Pattern Analysis and Applications
https://doi.org/10.1007/s10044-021-01012-9

ORIGINAL ARTICLE

Automatic stress analysis from facial videos based on deep facial action units recognition

Giorgos Giannakakis[1,3] · Mohammad Rami Koujan[2] · Anastasios Roussos[1,2] · Kostas Marias[1,4]

Received: 15 September 2020 / Accepted: 7 July 2021

Abstract

Stress conditions are manifested in different human body's physiological processes and the human face. Facial expressions are modelled consistently through the Facial Action Coding System (FACS) using the facial Action Units (AU) parameters. This paper focuses on the automated recognition and analysis of AUs in videos as quantitative indices to discriminate between neutral and stress states. A novel deep learning pipeline for automatic recognition of facial action units is proposed, relying on two publicly available annotated facial datasets for training, the UNBC and the BOSPHORUS datasets. Two types of descriptive facial features are extracted from the input images, geometric features (non-rigid 3D facial deformations due to facial expressions) and appearance features (deep facial appearance features). The extracted facial features are then fed to deep fully connected layers that regress AU intensities and robustly perform AU classification. The proposed algorithm is applied to the SRD'15 stress dataset, which contains neutral and stress states related to four types of stressors. We present thorough experimental results and comparisons, which indicate that the proposed methodology yields particularly promising performance in terms of both AU detection and stress recognition accuracy. Furthermore, the AUs relevant to stress were experimentally identified, providing evidence that their intensity is significantly increased during stress, which leads to a more expressive human face as compared to neutral states.

Keywords stress · facial action units · AU · Facial Action Coding System (FACS) · Convolutional Neural Networks · Deep Learning

1 Introduction

Emotion recognition is a crucial process for various disciplines such as medicine, psychology, neurosciences, human-computer interaction and computer vision. It aims at the interpretation and understanding of the human affective state. Body language is fundamental for non-verbal communication when analyzing human behaviour and its major expression is facial expressions. Some emotional states such as acute stress can be usually recognized from one's social circle, however, the reliable automatic stress estimation through computational models is not a straightforward process.

Methods that have been developed to estimate stress levels in humans include clinical interviews [1, 2], questionnaires/psychometric scales [3], biomarkers (e.g. cortisol) [4] and biosignals (e.g. EEG, ECG, EDA, EMG, respiration signals, etc.) [5, 6].

The human face conveys valuable information about one's emotional state [7, 8]. Charles Darwin perceived the

Giorgos Giannakakis and Mohammad Rami Koujan contributed equally to this article.

✉ Giorgos Giannakakis
ggian@ics.forth.gr

Mohammad Rami Koujan
mk538@exeter.ac.uk

Anastasios Roussos
troussos@ics.forth.gr

Kostas Marias
kmarias@ics.forth.gr

[1] Institute of Computer Science, Foundation for Research and Technology Hellas (FORTH), Heraklion, Greece

[2] College of Engineering, Mathematics and Physical Sciences, University of Exeter, Exeter, UK

[3] Institute of AgriFood and Life Sciences, University Research Centre, Hellenic Mediterranean University, Heraklion, Greece

[4] Department of Electrical and Computer Engineering, Hellenic Mediterranean University, Heraklion, Greece

Published online: 22 September 2021

significant contribution of facial expressions on human non-verbal communication and affect describing their modulation according to the perceived emotion in his work in 1872 [9]. The facial expressions are considered to be inferior to biosignals or biomarkers for emotion recognition tasks, since they are voluntary and can be manipulated or hidden [5]. However, over the last years, broad research has been conducted on identifying in- or semi-voluntary facial cues (such as facial micro-expressions, blinks, mouth micro-activity) which would provide reliable indices of affect and especially of stress estimation [10, 11].

A common practice in annotating and analyzing AUs has been through visual inspection by experts trained in the Facial Action Coding System (FACS). This procedure can be very tedious and time-consuming, since it is estimated that, for each minute of captured video, it is required approximately 1 hour to annotate all AU [12, 13]. Thus, there has been an extensive effort over the last years of automating AU recognition through computer vision and machine learning techniques.

Even though many researchers have been addressing the problem of reliable and accurate AU estimation, there are still few studies applying AU methodology towards the identification of stress. Because of that, there is currently a lack of a consistent guideline about constructing an AU stress model as well as the AU behaviour and AU involvement in stress conditions.

In this paper, a novel deep learning pipeline for AU recognition is proposed, which is trained based on two publicly available datasets with annotated AU by experts. The proposed pipeline is then applied to a thorough stress experimental dataset with 4 different types of stressors. The AUs most implicated during stress conditions are identified and selected, which leads to the development of a novel facial analysis model for the recognition of stress during different stressor conditions.

The main component of the proposed system consists of the pipeline for AU detection from facial images. This is based on two modules for deep feature extraction: a geometric and an appearance module. For the module of geometric feature extraction, we adopt a Convolutional Neural Network (CNN) that is able to represent changes in the 3D geometry of the subject's face due to solely facial expressions. For the module of appearance feature extraction, we propose a deep residual network that extracts robust and descriptive features of facial appearance, complementing the geometric features with fine-grained details. We subsequently use the extracted geometric and appearance features to perform AU estimation by adopting a fully-connected deep network on the combined features, being in accordance with the current state-of-the-art in similar classification tasks.

The novelty and the main contributions of this study can be summarized as follows:

1. We use deep learning networks, which to our knowledge have not been used before for stress detection based on AUs.
2. We propose a pairwise transformation for AU-based stress classification, which enables the establishment of a common reference taking into account the individual baselines of the participants dealing with the inter-subject variability.
3. Our incorporated deep networks were pretrained on large-scale in-the-wild datasets of human facial performances, making them very robust to challenging capture setups that might be encountered during test time.
4. The deep feature extractors we utilize offer real-time performance with few milliseconds extraction time per frame.

This work extends our preliminary research [14] by introducing more comprehensive approaches with respect to several different dimensions of our system. First, we have replaced the hand-crafted geometric and appearance features for AU detection that were based on more traditional feature extraction techniques with robust deep networks representations, building upon recent state-of-the-art computer vision methods. In addition, we have replaced the conventional machine learning methods for AU classification used in [14] with a fully-connected deep network, being also in accordance with the current state-of-the-art in similar classification tasks. Furthermore, we present a much more detailed description of all aspects of the proposed methodology. Finally, we present a more thorough experimental evaluation, including additional evaluations and comparisons and providing more sound evidence about the effectiveness and potential of our proposed framework.

2 Related work

There is a great research interest in the accurate recognition/analysis of facial action units (AU) for the reliable decoding into emotion. A recent review summarizes techniques of facial AU analysis, not concluding to specific guidelines but delineating good practices in facial action units analysis [12].

In the related literature, various emotion detection algorithms through facial AU have been proposed using different approaches. In [15], Ruiz et al. used Hidden (HTL) and Semi-Hidden-Task Learning (SHTL) in order to train the emotion models from annotated databases.

Over the last years, there has been a substantially growing interest in adopting deep learning techniques for various machine learning problems and facial AU recognition has not been an exception. However, deep learning techniques have been documented to provide only moderate

improvements [12], which may be attributed to the lack of substantially large datasets that are needed for training such systems. More recent works tackle this problem by adopting advanced hybrid deep learning methods such as LSTM and CNN [16], attention-based AU detection [17] or regional convolutional neural networks (R-CNN) [18], yielding promising results.

Regarding the research on facial behaviour during stress conditions, some recent studies are investigating facial cues [10] and [19, 20]. The literature on AU stress analysis is limited, and few studies are aiming to detect stress levels from FACS coding [14, 21, 22]. In [22], AU temporal evolution was investigated achieving an average stress recognition accuracy of 75% for person-independent and 93% for person-dependent analysis. In [21], facial landmarks were estimated from AAM models and AU were estimated from 4 types of artificial neural networks (FFNN, RBFNN, RNN and CNN) leading to a stress classification accuracy of 80.4% with the intersubject methodology. However, there are neither specific guidelines in which AU are implicated in stress state nor a consistent model describing the stress manifestation on specific facial expressions.

In our previous works, we had initial results on AU analysis during stress conditions using geometric and appearance features [14]. Besides, in our work [23] a real time-time facial expression recognition system was proposed based on a deep Convolutional Neural Network for estimating expression parameters of a 3D Morphable Model and applying effectively to stress conditions.

