Real-time Facial Expression Recognition "In The Wild" by Disentangling 3D Expression from Identity

Mohammad Rami Koujan[1], Luma Alharbawee[1], Giorgos Giannakakis[2], Nicolas Pugeault[1], Anastasios Roussos[2]

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

[2]Institute of Computer Science (ICS), Foundation for Research and Technology - Hellas (FORTH), Greece

***Abstract*— Human emotions analysis has been the focus of many studies, especially in the field of Affective Computing, and is important for many applications, e.g. human-computer intelligent interaction, stress analysis, interactive games, animations, etc. Solutions for automatic emotion analysis have also benefited from the development of deep learning approaches and the availability of vast amount of visual facial data on the internet. This paper proposes a novel method for human emotion recognition from a single RGB image. We construct a large-scale dataset of facial videos (FaceVid), rich in facial dynamics, identities, expressions, appearance and 3D pose variations. We use this dataset to train a deep Convolutional Neural Network for estimating expression parameters of a 3D Morphable Model and combine it with an effective back-end emotion classifier. Our proposed framework runs at 50 frames per second and is capable of robustly estimating parameters of 3D expression variation and accurately recognizing facial expressions from in-the-wild images. We present extensive experimental evaluation that shows that the proposed method outperforms the compared techniques in estimating the 3D expression parameters and achieves state-of-the-art performance in recognising the basic emotions from facial images, as well as recognising stress from facial videos.**

I. Introduction

As the Computer Vision (CV) and Machine Learning (ML) fields advance, the study of human faces progressively receives a notable attention due to its central role in a plethora of key applications. Human emotion recognition is an increasingly popular line of research, around which many research studies revolve. The aim of most of these studies is to automate the process of recognising a human's emotion from a captured image. Solving this problem successfully is immensely beneficial for a myriad of applications, e.g. human-computer intelligent interaction, stress analysis, interactive computer games, emotions transfer, most of which have been the focus of the Affective Computing field.

The recent availability of large benchmarks of facial expression (images and videos) and the fast development of deep learning approaches has lead to high performance in facial expression recognition for image data captured in both controlled and unconstrained conditions ("in the wild").

This paper presents a novel approach for recognizing human emotions from a *single* facial image. The proposed approach capitalises on the recent advancements in 3D face reconstruction from monocular videos and Convolutional Neural Networks (CNNs) architectures that proved effective in the CV field. Our method is driven by the idea of disentangling the subject's expression from identity with the aid of *3D Morphable Models (3DMM)* [3]. Given a *single* image, we regress a vector representing the *3D expression* of the depicted subject with the help of a novel Deep Convolutional Neural Network (DCNN), which we call *DeepExp3D*. This expression vector is ideal as a feature vector since it achieves various invariances (with respect to the individual's facial anatomy, 3D head pose and illumination conditions), and we show that an emotion classifier trained on this feature can recognise expressions reliably and robustly. Our contributions in this work can be summarized as[1]:

- Collection and annotation of a new large-scale dataset of human facial videos (6,000 in total), which we call **FaceVid**. With the help of an accurate model-based approach that we propose to use during training, each video is annotated with the per-frame: 1) facial landmarks, 2) 3D facial shape composed additively of identity and expression parts, 3) relative 3D pose of the head with respect to the camera.
- A robust deep convolutional neural network (CNN), termed as **DeepExp3D**, for regressing the expression parameters of a 3D Morphable Model of the facial shape from a single input image. Our network is robust to occlusions, illumination and view angle changes, and regresses the expression independently of the person's identity.
- We connect DeepExp3D with a classification module for the **Facial Expression Recognition** (FER) task from the estimated expression vectors, leading to an integrated framework for the robust recognition of facial expressions from single images.

Our trained DeepExp3D: **1)** outperforms state-of-the-art 3D face reconstruction methods in estimating the facial expression parameters, **2)** achieves state-of-the-art performance in FER from images and stress recognition from videos, **3)** can also be incorporated in other frameworks seeking to, e.g., recover the 3D geometry of facial image, image-to-image translation, facial reenactment, etc.

