



Personalized models for facial emotion recognition through transfer learning

Martina Rescigno¹ · Matteo Spezialetti¹ · Silvia Rossi¹ 

Received: 30 November 2019 / Revised: 4 July 2020 / Accepted: 21 July 2020 /

Published online: 13 August 2020

© The Author(s) 2020

Abstract

Emotions represent a key aspect of human life and behavior. In recent years, automatic recognition of emotions has become an important component in the fields of affective computing and human-machine interaction. Among many physiological and kinematic signals that could be used to recognize emotions, acquiring facial expression images is one of the most natural and inexpensive approaches. The creation of a generalized, inter-subject, model for emotion recognition from facial expression is still a challenge, due to anatomical, cultural and environmental differences. On the other hand, using traditional machine learning approaches to create a subject-customized, personal, model would require a large dataset of labelled samples. For these reasons, in this work, we propose the use of transfer learning to produce subject-specific models for extracting the emotional content of facial images in the valence/arousal dimensions. Transfer learning allows us to reuse the knowledge assimilated from a large multi-subject dataset by a deep-convolutional neural network and employ the feature extraction capability in the single subject scenario. In this way, it is possible to reduce the amount of labelled data necessary to train a personalized model, with respect to relying just on subjective data. Our results suggest that generalized transferred knowledge, in conjunction with a small amount of personal data, is sufficient to obtain high recognition performances and improvement with respect to both a generalized model and personal models. For both valence and arousal dimensions, quite good performances were obtained (RMSE = 0.09 and RMSE = 0.1 for valence and arousal, respectively). Overall results suggested that both the transferred knowledge and the personal data helped in achieving this improvement, even though they alternated in providing the main contribution. Moreover, in this task, we observed that the benefits of transferring knowledge are so remarkable that no specific active or passive sampling techniques are needed for selecting images to be labelled.

Keywords Facial emotion recognition · Convolutional neural networks · Transfer learning · Affective computing

✉ Silvia Rossi
silvia.rossi@unina.it

¹ Department of Electrical Engineering and Information Technology, University of Naples Federico II, Naples, Italy

1 Introduction

Emotions play a key role in how people think and behave. Emotional states affect how actions are taken and influence the decisions. Moreover, emotions play an important role in human–human communication and, in many situations, emotional intelligence, i.e. the ability to correctly appraisal, express, understand, and regulate emotions in the self and others [58], is crucial for a successful interaction. Affective computing researches aim to furnish computers with emotional intelligence [51] to allow them to be genuinely intelligent and support natural human-machine interaction (HMI). Emotion recognition has several applications in different areas such as marketing [18], safe and autonomous driving [22], mental health monitoring [17], brain-computer interfaces [65], social security [75], robotics [55].

Human emotions could be inferred by several modalities:

- physiological signals, including electroencephalogram (EEG), body temperature, electrocardiogram (ECG), electromyogram (EMG), galvanic skin response (GSR), respiration [62];
- speech [35];
- body gestures, including facial expressions [32].

Among them, the facial expression is one of the most natural information to use and is the main channel for nonverbal communication [30]. Moreover, RGB sensors for acquiring images of the face cost significantly less than others and do not need to be worn, which makes Facial Emotion Recognition (FER) a good candidate for commercial-grade systems. Traditional FER approaches are based on the detection of faces and landmarks, followed by the extraction of hand-crafted features, such as facial Action Units (AUs) [3]. Machine learning algorithms, such as Support Vector Machines (SVMs), are then trained and employed on these features. In contrast, deep learning approaches aim to provide an end-to-end learning mechanism, reducing the pre-processing of input images [33]. Among deep learning models, Convolutional Neural Networks (CNNs) are particularly suited for facial image processing and allow to highly reduce the dependence on physics-based models [78]. Moreover, recurrent nets, in particular Long Short Term Memory (LSTM), could take advantage of temporal features if the recognition is performed on videos instead of single images [9].

FER can be formulated as a classification or a regression problem. The distinction mainly depends on the emotional model employed for representing emotions. In categorical representations, emotions consist of discrete entities, associated with labels. Ekman observed six basic emotions (anger, disgust, fear, happiness, sadness, and surprise) characterized by distinctive universal signals and physiology [11]. Tomkins characterized nine biologically based affects: seven that can be described in couples (interest-excitement, enjoyment-joy, surprise-startle, distress-anguish, anger-rage, fear-terror, shame-humiliation), representing the mild and the intense representation, plus *dissmell* (a Tomkins' neologism) and disgust [71]. In contrast with discrete representations, dimensional models aim to describe emotions by quantifying some of their features over a continuous range. Russell developed a circumplex model suggesting that emotions can be represented over a two-dimensional circular space, where axes correspond to valence and arousal dimensions [57]. The circumplex model could be extended by adding axes, in order to describe complex emotions: Mehrabian proposed the Pleasure-Arousal-Dominance (PAD) model, where the Dominance-Submissiveness axis represents the how much one feels to control versus to be controlled by the emotion [45]; Tmka

and colleagues developed a four-dimensional model describing valence, intensity, controllability, and utility [72]. Between discrete and dimensional models, Plutchik and Kellerman proposed a hybrid three-dimensional model where 8 primary bipolar emotions are arranged in concentric circles (equivalent to a cone in space), such that inner circles contain more intense emotions [52]. Which model is more suitable for emotions representation and the universality of facial expressions of emotions are still debated [2, 12, 21, 23, 38, 41]. Our article steps back from these debates and focuses on the usability of the models. We decided to adopt the dimensional model, in particular the Valence-Arousal (VA) approach, because we observed that most of the datasets that rely on discrete emotions do not refer to the same emotion labels, thus making it difficult to conduct experiments that involve multiple datasets. Conversely, almost all available datasets based on the dimensional representations are compatible with each other, after normalizing values if needed, and the VA model represents their common denominator.

Independently of the model used to represent emotions, building a general FER model is still a challenge. The “one-fits-all” approach requires the predictor to be able to generalize acquired knowledge to unseen data, but this is prevented by subjects’ differences in anatomies, cultures, and variable environment setups [59]. A possible solution could be the creation of subject/environment-specific models, but, on the other hand, it would require a considerable amount of labelled data to train the predictor. Therefore, it could be infeasible in practice since emotional labelling is an expensive process that has to be performed by expert annotators [47].