3 Datasets and acquisition protocols

3.1 Training datasets for AU detection

In this study, we used two available facial datasets in order to train the AU models, the UNBC-McMaster Shoulder Pain Expression Archive Database (UNBC) [24] and the Bosphorus database (BOSPHORUS) [25].

The UNBC dataset contains 200 sequences across 25 subjects (48,398 images in total) annotated according to FACS code and their corresponding AU intensities in the scale of integer values $\{0, 1, .., 5\}$, where 0 corresponds to AU non-existence, whereas 5 corresponds to the maximum AU intensity. It contains annotated information for the following AUs: [6, 7, 9, 10, 12, 25, 26]. The BOSPHORUS dataset contains 105 subjects with a total of 4666 facial images annotated according to FACS coding and their corresponding AU intensity in the same scale $\{0, 1, .., 5\}$. It contains annotated information for the following AUs: [1, 2, 4, 5, 6, 7, 9, 10, 12, 14, 15, 17, 20, 23, 25, 26, 43].

Since in this kind of datasets there is a large discrepancy across different annotators on the criteria that a nonzero intensity is assigned to an existing AU, the intensity values $\{1, .., 5\}$ contain a significant amount of annotation noise. At the same time, the annotation of the intensity 0 has been made in a much more consistent way across different annotators, which makes the zero/nonzero separation significantly less noisy. For these reasons, we chose to binarize the AU annotations by thresholding the annotated AUs at the intensity value of 1, resulting in binary labels of 0 (non-existing AU with intensity label $= 1$) or 1 (existing AU with intensity label ≥ 1) per AU. In addition, in this work, we consider the following 15 AUs: [1, 2, 4, 5, 6, 7, 9, 10, 12, 14, 15, 17, 23, 25, 26], which are all included in at least one of the two datasets (with 7 AUs being included in both datasets).

Please note that, even though we use binarized AU labels as ground truth, our proposed deep classifier outputs continuous values between 0 to 1, which correspond to confidence scores about the presence of each AU in every input image. This representation is more effective for the subsequent step of stress analysis in videos. Please refer to Sect. 4.3 for further details.

3.2 Video acquisition protocol for stress recognition

An experimental protocol was designed and conducted to investigate facial response in stress conditions. Each of the participants was seated in front of a computer monitor. Videos were recorded with a PointGrey Grasshopper3 U3 camera. The camera was placed on a tripod at the back top of the monitor and at a distance of about 90 cm with its field of view covering the participant's face and possible movements during the experiment. The acquisition setup ensured conditions of ambient lighting, which minimized specular effects.

The experiment included neutral tasks (used as reference) and stressful tasks in which stress conditions were simulated and induced employing different types of stressors. These stressors were categorized into 4 different phases: *social exposure*, *emotional recall*, *mental workload tasks*, *stressful videos presentation*. The experimental tasks along with their duration and affective state are presented in Table 1.

The social exposure phase included an interview asking the participant to describe himself/herself. It has its origin in the stress due to exposure faced by an actor when he/she is at the stage. The reference for this phase was a neutral pose.

The emotion recall phase included stress elicitation by asking participants to recall and relive a stressful event from their past as if it was currently happening.

The mental tasks phase included cognitive load assessment through tasks such as the modified Stroop colour-word task (SCWT) [26], requiring participants to read colour names (red, green, and blue) printed in incongruous ink (e.g. the word RED appearing in blue ink). In the present task, difficulty was increased by asking participants to first read each word and then name the colour of the word. Another

Table 1 Experimental tasks employed in this study

Experimental task		Duration (min)	Affective State
Social exposure			
1.1	Neutral (reference)	1	N
1.2	Interview	2	S
Emotional recall			
2.1	Neutral (reference)	1	N
2.2	Recall anxious event	1	S
2.3	Recall stressful event	1	S
Stressful images/Stroop task			
3.1	IAPS stressful images	2	S
3.2	Stroop Colour-Word Test (SCWT)	2	S
Stressful videos			
4.1	Neutral (reference)	1	N
4.2	Calming video	2	R
4.3	Adventure video	2	S
4.4	Psychological pressure video	2	S

Intended affective state N:neutral, S:stress, R:relaxed)

task used in this phase was the presentation of unpleasant images from the International Affective Picture System (IAPS) [27] which were used as affect generating stimuli. Stimuli included images having stressful content such as human pain, drug abuse, violence against women and men, armed children. Each image was presented for 15 sec.

The stressful videos phase included the presentation of 2-minute video segments in attempting to induce low-intensity positive emotions (calming video), and stress (action scene from an adventure film, a scene involving heights to participants with moderate levels of acrophobia, a burglary/home invasion while the inhabitant is inside, car accidents, etc).

3.3 Stress recognition dataset (SRD'15)

The SRD'15 dataset was acquired following the aforementioned protocol and included 24 participants (7 women, 17 men) with age 47.3±9.3 years. For each participant, 11 tasks (3 neutral, 7 stressed and 1 relaxed states) were performed.

Videos had a sampling frequency of 60 fps with a resolution of 1.216x1.600 pixels, which was subsampled to 30 fps and a resolution of 608x800 pixels. A neutral condition was presented at the beginning of each phase of the experiment. This condition was used as a baseline for the subsequent stressful tasks. The study was approved by the North-West Tuscany ESTAV (Regional Health Service, Agency for the Technical Administrative Services of Wide Area) Ethical Committee. All participants provided informed consent. Data were recorded during the second data acquisition campaign of a research project aiming at the development of computational platform monitoring cardio-metabolic risk [28].

4 Methods

This study focuses on automatic stress recognition from facial AUs that are estimated from a deep learning pipeline trained in 2 annotated databases. A flow chart of the proposed system for automatic stress analysis from facial videos is depicted in Fig. 1. The input in our system is a facial video. Every frame of the video is fed into the proposed deep pipeline for AU detection. This pipeline outputs an estimate of the intensity of every AU, which can be considered as a per-frame AU feature vector. This sequence of AU feature vectors is then fed into the proposed module for stress recognition which in turn outputs a binary classification of the

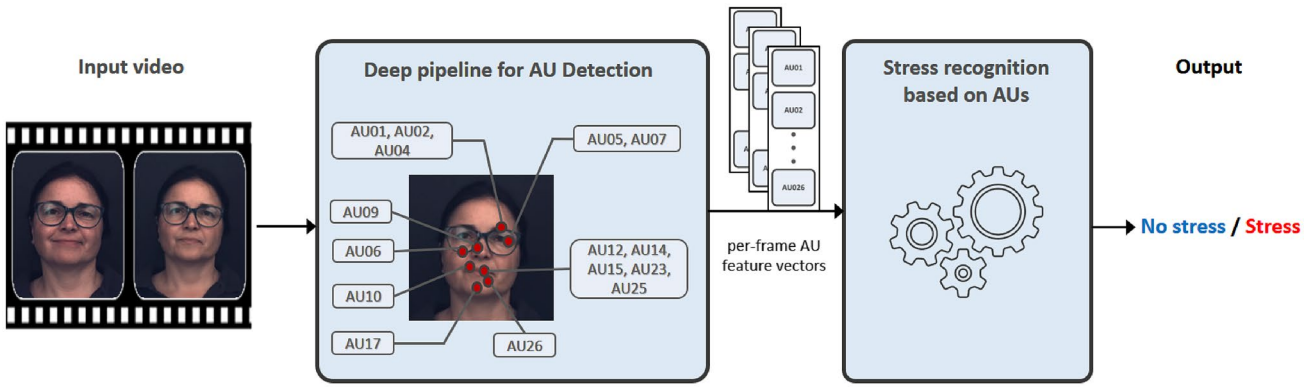

Fig. 1 Flow chart of the proposed system for automatic stress analysis from facial videos. The input video is fed into the proposed deep pipeline for AU detection (Sect. 4.2 and Fig. 2), which outputs a sequence of AU feature vectors (AU intensities). This sequence is then fed into the proposed module for stress recognition (Sect. 4.3), which outputs a binary classification of the subject's stress state in the input video

stress state that corresponds to the captured subject in the input video.

This section is organized as follows. After a description of the adopted Facial Action Coding System (Sect. 4.1), we present our deep pipeline for AU detection from images (Sect. 4.2), followed by our framework for automatic analysis of facial videos for stress classification (Sect. 4.3).

4.1 Facial action coding system (FACS)

The *Facial Action Coding System* (FACS) [29, 30] is a system that was initially developed by the Swedish anatomist Carl-Herman Hjortsjö [31] and updated by Ekman and Friesen in 1978 [30] and in 2002 [29], respectively. It systematically categorizes human facial muscle movements and expressions based on anatomic functions. The FACS consists of 32 distinct facial muscle movements named Action Units (AU). In addition, it defines 14 head movements actions and 11 eyes movements actions. In this study, 15 AUs are investigated (as the annotated training datasets included these AUs) in terms of association with stress types which are presented in Table 2.

4.2 Deep pipeline for AU detection

The proposed deep pipeline for AU detection is presented in Fig. 2. It consists of three main steps: the *preprocessing*, the *feature extraction* and the *AU classification* phase. The preprocessing phase includes the steps of face detection, face landmarking and 2D image registration (as described in Sect. 4.2.1). The feature extraction phase includes the deep geometric and appearance features extraction using DeepExp3D network and Deep Residual Network, respectively, (as described in Sect. 4.2.2). The AU classification phase utilizes the combined features of the previous phase

Table 2 Summary of the AU, their FACS name and muscular basis investigated in this study

A U	FACS name	Muscular basis
AU1	Inner brow raiser	frontalis (pars medialis)
AU2	Outer brow raiser	frontalis (pars lateralis)
AU4	Brow lowerer	depressor glabellae, depressor supercilii, corrugator supercilii
AU5	Upper lid raiser	levator palpebrae superioris, superior tarsal muscle
AU6	Cheek raiser	orbicularis oculi (pars orbitalis)
AU7	Lid tightener	orbicularis oculi (pars palpebralis)
AU9	Nose wrinkler	levator labii superioris alaeque nasi
AU10	Upper lip raiser	levator labii superioris, caput infraorbitalis
AU12	Lip corner puller	zygomaticus major
AU14	Dimpler	buccinator
AU15	Lip corner depressor	depressor anguli oris (triangularis)
AU17	Chin raiser	mentalis
AU23	Lip tightener	orbicularis oris
AU25	Lips part	depressor labii inferioris, or relaxation of mentalis or orbicularis oris
AU26	Jaw drop	masseter; relaxed temporalis and internal pterygoid

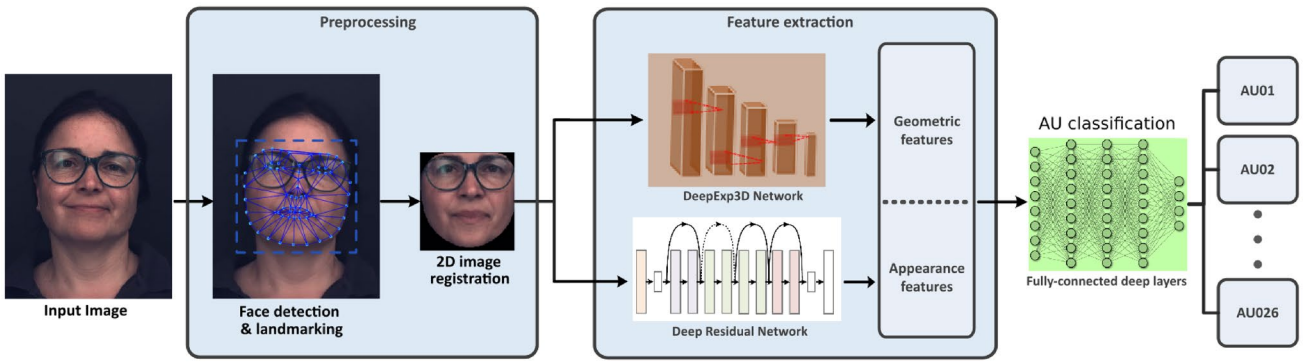

Fig. 2 Proposed deep pipeline for AU detection, consisting of the steps of preprocessing (face detection, landmarking and 2D image registration), feature extraction (deep geometric and appearance features) and deep AU classification

and performs AU classification using a deep fully connected neural network (as described in Sect. 4.2.3).

After the construction of the AU model, the stress model was established using the pairwise transformation (as described in Sect. 4.3.1), the most relevant to the problem under investigation was performed using the features selection methods mRMR, RF and Fischer Ratio (as described in Sect. 4.3.2). A more detailed description of each process is given in the following paragraphs.

4.2.1 Preprocessing

For every input image, the preprocessing phase consists of face detection, estimation of facial landmarks, as well as rigid image alignment. For face detection, we use the state-of-the-art detection method of [32]. For the landmark estimation, we exploit recent advances in deep learning and achieve highly-reliable, robust and computationally efficient results. In more detail, we utilize the so-called *3D-aware 2D landmarks* which we extract from every video frame (image) with the state-of-the-art*Cascade Multi-view Hourglass Model* of Deng et al. [33]. The localized 68 landmarks with this model correspond to projections of their corresponding 3D points on the image plane. This is important for our 3D-based feature extraction because, in high contrast to the commonly-used conventional 2D landmarks [34], this type of landmarks are directly associated with the real 3D geometry of the human face [33].

Afterwards, every input image is aligned to a template of size 224×224 that has the 68 facial landmarks on a mean facial shape arrangement. This alignment is achieved by applying Procrustes analysis [35] to find the 2D similarity transform that optimally aligns the landmarks extracted from the input image to the mean face landmarks. This similarity transform is then applied to the input image using Delaunay triangle-based affine warp [36]. This process offers a first normalization of the image data by removing variation due to in-plane 2D similarity transformations.

4.2.2 Feature extraction

In terms of feature representation, we use two types of features, namely geometric and appearance features, adopting deep learning frameworks in both cases:

Geometric features We want to use geometric features that robustly represent the changes in the 3D geometry of the subject's face due to facial expressions. Following the work of [23], we adopt a 3D-based representation of "pure" facial expression that is *invariant* to all other parameters that contribute in the formation of the input image (e.g. shape and appearance variation related to the subject's identity, relative 3D pose variation and illumination variations). To achieve that, we model the 3D face geometry using 3D Morphable Models (3DMMs) [37] and an additive combination of identity and expression variation [38–41]. In more detail, we represent a 3D facial shape $\mathbf{x} = [x_1, y_1, z_1, ..., x_N, y_N, z_N]^\top \in \mathbb{R}^{3N}$ using the following decomposition:

$$\mathbf{x} = \mathbf{x}(s^i, s^e) = \bar{\mathbf{x}} + \mathbf{U}_{id} s^i + \mathbf{U}_{exp} s^e \tag{1}$$

where $\bar{\mathbf{x}} \in \mathbb{R}^{3N}$ is a fixed mean shape of the 3D morphable model (corresponding to an average face). $\mathbf{U}_{id} \in \mathbb{R}^{3N \times n_i}$ is the identity orthonormal basis with $n_i = 157$ principal components ($n_i \ll 3N$) and $\mathbf{U}_{exp} \in \mathbb{R}^{3N \times n_e}$ is the expression orthonormal basis with $n_e = 28$ principal components ($n_e \ll 3N$). In addition, $s^i \in \mathbb{R}^{n_i}$ and $s^e \in \mathbb{R}^{n_e}$ are the identity and expression parameters of the morphable model, respectively. The identity part of the model ($\mathbf{U_{id}}$}) originates from the LSFM model [42] and the expression part of the model ($\mathbf{U}_{exp}$}) originates from the work of Zafeiriou et al. [38], who built it by registering the blendshapes model of Facewarehouse [43] with the LSFM model. In the adopted model (1), the 3D facial shape $\mathbf{x}$ is a function of both identity and expression coefficients ($\mathbf{x}(s^i, s^e)$), where expression variations are effectively represented as offsets from a given identity shape.

In addition to shape variation, the captured facial image is also affected by the relative 3D pose of the camera with respect to the facial shape as well as by the camera's intrinsics. Overall, this can be modelled as a camera projection function that takes as input the 3D facial shape and projects it on the image plane:

$$\mathbf{w}(s^i, s^e, \mathbf{R}, \mathbf{t}, \mathbf{c}_{\text{intr}}) = \mathcal{P}\big(\mathbf{x}(s^i, s^e), \mathbf{R}, \mathbf{t}, \mathbf{c}_{\text{intr}}\big) \tag{2}$$

where $\mathbf{w} \in \mathbb{R}^{2N}$ is the projected facial shape. Also, $\mathbf{R}$ and $\mathbf{t}$ are the relative 3D rotation and translation of the camera with respect to the 3D facial shape and $\mathbf{c}_{\text{intr}}$ are the intrinsic parameters of the camera.