II. Related work

There is a large body of work tackling the seven-class problem of static facial expression recognition defined by Ekman and Friesen [16]. Current facial expression recognition methods divide approximately into two groups: traditional handcrafted methods (appearance, geometric, dynamic and fusion) and deep learning models. Handcrafted methods

[1]Project page: https://github.com/mrkoujan/FER

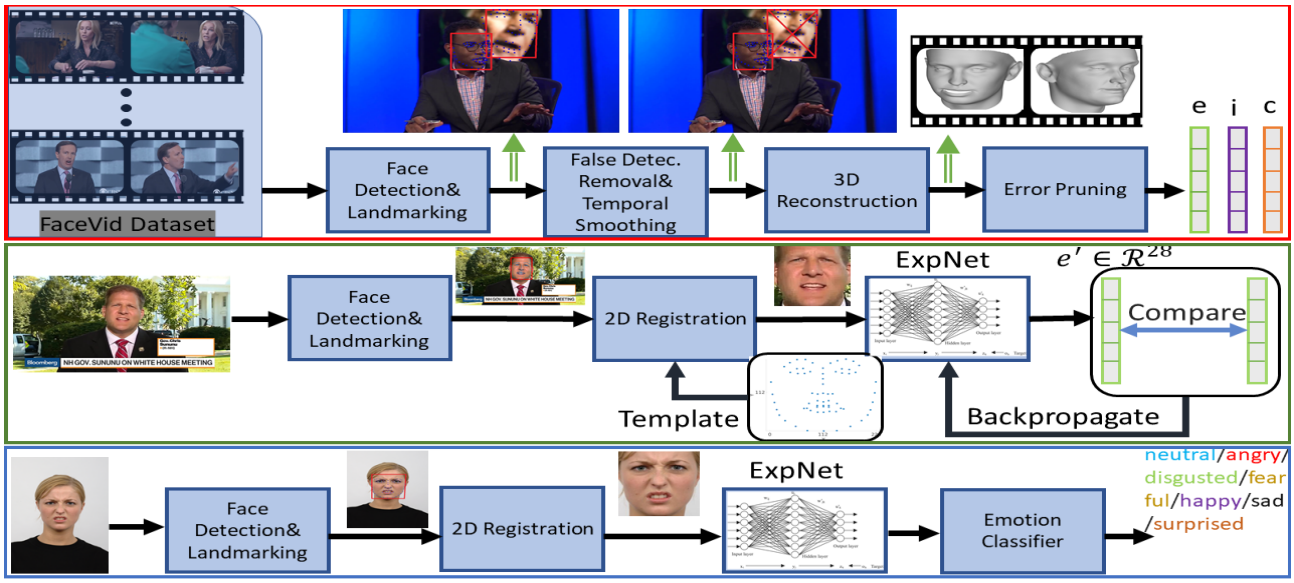

Fig. 1. Proposed framework of Facial Expression Recognition (FER) from images. Top: FaceVid annotation process (sec. III-A and III-B). Middle: training of **DeepExp3D** (sec. III-C). Bottom: final framework for FER (sec. III-D). Vectors **e**, **i**, **c** estimated in the annotation process (top) represent facial expression, identity and camera parameters, respectively.

have been widely adopted for FER and rely on features [33], [13]. Nevertheless, they have shown their restrictions in practical applications [35], [38]. Lately, deep learning, especially Convolutional Neural Networks (CNN), methods have proved competitive in many vision tasks, e.g, image classification, segmentation, emotion recognition, etc.

Xiao et al. [64] tackle the poor generalization of deep neural networks when enough data is not available by combining region of interest (ROI) and K-nearest neighbors (KNN) for facial expression classification. In [2], an attention model composed of a deep CNN learns the location of emotional expression in a cluttered scene, leading to an improved facial expression recognition. Liu et al. [44] proposed a deep learning approach trained on a geometric model of facial regions for facial expression analysis. Tang [60] proposed a CNN backed with a linear support vector machine (SVM) at the output and achieved the first place on the FER-2013 Challenge [28]. Liu et al. [45] suggested a facial expression recognition framework with 3D CNN and deformable action parts constraints to jointly localize specific facial action parts and recognize facial expressions. Peng [53] focused on a synthesis CNN to produce a non-frontal view from a single frontal face and Richardson et al. [54] transferred the face geometry from its image directly via a CNN based approach. The authors of [51] encode deep convolutional neural networks (DCNN) features with covariance matrices for facial expression recognition. In their paper, they show that covariance descriptors computed on DCNN features are more efficient than the standard classification with fully connected layers and softmax. For systematic and exhaustive surveys on automatic FER, we refer the reader to [19], [39].

In contrast to these approaches, we propose to estimate an intermediate 3D-based representation of "pure" facial expression that is *invariant* to all other parameters that contribute to the formation of the input image (shape and appearance variation related to the subject's identity, relative 3D pose variation, occlusions, strong illumination variations and other challenges of in-the-wild images). This means that, in contrast to the standard practice of most Deep Learning approaches in CV, we are not seeking to solve our problem (FER from single RGB images) in an "end-to-end" fashion; this would require a vast number of manually-annotated images in the level of facial expression classes, which would be laborious and prone to human annotation errors. Instead, we construct a large-scale video dataset and annotate all video frames individually with the expression parameters of a 3DMM. We are able to reliably automate this annotation process using an approach of 3D face reconstruction from videos that achieves high accuracy by exploiting the rich dynamic information that facial videos have. Using this dataset to learn to regress expression parameters from single RGB images, we massively simplify the problem of FER, since we use the expressions parameters as the features that feed our emotion classifier. These expression parameters are low-dimensional (28 dimensions) and exhibit a very wide range of invariance properties, therefore a classifier can be trained to recognise emotion classes with significantly less annotated data. This approach leads to a robust FER system that can deal with particularly challenging images.

The work that is most closely related to our approach is the so-called ExpNet [7], which uses a CNN to regress 3DMM-based expression coefficients from a single facial image. The proposed approach achieves vastly superior recognition performance (from 21% to 35% higher recognition accuracy on 5 different benchmarks, see Table III) with a significantly faster runtime (more than 4 times faster, see Table II). In contrast to ExpNet, our approach adopts a pre-processing step of 2D registration in the image space to a template

mean face, see Fig.1 middle. This significantly reduces the variability of input images and makes the estimations more robust and reliable [49]. Furthermore, the 3DMM that we use to model identity variation is the LSFM Model [4], which has been trained in 2 orders of magnitude more facial identities than the Basel Face Model [52] adopted in ExpNet, achieving much more accurate representation of the 3D shape of human faces [4].

III. Methodology

Figure 1 demonstrates an overview of the proposed framework. Motivated by the progress in the 3D facial reconstruction from images and the rich dynamic information accompanying videos of facial performances, we collected a large-scale dataset of facial videos from the internet (section III-B) and recovered the per-frame 3D geometry thereof with the aid of 3D Morphable Models (3DMMs) [3] of identity and expression (section III-A). The annotated dataset was used to train the proposed **DeepExp3D** network, in a supervised manner for regressing the expression coefficients vector $\mathbf{e}_f$ from a single input image $\mathbf{I}_f$ (section III-C). As a final step, a classifier was added to the output of the DeepExp3D to predict the emotion of each estimated facial expression, and was trained and tested on standard benchmarks for FER (section III-D).

A. 3D Face Reconstruction From Videos

1) Combined Identity and Expression 3D Face Modelling: Following several recent methods [66], [11], [36], [21], we model the 3D face geometry using 3DMMs and an additive combination of identity and expression variation. In more detail, let $\mathbf{x} = [x_1, y_1, z_1, ..., x_N, y_N, z_N]^T \in \mathbb{R}^{3N}$ be the vectorized form of a 3D facial shape consisting of N 3D vertices. We consider that any facial shape $\mathbf{x}$ can be represented using the following model of shape variation:

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

where $\bar{\mathbf{x}} \in \mathbb{R}^{3N}$ is the overall mean shape vector, given by $\bar{\mathbf{x}} = \bar{\mathbf{x}}_{id} + \bar{\mathbf{x}}_{exp}$, where $\bar{\mathbf{x}}_{id}$ and $\bar{\mathbf{x}}_{\mathbf{exp}}$ are the mean identity and mean expression shape vectors respectively. $\mathbf{U}_{id} \in \mathbb{R}^{3N \times n_i}$ is the orthonormal basis with $n_i = 157$ principal components ($n_i \ll 3N$) , $\mathbf{U}_{exp} \in \mathbb{R}^{3N \times n_e}$ is the orthonormal basis with the $n_e = 28$ principal components ($n_e \ll 3N$), and $\mathbf{i} \in \mathbb{R}^{n_i}$, $\mathbf{e} \in \mathbb{R}^{n_e}$ are the identity and expression parameters. In the adopted model (1), the 3D facial shape $\mathbf{x}$ is a function of both identity and expression coefficients ($\mathbf{x}(\mathbf{i},\mathbf{e})$). Additionally, the expression variations are effectively represented as offsets from a given identity shape.