In this work, we propose a transfer learning approach to work around the problem of training a high-capacity classifier over a small, subject-specific dataset. First, we trained a general-purpose CNN, namely AlexNet [36], over the AffectNet (Affect from the InterNet) database [47], and then we exploited the assimilated knowledge to perform fine-tuning over small personal datasets of images extracted from videos, obtained from the AMIGOS dataset [46]. Our approach belongs to the category of *transductive transfer learning* since the source and target domains are different but related (since AffectNet images are not acquired in a controlled environment such as for AMIGOS) and source and target tasks are the same [49]. Our results show that transfer learning in this domain helps in improving the emotion recognition performance with respect to both personal models (i.e., trained on a subject-specific data) and generalized models (i.e., trained on all the subjects). Moreover, they show that for the considered dataset, while the valence dimension is more generalizable from different subjects, arousal depends more on individual characteristics.

Additionally, we investigated if it would be possible to reduce the number of samples needed for fine-tuning the net. To this aim, we trained the net with sets of increasing size. We observed that a very small amount of personal data is needed to achieve quite good performance, regardless of the strategy employed to select, from the whole set, the samples to be labelled. Our test employed both active and passive sampling techniques developed for regression, namely Greedy Sampling (GS) [79] and Monte-Carlo Dropout Uncertainty Estimation (MCDUE) [73], to select samples characterized by higher uncertainty that have to be added to the training set. Tested approaches rely on the assumption that there is an unlabeled dataset from which some samples must be selected and labeled (pool-based sampling). Passive sampling techniques explore the input space of unlabeled samples, while active sampling approaches also take into account a previously trained regression model to estimate sample uncertainty. In particular, the GSx variant of GS is passive, while GSy, iGS, and MCDUE follow the active sampling paradigm. Our results show that the error of the trained model decreases very fast with respect to the size of the training set so that there are no significant differences between using random sampling and active/passive sampling.

The rest of the paper is structured as follows: Section 2 discusses related work and recent results in FER from literature; Section 3 describes our approach; Section 4 exhibits and discusses the results; Section 5 contains conclusions and future developments.

2 Related work

Quite a large effort has been spent in FER during the last decades, both by employing traditional machine learning and deep learning methods. The work by Ko and colleagues [33] offers a recent and comprehensive review of this topic. Here, we report some of the most significant methods and results, also summarized in Table 1.

Among the traditional machine learning approaches, Shan and colleagues [61] extracted Local Binary Pattern (LBP) to be used as features for SVM. The authors tested the method in a cross-dataset experiment, involving Cohn-Kanade (CK) [29], Japanese Female Facial Expression (JAFFE) [43], and M&M Initiative (MMI) [50] datasets and achieving 95.1% of accuracy (6 classes) on CK. In [10], the authors proposed the use of Kernel Subclass Discriminant Analysis (KSDA), classifying 7 basic and 15 compound emotions based on a Facial Action Coding System (FACS). Average accuracy was 88.7% for basic emotions and 76.9% for compound emotions. Suk and Prabhakaran [66] classified seven emotions with 86.0% accuracy employing SVM on data from extended Cohn-Kanade dataset (CK+) [42]. Features were obtained by using Active Shape Model (ASM) fitting landmarks and displacement between landmarks. The same dataset was used in [15], where 95.2% and 97.4% of accuracy was obtained by AdaBoost and SVM with boosted features, respectively. In [82], Multiple Kernel Learning (MKL) for SVM was used to combine multiple facial features, a histogram of oriented gradient (HOG), and local binary pattern histogram (LBPH). The authors evaluated the proposed approach on multiple datasets, outperforming many of the state-of-the-art algorithms.

FER was addressed with a deep learning approach by Jung and colleagues [27] that combined two deep networks that extract appearance features from images (using convolutional layers) and geometric features from facial landmarks (using fully connected layers), obtaining 97.3% of accuracy on CK+ data. In [28] authors showed that a hybrid CNN-RNN architecture can outperform simple CNNs. Hasani and Mahoor [19] combined two well-known CNN architectures [68, 69], in a 3D Inception-ResNet, followed by an LSTM layer, to extract both spatial and temporal features. Arriaga and colleagues [1] implemented a CNN, with four convolutional levels and a final level of global average pooling with a soft-max activation function. They classified images from The Facial Expression Recognition 2013 (FER-2013) database [16], achieving 66% of accuracy.

The previously mentioned approaches deal with the classification of categorical emotions. In general, the literature offers much more contributions regarding the classification of categorical emotions than the regression of emotional dimension. For this task, visual data are often employed also for improving the recognition power of multi-modal approaches [5, 64].

Using a dimensional model, Khorrani and colleagues [31] obtained a Root Mean Squared Error (RMSE) equal to 0.107 in predicting valence, by using a CNN-RNN architecture over the AV + EC2015 (Audio/Visual+Emotion Challenge) dataset [54]. In [47], authors employed the AlexNet CNN [36] to classify and regress emotions from AffectNet [47] dataset, that contains images collected by querying internet search engines, annotated both with the

Table 1 FER related works from literature. For each article, we report: employed datasets; adopted approach (both in terms of selected features and recognition algorithms); emotional model; numerical results expressed by means of accuracy, RMSE or Concordance Correlation Coefficient (CCC)

Reference	Datasets	Approach	Emotion Recognition	Results
Shan et al. (2009) [61]	CK, JAFFE, MMI	LBP + SVM	6 expressions 6 expressions + neutral	Accuracy 95.1% Accuracy 91.4%
Du et al. (2014) [10]	Images from 230 subjects	FACS + KSDA	7 basic expressions 15 compound expressions	Accuracy 88.7% Accuracy 76.9%
Suk and Prabhakaran (2014) [66]	CK+	ASM + SVM	7 expressions	Accuracy 86.0%
Ghimire and Lee (2013) [15]	CK+	Facial Landmarks + AdaBoost + SVM	6 expressions	Accuracy 97.4%
Zhang et al. (2015) [82]	CK+, MMI, GEMEP-FERA [74]	(HOG & LBPH) + MKL + SVM	6 expressions 7 expressions	Accuracy 95.5% (CK+) Accuracy 93.6% (CK+)
Jung et al. (2015) [27]	CK+	Facial Landmarks + Deep Networks	7 expressions	Accuracy 97.3%
Kahou et al. (2015) [28]	TDF [67], FER-2013	CNN-RNN	7 expressions	Accuracy 52.9%
Hasani and Mahoor (2017) [19]	MMI, CK+, GEMEP-FERA, DISFA [44]	CNN-LSTM	5 expressions (GEMEP--FERA) 6 expressions (MMI) 7 expressions (CK+)	Accuracy 77.4% Accuracy 77.5% Accuracy 93.2%
Arriaga et al. (2019) [1]	FER-2013	CNN	7 expressions	Accuracy 66.0%
Khorrami et al. (2016) [31]	AV + EC2015	CNN-RNN	Valence estimation	Valence RMSE 0.107
Mollahosseini et al. (2017) [47]	AffectNet	CNN	8 expressions Valence and arousal estimation	Accuracy 64.0% - 68.0% Valence RMSE 0.394 Arousal RMSE 0.402
Li et al. (2017) [40]	Aff-Wild	MM-Net	Valence and arousal estimation	Valence RMSE 0.134 Arousal RMSE 0.088
Chang et al. (2017) [4]	Aff-Wild	FATAUVA-Net	Valence and arousal estimation	Valence RMSE 0.123 Arousal RMSE 0.095
Hasani and Mahoor (2017) [20]	Aff-Wild	DRC-Net	Valence and arousal estimation	Valence RMSE 0.161 Arousal RMSE 0.094
Ng et al. (2015) [48]	FER-2013 EmotiW [8]	CNN (transfer learning)	7 expressions	Accuracy 55.6%
Chu et al. (2016) [7]	CK+, GEMEP-FERA, RU-FACS [3], GFT [60]	SVM (transfer learning)	5 expressions (GEMEP--FERA)	Accuracy 86.5%

Table 1 (continued)

Reference	Datasets	Approach	Emotion Recognition	Results
Feffer and Picard (2018) [13]	RECOLA	MoEs (supervised domain adaptation)	7 expressions (CK+) Valence and arousal estimation	Accuracy 96.4% Valence CCC 0.86 Arousal CCC 0.85

categorical and the dimensional models. They achieved accuracy between 64.0% and 68.0% in classification (on balanced sample sets) and an RMSE of 0.394 and 0.402 for valence and arousal, respectively. In the context of the Aff-Wild (Valence and Arousal In-The-Wild Challenge) [80], other methods have been proposed. MM-Net [40], a variation of the deep convolutional residual neural network, achieved RMSE values of 0.134 and 0.088 for valence and arousal, respectively. In FATAUVA-Net [4], AUs were learned from images, in a supervised manner, and then employed for estimating valence and arousal (RMSE equal to 0.123 and 0.095). Lastly, DRC-Net [20] was based on Inception-ResNet variants and achieved RMSE values of 0.161 (valence) and 0.094 (arousal). In this work, we started from the classification approach proposed by Mollahosseini and colleagues [47], since our aim is to take advantage of the knowledge learned from AffectNet. In this sense, the choice of AlexNet is motivated by the promising results achieved on such dataset.

However, to the best of our knowledge, the effect of transfer learning on FER has been poorly investigated and has mainly focused on categorical emotions recognition and AU detection [6, 7, 48, 81]. The work by Feffer and Picard [13] addressed the personalization of deep neural networks for valence and arousal estimation. The authors proposed the use of a Mixture-of-Experts (MoEs) technique [24]. In MoEs frameworks each expert represents a subnetwork trained on a subset of available data, thus tuned for a specific context. A gating network weights the contribution of experts in the inference step. In [13], a ResNet architecture was followed by an experts' subnetwork, where each expert corresponded to a different subject of the training set. A supervised domain adaptation approach [26] was adopted to fine-tuning the expert network to unseen subjects. Results obtained on the RECOLA dataset [53] demonstrated the efficacy of the proposed approach.

3 Materials and methods

3.1 Datasets preparation

We used two datasets of emotionally labeled images. AffectNet is a database of facial expressions *in the wild*, i.e. images are not acquired in a controlled environment of a laboratory. AffectNet contains more than 1 million samples collected by querying internet search engines, by using emotion-related keywords (Fig. 1 shows an example set of images). Images have been preprocessed to obtain bounding boxes of faces and rescaled to a common resolution. About half of the dataset has been manually annotated, both with respect to the categorical model (seven basic emotions) and the dimensional model (valence and arousal).



Fig. 1 Samples from AffectNet dataset

The remaining images are not annotated and are not provided in order to be used for future challenges. Hence, in this work, we employed just the annotated subset.

AMIGOS (A dataset for Multimodal research of affect, personality traits, and mood on Individuals and GrOupS) [46] is a dataset consisting of multimodal signals recording of 40 participants in response to emotional fragments of videos. Signals include EEG, ECG, GSR, frontal and full-body videos, depth camera videos. All participants took part in a first experimental session, in which they watched short video fragments (<250 s), while some of them participated in a second session, where they had to watch long videos (>14 min), individually or in groups. The dataset also contains emotional annotations both by three external annotators and by the participants themselves. We included in our study individual frontal videos from 10 participants recorded during the first experiment and, since participants could be unreliable at reporting on their own emotions or willing to hide them [63], we considered the related external annotations of valence and arousal. For each short video, we extracted a frame every 4 s, obtaining about 1000 frames for each subject. We manually removed frames where the participant's face was not visible. Each frame has then been preprocessed, by using a pre-trained Cascade Classifier (CC) [76] to extract the bounding box of the user's face and resize it (see Fig. 2). Labels associated with each frame have been computed as the average of external annotators' scores.

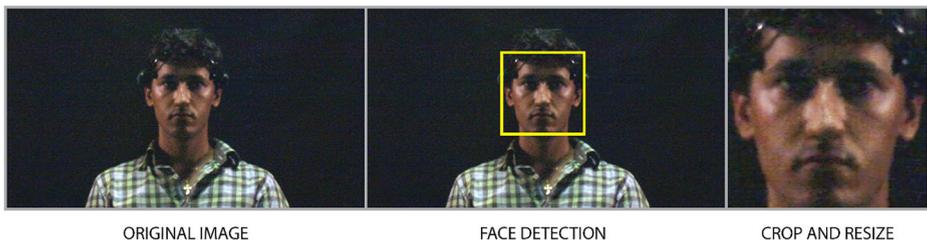


Fig. 2 Preprocessing steps shown for a sample frame from AMIGOS dataset

3.2 CNN architecture and training

AlexNet [36] is a famous CNN designed for general-purpose image classification and localization. It won the ImageNet LSVRC-2012 [56] with a large margin both in *Task 1* (Classification) and in *Task 2* (Localization and Classification). The neural network consists of five convolutional layers (the first, the second, and the fifth are followed by max-pooling layers) and three globally connected layers. AlexNet takes advantage of using Rectifier Linear Unit (ReLU), instead of hyperbolic tangent function (tanh) as activation, to introduce non-linearity. To reduce overfitting in the globally connected layers, hidden-unit dropout [77] is also employed. In its original form, AlexNet was developed to perform a classification between 1000 classes (a 1000-way softmax layer produces the output). Here, we built two twin CNNs, by adding a single unit with a linear activation function (to perform regression). CNNs were trained to estimate valence and arousal separately. AffectNet images have been split into training (60%) and validation (40%) sets. We selected mean square error (MSE) as loss function and trained the net by using Mini-Batch Gradient Descent (MBGD) [39], with batch size set to 16, learning rate equal to 0.001 and Nesterov momentum equal to 0.9. Training iterated for 50 epochs over the whole training set. The net achieved RMSE values of 0.279 and 0.242 for valence and arousal, respectively, over the validation set.

3.3 Fine-tuning

After the CNNs were trained over a large dataset, we could take advantage of two aspects: a) the features extraction capability of the nets (i.e. the pre-trained convolutional layers) and b) the current setting of the task-related (dense) layers that could be used as starting guess for the fine-tuning phase to ease the convergence. To this aim, we applied the following steps to each CNN. First, we split the net in its convolutional (including the flatten layer) and dense parts. From this point on, we operated just on the dense layers, treating them as a new net, except for the previously computed weights that were used as initialization for the fine-tuning training. This allowed to “freeze” the learning of the convolutional part and perform further experiments changing weights and bias just in the dense part. The frozen convolutional part is so used to obtain the features of each sample. Figure 3 shows the procedure.

3.4 Pool-based sampling techniques

3.4.1 Greedy sampling

GS [49] basic algorithm, GS_x, is based on the exploration of the feature space of the unlabeled pool. It is the only passive sampling approach we tested in this work. Its counterpart, GS_y, follows the same steps, but it is aimed to explore the output space of the pool (starting from a pre-trained regression model inferred by samples selected by means of GS_x). We employed a simplified version of the latter since a) our aim was to update the regression model only after the selection of all the samples that had to be queried and b) in the fine-tuning scenario the pre-trained regression model is already available. The algorithms steps are described in the pseudo-code shown in Table 2. The idea at the base of these methods is the exploration of the feature and the output spaces, respectively, by computing the Euclidean distances between samples. The initialization phase selects the starting element to be placed in the output set (labeled set), as the one closest to the centroid of the set of elements to be labeled. Then, the iterative phase

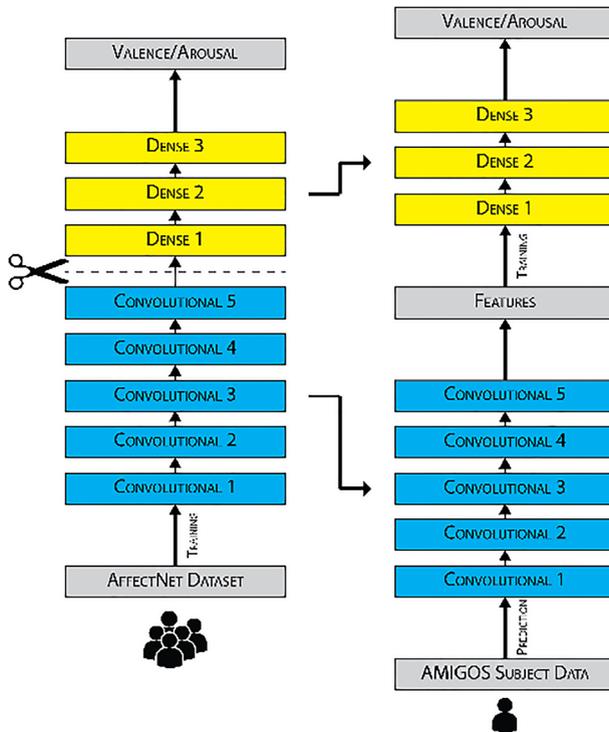


Fig. 3 Net splitting procedure for fine-tuning. Separating convolutional and dense layers allowed us to freeze the training of the first, maintaining the features extraction capability of the net. Note that two different networks are trained, one for valence and one for arousal

chooses, at each step, the element furthest from the labeled set, by computing the minimum distance from the labeled set of each candidate sample and selecting that with the maximum computed value (see Fig. 4 for a graphical representation of the strategy).

GSx and GSy can be combined in a further method (iGS) that takes advantage of the knowledge from both the feature and the output spaces (see Table 3 for the pseudo-code). iGS computes the minimum of the element-wise product of the distances calculated considering the feature and the output. This implies that the search is couple oriented, i.e. the best candidate would be the one that has the same labeled element very close in both spaces.

3.4.2 Monte-Carlo dropout uncertainty estimation

MCDUE [73] is based on a popular regularization technique for neural networks, namely Dropout [77]. Dropout randomly “turns off” some of the hidden layer neurons, so that they will output 0 regardless of their input. Dropout can be used as an active sampling technique too since it is a method for computing samples uncertain [14], starting from a pre-trained neural network. For each sample, this is obtained by iteratively (T times) muting a set of neurons (basing on a Dropout probability π) and predicting the regression output. Starting from the assumption that the Standard Deviation (STD) of the predictions could be used as a metric of the sample’s uncertainty, samples with the larger STD are selected to be queried. Table 4

Table 2 GSx and GSY pseudocode

Algorithm	GSx	(Reduced) GSY
Input	N unlabeled samples $U = \{x_n\}, n = 1 \dots N$ K, the maximum number of labels to query	N unlabeled samples $U = \{x_n\}, n = 1 \dots N$ K, the maximum number of labels to query $f()$, pre-trained regression function
Output	L, a set elements to be labeled ($ L = K$)	L, a set elements to be labeled ($ L = K$)
Steps	<pre>//Initialization L = ∅ x0 = element in U closest to the centroid of the set U → x0 → L //Iteration for k = 1 to K - 1 do ----Dx = [] ----for n = 1 to U do -----D = [] -----for m = 1 to L do -----D[m] = distance (U[n], L[m]) -----end -----Dx[n] = min(D) ----end ----xk = U[argmax(Dx)] ----U → xk → L end return L</pre>	<pre>//Initialization L = ∅ x0 = element in U s.t. f(x0) is the point closest to the centroid of f(U) U → x0 → L //Iteration for k = 1 to K - 1 do ----Dy = [] ----for n = 1 to U do -----D = [] -----for m = 1 to L do -----Dy[m] = distance (f(U[n]), f(L[m])) -----end -----Dy[n] = min(D) ----end ----xk = U[argmax(Dy)] ----U → xk → L end return L</pre>

describes the steps of the procedure. In our experiment, we set T and π to 50 and 0.5, respectively.

4 Results

In order to observe how our approach affects the performance of the recognition of a specific subject, we tested three conditions: in the first, used for comparison, the un-trained CNN (including the convolutional layers) was trained just on subject-specific data from AMIGOS, thus no transfer learning was performed (“No Transfer” label in Fig. 5); in the second condition, the net was pre-trained on AffectNet data, but no personal data were employed during training (“Transfer 0% Labeling” label in Fig. 5); in the third condition the net, pre-trained on AffectNet, was fine-tuned by using the whole available personal training set (“Transfer 100% Labeling” label in Fig. 5).

For each subject from AMIGOS, we performed training and testing in a k-fold cross-validation framework [34]. In k-fold cross-validation, the choice of the k parameter often represents a trade-off between bias and variance of the model. k is usually set to 5 or 10, but there is no formal rule [25, 37]. In our specific case, k = 5 and k = 10 would lead to a fold’s size of about 200 and 100 samples, respectively. We opted for k = 5 in order to obtain a test set better representing the variability in the underlying distribution. Results are averaged on the 10 selected subjects.

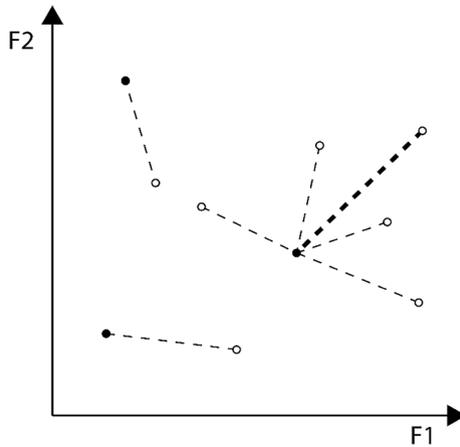


Fig. 4 Greedy iterative selection scheme: white and black circles stand for unlabeled and labeled samples, respectively. Samples are shown in a 2-dimensional space, representing the feature (or the output) space. Dashed lines indicate the minimum distance from each unlabeled point and the points from the labeled set. Tick line is the maximum minimum distance and corresponds to the next selected sample

Results show that there is a substantial difference between valence and arousal. For valence, the first test obtained poor results (average RMSE = 0.37), meaning that learning valence only by subjective data is a hard task. The pre-trained net (Transfer 0% Labeling) obtained better results (average RMSE = 0.14), suggesting that the network trained on the AffectNet database (the transferred knowledge) allowed us to create quite a good model able to generalize positive and negative valence among all the subjects. Conversely, for arousal, the net fed with subject-specific data achieved a better initial performance (average RMSE = 0.22), while the pre-trained net performed very poorly meaning that the correct recognition of arousal levels is more dependent on the specific subject and so more difficult to be generalized. Indeed, these differences can be explained by the intrinsic characteristics of valence and arousal and by the different nature of data used for the training. In general, learning from subject-specific data seems harder for valence than for arousal (see “No Transfer” scenario). Moreover, while moving from an *in the wild* context to a controlled environment does not negatively affect the learning process for valence, it is indeed quite confounding for arousal.

For both dimensions, the proposed method (Transfer 100% Labeling) improved recognition performances. This can be observed from the third scenario results, where quite good performances were obtained (RMSE = 0.09 and RMSE = 0.1 for valence and arousal, respectively). Overall results suggested that both the transferred knowledge and the personal data helped in achieving this improvement, even though they alternated in providing the main contribution.

To assess how many instances are needed by the fine-tuning process for developing an accurate personal model, for each fold, we considered different percentages of the training, from 5% (meaning that ~40 samples were used to fine-tuning) up to 90% (~720 samples) have been selected. Moreover, we were also interested in discovering if, in the context of FER, it would be convenient to adopt “smart” sampling techniques, in order to further reduce the number of samples. To this aim, we tested the different sampling techniques and we compared results with those obtained by random sampling the set of available instances.

Results obtained testing the considered sampling techniques and random sampling at several steps of the training set size, further highlight little differences between valence and

Table 3 iGS pseudocode

Algorithm	(Reduced) iGS
Input	N unlabeled samples $U = \{x_n\}$, $n = 1 \dots N$ K, the maximum number of labels to query $f()$, pre-trained regression function
Output	L, a set elements to be labeled ($ L = K$)
Steps	<pre> //Initialization L = \emptyset x0 = element in U s.t. $f(x_0)$ is the point closest to the centroid of $f(U)$ U \rightarrow x0 \rightarrow L //Iteration for k = 1 to K - 1 do ----D = [] ----for n = 1 to U do -----Dx = [] -----Dy = [] -----for m = 1 to L do -----Dx[m] = distance (U[n], L[m]) -----Dy[m] = distance ($f(U[n])$, $f(L[m])$) -----end -----D[n] = min(Dx * Dy) ----end ----xk = U[$\text{argmax}(D)$] ----U \rightarrow xk \rightarrow L end return L </pre>

arousal (Fig. 6). We observed a rapid decrease in the error in both dimensions, which makes it impossible to appreciate significant differences between sampling approaches, including random sampling. Surprisingly, the higher slope corresponds to valence, even if it started with a lower RMSE in the “Transfer 0% Labeling” case, with respect to arousal.

Finally, in Fig. 6, the presence of an elbow point is evident, both in valence (20%) and in arousal (30%), meaning that querying from ~ 240 to ~ 320 personal samples is enough to obtain quite good results with RMSE values that are comparable with the results obtained in the Transfer 100% Labeling case (~ 800 instances). These results are in support of the creation of personal models for FER since the improvement in performance is obtained with a limited required number of labelled data.

Table 4 MCDUE pseudo-code

Algorithm	MCDUE
Input	N unlabeled samples $U = \{x_n\}, n = 1 \dots N$ K, the maximum number of labels to query $f()$, pre-trained neural network T, iteration of the main loop
Output	L, a set elements to be labeled ($ L = K$)
Steps	<pre> STD=[] for n = 1 to N do ----OUT=[] ----for t = 1 to T do -----f()\equivDropout(f(),π) -----OUT[t]=f(U[n]) ----end ----STD[n]=std(OUT) end L = U(argsort(STD, descend)) L = L[1:K] return L </pre>

5 Conclusions

In this paper, we addressed the FER challenge, proposing a transfer learning approach for regression, to exploit information learned by a CNN (AlexNet) on a large *in the wild* dataset (AffectNet) and then create a subject-specific model.

In summary, the contribution of this work is the following:

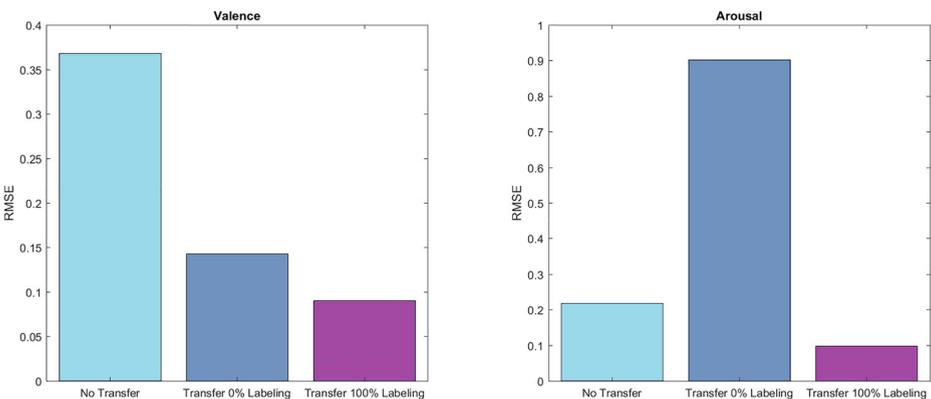


Fig. 5 Average transfer learning results (RMSE) between subjects for valence and arousal, in three different transfer-learning settings

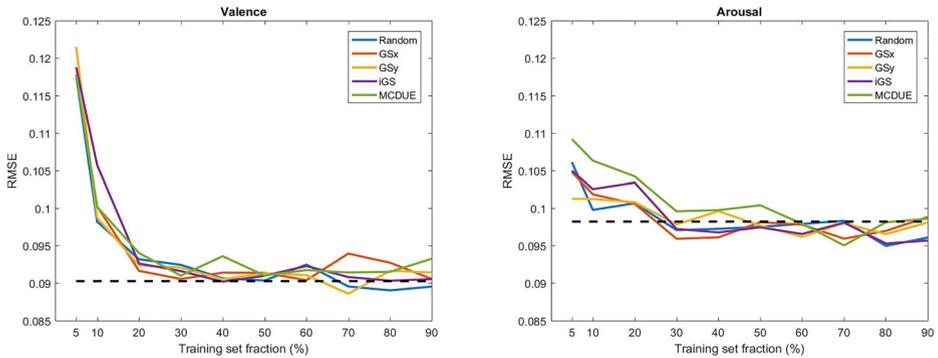


Fig. 6 Average learning results between subjects at different training set sizes, for valence and arousal. The dashed lines represent the RMSE value obtained by using the whole training set

- the transfer learning and fine-tuning approach, that has been proven to be very effective on CNNs architectures [70], has been tested in the FER domain where, to the best of our knowledge, it has been poorly explored, especially in relation with dimensional emotional models [6, 7, 13, 48, 81];
- we conducted experiments in order to quantify the performance of transfer learning with respect to personal and generalized models;
- we conducted experiments to quantify the amount of personal data required for tuning the net;
- the impact of smart sampling techniques [73, 79] for the personal dataset has been evaluated.

The results suggested that both transfer learning and subject-specific data are needed, as demonstrated by the fact that the value of RMSE obtained with transfer learning is far better both than the value obtained by simply evaluating the pre-trained network on AffectNet and than the value obtained by training the network, with random initialization of weights, just on the user's images. Interestingly, valence and arousal exhibited quite different behaviors, probably due to intrinsic differences between them. Arousal demonstrated to be less generalizable between subjects, but easier to detect by using few personal data, with respect to valence. This difference between valence and arousal could also be affected by the differences between datasets (*in the wild* with respect to controlled environments). Moreover, by considering these results together with the RMSE slope in performance, while changing the number of personal samples in the training, it is evident that for arousal very few samples are enough to fine-tune the network. This confirms the need but also the feasibility of training personal models for efficient Facial Emotion Recognition. Moreover, this different behavior and impact on the emotion recognition performance of the valence and arousal could be highlighted only by considering a dimensional approach instead of a categorical one.

Finally, in order to learn a solid single-subject model with minimum demand for new target data, we evaluate different active learning algorithms for regression. The experiment showed that our approach can significantly improve recognition performance with a limited number of target samples, regardless of the sampling techniques employed for querying samples. In particular, it would be sufficient to have a number of samples between ~ 240 and ~ 320 , which is very low if we consider that the annotation in the subjective dataset was performed once for an entire videoclip. Moreover, taking into account that the dataset used external annotations, the process could be automatized, by using pre-annotated stimuli, thus sparing the user to

annotate data by himself. This could be extremely useful in situations where the number of images accessed by a single person is few or one wants to optimize interaction with the user.

Acknowledgments This work has been partially supported by Italian MIUR within the POR Campania FESR 2014-2020 AVATEA “Advanced Virtual Adaptive Technologies e-Health” research project.

Funding Information Open access funding provided by Università degli Studi di Napoli Federico II within the CRUI-CARE Agreement.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

1. Arriaga O, Valdenegro-Toro M, Plöger PG (2019) Real-time convolutional neural networks for emotion and gender classification. In: Proceedings of the 2019 European symposium on artificial neural networks, computational intelligence. ISBN 978-287-587-065-0
2. Barrett LF, Adolphs R, Marsella S, Martinez AM, Pollak SD (2019) Emotional expressions reconsidered: challenges to inferring emotion from human facial movements. *Psychol Sci Public Interest* 20(1):1–68
3. Bartlett MS, Littlewort G, Frank MG, Lainscsek C, Fasel IR, Movellan JR (2006) Automatic recognition of facial actions in spontaneous expressions. *J Multimed* 1(6):22–35
4. 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. In: Proceedings of the IEEE conference on computer vision and pattern recognition workshops, pp 17–25
5. Chao L, Tao J, Yang M, Li Y, Wen Z (2015) Long short term memory recurrent neural network based multimodal dimensional emotion recognition. In: Proceedings of the 5th international workshop on audio/visual emotion challenge, pp 65–72
6. Chen J, Liu X, Tu P, Aragonés A (2013) Learning person-specific models for facial expression and action unit recognition. *Pattern Recogn Lett* 34(15):1964–1970
7. Chu WS, De la Torre F, Cohn JF (2016) Selective transfer machine for personalized facial expression analysis. *IEEE Trans Pattern Anal Mach Intell* 39(3):529–545
8. Dhall A, Ramana Murthy O, Goecke R, Joshi J, Gedeon T (2015) Video and image based emotion recognition challenges in the wild: EmotiW 2015. In: Proceedings of the 2015 ACM on international conference on multi-modal interaction, pp 423–426
9. Donahue J, Hendricks AL, Guadarrama S, Rohrbach M, Venugopalan S, Saenko K, Darrell T (2015) Long-term recurrent convolutional networks for visual recognition and description. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 2625–2634
10. Du S, Tao Y, Martinez AM (2014) Compound facial expressions of emotion. *Proc Natl Acad Sci* 111(15):E1454–E1462
11. Ekman P (1992) An argument for basic emotions. *Cognit Emot* 6(3–4):169–200
12. Ekman P, Keltner D (1997) Universal facial expressions of emotion. In: Segerstrale U, Molnar P (eds) *Nonverbal communication: where nature meets culture*, pp 27–46
13. Feffer M, Picard RW (2018) A mixture of personalized experts for human affect estimation. In: International conference on machine learning and data mining in pattern recognition, pp 316–330
14. Gal Y (2016) *Uncertainty in deep learning*. University of Cambridge, Cambridge
15. Ghimire D, Lee J (2013) Geometric feature-based facial expression recognition in image sequences using multi-class adaboost and support vector machines. *Sensors* 13(6):7714–7734

16. Goodfellow IJ, Erhan D, Carrier PL, Courville A, Mirza M, Hamner B, Cukierski W et al (2013) Challenges in representation learning: a report on three machine learning contests. In: International conference on neural information processing, pp 117–124
17. Guo R, Li S, He L, Gao W, Qi H, Owens G (2013) Pervasive and unobtrusive emotion sensing for human mental health. In: Proceedings of the 7th international conference on pervasive computing Technologies for Healthcare, Venice, Italy, 5–8 May 2013, pp 436–439
18. Harris JM, Ciorciari J, Gountas J (2018) Consumer neuroscience for marketing researchers. *J Consum Behav* 17(3):239–252
19. Hasani B, Mahoor MH (2017) Facial expression recognition using enhanced deep 3D convolutional neural networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition workshops, pp 30–40
20. Hasani B, Mahoor MH (2017) Facial affect estimation in the wild using deep residual and convolutional networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition workshops, pp 9–16
21. Izard CE (2007) Basic emotions, natural kinds, emotion schemas, and a new paradigm. *Perspect Psychol Sci* 2(3):260–280
22. Izquierdo-Reyes J, Ramirez-Mendoza RA, Bustamante-Bello MR, Pons-Rovira JL, Gonzalez-Vargas JE (2018) Emotion recognition for semi-autonomous vehicles framework. *International Journal on Interactive Design and Manufacturing (IJIDeM)* 12(4):1447–1454
23. Jack RE, Garrod OG, Yu H, Caldara R, Schyns PG (2012) Facial expressions of emotion are not culturally universal. *Proc Natl Acad Sci* 109(19):7241–7244
24. Jacobs RA, Jordan MI, Nowlan SJ, Hinton GE (1991) Adaptive mixtures of local experts. *Neural Comput* 3(1):79–87
25. James G, Witten D, Hastie T, Tibshirani R (2013) An introduction to statistical learning. Springer, New York
26. Jiang J (2008) A literature survey on domain adaptation of statistical classifiers. Technical report, University of Illinois at Urbana-Champaign
27. Jung H, Lee S, Yim J, Park S, Kim J (2015) Joint fine-tuning in deep neural networks for facial expression recognition. In: Proceedings of the IEEE international conference on computer vision, pp 2983–2991
28. Kahou ES, Michalski V, Konda K, Memisevic R, Pal C (2015) Recurrent neural networks for emotion recognition in video. In: Proceedings of the 2015 ACM on international conference on multimodal interaction, pp 467–474
29. Kanade T, Cohn JF, Tian Y (2000) Comprehensive database for facial expression analysis. In: Proceedings fourth IEEE international conference on automatic face and gesture recognition (cat. No. PR00580), pp 46–53
30. Kaulard K, Cunningham DW, Bühlhoff HH, Wallraven C (2012) The MPI facial expression database—a validated database of emotional and conversational facial expressions. *PLoS One* 7(3):e32321
31. Khorrami P, Le Paine T, Brady K, Dagli C, Huang TS (2016) How deep neural networks can improve emotion recognition on video data. In: 2016 IEEE international conference on image processing (ICIP), pp 619–623
32. Kleinsmith A, Bianchi-Berthouze N (2012) Affective body expression perception and recognition: a survey. *IEEE Trans Affect Comput* 4(1):15–33
33. Ko B (2018) A brief review of facial emotion recognition based on visual information. *Sensors* 18(2):401
34. Kohavi R (1995) A study of cross-validation and bootstrap for accuracy estimation and model selection. *Ijcai* 14(2):1137–1145
35. Koolagudi SG, Rao KS (2012) Emotion recognition from speech: a review. *International journal of speech technology* 15(2):99–117
36. Krizhevsky A, Sutskever I, Hinton GE (2012) Imagenet classification with deep convolutional neural networks. In: Advances in neural information processing systems, pp 1097–1105
37. Kuhn M, Johnson K (2013) Applied predictive modeling. Springer, New York
38. Lench HC, Flores SA, Bench SW (2011) Discrete emotions predict changes in cognition, judgment, experience, behavior, and physiology: a meta-analysis of experimental emotion elicitations. *Psychol Bull* 137(5):834–855
39. Li M, Zhang T, Chen Y, Smola AJ (2014) Efficient mini-batch training for stochastic optimization. In: Proceedings of the 20th ACM SIGKDD international conference on knowledge discovery and data mining, pp 661–670
40. 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. In: Proceedings of the IEEE conference on computer vision and pattern recognition workshops, pp 1–8

41. Lindquist KA, Siegel EH, Quigley KS, Barrett LF (2013) The hundred-year emotion war: are emotions natural kinds or psychological constructions? Comment on Lench, Flores, and Bench (2011). *Psychol Bull* 139(1):255–263
42. Lucey P, Cohn JF, Kanade T, Saragih J, Ambadar Z, Matthews I (2010) The extended cohn-kanade dataset (ck+): a complete dataset for action unit and emotion-specified expression. In: 2010 IEEE computer society conference on computer vision and pattern recognition-workshops, pp 94–101
43. Lyons M, Akamatsu S, Kamachi M, Gyoba J (1998) Coding facial expressions with gabor wavelets. In: Proceedings third IEEE international conference on automatic face and gesture recognition, pp 200–205
44. Mavadati SM, Mahoor MH, Bartlett K, Trinh P, Cohn JF (2013) Disfa: a spontaneous facial action intensity database. *IEEE Trans Affect Comput* 4(2):151–160
45. Mehrabian A (1996) Pleasure-arousal-dominance: a general framework for describing and measuring individual differences in temperament. *Curr Psychol* 14(4):261–292
46. Miranda-Correa JA, Abadi MK, Sebe N, Patras I (2018) AMIGOS: a dataset for affect, personality and mood research on individuals and groups. *IEEE Trans Affect Comput*
47. Mollahosseini A, Hasani B, Mahoor MH (2017) Affectnet: a database for facial expression, valence, and arousal computing in the wild. *IEEE Trans Affect Comput* 10(1):18–31
48. Ng HW, Nguyen VD, Vonikakis V, Winkler S (2015) Deep learning for emotion recognition on small datasets using transfer learning. In: Proceedings of the 2015 ACM on international conference on multimodal interaction. ACM, pp 443–449
49. Pan SJ, Yang Q (2009) A survey on transfer learning. *IEEE Trans Knowl Data Eng* 22(10):1345–1359
50. Pantic M, Valstar M, Rademaker R, Maat L (2005) Web-based database for facial expression analysis. In: 2005 IEEE international conference on multimedia and expo, p 5
51. Picard RW (1999) Affective computing for HCI. In: *HCI* (1), pp 829–833
52. Plutchik R, Kellerman H (1980) Theories of emotion. Academic, New York
53. Ringeval F, Sonderegger A, Sauer J, Lalanne D (2013) Introducing the RECOLA multimodal corpus of remote collaborative and affective interactions. In: 2013 10th IEEE international conference and workshops on automatic face and gesture recognition (FG), pp 1–8
54. Ringeval F, Schuller B, Valstar M, Jaiswal S, Marchi E, Lalanne D, Cowie R, Pantic M (2015) Av+ ee 2015: the first affect recognition challenge bridging across audio, video, and physiological data. In: Proceedings of the 5th international workshop on audio/visual emotion challenge, pp 3–8
55. Rossi S, Ercolano G, Raggioli L, Savino E, Ruocco M (2018) The disappearing robot: an analysis of disengagement and distraction during non-interactive tasks. In: 2018 27th IEEE international symposium on robot and human interactive communication (RO-MAN), pp 522–527
56. Russakovsky O, Deng J, Su H, Krause J, Satheesh S, Ma S, Huang Z, Karpathy A, Khosla A, Bernstein M, Berg AC, Fei-Fei L (2015) Imagenet large scale visual recognition challenge. *Int J Comput Vis* 115(3):211–252
57. Russell J (1980) A circumplex model of affect. *J Pers Soc Psychol* 39(6):1161–1178
58. Salovey P, Mayer JD (1990) Emotional intelligence. *Imagin Cogn Pers* 9(3):185–211
59. Sariyanidi E, Gunes H, Cavallaro A (2014) Automatic analysis of facial affect: a survey of registration, representation, and recognition. *IEEE Trans Pattern Anal Mach Intell* 37(6):1113–1133
60. Sayette MA, Creswell KG, Dimoff JD, Fairbairn CE, Cohn JF, Heckman BW, Kirchner TR, Levine JM, Moreland RL (2012) Alcohol and group formation: a multimodal investigation of the effects of alcohol on emotion and social bonding. *Psychol Sci* 23(8):869–878
61. Shan C, Gong S, McOwan PW (2009) Facial expression recognition based on local binary patterns: a comprehensive study. *Image Vis Comput* 27(6):803–816
62. Shu L, Xie J, Yang M, Li Z, Li Z, Liao D, Xu X, Yang X (2018) A review of emotion recognition using physiological signals. *Sensors* 18(7):2074
63. Soleymani M, Pantic M (2012) Human-centered implicit tagging: overview and perspectives. In: 2012 IEEE international conference on systems, man, and cybernetics (SMC), pp 3304–3309
64. Soleymani M, Asghari-Esfeden S, Pantic M, Fu Y (2014) Continuous emotion detection using EEG signals and facial expressions. In: 2014 IEEE international conference on multimedia and expo (ICME), pp 1–6
65. Spezialetti M, Cinque L, Tavares JMR, Placidi G (2018) Towards EEG-based BCI driven by emotions for addressing BCI-illiteracy: a meta-analytic review. *Behav Inform Technol* 37(8):855–871
66. Suk M, Prabhakaran B (2014) Real-time mobile facial expression recognition system—a case study. In: Proceedings of the IEEE conference on computer vision and pattern recognition workshops, pp 132–137
67. Susskind J, Anderson A, Hinton G (2010). The Toronto face database. Technical report, UTML TR 2010-001, University of Toronto.
68. Szegegy C, Liu W, Jia Y, Sermanet P, Reed S, Anguelov D, Erhan D, VanHoucke V, Rabinovich A (2015) Going deeper with convolutions. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 1–9

69. Szegedy C, Ioffe S, Vanhoucke V, Alemi AA (2017) Inception-v4, inception-resnet and the impact of residual connections on learning. In: Thirty-First AAAI Conference on Artificial Intelligence
70. Tan C, Sun F, Kong T, Zhang W, Yang C, Liu C (2018) A survey on deep transfer learning. In: International conference on artificial neural networks. Springer, Cham, pp 270–279
71. Tomkins SS (2008) Affect imagery consciousness: the complete edition: two volumes. Springer publishing company, New York
72. Trnka R, Lačev A, Balcar K, Kuška M, Tavel P (2016) Modeling semantic emotion space using a 3D hypercube-projection: an innovative analytical approach for the psychology of emotions. *Front Psychol* 7: 522
73. Tsymbalov E, Panov M, Shapeev A (2018) Dropout-based active learning for regression. In: International conference on analysis of images, social networks and texts, pp 247–258
74. Valstar MF, Jiang B, Mehu M, Pantic M, Scherer K (2011) The first facial expression recognition and analysis challenge. In: IEEE international conference on automatic face and gesture recognition and workshops (FG'11), pp 921–926
75. Verschuere B, Crombez G, Koster E, Uzieblo K (2006) Psychopathy and physiological detection of concealed information: a review. *Psychol Belg* 46:99–116
76. Viola P, Jones M (2001) Rapid object detection using a boosted cascade of simple features. In: *CVPR* (1), vol 1, pp 511–518 3
77. Wager S, Wang S, Liang PS (2013) Dropout training as adaptive regularization. In: *Advances in neural information processing systems*, pp 351–359
78. Walecki R, Rudovic O, Pavlovic V, Schuller B, Pantic M (2017) Deep structured learning for facial expression intensity estimation. *Image Vis Comput* 259:143–154
79. Wu D, Lin CT, Huang J (2019) Active learning for regression using greedy sampling. *Inf Sci* 474:90–105
80. Zafeiriou S, Kollias D, Nicolaou MA, Papaioannou A, Zhao G, Kotsia I (2017) Aff-wild: valence and arousal 'In-the-Wild' challenge. In: *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pp 34–41
81. Zen G, Porzi L, Sangineto E, Ricci E, Sebe N (2016) Learning personalized models for facial expression analysis and gesture recognition. *IEEE Transactions on Multimedia* 18(4):775–788
82. Zhang X, Mahoor MH, Mavadati SM (2015) Facial expression recognition using lp-norm MKL multiclass-SVM. *Mach Vis Appl* 26(4):467–483

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