In order to extract "pure" facial expression features, we seek to estimate the expression vector s^e by disentangling it from the rest unknowns, i.e. the subject's identity vector s^i as well as the 3D pose parameters ($\mathbf{R}, \mathbf{t}$) and camera intrinsics ($\mathbf{c}_{\text{intr}}$). For that, we feed the Procrustes-aligned input images to the *DeepExp3D* network that we proposed in [23], a robust and efficient CNN that regresses the expression parameters s^e, having being trained on 5,000 videos with 1,500 different identities and around 9 million frames.

Appearance features The geometric features capture most of the deformations of the face during facial expressions. However, there might be some very fine details (e.g. wrinkles) or facial events (delicate mouth or eyes motions) that are important in terms of expressions but are not able to be captured by the 3D expression vector. For this reason, we expand our feature representation by adding appearance features. We propose a deep network for this step too. We use the same Procrustes-aligned input image and train a CNN

that offers robust facial feature representation. To achieve that, we adopt the ResNet50 [44] network architecture after removing the last fully-connected layer, resulting in a feature vector of size 2048 at the output of the last convolutional layer. We train this network on the facial images of the VGG-face2 dataset [45]. Adam optimiser was used [46] during the training with learning rate of 0.0001, $\beta_1 = 0.9$, $\beta_2 = 0.999$, and batch size 32.

The proposed CNN for extracting appearance features is computationally efficient and further improves the expressiveness of our feature vector representation.

4.2.3 Action unit classification

For the module of AU classification, we also adopt a deep network. Our architecture consists of fully connected layers that take as input the combined feature vector (consisting of geometric and appearance features) and returns a multi-channel binary classification, with a binary label about the existence or non-existence of each AU. In more detail, we use one Fully-Connected Layer (FCL) for each feature vector (geometry vs appearance) with 16 and 256 neurons, respectively, concatenate the results and pass them to an output FCL with 15 neurons representing the 15 targeted AUs. Rectified linear units (ReLU) activation with dropout was applied after each FCL and sigmoid nonlinearity was used at the output. The binary cross-entropy with logits loss function was utilized to train this network.

4.3 Stress recognition based on facial AUs

After having developed a robust and reliable pipeline for AU detection, we apply it to our experimental facial videos to analyse the subjects stress levels. Apart from the AU classification itself, the proposed AU classification network offers also a reliable confidence score about the presence of each AU in every input image. For the purposes of stress analysis, we use the vector with the scores per AU as an *AU feature vector*, which provides a reliable estimate of the *intensity* of every AU. We use AUs for stress analysis and recognition because the AUs present different patterns between neutral and stress states. In most cases, the AU's intensity is significantly higher during stress conditions as compared to a neutral state indicating a more expressive face. As an example, Fig. 3 visualizes typical timeseries of an AU (AU17) for a neutral and a stressful task (adventure video). We observe that the stress state demonstrates significantly increased AU intensities.

4.3.1 Pairwise transformation and normalization

As ordinal regression is performed, it is appropriate to take into account each participant's personalized values on neutral state. This period corresponds to each subject's baseline and using the mapping transformation to rankings [47] generates a common reference to each feature across subjects providing data normalization.

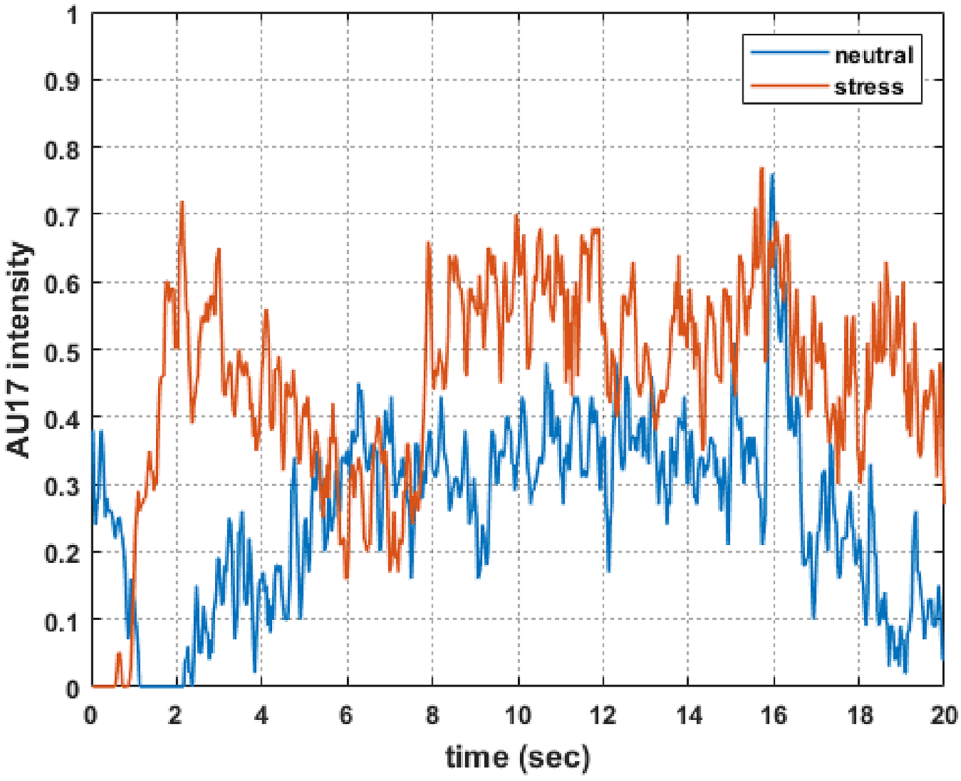

Fig. 3 Plot showing the AU17 intensity temporal evolution over 20 sec of one subject. A neutral state (task 4.1) denoted in blue line and the adventure video (task.4.3) denoted in red line

In this case, the problem of stress detection can be viewed as a ranking problem. In order to transform it into a 2-class classification problem (classes: no stress vs stress), we use the pairwise transformation introduced in [47, 48]. The pairwise transformation which maps the features matrix X (described in 4.2.2) and the class labels Y is described by equation:

$$T : \left\{ \begin{array}{l} X' = X(t_i) - X(t_j) \\ Y' = \text{sign}\{Y(t_i) - Y(t_j)\} \end{array} \right\}, \forall \text{ corresponding } i, j$$

where i, j refer to the indices of neutral and stress states, respectively, with all possible pairs of a specific subject of the feature matrix. The overall transformation procedure is described in Algorithm 1.

Algorithm 1: Pairwise transformation used in this study

Input:
X – feature matrix [cases x features]
Y – classes [1: non-stress, 2: stress]
Output:
X' – pairwise transformed feature matrix
Y' – classes [-1,1]
for each extracted data **do**
X_1, X_2 feature vectors of class Y_1, Y_2 respectively
Find indices i, j of all permutations without repetition of X_1, X_2
for each pair i, j **do**
$X' = X_1(i) - X_2(j)$
if $Y_i > Y_j$ then $Y' = 1$
if $Y_i < Y_j$ then $Y' = -1$
end
end

This transformation creates preference pairs of feature vectors $X(i) - X(j) = [f_1(i) - f_1(j), \ldots, f_M(i) - f_M(j)]$ and their labels $\text{sign}\{Y(i) - Y(j)\}$. If $Y(i) > Y(j)$ then $X(i) \succ X(j)$ and this preference pair is a positive instance, otherwise, it is a negative instance $X(i) \prec X(j)$. The preference pairs and their corresponding labels after transformation can be considered as instances and labels in a new classification problem which then can be performed with traditional classification schemes. This step is significant for the subsequent analysis as it addresses the inter-subject variability taking into account the baseline of each subject of the neutral tasks.

4.3.2 Feature selection and relevance

An issue of great importance is to identify the most relevant/important AUs to the stress conditions which will provide experimental evidence in the scientific research of stress and at the same time offer robust features for our proposed AU-based stress recognition from videos.

In order to achieve increased transparency and intuition in the results of this part of the methodology, we choose to analyse the estimated AU intensities with traditional machine learning techniques rather than deep learning. In more detail, we use Support Vector Machines (SVMs), which remain a widely-used choice for AU-based image analysis [12]. In terms of the SVM kernel, we use the non-linear Radial basis function (RBF) kernel.

The most relevant retained subset was determined by minimizing the misclassification error using 10-fold SVM discrimination accuracy between neutral and stress states. Filter and embedded selection and ranking methods [49] were used in this study and specifically the *minimum Redundancy Maximum Relevance* (mRMR), *Random Forest* (RF) and *Fisher ratio*.

The mRMR algorithm [50] evaluates the features' importance ranking based on maximal relevance and minimum redundancy optimizing in terms of the Mutual Information Quotient (MIQ) criterion [51]. Assume that $m-1$ features are selected from X features and these features form the subset S_{m-1} for selecting the next best feature, it can be calculated by equation:

$$\max_{x_j \in x - S_{m-1}} \left[I(x_j;c) - \frac{1}{m-1} \sum_{x_j \in S_{m-1}} I(x_j;x_i) \right],$$

where the class labels is C and $I(x; y)$ is the mutual information (MI) function defined as:

$$I(x;y) = \iint p(x,y) \log \frac{p(x,y)}{p(x)p(y)} dxdy$$

The algorithm selects and ranks the most relevant to the class label and the least redundancy with the previously selected features.

The Fisher ratio algorithm [52] uses the statistical properties of the classes distributions, pursuing to maximize the between-classes variance whilst at the same time minimize the within-class variance:

$$J(w) = \frac{\sigma^2_{between}}{\sigma^2_{within}} = \frac{(\mu_1 - \mu_2)^2}{\sigma_1^2 + \sigma_2^2}$$

In Fisher's method, the features whose mean difference from the total data mean is higher in various classes receive more ranking weight.

4.3.3 SVM hyperparameters optimization

In order to provide robust SVM models, we tune two SVM hyperparameters, namely the parameters C and ϵ: The parameter C is the penalty for the error term and determines the trade-off between smooth decision boundary and classifying the training points correctly. A big value of C provides a great penalty for misclassification and thus will lead to a

narrower margin and fewer support vectors, while a small value of C will lead to a more smooth decision boundary. The parameter ϵ is the stopping criterion tolerance. High values of ϵ let more training epochs so as the model tries to exactly fit the training data set that may lead to an overfitting phase, while low values of ϵ provide tolerance in classification procedure in favour of generalizability.

We created a grid of hyperparameters in order to test and identify their optimal combination (global optimization). We used the following values for the grid creation $C = 10^{[-7:0.5:2]}$ and $\epsilon = 10^{[-3:1:1]}$.

5 Experimental results

5.1 Evaluation of AU detection accuracy

First, we evaluate our proposed deep pipeline for AU detection (Fig. 2) using the UNBC and BOSPHOTUS datasets. We compare it with the AU detection method of OpenFace [53]. We also compare our pipeline with the corresponding **pipeline of our conference paper** [14], which was consisting of the same conceptual steps but was using hand-crafted features and traditional classification techniques instead of state-of-the-art deep networks.

Furthermore, we also include in these comparisons an **intermediate pipeline**, which uses the deep AU classification module of our proposed pipeline but feeds it with the hand-crafted features of our conference paper [14]. Including this intermediate pipeline can be thought of as a first ablation study, which provides experimental evidence that all steps of the proposed deep pipeline are necessary and contribute in the gradual improvement of the AU detection accuracy. In more detail, this intermediate pipeline is an ablation of the proposed pipeline, as it has been constructed by removing the deep feature extraction from our pipeline and replacing it with a baseline feature extraction with hand-crafted features.

For the evaluation of all methods, we adopt train-test splits of 80% versus 20%, combined with 5-fold cross-validation, while ensuring that frames coming from the same subject do not exist in both training and testing folds at the same time. In terms of classification performance, we adopt the standard classification accuracy defined as:

$$Acc = \frac{TP + TN}{TP + TN + FN + FP} \tag{3}$$

where TP, TN, FP and FN are the number of true positives, true negatives, false positives and false negatives, respectively.

The classification accuracy results for each AU and each pipeline are reported in Table 3. First of all, we observe

Table 3 AU classification accuracy measures using 5-fold cross validation on UNBC & BOSPHORUS.

	AU01	AU02	AU04	AU05	AU06	AU07	AU09	AU10	AU12	AU14	AU15	AU17	AU23	AU25	AU26	MEAN
OpenFace [53]	0.74	0.74	0.87	0.75	0.75	0.81	0.84	0.80	0.76	0.88	0.75	0.82	0.78	0.67	0.68	0.78
Pipeline of [14]	0.84	0.85	0.81	0.84	0.74	0.67	0.94	0.89	0.82	0.88	0.91	0.83	0.90	0.80	0.79	0.83
Intermediate pipeline (ablation)	0.85	0.84	0.84	0.85	0.76	0.69	0.94	0.90	0.84	0.90	0.92	0.85	0.91	0.82	0.81	0.85
Proposed pipeline	**0.99**	**0.99**	**0.98**	**0.98**	**0.89**	**0.89**	**0.99**	**0.99**	**0.89**	**0.98**	**0.99**	**0.98**	**0.99**	**0.91**	**0.92**	**0.96**

Compared methods: **1)** OpenFace AU detection [53] **2)** Pipeline of [14]: hand-crafted features & SVM classification, **3)** Intermediate pipeline: hand-crafted features & **deep** AU classification (ablation of proposed pipeline by replacing deep features with baseline hand-crafted ones), **4)** Proposed pipeline: **deep** features & **deep** AU classification

that the method of OpenFace yields the lowest classification accuracy measures. We observe that the proposed deep pipeline outperforms the pipeline of [14], yielding consistently a much higher classification accuracy. This is attributed to the adoption of state-of-the-art deep networks that yield much more robust and reliable performance. In addition, we observe that the intermediate pipeline, as compared to the pipeline of [14], yields better average accuracy, as well as a better accuracy per AU in almost all the cases. This is due to the fact that replacing the SVM classification with deep fully connected layers offers a better classification performance. In addition, we observe that the proposed deep pipeline yields also a significantly improved performance than the intermediate pipeline, which suggests that replacing the hand-crafted features with deep features yields a much more robust and descriptive feature representation of facial geometry and appearance.

Table 3 provides also some experimental evidence about how the performance of the proposed pipeline varies across different AUs. We observe that the proposed pipeline manages to achieve extremely high classification rates in most AUs (reaching 98% and 99% in most cases). The only exceptions are the Action Units AU6 (Cheek raiser), AU7 (Lid tightener) and AU12 (Lip corner puller), for which the accuracy of our method is relatively lower (89%). These correspond to challenging cases, which create problems to the other tested methods too: AU6 and AU7 correspond movements of the "orbicularis oculi" muscle, which can be mostly detected by subtle movements around the eye region, which sometimes is a difficult task. In addition, AU12 corresponds to a motion of the mouth region, which is sometimes difficult to differentiate from other AUs such as AU14 (Dimpler).

Ablation study of deep feature extraction: We also conduct a thorough ablation study of the feature extraction module of our pipeline for AU detection. This module consists of deep networks that estimate both geometric and appearance features and this experiment evaluates the importance and contribution of both types of deep features in the ability of our pipeline to reliably detect AUs. With reference to the proposed deep pipeline for AU detection (Fig. 2), we include comparisons with: a) a version of our pipeline where we remove the geometric feature extraction with DeepExp3D and we keep the **appearance features only**, as well as b) a version of our pipeline where we remove the appearance feature extraction with the Deep ResNet and we keep the **geometric features only**. For these versions, we follow exactly the same experimental protocol for measuring the AU classification accuracy using 5-fold cross-validation on UNBC and BOSPHORUS datasets. The results are reported in Table 4.

In Table 4, we observe that, when we remove the deep geometric features from the proposed pipeline (first row), the AU classification accuracy drops significantly (from a mean value of 96% to 70%). In addition, when we remove the deep appearance features (second row), the classification accuracy also drops by a non-negligible amount (from a mean value of 96% to 93%). We conclude that the larger part of the success of our AU classification pipeline is attributed to the adopted deep geometric features, which have the ability to describe most of the facial events in a robust and subject-independent manner. In addition to that, incorporating the deep appearance features offers a further boost to the accuracy performance, as they can describe some fine details that might have been missed by the geometric features. In other words, our full pipeline offers the highest classification accuracies by combining the strengths of both geometric and appearance features to address the AU classification problem.

Comparison of different alternatives in training our pipeline. Our deep pipeline for AU detection is not trained in an end-to-end fashion. This is due to the fact that obtaining facial images with reliable AU annotations is very time-consuming and laborious and the available images with AU annotations are not enough for an effective end-to-end training. To support the aforementioned arguments with experimental evidence, we conduct an experiment that trains our deep pipeline for AU detection (Fig. 2) in an end-to-end

Table 4 Ablation study of the proposed pipeline for AU classification, in terms of deep feature extraction. We report AU classification accuracy measures using 5-fold cross validation on UNBC & BOSPHORUS. Compared methods: **1)** Proposed pipeline with **appearance** features **only**, **2)** Proposed pipeline with **geometric** features **only**, **3)** Proposed pipeline with **full version** of deep feature extraction (**appearance and geometric**)

	AU01	AU02	AU04	AU05	AU06	AU07	AU09	AU10	AU12	AU14	AU15	AU17	AU23	AU25	AU26	MEAN
Appearance only	0.67	0.69	0.71	0.63	0.66	0.72	0.74	0.76	0.69	0.68	0.74	0.77	0.73	0.68	0.70	0.70
Geometric only	0.95	0.94	0.95	0.92	0.88	0.88	0.96	0.97	0.89	0.96	0.95	0.96	0.94	0.90	0.91	0.93
Full features (appear. & geom.)	**0.99**	**0.99**	**0.98**	**0.98**	**0.89**	**0.89**	**0.99**	**0.99**	**0.89**	**0.98**	**0.99**	**0.98**	**0.99**	**0.91**	**0.92**	**0.96**

We report AU classification accuracy measures using 5-fold cross validation on UNBC & BOSPHORUS. Compared methods: **1)** Proposed pipeline with **appearance** features **only**, **2)** Proposed pipeline with **geometric** features **only**, **3)** Proposed pipeline with **full version** of deep feature extraction (**appearance and geometric**)

fashion. We have used the same training data with AU annotations and followed the same 5-fold cross-validation process with the case where we train our proposed pipeline without end-to-end training. The classification accuracy results for each AU are reported in Table 5. In this table, we observe that end-to-end training yields consistently worse results, with the mean classification accuracy dropping from 96% to 61%. As explained above, this can be attributed to the lack of a sufficient amount of training data with AU annotations, which leads the end-to-end training process to overfit the training samples.

5.2 Evaluation of the computational efficiency of AU detection

We also evaluate the computational efficiency of the proposed deep pipeline for AU detection and compare it with the corresponding pipeline of our conference paper [14]. For the comparisons, we use a machine with an Nvidia Tesla V100 GPU and an Intel(R) Xeon(R) CPU E5-1660 v4@3.20GHz.

As already mentioned, both the proposed pipeline and the pipeline of the [14] consist of the same three steps (image preprocessing, feature extraction and AU classification). Table 6 provides the average runtimes per input image, broken down per step as well as in terms of the total runtime of the corresponding pipeline. We observe that in both cases, the step of AU classification is the least time-consuming, having a small share on the overall runtime. The larger difference in runtimes between the two pipelines comes from the feature extraction step. This is attributed to the adoption of deep networks, which are especially efficient during this type of inferences, especially when the input is an image. We also observe that the deep pipeline is slower in the step of AU classification, but this difference is compensated by the much faster performance in the other two steps. This difference is due to the fact that the SVM classification used in the pipeline of [14] is equivalent to a single layer, whereas the proposed pipeline adopts multiple fully-connected layers to achieve a much more robust classification performance. Overall, one can clearly see that the proposed pipeline is significantly faster, with a total runtime per image that is more than 40 times smaller than the runtime of the pipeline of [14]. In addition, the average runtime of the proposed pipeline (36ms) corresponds to a framerate of about 28 frames per second on average, which constitutes a real-time performance

Table 6 Comparison of runtimes for AU detection. Average runtimes (in milliseconds) per input image for every step as well as in total for the overall pipelines

Step	Pipeline of [14]	Proposed deep pipeline
Preprocessing	187.13	**20.10**
Feature extraction	1325.24	**13.90**
AU classification	**0.126**	2.00
Total	1512.50	**36.00**

5.3 Statistical analysis of AUs in the stress dataset SRD'15

As described in Sect. 4.3, we have applied our deep pipeline for AU detection in the SRD'15 dataset and have estimated AU intensities for every frame of every video. Using the extracted AU intensities, the SRD'15 dataset was checked for normality for each AU and each task according to the Kolmogorov-Smirnov (KS) test. In most cases, data samples under consideration follow the normal distribution. An initial statistical evaluation (dependent samples t test or the corresponding nonparametric Wilcoxon signed-rank test, respectively) was performed and the results of selected features are presented in Table 7.

Increased AU intensities can be observed along all stressful tasks meaning that the face tends to be more "expressive", i.e. manifesting more intense AU during stress

Table 5 Comparison of different alternatives in training the proposed deep pipeline for AU classification.

	AU01	AU02	AU04	AU05	AU06	AU07	AU09	AU10	AU12	AU14	AU15	AU17	AU23	AU25	AU26	MEAN
With end-to-end training	0.60	0.58	0.67	0.62	0.58	0.66	0.69	0.55	0.57	0.63	0.59	0.54	0.59	0.68	0.61	0.61
Without end-to-end training	**0.99**	**0.99**	**0.98**	**0.98**	**0.89**	**0.89**	**0.99**	**0.99**	**0.89**	**0.98**	**0.99**	**0.98**	**0.99**	**0.91**	**0.92**	**0.96**

We report AU classification accuracy measures using 5-fold cross-validation on UNBC & BOSPHORUS. Compared methods: **1)** Proposed pipeline **with end-to-end training**, **2)** Proposed pipeline **without end-to-end training** (selected strategy)

Table 7 Summary of AU statistics along experimental tasks presenting significant differences during stress conditions

A U	Interview		Recall anxious event		Recall stressful event		IAPS		Stroop CWT		Adventure video		Psychological pressure video	
	p	diff	p	diff	p	diff	p	diff	p	diff	p	diff	p	diff
AU01	0.000	↑					0.004	↑	0.000	↑	0.017	↑	0.006	↑
AU02	0.000	↑					0.016	↑	0.000	↑			0.035	↑
AU04	0.029	↓	0.009	↑			0.000	↑			0.001	↑	0.001	↑
AU05	0.000	↑			0.015	↑	0.005	↑	0.006	↑				
AU06	0.000	↑							0.001	↑				
AU07	0.000	↑	0.02	↑					0.003	↑				
AU09	0.000	↑							0.000	↑	0.005	↑		
AU10	0.000	↑							0.001	↑				
AU12	0.000	↑							0.006	↑				
AU14					0.012	↑					0.041	↑		
AU15	0.000	↑							0.019	↑				
AU17	0.000	↑					0.007	↑	0.000	↑	0.01	↑	0.016	↑
AU20	0.000	↑					0.026	↑	0.000	↑				
AU23	0.000	↑					0.003	↑	0.000	↑				
AU25	0.000	↑					0.011	↑	0.000	↑				
AU26	0.000	↑					0.017	↑	0.000	↑			0.049	↑
AU45	0.001	↑											0.023	↓

↑/↓ significant increase/decrease during stress conditions

ns: non-significant difference

conditions. Only during the emotional recall phase, there were not any significant widespread differences.

5.4 AU involvement in stress conditions

The AUs most implicated in stress conditions were also investigated using the SRD'15 dataset. Towards this end, the mRMR and random forest (RF) were employed. The top-ranked features were inserted iteratively in the feature subset, evaluating each candidate subset's performance in terms of a 10-fold SVM classification accuracy used as the objective function (Eq. (3)). The results are presented in Table 8.

This procedure revealed that for the AU stress detection problem under investigation, a subset of 5 or 6 most relevant features may differentiate effectively the two states. It should be noted that there is a relatively consistent selection of relevant features along the 3 algorithms used.

5.5 Evaluation of AU-based stress recognition

The data used for stress detection were the estimated AU intensities in all frames of the facial videos of SRD'15. The videos were grouped according to the label of the task in 2 groups (no stress vs stress state).

The most relevant features were assessed using Machine Learning classification schemes in terms of their ability to discriminate between the two classes (no stress, stress) for all experimental phases. A 10-fold cross-validation scheme was used utilizing the classifiers k nearest neighbours

Table 8 Deep features Relevant AU implicated in stress conditions using mRMR, RF and Fischer ratio algorithms and their corresponding objective function accuracy

Algorithm	Relevant AU	Classification accuracy (10-fold)
mRMR	AU06, AU26, AU17, AU25, AU10, AU23, AU12, AU15, AU07, AU05, AU14, AU01	0.79
RF	AU06, AU26, AU12, AU07, AU25, AU09, AU01, AU04, AU15, AU10, AU17, AU23, AU05, AU14	0.80
Fischer ratio	AU06, AU12, AU25, AU26, AU14, AU07, AU17, AU23, AU15, AU04, AU05, AU10	0.78

(KNN), Generalized Linear Model (GLM), Naïve Bayes (NVB), Linear Discriminant Analysis (LDA), Support Vector Machines (SVM).

The classifier parameters used in this study are as follows. The KNN parameters (neighbours=10, euclidean distance), the GLM parameters (quadratic model, binomial distribution, logit link function), the NVB parameters (normal distribution, normal kernel smoother) and the LDA parameters (linear discrimination, Gamma regularization parameter=0) and the SVM parameters are (linear kernel function, box constraint=0.358, kernel scale=0.0054).

The classification results are presented in Table 9. The results of Table 9 indicate that the SVM outperforms all other classification schemes with a classification accuracy of 81.1%. Besides, the stress detection was checked with the OpenFace AU detection [53] in order to compare the performance of the proposed algorithm. It is observed that the OpenFace AU detection presents its best performance using the GLM classifier achieving a best 10-fold cross-validation accuracy of 69.3% which is inferior to the results achieved with the proposed pipeline.

Furthermore, we evaluate the system's performance on a subject basis. In this case, we consider the tasks (neutral or stress) of a specific subject as a batch of data and the split into folds was performed based on these batches. A 10-fold cross-validation scheme was again utilized for the same classifiers and the results are shown in Table 10. The results present inferior performance in relation to the dataset-wise approach with a best-achieving performance of 67.4% using the LDA classifier.

It should be taken into account that the method is subject-independent because of the normalization that is performed in the step of face alignment and the associated removal of the rigid information. Besides, except for the interview task, the other tasks do not include intense facial expressions.

From Tables 9 and 10, we also observe that the proposed pipeline achieves a balanced behaviour, since in almost all cases the classification accuracy, sensitivity and specificity rates are very close to each other.

The system performance was also checked along the individual experimental tasks of each phase to assess its discriminatory ability on the different stress types. The classification results, following again a 10-fold cross-validation approach, are presented in Table 11. It can be observed that social exposure (through the interview process) presents the best performance achieving a mean accuracy across classifiers of 94%, followed by the SCWT task with a mean accuracy of 82% and the stressful images stimuli with a mean accuracy of 64.6%. This can be attributed to the fact that the interview and SCWT tasks include increased interaction and speech of the participant, which consequently leads to a more expressive face. It should be noted that the breakdown to individual experimental tasks reduces the cases number available for training, thus limits the power of the analysis.

6 Discussion

This paper proposes a framework of facial Action Units (AU) models for stress recognition. The framework was initially introduced using more traditional, hand-crafted geometric and appearance features combined with machine learning techniques [14]. In this study, it was extended to deep feature representations and deep classifiers.

The system was trained on two publicly available datasets, the UNBC and BOSPHORUS, fusing them in a combined

Table 10 Stress classification results of our proposed pipeline, using **subject-based** 10-fold cross-validation on the SRD'15 dataset.

Classifier	Classification Accuracy (%)	Sensitivity (%)	Specificity (%)
KNN	65.7	66.7	65.0
GLM	66.7	66.7	66.6
NVB	66.1	66.1	66.0
LDA	**67.4**	67.5	67.4
SVM	67.0	67.1	67.0

The results of 5 different classifiers are reported

Table 9 Stress classification comparisons on SRD'15 dataset, using dataset-wise 10-fold cross validation.

Classifiers	Proposed pipeline			OpenFace		
	Accuracy (%)	Sensitivity (%)	Specificity (%)	Accuracy (%)	Sensitivity (%)	Specificity (%)
KNN	76.4	74.3	79.0	67.0	74.0	56.2
GLM	69.5	69.7	73.1	**69.3**	74.7	61.3
Naïve Bayes	68.2	68.6	68.2	56.4	82.7	45.2
LDA	77.5	77.5	77.8	66.0	69.1	59.4
SVM	**81.1**	82.1	80.5	63.2	73.9	48.9

The proposed deep pipeline for AU detection is compared with using OpenFace for AU detection [53]. In both cases, the results of 5 different classifiers are reported

Table 11 Breakdown of stress classification results of our pipeline on SRD'15 per stressful experimental task (Table 1).

Classifiers	Classification accuracy				
	Interview (Task 1.2)	Emotional recall (Tasks 2.2, 2.3)	Stressful images (Task 3.1)	SCWT (Task 3.2)	Stressful videos (Tasks 4.3, 4.4)
KNN	96.0	63.3	68.5	88.0	47.2
GLM	92.0	56.2	56.5	69.0	36.7
Naïve Bayes	82.0	53.6	48.0	84.0	47.2
LDA	100.0	76.9	78.5	92.0	81.4
SVM	100.0	70.7	71.5	77.0	63.7

A dataset-wise 10-fold cross validation is used and the results of 5 different classifiers are reported

model. It was shown that in most cases the combined model presents better performance in comparison to the individual performance of each model. We believe that the inclusion of more datasets in the model would probably further enhance its generalizability and this constitutes an interesting direction for future research.

When the model was applied to the stress dataset (SRD'15), it was deduced that the AUs that implicated and are more relevant to stress conditions according to automatic ranking/relevance algorithms such as the Random Forest were the AU06, AU26, AU12, AU07, AU25, AU09, AU01, AU04, AU15, AU10, AU17, AU23, AU05, AU14. It was interesting that there were relatively consistent features that were selected in all three ranking algorithms used.

An interesting conclusion is that stressful tasks lead to significant increased AU intensities, i.e. a more "expressive" face. This is in accordance with the increased head motility during stress conditions [54] and reduced motility according to the levels of depression severity [55]. Notably, the best classification measures are observed, as expected, on experiment phases where the participant was asked to be more communicative, such as interview and Stroop Colour Word task.

Results indicate that during stress conditions, there are specific AUs that are differentiated from the normal state. There are some facial areas that manifest increased motility, leading to what is considered micro-expressions which is sometimes the result of nervousness and irritability. The classification accuracy between neutral and stress states using this study's pipeline reached 81.1%. The dataset that was investigated in this study had a frame rate of 30fps. Arguably, a better temporal resolution would reveal more cues coming from micro-expressions that in this resolution remain hidden, which constitutes another interesting direction of future research. As a final conclusion, this study reveals that even though physiological processes (such as biosignals, hormones, etc) are more reliable, a non-invasive approach based on solely facial videos like the one proposed here constitutes an interesting alternative, as it achieves promising performance.

References

1. Wethington E, Brown GW, Kessler RC (1995) Interview measurement of stressful life events. Meas Stress: A Guide for Health Soc Sci 59–79
2. Dohrenwend BP, Raphael KG, Schwartz S, Stueve A, Skodol A (1993) The structured event probe and narrative rating method for measuring stressful life events. Free Press, pp 174–199
3. Aigrain J, Spodenkiewicz M, Dubuiss S, Detyniecki M, Cohen D, Chetouani M (2016) Multimodal stress detection from multiple assessments. IEEE Trans Affect Comput 9(4):491–506
4. Chrousos GP (2009) Stress and disorders of the stress system. Nat Rev Endocrinol 5(7):374
5. Giannakakis G, Grigoriadis D, Giannakaki K, Simantiraki O, Roniotis A, Tsiknakis M (2019) Review on psychological stress detection using biosignals. IEEE Trans Affect Computing
6. Giannakakis G, Marias K, Tsiknakis M (2019) A stress recognition system using hrv parameters and machine learning techniques. In: 2019 8th international conference on affective computing and intelligent interaction workshops and demos (ACIIW). IEEE, pp 269–272
7. Weber R, Barrielle V, Soladie C, Seguier R (2018) Unsupervised adaptation of a person-specific manifold of facial expressions. IEEE Trans Affect Comput
8. Henriquez P, Matuszewski BJ, Andreu-Cabedo Y, Bastiani L, Colantonio S, Coppini G, D'Acunto M, Favilla R, Germanese D, Giorgi D et al (2017) Mirror mirror on the wall... an unobtrusive intelligent multisensory mirror for well-being status self-assessment and visualization. IEEE Trans Multimed 19(7):1467–1481
9. Darwin C (1872) The expression of the emotions in man and animals. John Marry, London, UK
10. Giannakakis G, Pediaditis M, Manousos D, Kazantzaki E, Chiarugi F, Simos P, Marias K, Tsiknakis M (2017) Stress and anxiety detection using facial cues from videos. Biomed Signal Process Control 31:89–101
11. Korda AI, Giannakakis G, Ventouras E, Asvestas PA, Smyrnis N, Marias K, Matsopoulos GK (2021) Recognition of blinks activity patterns during stress conditions using cnn and markovian analysis. Signals 2(1):55–71
12. Martinez B, Valstar MF, Jiang B, Pantic M (2017) Automatic analysis of facial actions: a survey. IEEE Trans Affect Comput
13. Donato G, Bartlett MS, Hager JC, Ekman P, Sejnowski TJ (1999) Classifying facial actions. IEEE Trans Pattern Anal Mach Intell 21(10):974–989
14. Giannakakis G, Koujan MR, Roussos A, Marias K (2020) Automatic stress detection evaluating models of facial action units. In: 2020 15th IEEE international conference on automatic face and gesture recognition (FG 2020), pp 817–822

15. Ruiz A, Van de Weijer J, Binefa X (2015) From emotions to action units with hidden and semi-hidden-task learning. In: Proceedings of the IEEE international conference on computer vision, pp 3703–3711
16. Chu W-S, De la Torre F, Cohn JF (2017) Learning spatial and temporal cues for multi-label facial action unit detection. In: 2017 12th IEEE international conference on automatic face & gesture recognition (FG 2017). IEEE, pp 25–32
17. Shao Z, Liu Z, Cai J, Ma L (2018) Deep adaptive attention for joint facial action unit detection and face alignment. In: Proceedings of the European conference on computer vision (ECCV), pp 705–720
18. Ma C, Chen L, Yong J (2019) Au r-cnn: encoding expert prior knowledge into r-cnn for action unit detection. Neurocomput 355:35–47
19. Bevilacqua F, Engstrom H, Backlund P (2018) Automated analysis of facial cues from videos as a potential method for differentiating stress and boredom of players in games. Int J Comput Games Technol
20. Daudelin-Peltier C, Forget H, Blais C, Deschˆenes A, Fiset D (2017) The effect of acute social stress on the recognition of facial expression of emotions. Sci Rep 7(1):1036
21. Gavrilescu M, Vizireanu N (2019) Predicting depression, anxiety, and stress levels from videos using the facial action coding system. Sensors 19(17):3693
22. Viegas C, Lau S-H, Maxion R, Hauptmann A (2018) Distinction of stress and non-stress tasks using facial action units. In: Proceedings of the 20th international conference on multimodal interaction: Adjunct, pp 1–6
23. Koujan MR, Alharbawee L, Giannakakis G, Pugeault N, Roussos A (2020) Real-time facial expression recognition "in the wild" by disentangling 3d expression from identity. In: IEEE international conference on automatic face and gesture recognition (FG 2020)
24. Lucey P, Cohn JF, Prkachin KM, Solomon PE, Matthews I (2011) Painful data: the unbc-mcmaster shoulder pain expression archive database. In: Automatic face & gesture recognition and workshops (FG 2011), 2011 IEEE international conference On. IEEE, pp 57–64
25. Savran A, Alyüz N, Dibeklioğlu H, Çeliktutan O, Gökberk B, Sankur B, Akarun L (2008) Bosphorus database for 3d face analysis. In: European workshop on biometrics and identity management. Springer, pp 47–56
26. Stroop JR (1935) Studies of interference in serial verbal reactions. JExp Psychol 18(6):643
27. Lang PJ, Bradley MM, Cuthbert BN et al (1997) International affective picture system (iaps): technical manual and affective ratings. NIMH Center Study Emot Atten 1:39–58
28. Andreu Y, Chiarugi F, Colantonio S, Giannakakis G, Giorgi D, Henriquez P, Kazantzaki E, Manousos D, Marias K, Matuszewski BJ, Pascali MA, Pediaditis M, Raccichini G, Tsiknakis M (2016) Wizemirror - a smart, multisensory cardio-metabolic risk monitoring system. Comput Vision Image Underst 148:3–22
29. Ekman P (2002) Facial action coding system (FACS). A human face
30. Ekman P, Friesen W (1978) Facial action coding system (FACS): manual. Consulting Psychol Press Palo Alto
31. Hjortsjo C-H (1969) Man's face and mimic language. Studentlitteratur
32. Najibi M, Samangouei P, Chellappa R, Davis LS (2017) Ssh: single stage headless face detector. In: Proceedings of the IEEE international conference on computer vision, pp 4875–4884
33. Deng J, Zhou Y, Cheng S, Zaferiou S (2018) Cascade multi-view hourglass model for robust 3d face alignment. In: FG
34. Matthews I, Baker S (2004) Active appearance models revisited. Int J Comput Vis 60(2):135–164
35. Cootes TF, Taylor CJ (2004) Statistical models of appearance for computer vision. Technical report, University of Manchester
36. Watson D (2013) Contouring: a guide to the analysis and display of spatial data, vol. 10. Elsevier
37. Blanz V, Vetter T (1999) A morphable model for the synthesis of 3d faces. In: Proceedings of the 26th annual conference on computer graphics and interactive techniques. ACM Press/Addison-Wesley Publishing Co., pp 187–194
38. Zafeiriou S, Chrysos GG, Roussos A, Ververas E, Deng J, Trigeorgis G (2017) The 3d menpo facial landmark tracking challenge. In: ICCV, pp 2503–2511
39. Deng J, Roussos A, Chrysos G, Ververas E, Kotsia I, Shen J, Zafeiriou S (2018) The menpo benchmark for multi-pose 2d and 3d facial landmark localisation and tracking. IJCV
40. Koujan MR, Roussos A (2018) Combining dense nonrigid structure from motion and 3d morphable models for monocular 4d face reconstruction. In: CVMP
41. Gecer B, Ploumpis S, Kotsia I, Zafeiriou S (2019) Ganfit: generative adversarial network fitting for high fidelity 3d face reconstruction. arXiv preprint arXiv:1902.05978
42. Booth J, Roussos A, Ponniah A, Dunaway D, Zafeiriou S (2018) Large scale 3d morphable models. IJCV
43. Cao C, Weng Y, Zhou S, Tong Y, Zhou K (2014) Facewarehouse: a 3dfacial expression database for visual computing. IEEE Trans Vis Comp Gr 20(3):413–425
44. He K, Zhang X, Ren S, Sun J (2016) Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 770–778
45. Cao Q, Shen L, Xie W, Parkhi OM, Zisserman A (2018) Vggface2: a dataset for recognising faces across pose and age. In: International conference on automatic face and gesture recognition
46. Kingma D, Ba J (2014) Adam: a method for stochastic optimization. International conference on learning representations
47. Herbrich R, Graepel T, Obermayer K (1999) Support vector learning for ordinal regression. In: 1999 ninth international conference on artificial neural networks ICANN 99, vol 1, pp 97–1021. https://doi.org/10.1049/cp:19991091
48. Fürnkranz J, Hüllermeier E (2003) Pairwise preference learning and ranking. In: European conference on machine learning. Springer, pp 145–156
49. Chandrashekar G, Sahin F (2014) A survey on feature selection methods. Comput Elect Eng 40(1):16–28
50. Ding C, Peng H (2005) Minimum redundancy feature selection from microarraygene expression data. J Bioinf Comput Biol 3(02):185–205
51. Gulgezen G, Cataltepe Z, Yu L (2009) Stable and accurate feature selection. In: Joint European conference on machine learning and knowledge discovery in databases. Springer, pp 455–468
52. Gu Q, Li Z, Han J (2012) Generalized fisher score for feature selection. arXivpreprint https://arxiv.org/abs/1202.3725
53. Baltrusaitis T, Zadeh A, Lim YC, Morency L-P (2018) Openface 2.0: facial behavior analysis toolkit. In: 2018 13th IEEE international conference on automatic face & gesture recognition (FG 2018). IEEE, pp 59–66
54. Giannakakis G, Manousos D, Chaniotakis V, Tsiknakis M (2018) Evaluation of head pose features for stress detection and classification. In: 2018 IEEE EMBS international conference on biomedical & health informatics (BHI), pp 406–409
55. Anis K, Zakia H, Mohamed D, Jeffrey C (2018) Detecting depression severity by interpretable representations of motion dynamics. In: 2018 13th IEEE international conference on automatic face & gesture recognition (FG2018), pp 739–745