The identity part of the model, $\{\bar{\mathbf{x}}_{id}, \mathbf{U}_{\mathbf{id}}\}$, originates from the LSFM [4] built from approximately 10,000 scans of different people, the largest 3DMM ever constructed, with varied demographic information. In addition, the expression part of the model, $\{\bar{\mathbf{x}}_{exp}, \mathbf{U}_{exp}\}$ originates from the work of Zafeiriou et al. [66], who built it using the blendshapes model of Facewarehouse [6] and adopting Nonrigid ICP [9] to register the blendshapes model with the LSFM model.

To create effective pseudo-ground truth, we need to perform 3D face reconstruction on an especially large-scale video dataset that is both efficient and accurate. For this reason, we choose to fit the adopted 3DMM model on the sequence of facial landmarks over each video of the dataset. Since this process is intended for the creation of pseudo-ground truth on a large collection of videos, we are not constrained by the need of online performance. Therefore, we adopt the approach of [11] (with the exception of the initialization stage, as described next), which is a batch approach that takes into account the information from all video frames simultaneously and exploits the rich dynamic information usually contained in facial videos. This is an energy minimization approach to fit the combined identity and expression 3DMM model on facial landmarks from all frames of the input video simultaneously. We utilise the so-called 3D-aware 2D landmarks which we extract with [12]. The localised 68 landmarks with this method correspond to projections of their corresponding 3D points on the 3D face. More details are given in the Supplementary Material.

2) Initialization Stage of Estimating Camera Parameters: In this stage of the 3D video reconstruction proposed in [11], the camera parameters are estimated using rigid Structure from Motion (SfM). This works reliably for facial videos with substantial head rotation, since it creates the required variation in the relative 3D pose that is typically needed in SfM. However, in cases of videos with almost no or very little head rotation (e.g. a video of a person looking straight at the camera and talking), SfM yields a very unstable estimation of the camera parameters, due to the ambiguities caused when viewing the scene from almost the same view point. To overcome this limitation and exploit much wider types of facial videos, we adopt a substantially different approach in this stage, which utilizes earlier the adopted 3D face model and effectively constraints the problem, yielding not only robust but also accurate estimations.

In more detail, similar to [11], our initialization stage assumes that the shape to be recovered remains rigid over the whole video. This assumption is over-simplistic but is adequate for an accurate estimation of camera parameters, since the deformations in human faces can be reliably modelled as localized deviations from a rigid shape. However, in contrast to [11], we do not seek to estimate the full degrees of freedom of the 3D facial shape (i.e. every coordinate of every point of the 3D shape being a separate independent parameter); instead we significantly reduce the allowed degrees of freedom by imposing the constraint that it is synthesised using the 3D face model (1). This makes our camera estimations much more robust. Please refer to the supplementary materials for more details about the implementation of this stage.

B. Ground Truth Creation from a Large-scale Videos Dataset

This section describes how we process a very large-scale dataset to construct pseudo-ground truth, which was used to train the **DeepExp3D**, a robust CNN capable of regressing the 3DMM expression parameters from a single RGB image.

Dataset	# images	# subjects	Emotions	Elicitation	Resolution
RadFD[37]	8040	67	7 B+ 1 N	posed	681×1024
KDEF[47]	4900	70	6 B+ 1 N	posed	562×762
RAF-DB[41]	29672	N/A	6 B+ 1 N, 12 C	posed& spontaneous	web images
CFEE[15]	5060	230	6 B+ 1 N, 15 C	posed	1000×750
CK+[46]	327 seq. (10 to 60 frames/seq.)	210	7 B+ 1 N	posed& spontaneous	640×480

TABLE I

PUBLIC DATABASES OF EMOTIONS UTILISED IN THIS PAPER. B, N, C STAND FOR BASIC, NEUTRAL AND COMPOUND EMOTIONS, RESPECTIVELY.

We start from a collection of 6,000 RGB videos with 12 million frames in total and 1,700 unique identities. