., Zhang, X., Ren, S. & Sun, J. Deep residual learning for image recognition. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition* 770–778 (2016).
9. Lin, L. I.-K. A concordance correlation coefficient to evaluate reproducibility. *Biometrics* **45**, 255–268 (1989).
10. Ringeval, F. et al. AV+ EC 2015: The first affect recognition challenge bridging across audio, video, and physiological data. In *Proc. 5th International Workshop on Audio/Visual Emotion Challenge* (ACM, 2015).
11. Valstar, M. et al. AVEC 2016: Depression, mood, and emotion recognition workshop and challenge. In *Proc. 6th International Workshop on Audio/Visual Emotion Challenge* (ACM, 2016).
12. Ringeval, F. et al. Avec 2017: Real-life depression, and affect recognition workshop and challenge. In *Proc. 7th Annual Workshop on Audio/Visual Emotion Challenge* (ACM, 2017).
13. Ringeval, F. et al. Avec 2018 workshop and challenge: Bipolar disorder and cross-cultural affect recognition. In *Proc. 2018 on Audio/Visual Emotion Challenge and Workshop* (ACM, 2018).
14. Kollias, D., Cheng, S., Ververas, E., Kotsia, I. & Zafeiriou, S. Deep neural network augmentation: generating faces for affect analysis. *Int. J. Comp. Vision* **128**, 1455–1484 (2020).
15. Jang, Y., Gunes, H. & Patras, I. Registration-free face-ssd: Single shot analysis of smiles, facial attributes, and affect in the wild. *Comput. Vision Image Understanding* **182**, 17–29 (2019).
16. Bulat, A. & Tzimiropoulos, G. How far are we from solving the 2D & 3D face alignment problem? (And a dataset of 230,000 3D facial landmarks). In *Proc. IEEE International Conference on Computer Vision* (2017).
17. Vaswani, A. et al. Attention is all you need. In *Advances in Neural Information Processing Systems* 5998–6008 (2017).
18. Hinton, G., Vinyals, O. & Dean, J. Distilling the knowledge in a neural network. In *NIPS Deep Learning and Representation Learning Workshop* (NIPS, 2015).
19. Furlanello, T., Lipton, Z., Tschannen, M., Itti, L. & Anandkumar, A. Born-again neural networks. In *International Conference on Machine Learning* 1602–1611 (2018).
20. Mitenkova, A., Kossaifi, J., Panagakis, Y. & Pantic, M. Valence and arousal estimation in-the-wild with tensor methods. In *14th IEEE International Conference on Automatic Face & Gesture Recognition* (2019).
21. Kollias, D. et al. Deep affect prediction in-the-wild: Aff-Wild database and challenge, deep architectures, and beyond. *Int. J. Comput. Vision* **127**, 907–929 (2019).
22. Russell, J., J.A, B. & Fernandez-Dols, J. Facial and vocal expressions of emotions. *Annu. Rev. Psychol.* **54**, 329–349 (2003).
23. Grimm, M. & Kroschel, K. in *Robust Speech* (eds Grimm, M. & Kroschel, K.) Ch. 16 (IntechOpen, 2007).
24. Marsh, A. A., Kozak, M. N. & Ambady, N. Accurate identification of fear facial expressions predicts prosocial behavior. *Emotion* **7**, 239–251 (2007).
25. Clark, C. M., Gosselin, F. & Goghari, V. M. Aberrant patterns of visual facial information usage in schizophrenia. *J. Abnorm. Psychol.* **122**, 513–519 (2013).
26. Kring, A. M. & Elis, O. Emotion deficits in people with schizophrenia. *Annu. Rev. Clin. Psychol.* **9**, 409–433 (2013).
27. Bishay, M., Palasek, P., Priebe, S. & Patras, I. Schinet: Automatic estimation of symptoms of schizophrenia from facial behaviour analysis. *IEEE Trans. Affective Comput.* https://doi.org/10.1109/TAFFC.2019.2907628 (2019).
28. Caligiuri, M. P. & Ellwanger, J. Motor and cognitive aspects of motor retardation in depression. *J. Affective Disord.* **57**, 83–93 (2000).
29. Dibeklioğlu, H., Hammal, Z. & Cohn, J. F. Dynamic multimodal measurement of depression severity using deep autoencoding. *IEEE J. Biomed. Health Inform.* **22**, 525–536 (2017).
30. Harms, M. B., Martin, A. & Wallace, G. L. Facial emotion recognition in autism spectrum disorders: a review of behavioral and neuroimaging studies. *Neuropsychol. Rev.* **20**, 290–322 (2010).
31. Rudovic, O., Lee, J., Dai, M., Schuller, B. & Picard, R. W. Personalized machine learning for robot perception of affect and engagement in autism therapy. *Sci. Robot.* **3**, eaao6760 (2018).
32. Boy, N., Lidén, K. & Jacobsen, E. K. *Societal Ethics of Biometric Technologies* (SOURCE Societal Security Network, 2018).
33. Cowie, R. in *The Oxford Handbook of Affective Computing* 334–348 (Oxford Univ. Press, 2015).
34. Gendron, M., Crivelli, C. & Barrett, L. F. Universality reconsidered: diversity in making meaning of facial expressions. *Curr. Directions Psychol. Sci.* **27**, 211–219 (2018).
35. Bryant, D. & Howard, A. A comparative analysis of emotion-detecting ai systems with respect to algorithm performance and dataset diversity. In *Proc. 2019 AAAI/ACM Conference on AI, Ethics, and Society* 377–382 (2019).
36. Rhue, L. Racial influence on automated perceptions of emotions. Preprint at https://doi.org/10.2139/ssrn.3281765 (2018).
37. Pantic, M. & Bartlett, M. S. in *Face recognition* (eds Delac, K. & Grgic, M.) Ch. 21 (IntechOpen, 2007).
38. Smith, F. W. & Rossit, S. Identifying and detecting facial expressions of emotion in peripheral vision. *PLoS ONE* **13**, e0197160 (2018).
39. Gastaldi, X. Shake–Shake regularization. Preprint at https://arxiv.org/abs/1705.07485 (2017).
40. Paszke, A. et al. PyTorch: an imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems* 32 (eds Wallach, H. et al.) 8024–8035 (Curran Associates, 2017).
41. Kingma, D. P. & Ba, J. Adam: A method for stochastic optimization. Preprint at https://arxiv.org/abs/1412.6980 (2014).

Author contributions

The code was written by A.T., J.K. and A.B., and the experiments were conducted by A.T. and J.K. The manuscript was written by A.T., J.K., A.B. and M.P.; G.T. helped with discussions regarding the face-alignment network. M.P. supervised the entire project.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s42256-020-00280-0.

Correspondence and requests for materials should be addressed to A.T.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

nature research

Corresponding author(s): Antoine Toisoul

Last updated by author(s): Jan 8, 2020

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see Authors & Referees and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed
☒	☐ The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☒	☐ A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☒	☐ The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☒	☐ A description of all covariates tested
☒	☐ A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☒	☐ A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☒	☐ For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐ For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐ For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☒	☐ Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	No software was used for data collection.
Data analysis	We employed the open source python libraries Pytorch, Numpy, OpenCV and scikit-image.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

The dataset analysed during the current study is available in the AFEW-VA repository, https://ibug.doc.ic.ac.uk/resources/afew-va-database/.
The dataset analysed during the current study is available in the SEWA repository, https://db.sewaproject.eu/.
The dataset analysed during the current study is available in the AffectNet repository, http://mohammadmahoor.com/affectnet/.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☐ Life sciences ☒ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Behavioural & social sciences study design

All studies must disclose on these points even when the disclosure is negative.

Study description	We propose an artificial intelligence algorithm for dimensional affect estimation (continuous valence and arousal) from images of facial display recorded in naturalistic conditions. The algorithm is evaluated quantitatively and compared with existing methods.
Research sample	Our study was performed on three publicly available datasets : AFEW-VA (https://ibug.doc.ic.ac.uk/resources/afew-va-database/), AffectNet (http://mohammadmahoor.com/affectnet/) and SEWA (https://db.sewaproject.eu/).
Sampling strategy	We followed the established procedure for these datasets and employed the train, validation and test sets provided by each dataset in order to be able to compare our method with previous work.
Data collection	No data collection has been performed in this study.
Timing	No data collection has been performed in this study.
Data exclusions	No data were excluded from the analyses.
Non-participation	No participants were dropped out/declined participation.
Randomization	Participants were not allocated into experimental groups.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology
☒	☐ Animals and other organisms
☒	☐ Human research participants
☒	☐ Clinical data

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging