Journal of Ambient Intelligence and Humanized Computing (2021) 12:57–73
https://doi.org/10.1007/s12652-020-02851-w

ORIGINAL RESEARCH

Simultaneous prediction of valence / arousal and emotion categories and its application in an HRC scenario

Sebastian Handrich[1] · Laslo Dinges[1] · Ayoub Al-Hamadi[1] · Philipp Werner[1] · Frerk Saxen[1] · Zaher Al Aghbari[2]

Received: 15 July 2020 / Accepted: 5 November 2020 / Published online: 15 January 2021

Abstract
We address the problem of facial expression analysis. The proposed approach predicts both basic emotion and valence/arousal values as a continuous measure for the emotional state. Experimental results including cross-database evaluation on the AffectNet, Aff-Wild, and AFEW dataset shows that our approach predicts emotion categories and valence/arousal values with high accuracies and that the simultaneous learning of discrete categories and continuous values improves the prediction of both. In addition, we use our approach to measure the emotional states of users in an Human-Robot-Collaboration scenario (HRC), show how these emotional states are affected by multiple difficulties that arise for the test subjects, and examine how different feedback mechanisms counteract negative emotions users experience while interacting with a robot system.

Keywords Facial expression · Valence · Arousal · HRC · AffectNet · AFEW · Aff-Wild

1 Introduction

Human-Robot Cooperation (HCR) is a vital approach to increase the efficiency of industrial workflows. Robots can perform stressful, strenuous or difficult tasks precisely and repeatedly, while humans contribute to flexibility and awareness of certain situations (Darvish et al. 2018). The development of the last years has led away from robots as autonomous systems towards cooperative partners who share a working space and work cooperatively with human employees. The main requirements for a productive HCR environment are security and acceptance of robots (Bröhl et al. 2019). This is especially true when large robots take on heavy physical tasks while operating in close vicinity to humans.

First of all, reliable control mechanisms are of importance to ensure a smooth cooperation of humans and robots (Ao et al. 2017). To ensure the safety of workers and thus increase confidence in HCR systems, approaches to avoid collisions between humans and robots are essential. Pellegrinelli et al. (2016) propose a statistical based approach, where the robot trajectories are optimized according to volumes that represent the probability of the presence of a workers arm. The path with minimal probability is then selected, while at the same time the robot speed is reduced in accordance with the expected risk of collision. Another approach by Anvaripour et al. (2019) uses force myography data and neural networks to predict human movements and increase HCR.

Apart from objective security issues, the attitude of workers based on experience but also on subjective reasons is vital for a smooth human robot cooperation. Inordinate fear related to possible physical threats might cause a negative attitude and hinder a successful cooperation. This is also true for subjective impressions of impeding (robot might be perceived as slow or unreliable) as well as social aspects (robot might surrogate human workers). Bröhl et al. (2019) discuss cultural deviations of the acceptance of robots in Germany, Japan, USA and China. They found that robot actions must be predictable based on human perception and

This work is funded by the German Federal Ministry of Education and Research (BMBF), projects 3DIMiR (03ZZ0459C), HuBA (03ZZ0470) and RoboAssist (03ZZ0448G-L), RoboLab(03ZZ0402B) and the IAIS project (ZS/2017/10/88785).

✉ Sebastian Handrich
sebastian.handrich@ovgu.de

Ayoub Al-Hamadi
ayoub.al-hamadi@ovgu.de

[1] Otto-von-Guericke University Magdeburg, Magdeburg, Germany

[2] Department of Computer Science, University of Sharjah, Sharjah, UAE

that acceptance can be increased if robot behavior respects cultural characteristics of the respective country. Workers in western countries prefer direct communication with technical systems, but indirect communication in eastern countries. The professional relevance of HCR also affects acceptance, which is, for example, low in China due to the lack of automation in industrial processes. Furthermore, the fear of losing contact with colleagues is much higher in Germany than in the US or Japan.

Human emotions can indicate threats or problems that may arise when sharing a workspace with an industrial robot and could also be a sign of subjective aspects such as fatigue, discomfort or fear. This is why automatic recognition of emotions – which are often revealed by facial expressions (Wegrzyn et al. 2017; Höfling et al. 2020), speech (Huang et al. 2019), eye-gaze (Krishnappa Babu and Lahiri 2020), contextual informations (Salido Ortega et al. 2020), or body movements (Ahmed et al. 2020) can be helpful in evaluating and improving the cooperation between humans and robots in HCR environments. Image processing techniques allow to analyze such facial expressions in order to predict the underlying emotions and deduce human intention or condition (Jeon 2017; Samara et al. 2019; Werner et al. 2019). Approaches of facial expression analysis require a successfully detected face to begin with. Thereafter, typically a set of facial landmarks is extracted, followed by the calculation of so called (AUs) which encode movements of facial muscles (Wegrzyn et al. 2017; Vinkemeier et al. 2018; Werner et al. 2017). These AUs are then delivered to a classifier predicting discrete emotion classes. Alternatively, the AUs are mapped (Al-Hamadi et al. 2016) into continuous emotion state spaces which have been developed with respect to human cognition (Kragel and LaBar 2016). Examples for this are the Pleasure-Arousal-Dominance space (PAD) by Mehrabian and Russell (1974) and the two dimensional Circumplex model by Russell (1980) with its dimensions valence (attractiveness / averseness) and arousal (intensity). Despite being sometimes depicted this way, there is not necessarily a unique mapping between discrete emotion classes and the dimensions of the continuous state space models (Hoffmann et al. 2012). It therefore is reasonable to use both discrete emotion classes and continuous values in order to better cover the latent emotional state space.

Chang et al. (2017) used a Convolutional Neural Network to compute AUs, which are then used to predict valence and arousal. Khorrami et al. (2015) employed an holistic approach predicting the valence / arousal values directly from (normalized) face images. AUs are then extracted implicitly by the Convolutional-Neural-Network (CNN). Mollahosseini et al. (2019) compared an approach based on HoG features and (SVR) to a CNN based one (AlexNet). The latter achieved better results for both predicting the emotion and the (V/A) values. To predict V/A values in video sequences, Hasani and Mahoor (2017) combined a CNN with a recurrent neural-network (RNN) to extract temporal features. Zhang et al. (2020) pursued a similar approach, using a pre-trained network to predict emotions and personalities. Li et al. (2017) trained a bidirectional RNN, which also includes future frames for predictions. In Chu et al. (2017), the authors propose to extract both spatial representations using a CNN and also temporal representations from the data using a Long-Short-Term-Memory (LSTM) model. By fusing the output of the CNN and LSTM model, their approach is able to predict Action Units with a superior performance. However, they do not perform any further emotion classification. In their approach called EmotionalDAN, Tautkute et al. (2018) extend the loss function by incorporating facial landmarks in the learning process. They report high accuracies results for the CK+, JAFFE and ISED dataset but while trained on AffectNet – do not report any results on this dataset.

Apart from optical sensors, biophysiological measurements can be used for emotion analysis. Seo et al. (2019) used machine learning to predict boredom from electroencephalogram data (EEG). Savur et al. (2019) used EEG, Electrocardiogram (ECG), Electromyography (EMG), Galvanic Skin Response (GSR), Heart Rate (HR), Heart Rate Bariability (HRV), and pupil dilation to adjust robot speed within HCR-scenarios. Höfling et al. (2020) compared biophysiological measurements with the output of commercially available facial analysis software (FaceReader). They found that while facial EMG and skin conductance correlate strongly with the subjects' self–assessments, especially facial expressions with low or negative arousal are hard to detect by optical face analysis. However, despite being strong signals, biophysiological measurements are hardly applicable in realistic HRC scenarios as the signals are significantly interfered by human movements and measurement equipment hinder the subjects actions.

A crucial prerequisite in order to train classifiers that work beyond lab-conditions is the availability of datasets that contain real-case samples with a high variation of subjects, light conditions, head poses and none cooperative features (e.g. partial occlusion). However, many datasets are limited to only a few subjects and are acquired within a controlled environment. We therefore evaluate three current in-the-wild datasets in order to investigate their applicability. In the next section, we propose a fast deep learning based approach to predict continuous valence and arousal values as well as discrete emotion classes. In the experimental section, we compare our approach to other state-of-the-art approaches. Finally, we apply our approach in an HCR scenario to predict the emotional state of human workers and examine how different feedback mechanisms counteract negative emotions users may experience while interacting with a robot system.

2 Prediction of valence/arousal and basic emotions

We aim to predict the emotional state of a person based on its facial expression, i.e. to predict both discrete emotion categories (neutral, joy, sad, surprise, fear, anger, disgust) and continuous (V/A) values. Typically, the analysis of facial expressions consists of several steps (as face and landmark detection, feature extraction and classification). The major disadvantage of those approaches is that they cannot be trained in an end-to-end manner. We, therefore, propose a network architecture which detects the users face and predicts basic emotion categories as well as (V/A) directly from the RGB image.

Network architecture and loss function The proposed network (Fig. 1) is based on the tiny YOLOv3 architecture (Redmon and Farhadi 2017) which is used in the domain of object detection. The network takes the complete RGB image as input. The output is predicted directly from the input image without using any pre-detections like facial landmarks or AUs. Unlike other object detection networks (e.g. Mask R-CNN), the network does not first generate a list of region proposals (ROIs), which then have to be processed individually. Instead, the input image is divided into an $S \times S$ grid. For each grid element, a set of K vectors

$$V = \{x, y, w, h, p_r, \mathbf{p}_c, v', a'\}_K \tag{1}$$

is predicted. Here, $[x, y, w, h]$ is the center and the size of the predicted bounding box, relative to the dimensions of the input image. p_r is a confidence score indicating the probability that the box $[x, y, w, h]$ actually contains a face (see Redmon and Farhadi 2017 for further details). The vector $\mathbf{p}_c \in \mathcal{R}^7$ denotes the class probabilities of each discrete emotion category and v', a' are predictions for the V/A values. Both v' and a' are limited to the range [0.1] using logistic activation (Eq. 6), so the actual V/A values with range $[-1, 1]$ are obtained as $[v, a] = 2([v', a'] - 0.5)$.

The loss function consists of 4 loss terms (Eq.2). L_{box} and L_{pr} are the losses for the bounding box and the confidence scores (see Redmon and Farhadi 2017). For the class probabilities, L_c, softmax cross entropy loss (Eq. 3).

$$L = L_{\text{box}} + L_{pr} + L_c + L_{v,a}. \tag{2}$$

$$L_c = -\sum_i y_i log\big(p_c(i)\big), \text{ w. } p_c(i) = \text{softmax}\big(\beta\sigma_i\big). \tag{3}$$

For $L_{v,a}$, the loss of the V/A values, we also use cross entropy loss (Eq. 4) rather than MSE due to the logistic activation. In order to do so, true valence and arousal values must be scaled to the range [0,1] (Eq. 5).

$$L_{v,a} = -y_t log\big(y_p\big) - \big(1 - y_t\big) log\big(1 - y_p\big), \text{ with} \tag{4}$$

$$y_t = \left[\frac{1}{2}\big(v_t + 1\big), \frac{1}{2}\big(a_t + 1\big)\right] \tag{5}$$

$$y_p = \big[\sigma(v'), \sigma(a')\big] \tag{6}$$

$$\frac{\partial L_{v',a'}}{\partial v, a} = \frac{\partial L_{v',a'}}{\partial y_p} \frac{\partial y_p}{\partial v', a'} = y_p - y_t \tag{7}$$

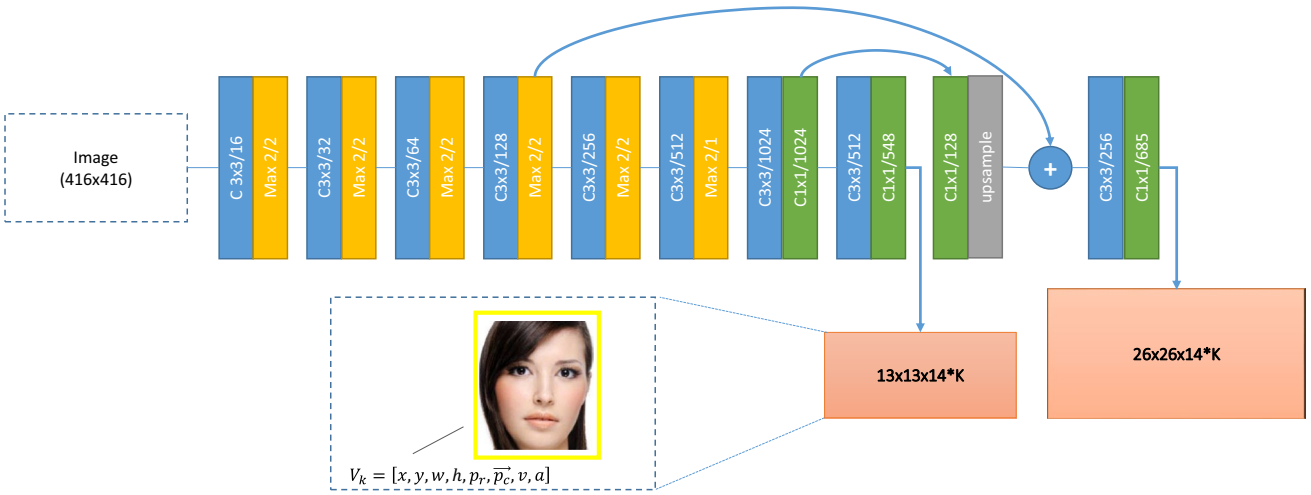

Fig. 1 Network architecture for emotion classification. A modified tiny Yolov3 model is used in order to predict discrete emotion labels and continuous valence/arousal values. The network combines features from layers of different spatial resolutions in order to better classify faces of different sizes

As stated above, for each grid element K vectors V are predicted. The reason for this is that the face bounding boxes may vary greatly in size depending on the camera distance of the person which is difficult to train with only one parameter set per grid element. Instead, a set of K predefined anchor boxes $A = \{[0, 0, w, h]\}_k$ is used, following the approach in Redmon and Farhadi (2017). To obtain these anchors, first, all face bounding boxes of the training set were clustered using k-means for different numbers of clusters $N_C = \{1, 2, 3, 5, 10\}$. We then computed the average intersection–over-union (IOU) between the face bounding boxes and the closest centroid for each number of clusters. We observed that for $N_C = 3$ the average IOU already exceeds a value of 0.8 and, therefore, used three anchor boxes and intialized them with the centroids for $N_C = 3$. While more cluster centroids would increase the average IOU even further, it would also increase the model complexity, making training process more difficult. During training, we first find the anchor box $A_{\hat{K}}$ which overlaps the most with the ground truth face bounding box B_t (using $\hat{K} = \text{argmax}\, IOU(B_t, A_K)$) and only compute gradients for the box parameters $[x, y, w, h]_{K=\hat{K}}$. Unlike for the box parameters, gradients for the V/A values and the class probabilities are computed independently of the anchors.

The network is pre-trained on the COCO dataset (Lin et al. 2014) and has 7,7M trainable parameters. The small number of parameters and the fact that the complete input image is processed in a single forward pass, make our approach very fast and capable for real time applications. This architecture allows us to learn either discrete emotion classes, V/A values or both at the same time, when available.

Datasets We use three different datasets for training and evaluation.

- *AffectNet*: With 450,000 manually annotated image samples, and almost the same number of different subjects, AffectNet is currently the most comprehensive 'in-the-wild' database for emotions (Mollahosseini et al. 2019). The ground truths cover 7 basic emotion categories as well as V/A labels.
- *Aff-Wild*: A video database containing 298 'in-the-wild' video sequences and labels for V/A for each frame (Kollias et al. 2019), but no labels for the discrete emotions.
- *AFEW-VA*: Consists of 600 short video sequences, extracted from movies (Kossaifi et al. 2017). While the videos in Aff-Wild take several minutes, AFEW-VA only contains short sequences of 10 to less than 200 frames (<30k frames in total). Thus, AFEW-VA is a small database, making it necessary to employ k-fold cross validation for a robust evaluation.

Each dataset has its benefits and drawbacks. As a video database, Aff-Wild consists of a large number of images, which, however, are highly correlated. In contrast, AffectNet has a high variance. The database contains a large number of images showing different head rotations, occluded faces, persons with different ages and skin colors and other image distortions, like e.g. text. Fig. 2 shows the V/A ground truth distribution for each database. AffectNet almost completely covers the V/A space, but - in contrast to Aff-Wild - contains only a few samples of neutral valence and strong negative arousal (boredom/sleepiness). Furthermore, the V/A distribution is similar to a circle, i.e. there are no samples with both strong valence and arousal (e.g. $v, a = [1, 1]$). We assume this is due to the annotation process and the used annotation tools. Thus, by combining different datasets, we do not only increase the variance of the input images, but also cover the V/A space in a more complete way.

Data augmentation Several augmentation strategies are used to further increase the variance of the training data. We apply standard geometric augmentations like random cropping, jittering and image flipping. About 25% of all images are converted to gray-scale images and also randomized HSV color changes are applied. Rotation was not used. Furthermore, randomly selected image parts within the face bounding box were occluded with random images from the Pascal VOC dataset (Everingham et al. 2015). This technique has been used in object detection, identification, and facial landmark detection and was also successfully applied in the field human pose estimation (Sárándi et al. 2018). Occluding random image parts acts as a regularizer by forcing the network to learn V/A values and emotion classes not only based on the most meaningful facial area (e.g. eyebrows) for a particular emotion but also include less meaningful areas (e.g. cheeks). This is particularly important if the respective facial area is not visible.

3 Evaluation of the proposed method

We run several experiments – using the AffectNet, Aff-Wild, and AFEW-VA datasets – to evaluate our approach for predicting both discrete emotion classes and continuous (V/A) values. Firstly, we describe experiments using training and test sets from the same dataset and compare our approach with other state-of-the-art approaches. Thereafter, we report results for cross-database evaluation and fusion of datasets. Using the AffectNet dataset which contains both labels for discrete emotion classes and V/A values, we then examine whether the simultaneous training of emotion classes and V/A leads to an improved classification.

Since the test set of AffectNet is not yet available, we spare 30% of the training samples for testing. For all datasets, training sample images were resized to 192x192 pixels and augmented as described in Sect.2. The network was trained for 50 epoches with a batch size of 32. The learning

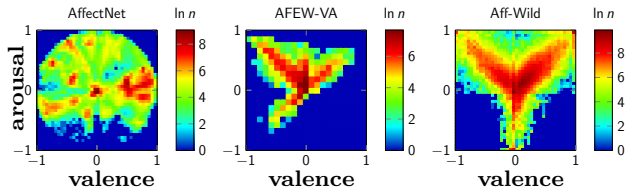

Fig. 2 Valence/arousal ground truth distribution of the **a** AffectNet **b** AFEW-VA and **c** Aff-Wild dataset

rate was initially set to 0.001 and was reduced by a factor of 0.1 whenever the training error did not decrease within 5 epoches. This happend the first time after 25 epoches. For implementation, we adapted the Darknet Framework (Redmon and Bochkovskiy 2018) to our needs. Using a single Geforce GTX 1080 TI, our approach is able to process input images with 120 fps (inference step).

Prediction of discrete emotion classes: In a first experiment, we evaluated the performance of the emotion class prediction. Since Aff-Wild and AFEW-VA have no discrete class labels, only AffectNet was used here. The results are shown in Table 1. Our approach achieved an accuracy of 75%. However, accuracy may be misleading due to highly imbalanced test and training sets. In this regard, the F_1 measure might be more suitable to compare different approaches. As expected, the proposed approach outperforms the SVM approach with HoG features of Mollahosseini et al. (2019). Compared to other deep learning based approaches, we achieve a slightly higher value for both accuracy and F_1 measure. Hereof, it should be noted that our approach does not only predict the emotion label, but also detects the face and additionally predicts V/A values.

Many of the observed class confusions occurred between an emotion and the *neutral* class. This can be explained by the fact that AffectNet contains also many borderline samples with moderate facial expressions. Also *fear* and *surprise* are confused quite often. This is expected as both emotions are close in the V/A space and expressed with similar facial expressions. However, some misclassifications were a result of incorrect labels, as shown in Fig. 3. In fact, the annotators agreed on only 60.7% of the images in the AffectNet test subset (Mollahosseini et al. 2019).

Valence / arousal values: We first evaluated the prediction of V/A values using respectively one of the three datasets for training and testing. Following Mollahosseini et al. (2019), we computed root-mean-square error (RMSE) and the Concordance Correlation Coefficient (CCC) measure. The results for AffectNet are shown in Table 1. With a RMSE of 0.282 for valence and 0.237 for arousal, the proposed approach outperforms the CNN approach by Mollahosseini et al. (2019) and the SVR based one. For Aff-Wild, the standard Aff-Wild training and test sets were used. The results and comparisons to other state-of-the-art approaches are shown in Table 2. The proposed approach shows good, but not best results in sense of RMSE and CCC. In fact, the proposed approach is outperformed by approaches that include additional features like facial landmarks or temporal features. When no additional features are used, however, the proposed approach outperforms the approaches reported by Kollias et al. (2019) and shows a slightly reduced error for the arousal estimation compared to the remaining approaches.

For AFEW-VA, we used a 5-fold cross-validation as proposed by Kossaifi et al. (2017). We trained five models on four of the folds using the remaining one for testing. Thereby, we ensured that the test fold shares no subject with any of the train folds. Results are shown in Table 3. As one can see, the deep learning approaches (Kossaifi et al. 2017) are outperformed by SVR and Random Forest based approaches, which extracts diverse handcrafted features. This is expected since the AFEW-VA database contains just about 31,000 images, which is little in the context of deep learning. However, trained on AFEW-VA, our approach still performs better than state-of-the-art DCNN and FT-DCNN. Furthermore, RMSE of valence is equal to the best approach.

Simultaneous class and valence / arousal prediction: To evaluate, whether the simultaneous training of the emotion class and the valence /arousal values is of any advantage, we firstly trained separate models for class and V/A prediction, then a single model, which learned to predict both at the same time. In both cases, AffectNet was used for training and testing. The results are also reported in Table 1. We found that – using the simultaneous approach – CCC for both valence and arousal increased by 0.019 and 0.05, respectively, while the classification accuracy improved by 4%. This clearly shows that the simultaneous learning of V/A values and discrete emotion categories is beneficial.

Cross-database evaluation To examine how well our approach generalizes to unseen data and because the number of approaches evaluated on the AffectNet dataset is still somewhat limited, we also evaluated our approach on the Aff-Wild and AFEW-VA dataset.

In Fig. 4, qualitative results for AFEW-VA and Aff-Wild depicting predicted and true V/A values are shown when AffectNet was used as the training set. For Aff-Wild (top), V/A values heavily fluctuate around the relatively steady labels. Note that the higher prediction of valence after $t = 25$, was caused by interpreting the facial expression more as fear or surprise than anger, since the model was trained on AffectNet. Inspecting the images, we believe this is a reasonable prediction. For AFEW-VA (Fig. 4 bottom), the predictions are quite close to the actual values.

Quantitative cross-database results of valence/arousal predictions are listed in Table 2 (Aff-Wild) and Table 3 (AFEW-VA) and show how a model which we trained on

the full AffectNet training set performed on Aff-Wild resp. AFEW-VA. Additionally, Table 4 shows how all models trained on one, two or three datasets perform on each of the three datasets. For this, 70% of all samples of each dataset were used for training and 30% for testing:

AffectNet Using the same dataset for training and testing, best results in sense of ICC and RMSE were achieved for AffectNet. When using AFEW-VA, Aff-Wild or a combination of both for training, the achieved results were clearly worse, as shown in Table 4. This was expected, since most regions of AffectNet were not or only poorly covered by the other datasets.

Aff-Wild In general, the cross-database experiment showed similar RMSE and CCC errors and, therefore, generalized well. However, it could not achieve the lowest errors of approaches that were directly trained on Aff-Wild. Although AffectNet covers the V/A space of Aff-Wild better than vice versa, the distribution of V/A values is very different (compare Fig. 2a and c).

AFEW-VA Table 4 shows that good results were obtained when using AffectNet for training. Nevertheless, the V/A-space of AFEW-VA also differs significantly from that of AffectNet. For example, samples of AFEW-VA with crying subjects are located in the same V/A region as angry subjects. As a consequence, the performance of our proposed CNN slightly decreases in sense of RMSE when trained on AffectNet instead of AFEW-VA. However, valence / arousal values that are not close to the expected value are generally predicted better, which explains the better ICC results (see Table 3). The reasons might be the higher number of samples within AffectNet with strong V/A intensities, as well as the general higher number of training samples.

Multiple datasets As shown in Table 4, best results in sense of RMSE were achieved using the training samples of all three datasets. Best ICC results were achieved with three or two databases, expect for cases where AffectNet was used for testing. This shows, that the high number of AffectNet samples, subjects, light conditions etc.

Fig. 3 Qualitative results for predicting discrete emotion classes (AffectNet) Top: True positives for classes anger,disgust,joy,neutral,sad and surprise. Middle: False positives due to incorrect labels in the dataset. All samples were predicted as fear but have incorrect labels anger, disgust, etc.. Bottom: False negatives due to incorrect labels. All samples have label fear, but were predicted as anger, disgust, joy, neutral, sad and surprise

Table 1 Quantitative results for emotion classes and valence/arousal on AffectNet

Approach	Class		Valence		Arousal	
	Acc	F_1	CCC	RMSE	CCC	RMSE
AlexNet[4]	72	0.57	–	–	–	–
SVM[4]	60	0.37	–	–	–	–
MSCognitive[4]	68	0.51	–	–	–	–
SVR[4]	–	–	0.340	0.494	0.199	0.400
AlexNet[4]	–	–	0.541	0.394	0.450	0.402
Proposed[1]	75	0.58	–	–	–	–
Proposed[2]	–	–	0.826	0.282	0.556	0.237
Proposed[3]	79	0.61	0.845	0.269	0.606	0.228

[1]Trained with class labels only

[2]trained with valence/arousal values only

[3]trained on class labels and valence/arousal simultaneously

[4]Mollahosseini et al. (2019)

Table 2 Quantitative results for valence/arousal prediction on the Aff-Wild dataset

Approach	Valence		Arousal	
	CCC	RMSE	CCC	RMSE
Shallow I.ResNet[4]	0.03	0.41	0.19	0.33
I.ResNet& LSTM[4]	0.04	0.40	0.29	0.30
Deep I.-ResNet[4]	0.04	0.40	0.17	0.33
CNN-M[5]	0.15	0.36	0.10	0.37
VGG Face[1], 5	0.51	**0.32**	**0.33**	**0.28**
VGG-16[5]	0.40	0.36	0.30	0.33
ResNet-50[5]	0.43	0.33	0.30	0.33
PersEmoN[6]	–	0.36	–	0.33
MM-Net[7]	–	0.36	–	0.30
Proposed[2]	0.34	0.37	0.20	0.30
Proposed[3]	**0.53**	0.35	0.10	0.36

[1]Using additional landmarks

[2]Trained on Aff-Wild

[3]Trained on AffectNet

[4]Hasani and Mahoor (2017)

[5]Kollias et al. (2019)

[6]Zhang et al. (2020)

[7]Li et al. (2017)

Bold values indicate best values

overcomes the drawback caused by the strong deviation between the datasets in V/A space. This proofs that a combination of AffectNet with Aff-Wild or AFEW-VA leads to an improved coverage of the V/A space and allows to train deep learning models which are more robust regarding real world problems.

Figure 5 shows the V/A histograms for cross-database experiments. The left diagonal show all experiments, where the training and test set where taken from the same database. In all other cases the model was trained on the training set of one and tested on the test set of another database. The coverage of the V/A-space is related to achieved results in sense of ICC. While most areas of AFEW-VA are still covered using the other datasets for training (third row), the distribution of Aff-Wild and AffectNet is clearly different due to missing connection to basic emotions within Aff-Wild labels. Since the model trained on AffectNet allows the highest differentiation of V/A predictions, we believe AffectNet is the best choice in the context of HCR scenarios.

4 HRC scenario

Apart from the evaluation on standard datasets, we use our system for emotion recognition in an human-robot-collaboration scenario (HCR). It was developed in collaboration with the Fraunhofer Institute in Chemnitz and the German car manufacturer Opel AG and models a realistic industry-related task: The assemblage of a front axle beam. To fulfill this task, human workers need to collaborate with an industrial robot in a shared working space. We employ this scenario to address the following questions: a) To what extent do human workers show negative emotions, i.e. emotions with negative valence , while sharing a workspace with a

Table 3 Evaluation (ICC(3,1) and RMSE) on AFEW-VA dataset

Approach	Feat.	Valence			Arousal		
		RMSE	CORR	ICC	RMSE	CORR	ICC
SVR[3]	S	0.28	0.29	0.21	0.24	0.43	0.36
SVR[3]	D	0.27	0.37	0.29	0.23	0.38	0.32
RF[3]	S	0.27	0.36	0.27	0.23	0.41	0.30
RF[3]	D	0.27	0.41	0.15	0.23	0.45	0.20
DCNN[3]	I	0.41	0.17	–	0.46	0.25	–
FT-DCNN[3]	I	0.37	0.26	–	0.39	0.31	–
BoW[3]	D	0.29	0.12	0.07	0.25	0.25	0.19
OR[3]	S	–	0.25	0.20	–	0.28	0.23
MKL[3]	S+D	**0.26**	0.40	0.27	**0.22**	0.45	0.34
Proposed[1]	I	**0.26**	0.39	0.32	0.25	0.29	0.21
Proposed[2]	I	0.28	**0.58**	**0.57**	0.26	**0.46**	**0.44**

Features: S = Norm-shape, D = Hybrid-DCT, I = RGB-Images

[1]Proposed, trained on AFEW-VA, 5-fold cross-validation

[2]Cross-database: Proposed model trained on complete AffectNet train set and tested on AFEW-VA

[3]Kossaifi et al. (2017)

Bold values indicate best values

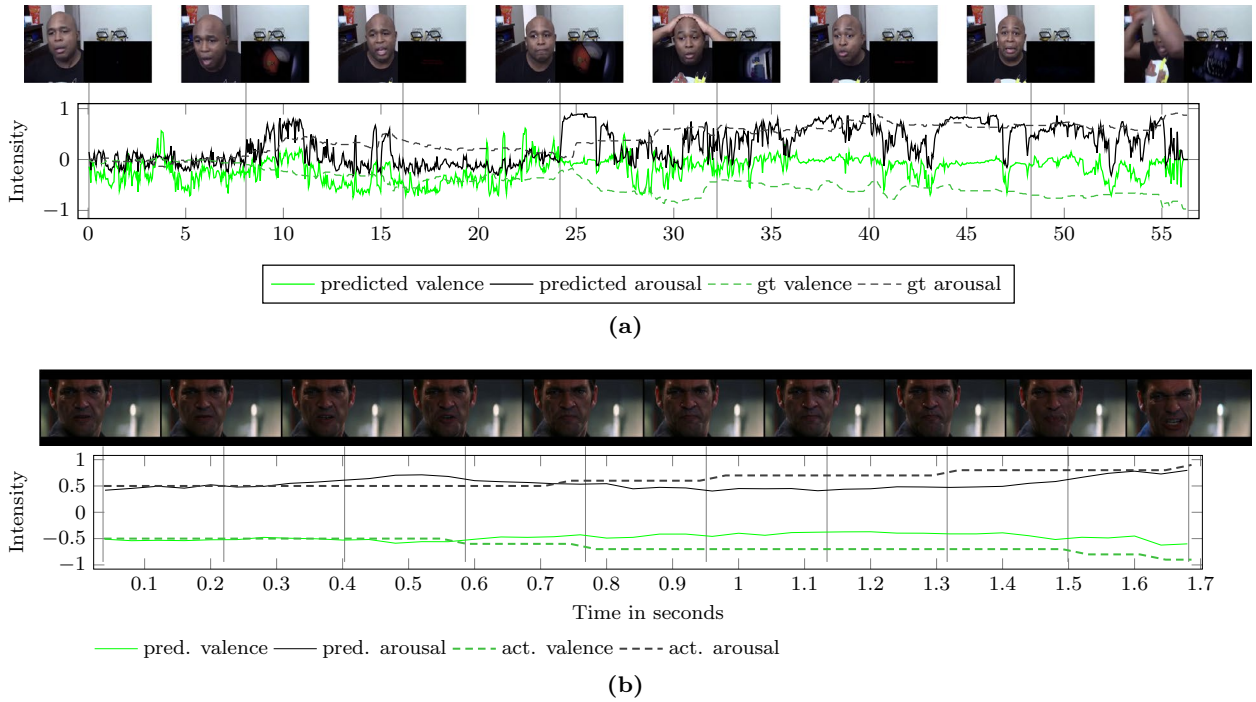

Fig. 4 Qualitative result for cross-database valence/arousal prediction for Aff-Wild (**top**) and AFEW-VA (**bottom**) using the proposed CNN trained on AffectNet

larger robot b) How do further difficulties like time pressure, control errors, and insufficient informations affect the emotional states and c) How do user feedback mechanisms reduce negative emotions and thus help to increase the acceptance of robots in industry-related environments. In the following, we describe the scenario and the experimental design in more detail, define hypotheses about expected observations, present possible measures to reduce negative emotions and evaluate their effectiveness in this scenario.

General scenario description In this scenario, the task of a human worker (user) is to attach a set of screws (8) to a front axle beam (FAB) at predefined positions. Due to its heaviness, the user cannot move the front axle beam by himself, but is dependent on the help of the robot.

The whole scenario is split into five sub-scenarios which we refer to as Baseline, S1, S2, S3, and S4. The general procedure in each scenario consists of five phases and is as follows (Fig. 6): Firstly, the robot fetches the front axle beam from a shelf (phase 1). The robot then transports the FAB into the shared working space between the robot and the user (phase 2). The user now enters the shared working space (phase 3). He has the option of using gesture control to adjust the position of the robot in order to achieve a comfortable position for attaching the screws (phase 4). Since the front axle beam is attached to the robot arm the entire time, we substituted the screws with colored markers, which have to be attached at eight predefined positions and are manually removed after each cycle. After the user has left the shared workspace, the robot puts the fully assembled front axle beam back on a shelf (phase 5). When moving, the robot uses different, predefined paths and different speeds. This was done to make it harder for the user to anticipate the robots next movements. For safety reasons, the speed of the robot had to be limited to 5 km/h. In each scenario, this general procedure is repeated 4 times (9 for baselines). We refer to this repetitions as cycles. This general procedure is modified in scenarios S1, S2, S3 as described below.

Scenario modifications: We assume that the presence of the larger robot alone, as well as the heavy workpiece attached to it and the movements in the immediate vicinity of the worker, already make the worker feel uncomfortable and induce emotions with negative valence values. In order to simulate other situations that are likely to occur in such a scenario and could weaken the user's acceptance of the robot system, we modify scenarios S1, S2 and S3 by creating further difficulties for the user (Table 5).

The first difficulty is an incorrect working gesture control. The robot arm usually follows the user's hand movements. In the event of an error, however, we invert the direction in which the robot is moving. After some time has passed, the gesture control works again as usual. This error violates several basic conditions for the acceptance of technical systems: its reliability and correct functioning, its intuitive usability, and the permanent controllability of a technical system by the user. For a second difficulty, we halve the time available to the user to assemble the screws (time pressure). Normally, the

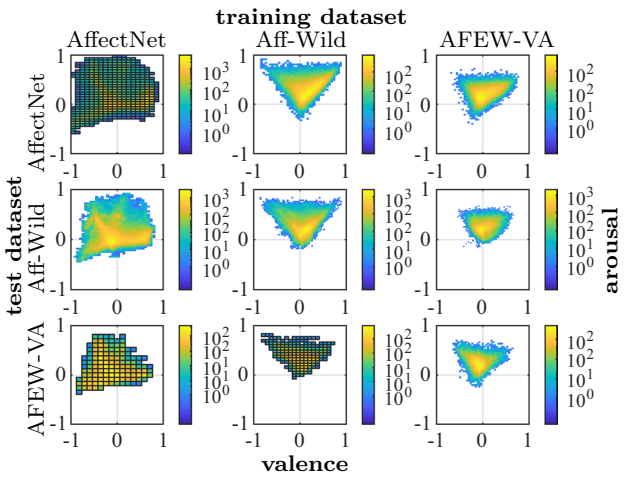

Fig. 5 Histograms of predicted valence(x-axis) and arousal (y-axis) values for the AffectNet, Aff-Wild and AFEW-VA datasets. Training and test sets are aligned as columns and row, respectively

user has T seconds to attach the screws to the front axle beam, where T is measured individually for each user by the experimenter during the baseline scenario. In case of time pressure the robot transports the front axle beam back to the shelf after $T/2$ seconds. The user is not informed about the reduction of the available time, i.e. the technical system does not provide the information necessary for a successful collaboration. From the user's point of view, it therefore appears as if the technical system is not working correctly. This and the availability of all needed information, however, is also a basic requirement for the acceptance of technical systems. We let the errors appear in cycle 1 and 3 of a scenario. The faulty gesture control occurs in S1 and the reduced available time in S3. In scenario S2, both errors occur simultaneously. No errors occur in S4.

Feedback systems: In order to investigate approaches which reduce negative effects on human workers caused by such HCR related aggravations, three basic visual feedback systems were implemented. The first feedback system is an LED strip installed on the base of the robot. In the event of an error, the LEDs flash with a red light. When the robot stops, the LEDs light up green permanently. If the robot is about to move, a rotating light indicates the next direction of rotation of the robot. Thus, by observing the LED lights, the user no longer has to guess the movement of the robot but is informed about the robots state and its next actions. If the gesture control fails, the reason for the failure and the time required for reinitialization are displayed to the user. The user is thus informed about the internal state of the technical system and also knows that it is not an operating error on his part. In the event of a reduced available time, the remaining time is displayed to the user. This way, he receives the information required to adapt his own behavior for a successful collaboration with the technical system. The latter two informations are displayed to the user on a monitor next to him.

Experimental design: The test group consisted of 48 subjects. Since reactions of workers which are not familiar with HCR are the primary research object, all 48 subjects were laymen and have no previous experience regarding cooperative work with industrial robots. Immediately before the beginning of the experiments, the subjects were

Table 4 Results on AffectNet(A), Aff-wild(B), and AFEW-VA (C) using different combinations of training sets

Set	Valence			Arousal		
↓train\ test	A	B	C	A	B	C
ICC(3,1)						
A	**0.83**	0.36	0.42	**0.60**	0.02	0.43
B	0.29	0.29	0.26	0.13	0.22	0.08
C	0.33	0.13	0.34	0.10	0.08	0.13
C + B	0.36	0.28	0.44	0.15	0.21	0.24
B + A	0.80	0.32	0.41	0.49	**0.29**	0.32
C + A	0.82	0.36	0.54	0.55	0.16	**0.48**
A + B + C	**0.83**	**0.37**	**0.54**	0.58	0.27	0.40
RMSE						
A	0.28	0.39	0.34	0.24	0.42	0.32
B	0.47	0.30	0.41	0.37	0.37	0.34
C	0.48	0.33	0.30	0.32	0.38	0.31
C + B	0.47	0.31	0.31	0.34	0.37	0.30
B + A	0.30	0.31	0.33	0.25	0.35	0.29
C + A	0.28	0.34	0.27	0.24	0.37	**0.26**
A + B + C	**0.28**	**0.29**	**0.26**	**0.23**	**0.36**	0.28

For all three datasets, we split all samples, using 70% for training and the remaining 30% for testing

Bold values indicate best values

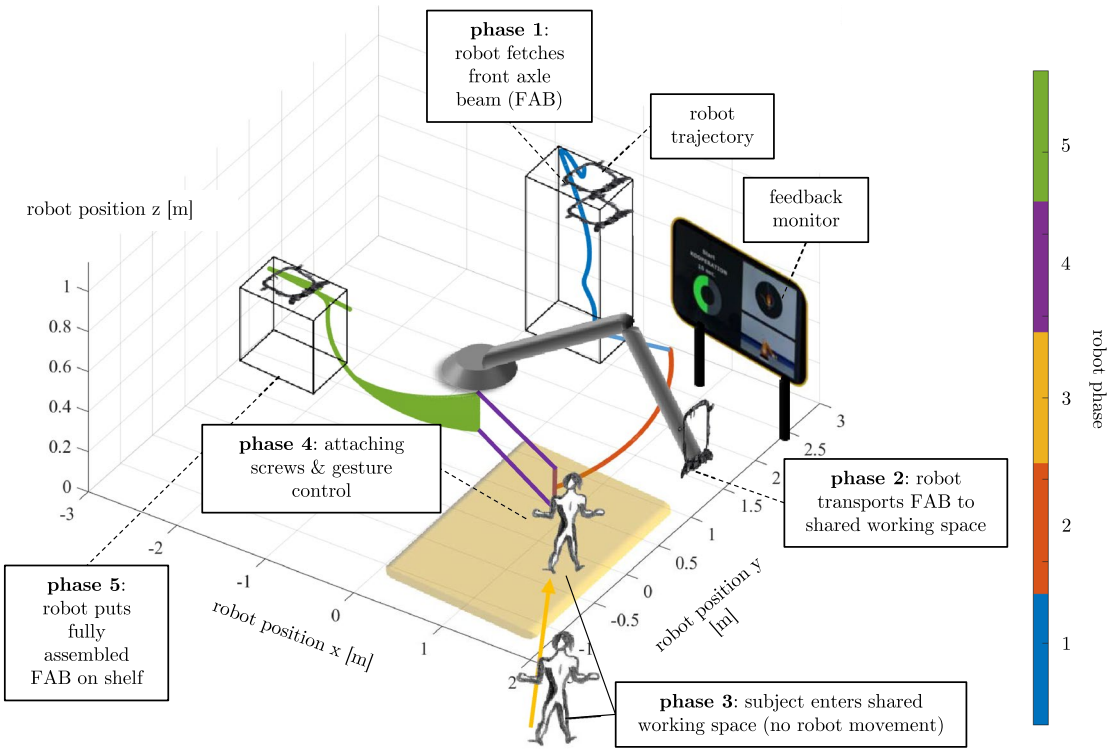

Fig. 6 Schematic overview of the HRC scenario including the robot arm trajectory for the five different phases

informed that an additional remuneration depends on how well they are able to accomplish their task. The subjects were split into two groups of equal size. One group received no automatic feedback from the system during the entire experiment, while the other group was informed about the current state of the robot using the feedback mechanisms described above. From here on, we refer to these two groups as NF-group (No feedback) and F-group, respectively. Each subject undergoes the baseline scenario and scenarios S1-S4. The baseline scenario is always the first scenario, while the other scenarios are in a random order. To be more precise: The order of the scenarios results from all possible permutations of the scenarios S1, S2, S3 and S4 (24 in total). Thus, any bias effects due to habituation or exhaustion are avoided. During the baseline scenario, examiners may also give advice to ensure that all subjects fulfill their task correctly. However, subjects in preliminary experiments tended to concentrate on the examiner behind them, and were thus influencing the sensor data and the emotions expressed. This happened especially in the event of an error. To avoid this, all subjects were isolated during the scenarios S1–S4.

The experimental setup is placed within a demonstration cell located at the Fraunhofer Institute Chemnitz and reflects typical industry-related environment conditions (including light conditions and noise from neighbored sections of the hall). Throughout the entire experiment, the users' faces were captured with a standard Logitech Webcam and their facial expressions were analyzed using the proposed approach. We first tried to place the camera at a stationary position but were unable to find a location that would allow the user to be recorded permanently and was not within the robot's operating range. We therefore decided to use a body-worn camera aiming the users face. For this purpose a self-constructed device was used (Fig. 7). Furthermore, human

Table 5 Overview of the experimental trials

Name	#	Desc	Time [min]
Interview	1	Inquire about demography and relation to technology. General instructions.	15
Baseline (S0)	9	Exercising / instructions	15
S1	4	Gesture error	6
Pause	1		4
S2	4	Time pressure and gesture error	6
Pause	1		4
S3	4	Time pressure	6
Pause	1		4
S4	4	Normal	6

poses and voices were recorded, but this is beyond this work. A preliminary evaluation shows, however, that estimating emotions from pose and speech was not successful. The is mainly due to the fact that the subjects hardly said anything as they had to be isolated from the examiner. The subjects also did not show any emotional reactions regarding their pose. We believe this is because the actions shown by the subjects are pure control actions in order to fulfill the assembly task but not any kind of movements/gesture typically shown in a communication setting. However, we would like to emphasize that these are only preliminary results that are not actually part of this work.

Hypotheses The following hypotheses were set up prior to the experiment.

- $\mathcal{H}_1$: Standard datasets often contain samples with strong facial expressions sometimes played by actors. In contrast, we do not expect such strong facial expressions and strong emotions in the HRC experiment, but rather emotions that are close to the neutral range. For such weak expressions, the categorization in basic emotions is insufficient. We therefore focus on measuring valence and arousal values, which we expect to be low in general.
- $\mathcal{H}_2$: The errors (gesture control error and time pressure) have a negative effect on the emotional states of the users. We therefore expect lower valence values in scenario S1, S2, and S3 compared to scenario S4 for the subjects without any feedback.
- $\mathcal{H}_3$: The feedback mechanisms have a positive effect on the emotional states of the users (F-group). On average, we therefore expect a higher valence compared to users without any feedback (NF group).
- $\mathcal{H}_4$: In scenario S2 two errors occur simultaneously. Their effects are supposed to be compensated by the feedback mechanisms. Since users in the NF-group do not receive any feedback, we - compared to all other scenarios - expect the biggest difference in valence between the NF-group and F-group in scenario S2.
- $\mathcal{H}_5$: In scenario S4 no errors occur and regarding this fact it does not differ from the baseline scenario. So, apart from the display of the robot direction, there are no other feedback mechanisms. We therefore expect the difference in valence between the NF- and F group to be rather small and similar to that of the baseline scenario.

In order to test these hypotheses, we compare the average valence of the NF-group with that of the F-group at different times in each scenario. As described above, each scenario consists of four cycles (nine for baseline) and each cycle consists of five phases. When calculating the average valence values within a group, it must be considered that the duration of individual phases can differ among users and that for a given time t users may therefore be in different phases and cycles of the scenario. With simple averaging, this would lead to a higher weighting of users with longer phases and valence values from different phases would be averaged, respectively. To avoid this, we split each phase into ten bins of equal size and calculate the mean valence $\bar{v}_{b,u}$ for each bin b and user u. This results into $N_b = 450$ bins for the baseline and $N_b = 200$ bins for the scenarios. The valence is then averaged over all users in the NF-group and F-group, respectively.

$$\bar{v}_{NF}(b) = \frac{1}{|NF|} \sum_{u \in NF} \left(\bar{v}_{b,u}\right), \tag{8}$$

$$\bar{v}_{F}(b) = \frac{1}{|F|} \sum_{u \in F} \left(\bar{v}_{b,u}\right). \tag{9}$$

Experimental Evaluation To give a first overview of the experimental results, Fig. 8 shows the distributions of the measured valence / arousal values for the NF- and F-group. It can be seen that the subjects of the NF- and F-group show similar values on the y axis (arousal) but have different ones on the x axis (valence). In scenario S4, where no error occurred and therefore no feedback mechanisms were used, the V/A distributions of the NF- and F-group are almost identical. In the event of an error (scenario S1 and S3), the histogram of the NF groups shows a shift towards more negative valence values whereas the subjects of the F-group show more positive valence values. This effect is greatest in scenario S2, in which the two errors occurred in parallel.

Figure 9, a-e shows the measured average valences $\bar{v}_{NF}$ and $\bar{v}_{F}$ in each scenario (baseline, S1-S4) of the NF group (no feedback, blue graph) and F-group (with feedback, red graph), respectively. The abscissa represents the bins as described above. The end of a cycle is indicated by the dotted lines. The blue and red numbers are the average valences of the two groups in each cycle. For the entire sub-scenario, the average valence of both groups are denoted as v_{NF} and v_{F} and are shown above each plot and

Fig. 7 Construction for a body-worn camera to ensure the permanent analysis of subjects facial expressions

separately in Fig. 9g. Their difference is $\Delta v = v_F - v_{NF}$. In the following, we will discuss this Figure in detail:

(i) The first general observation is the small range of measured valence values. For scenarios 1-4, the valences are in the interval [-0.1, 0.1] and therefore rather reflect the neutral range. However, this corresponds to our expectations (hypothesis $\mathcal{H}_1$), since in an HRC scenario it is reasonable to assume that when interacting with a robot test subjects do not show excessive facial expressions and experience intense emotions such as great anger or deep grief.

(ii) The second general observation is the somewhat noisy looking signal. There are two reasons for this: Firstly, the signal represents the mean of all subjects in a group. While some subjects show facial expressions in certain situations, other subjects do so with a delay or not at all. Secondly, facial expression analysis in general measures rather short-term expressions and reactions rather than longterm emotions which would result in a more smooth signal. Nevertheless, there are significant differences and similarities between the two groups, which we will discuss below.

(iii) High valence in baseline scenario: Compared to all other scenarios, the test subjects of both the F-group and NF-group showed the highest valence during the baseline scenario (e.g. $v_{NF} = 0.056$ vs. $[0.003, -0.012, -0.003, 0.014]$). We suspect several reasons for this. Firstly, the subjects were still allowed to interact with the experimenter in the baseline scenario, for example to receive information on how to operate with the robot. This resulted in facial expressions with positive valence. In fact, inspecting the raw video data shows that the test subjects often smiled at the experimenter after he had given advices. To prevent this, the test subjects were isolated from the experimenter in all other scenarios. A second possible cause is the fact that the baseline scenario is always the first scenario. Compared to it, subjects in later scenarios show signs of habituation and fatigue more often (lower valence). This assumption is supported by the observation that a decrease in valence can already be observed in both groups within the baseline when the subjects get used to the scenario (Fig.9a).

(iv) Similarities between baseline and S4: The baseline scenario and scenario S4 have in common that no errors occur within the scenarios (i.e. no gesture control error and no time pressure). As a result, no feedback mechanisms are used that could affect the emotional state of the subjects in the F-group. One would therefore expect that the valence differences between the F- and NF-group in scenario S4 and the baseline are very small. The valence differences actually measured are $\Delta v = -0.003124$ for the baseline scenario and $\Delta v = -0.000833$ for scenario S4. These valence differences are close to each other and are significantly smaller compared to all other scenarios (in which feedback mechanisms are used). In fact, there are cycles in which the NF-subjects have a higher valence than the subjects of the F-group. This is only observed in S4 and the baseline scenario and does not occur in any other scenario. We interpret this observation as a confirmation of our hypothesis $\mathcal{H}_5$. Furthermore, this shows that there is no general difference between the F- and NF-group. Otherwise one would have measured a constant offset between the two groups of subjects which is not the case. While the valence difference Δv in S4 is similar to that of the baseline scenario, the absolute valence values differ significantly ($[v_{NF}, v_F] = [0.056\ 0.053]$

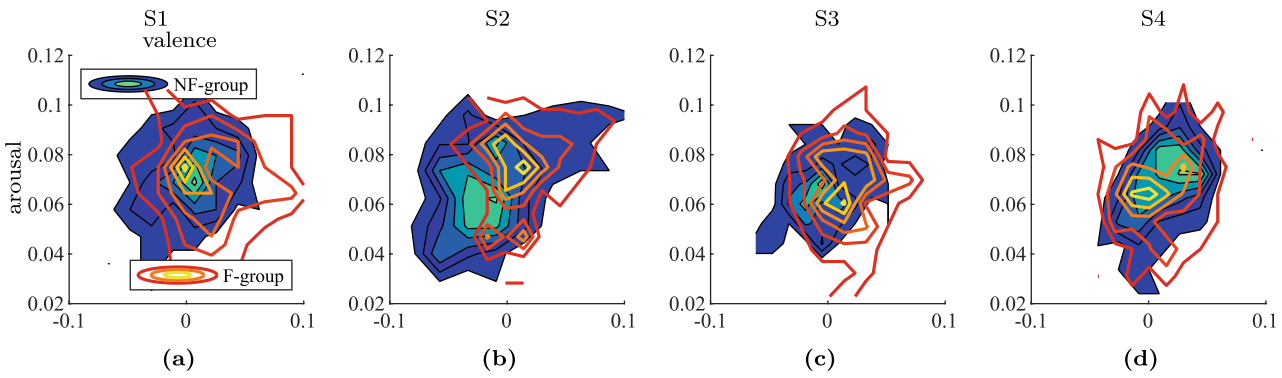

Fig. 8 Distribution of valence / arousal values in scenarios S1-S4 for the NF- and F-group. If no errors occur (S4), both distributions are almost congruent. If an error occurs, the subjects in the NF-group show a slightly more negative valence (S1+S3). The effect is largest when two errors occur in parallel (S2)

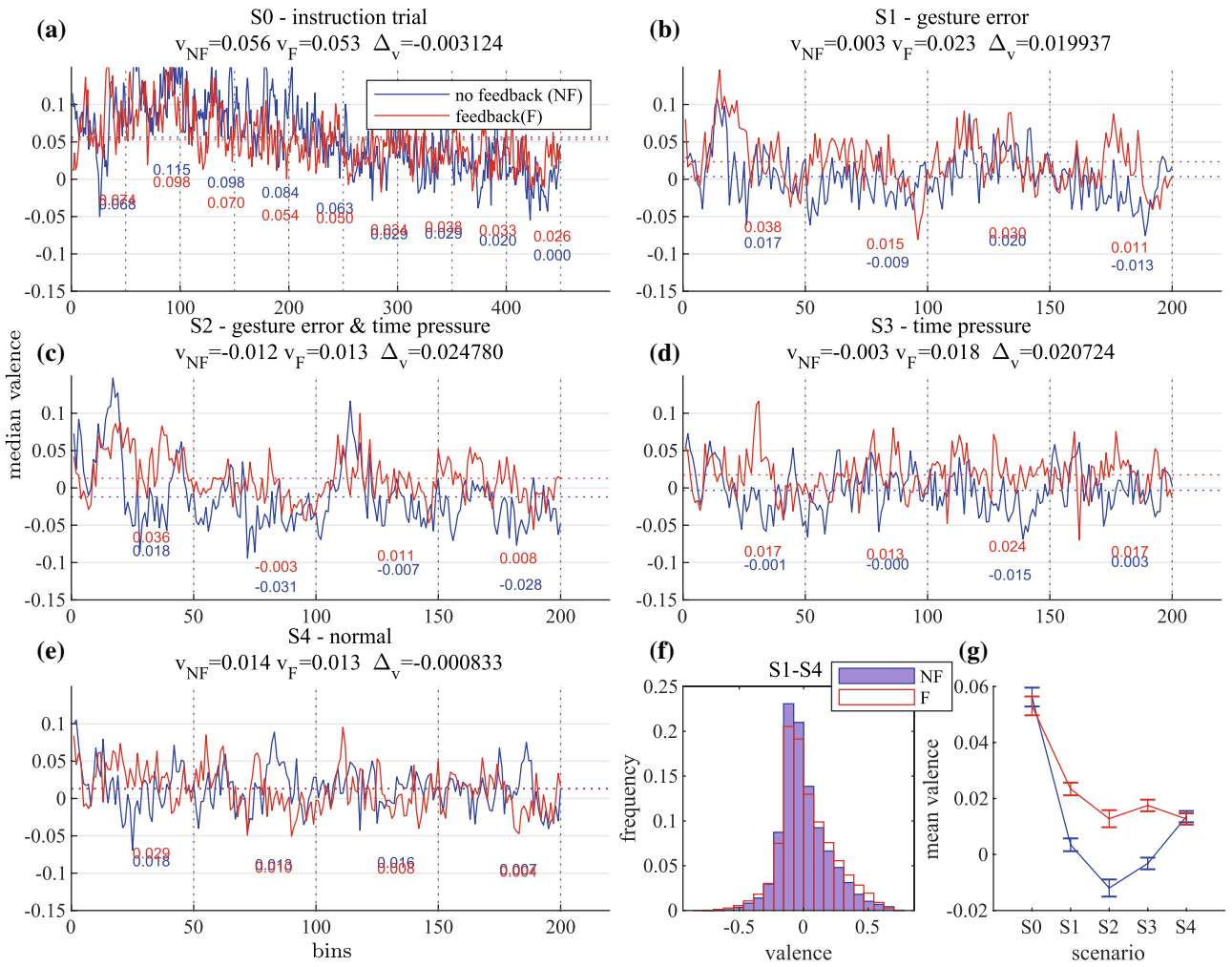

Fig. 9 Impact of scenario errors and user feedback mechanisms. **a-e:** Mean valence for subjects without any feedback (NF-group, blue graph) and subjects with feedback (F-group, red graph). In scenarios with either gesture control error (S1), time pressure (S3) or both (S2) the mean valence v_{NF} of the NF-group drops compared to scenario S4, where no error occurred. This does not apply to subjects of the F-group, whose valence is higher than that of the NF-group due to the received feedback information. **f:** Valence distribution of the NF and F-group. **g:** Mean valences (v_{NF},v_F) and standard errors for NF and F-group

(baseline) vs. [0.014 0.013] (S4)). Possible explanations for this were given in (iii).

(v) Emotional states are affected by occurring errors: The average valence of users who received no feedback is $v_{NF} = 0.014$ in scenario S4. In this scenario there are no errors (gesture control error and time pressure). The expectation is that those errors will lead to more negative emotions and lower the average valence (hypothesis $\mathcal{H}_2$). When the gesture control fails (S1), the mean valence of the NF group drops to $v_{NF} = 0.003$. Similarly, the mean valence in the scenario with time pressure (S3) drops to $v_{NF} = -0.003$. In the scenario in which both errors occur in parallel (S2), the mean valence of the NF subjects as expected drops even more to $v_{NF} = -0.012$. We interpret this as confirmation of our hypothesis $\mathcal{H}_2$. It means that the experimental design and the integrated errors are in principle suitable for inducing emotions with negative valence. Note that we only considered members of the NF-group here because the F-group experiences feedback mechanisms that are supposed to counteract negative emotions.

(vi) Effects of feedback mechanisms: As one can see in Fig. 9b-d, the mean valences of the F-group in scenarios with errors and feedback mechanisms is $v_F = 0.023$ (S1), $v_F = 0.013$ (S2) and $v_F = 0.018$ (S3). The valences are therefore higher than the corresponding ones of the NF-group both on average and in each cycle. ($v_{NF} = [0.003 \ -0.012 \ 0.003]$). As expected, this is not the case in scenario S4. Since no errors occur here, both groups receive no feedback information, so that the mean valence of both

groups is approximately the same ($v_{NF} = 0.014$ vs. $v_F = 0.013$). This means that while the valences of the NF group decrease due to the errors, the valences of the F-group remain at a comparatively high level due to the feedback mechanisms. We interpret this as confirmation of hypothesis $\mathcal{H}_3$, which states that the feedback mechanisms have a positive effect on the emotional states of the subjects.

A stronger form of hypothesis $\mathcal{H}_3$ is that the feedback mechanisms not only have a positive effect but compensate for most of the valence loss caused by the errors. In this case, one would expect that the valence difference between the F and NF-group would be larger, the more the valence of the NF group was reduced by the errors that occurred ($\mathcal{H}_4$). The measured valence differences are $\Delta v = [0.0199, 0.0248, 0.02072, -0.0008]$ for scenario S1, S2, S3, and S4, respectively. This means that $\Delta v(S2) > \Delta v(S1) = \Delta v(S3) > \Delta v(S4)$ and thus Δv(two errors) $> \Delta v$ (1 error) $> \Delta v$(0 errors). That is, the more errors reducing the valence, the larger the valence differences between the NF- and F-group due to the feedback mechanisms. In its strongest form, the hypothesis would be that the feedback mechanisms fully compensate for the loss of valence. In fact, it can be observed (Fig. 9c,e) that due to the feedback mechanisms even in the scenario with two errors (S2) the valence of the F group ($v_F = 0.013$) does not drop below the valence of the scenario without any errors (S4, $v_F = 0.013$).

In Fig. 10 we plot the robot xy-position (top view) against the shown valence values to show the influence of the robot position and phase on the emotional state of the test subjects. In phase 2 (denoted as P2 in Fig. 10), when the robot approaches the shared working area, the valence of the test subjects decreases. This applies to both groups of subjects (NF and F). In phase 4, subjects and robot work in the shared working area and the errors described above occur. Here, the subjects without feedback mostly show clearly negative valence values, while the valence of subjects with feedback mostly remains in the positive and neutral range. In phase 5, when the robot moves the front axle beam back onto the shelf, subjects from the NF-group still show mostly negative valence. It is reasonable to assume that this is due to the aftermath of the negative emotions experienced in phase 4.

So far, mean valence values were considered. As commonly known, this can be misleading as the mean is susceptible to outliers. We therefore also provide the median values (Table 6) and the distribution of the valence values (Fig. 9f). For this, the valence values of each user $\bar{v}_{b,u}$ were considered and not their mean values $\bar{v}_{NF}$ and $\bar{v}_F$. From the histogram plot in Fig. 9f it can be seen that emotions with neutral or negative valence are shown more often by users of the NF-group, while emotions with positive valence are more likely to be shown by users of the F-group. This is in line with the previous observations that the errors induce emotions with negative valence, which is compensated for the subjects of the F-group through the feedback mechanisms. From the median values (Table 6) too, it can be observed for the NF-group that the most negative valence occurs in the scenario with two errors (S2) and the highest valence in S4, the scenario without errors (apart from the baseline scenario). Furthermore, the differences to the F-group mentioned above are also evident in the median values.

Table 6 (bottom) shows the median values of the arousal measurement. In contrast to valence, we did not find any significant difference between the F- and NF-group. This may be due to the fact that the test subjects show no changes in the arousal or that the experimental setup is not suitable for inducing emotions that differ in their arousal. Here, one has to consider that strongly different emotions such as joy or anger are characterized by similar arousal values and only differ by their valence. However, we believe it is more likely that the proposed method is not sensitive enough to measure the low arousal values in a real-world scenario. In fact, the proposed method, although comparable to other state of the art methods, achieved lower ICC (3.1) values for arousal on the Aff-Wild, AFEW-VA, and AffectNet dataset when compared to valence (Table 4). The major problem here is the small number of examples with low or negative arousal (Fig. 2), even with several datasets combined. Nevertheless, we believe that valence is the more important measure when it comes to assessing a positive or negative attitude towards a robot and evaluating the acceptance in a collaborative environment.

5 Conclusion

In this work, we addressed the problem of facial expression analysis to deduce the emotional state of subjects in an industry related Human-Robot Cooperation (HRC) scenario. Therefore, we proposed an deep learning approach that is based on the YOLO architecture (you-only-look-once) and has been trained and tested on several comprehensive in-the-wild datasets.

The network has been designed to simultaneously predict the face bounding boxes, basic emotions and V/A values. Compared to other state-of-the-art models, our CNN is small and consists of only ten convolutional layers. As a consequence, it is suitable for real-time applications even with multiple cameras. We evaluated the capability of the proposed network to predict valence/arousal values and discrete emotion classes on the AffectNet database. For V/A predictions, we also performed a cross-database evaluation on the Aff-Wild and AFEW-VA dataset and found that our approach generalizes well and produces reliable predictions

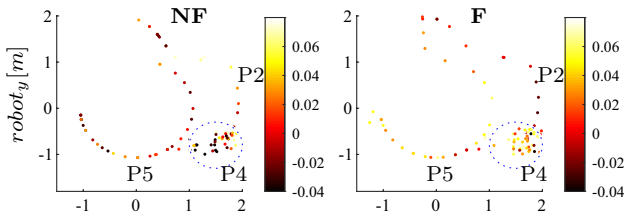

Fig. 10 Predicted valence vs. xy-robot positions / robot phases (top view). The dashed circle indicate phase 4, when the user and robot share a working space. Left: Subjects without feedback (NF-group) show mostly negative valence values in and after phase 4, while the valence of subjects with feedback (F) remains mostly positive (right)

Table 6 Mean, standard error, and median for valence and arousal of the NF and F group in scenario S0–S4

Scenario	NF		F	
	$\varnothing(S_E)$	$Q_{0.5}$	$\varnothing(S_E)$	$Q_{0.5}$
Valence				
S_0	0.056 (0.0022)	0.025	0.053 (0.0016)	0.033
S_1	0.004 (0.0023)	− 0.02	0.023 (0.0025)	0.022
S_2	− 0.012 (0.0031)	− 0.038	0.012 (0.0021)	0.009
S_3	− 0.003 (0.0021)	− 0.01	0.017 (0.0019)	0.018
S_4	0.013 (0.002)	0.00	0.012 (0.002)	0.008
Arousal				
S_0	0.07 (0.0009)	0.06	0.079 (0.0008)	0.058
S_1	0.071 (0.0013)	0.069	0.068 (0.0012)	0.053
S_2	0.067 (0.0014)	0.062	0.073 (0.0013)	0.068
S_3	0.064 (0.0011)	0.067	0.066 (0.0013)	0.049
S_4	0.067 (0.0013)	0.064	0.065 (0.0014)	0.043

even beyond laboratory conditions when trained on AffectNet. This is crucial to ensure that the model is also capable to handle unseen faces, camera views and light conditions of the investigated HCR scenario. Using multitask learning lead to a further improvement both for the detection of discrete basic emotions and for the determination of valence and arousal values. We also found that regression of valence and arousal is more suitable than emotion classification in case of real life expressions, which might be moderate compared to expressions which are performed by actors or acquired under lab conditions. Even though the training samples of AffectNet already show a wide range of variations in lighting, pose and more, we further increase variation by data augmentation including random occlusion that reduced overfitting and the prediction error. Furthermore, we showed how results on Aff-Wild and AFEW-VA benefit from fusing training samples from multiple datasets.

Finally, we applied the proposed approach to analyze facial expressions in an industrial HCR scenario, where human workers have to collaborate with a large industrial robot in order to fulfill an assemblage task.

Various difficulties, which we assume are likely to occur in the real world, were incorporated into the scenario in order to induce negative emotions in the test subjects. This includes control errors (subject fails to adjust heights of robot arm) and/or time pressure (granted time for task was suddenly reduced). In addition, a feedback system was integrated, which on the one hand informs test subjects about the next robot movements and on the other hand informs them of the internal state of the robot system in the event of difficulties, so that successful cooperation with the system and completion of the assemblage task remains possible.

First of all, we found that human workers show only slightly negative valence values and typically no high arousal when they share a workspace with a large robot (likelihoods for fear, anger, surprise or sadness were even more minor). One ascertained reason was the – due to safety reasons – limited robot speed, another a high level of trust in the overall system. Furthermore, we found that the induced difficulties as expected did not cause strong expressions of negative emotions which would indicate fear, anger or frustration. This applies for both, the manually examination of the captured facial expressions and the CNN based predictions. However, for such aggravated conditions, subjects who received no feedback information from the system showed a lower mean and median value of valence and thus experienced more negative emotions.

Finally, we found that – inducing the same difficulties as before – subjects who received feedback from the system and therefore knew its internal state, showed a significantly higher valence. This shows that the provided feedback information is sufficient to reduce negative emotions. We assume that by knowing the internal state, especially in the event of an error, the system remains more controllable for the user which is a necessary condition for successful cooperation between humans and technical systems in order to strengthen human trust in a such system and thus increase their acceptance. Since the valence values shown are rather small and changes only occur on average, we do not believe that it is possible, for example, to set a fixed threshold value under which feedback mechanisms can be used to specifically compensate for negative emotions experienced in an actual industrial environment. It would also involve labor law problems, as it would require permanent monitoring of workers and an assessment of their emotional state. Nevertheless, we could show that the proposed classification system is suitable for detecting even minor changes in facial expressions and that – by comparing two groups – the proposed approach can be used to evaluate the effectiveness of mechanisms for compensating negative emotions experienced in realistic HRC scenarios. However, one has to keep in mind that the approach – like all systems for facial

expression analysis – has been trained on labels which define the observed range of facial expressions. Hence, the CNN rather predict the shown facial expressions than the hidden subjective emotions (which might deviate).

Limitations of the described HCR study are given by the fact that arousal values are less sensitive to slight variations of facial expressions. In particular, negative arousal values are hard to predict due to the very few available training data. In addition, the body-worn camera is a self-constructed one-off production and has not been designed for everyday use. Hence, it is practically limited for experiment trials.

In future work, we intend to use the proposed approach to predict whether a user agrees with the current actions, suggestions, and decisions of a technical system in an HCR scenario. This includes self-initiated actions of a mobile assistance robot that predicts whether a user may need help and, if so, approaches. This can be done from far using pose estimation and verified in the users' vicinity by analyzing its facial expression.

Funding Open Access funding enabled and organized by Projekt DEAL..

Compliance with ethical standards

Conflict of interest The authors declare they have no conflict of interest.

References

Ahmed F, Bari ASMH, Gavrilova ML (2020) Emotion recognition from body movement. IEEE Access 8:11761–11781. https://doi.org/10.1109/ACCESS.2019.2963113

Al-Hamadi A, Saeed A, Niese R, Handrich S, Neumann H (2016) Emotional trace: mapping of facial expression to valence-arousal space. Br J Appl Sci Technol 16(6):1–14. https://doi.org/10.9734/BJAST/2016/27294

Anvaripour M, Khoshnam M, Menon C, Saif M (2019) Safe human robot cooperation in task performed on the shared load. Proc IEEE Int Conf Robot Autom. https://doi.org/10.1109/ICRA.2019.8794176

Ao D, Song R, Gao J (2017) Movement performance of human-robot cooperation control based on EMG-driven hill-type and proportional models for an ankle power-assist exoskeleton robot. IEEE Trans Neural Syst Rehab Eng 25(8):1125–1134. https://doi.org/10.1109/TNSRE.2016.2583464

Bröhl C, Nelles J, Brandl C, Mertens A, Nitsch V (2019) Human robot collaboration acceptance model: development and comparison for Germany, Japan, China and the USA. Int J Soc Robot 11(5):709–726. https://doi.org/10.1007/s12369-019-00593-0

Chang WY, Hsu SH, Chien JH (2017) FATAUVA-Net: an integrated deep learning framework for facial attribute recognition, action unit detection, and valence-arousal estimation. IEEE Conf Comput Vis Pattern Recognit Workshops. https://doi.org/10.1109/CVPRW.2017.246

Chu WS, De la Torre F, Cohn JF (2017) Learning spatial and temporal cues for multi-Label facial action unit detection. IEEE Int Conf Autom Face Gesture Recogn. https://doi.org/10.1109/FG.2017.13

Darvish K, Wanderlingh F, Bruno B, Simetti E, Mastrogiovanni F, Casalino G (2018) Flexible humanrobot cooperation models for assisted shop-floor tasks. Mechatronics 51:97–114. https://doi.org/10.1016/j.mechatronics.2018.03.006

Everingham M, Eslami SMA, Van Gool L, Williams CKI, Winn J, Zisserman A (2015) The Pascal visual object classes challenge: a retrospective. Int J Comput Vis 111(1):98–136. https://doi.org/10.1007/s11263-014-0733-5

Hasani B, Mahoor MH (2017) Facial affect estimation in the wild using deep residual and convolutional networks. IEEE Conf Comput Vis Pattern Recognit Workshops. https://doi.org/10.1109/CVPRW.2017.245

Hoffmann H, Scheck A, Schuster T, Walter S, Limbrecht K, Traue HC, Kessler H (2012) Mapping discrete emotions into the dimensional space: an empirical approach. IEEE Int Conf Syst Man Cybern. https://doi.org/10.1109/ICSMC.2012.6378303

Höfling TTA, Gerdes ABM, Föhl U, Alpers GW (2020) Read my face: automatic facial coding versus psychophysiological indicators of emotional valence and arousal. Front Psychol 11:1388. https://doi.org/10.3389/fpsyg.2020.01388

Huang Y, Tian K, Wu A, Zhang G (2019) Feature fusion methods research based on deep belief networks for speech emotion recognition under noise condition. J Ambient Intell Humaniz Comput 10(5):1787–1798. https://doi.org/10.1007/s12652-017-0644-8

Jeon M (2017) Emotions and affect in human factors and human-computer interaction, 1st edn. Academic Press Inc, Orlando

Khorrami P, Paine TL, Huang TS (2015) Do deep neural networks learn facial action units when doing expression recognition? IEEE Int Conf Comput Vis Workshop (ICCVW). https://doi.org/10.1109/ICCVW.2015.12

Kollias D, Tzirakis P, Nicolaou MA, Papaioannou A, Zhao G, Schuller B, Kotsia I, Zafeiriou S (2019) Deep affect prediction in-the-wild: aff-wild database and challenge, deep architectures, and beyond. Int J Comput Vis 127(6–7):907–929. https://doi.org/10.1007/s11263-019-01158-4

Kossaifi J, Tzimiropoulos G, Todorovic S, Pantic M (2017) AFEW-VA database for valence and arousal estimation in-the-wild. Image Vis Comput 65:23–36. https://doi.org/10.1016/j.imavis.2017.02.001

Kragel PA, LaBar KS (2016) Decoding the nature of emotion in the brain. Trends Cognit Sci 20(6):444–455. https://doi.org/10.1016/j.tics.2016.03.011

Krishnappa Babu PR, Lahiri U (2020) Classification approach for understanding implications of emotions using eye-gaze. J Ambient Intell Humaniz Comput 11(7):2701–2713. https://doi.org/10.1007/s12652-019-01329-8

Li J, Chen Y, Xiao S, Zhao J, Roy S, Feng J, Yan S, Sim T (2017) Estimation of affective level in the wild with multiple memory networks. IEEE Conf Comput Vis Pattern Recognit Workshops (CVPRW). https://doi.org/10.1109/CVPRW.2017.244

Lin TY, Maire M, Belongie S, Hays J, Perona P, Ramanan D, Dollár P, Zitnick CL (2014) Microsoft COCO: common objects in context.

European conference on computer vision. Springer, Berlin, pp 740–755. https://doi.org/10.1007/978-3-319-10602-1_48
Mehrabian A, Russell JA (1974) An approach to environmental psychology. MIT Press, Cambridge, p 266
Mollahosseini A, Hasani B, Mahoor MH (2019) AffectNet: a database for facial expression, valence, and arousal computing in the wild. IEEE Trans Affect Comput 10(1):18–31. https://doi.org/10.1109/TAFFC.2017.2740923
Pellegrinelli S, Moro FL, Pedrocchi N, Molinari Tosatti L, Tolio T (2016) A probabilistic approach to workspace sharing for humanrobot cooperation in assembly tasks. CIRP Ann Manuf Technol 65(1):57–60. https://doi.org/10.1016/j.cirp.2016.04.035
Redmon JC, Bochkovskiy A (2018) Darknet framework. Retrieved from https://github.com/AlexeyAB/darknet
Redmon J, Farhadi A (2017) YOLO9000: better, faster, stronger. IEEE Conf Comput Visi Pattern Recognit. https://doi.org/10.1109/CVPR.2017.69010.1109/CVPR.2017.690
Russell JA (1980) A circumplex model of affect. J Personal Soc Psychol 39(6):1161–1178. https://doi.org/10.1037/h0077714
Salido Ortega MG, Rodríguez LF, Gutierrez-Garcia JO (2020) Towards emotion recognition from contextual information using machine learning. J Ambient Intell Humaniz Comput 11(8):3187–3207. https://doi.org/10.1007/s12652-019-01485-x
Samara A, Galway L, Bond R, Wang H (2019) Affective state detection via facial expression analysis within a humancomputer interaction context. J Ambient Intell Humaniz Comput 10(6):2175–2184. https://doi.org/10.1007/s12652-017-0636-8
Sárándi I, Linder T, Arras KO, Leibe B (2018) Synthetic occlusion augmentation with volumetric heatmaps for the 2018 eccv posetrack challenge on 3d human pose estimation. In: ECCV.
Savur C, Kumar S, Sahin F (2019) A Framework for monitoring human physiological response during human robot collaborative task. IEEE Int Conf Syst Man Cybern. https://doi.org/10.1109/SMC.2019.8914593
Seo J, Laine TH, Sohn KA (2019) Machine learning approaches for boredom classification using EEG. J Ambient Intell Humaniz Comput 10(10):3831–3846. https://doi.org/10.1007/s12652-019-01196-3
Tautkute I, Trzcinski T, Bielski A (2018) I know how you feel: emotion recognition with facial landmarks. IEEE/CVF Conf Comput Vis Pattern Recognit Workshops (CVPRW). https://doi.org/10.1109/CVPRW.2018.00246
Vinkemeier D, Valstar M, Gratch J (2018) Predicting folds in poker using action unit detectors and decision trees. IEEE Int Conf Autom Face Gesture Recognit. https://doi.org/10.1109/FG.2018.00081
Wegrzyn M, Vogt M, Kireclioglu B, Schneider J, Kissler J (2017) Mapping the emotional face. How individual face parts contribute to successful emotion recognition. PLOS One 12(5):e0177239. https://doi.org/10.1371/journal.pone.0177239
Werner P, Handrich S, Al-Hamadi A (2017) Facial action unit intensity estimation and feature relevance visualization with random regression forests. Seven Int Conf Affect Comput Intell Interact (ACII). https://doi.org/10.1109/ACII.2017.8273631
Werner P, Lopez-Martinez D, Walter S, Al-Hamadi A, Gruss S, Picard R (2019) Automatic recognition methods supporting pain assessment: a survey. IEEE Trans Affect Comput. https://doi.org/10.1109/TAFFC.2019.2946774
Zhang L, Peng S, Winkler S (2020) PersEmoN: a deep network for joint analysis of apparent personality, emotion and their relationship. IEEE Trans Affect Comput. https://doi.org/10.1109/TAFFC.2019.2951656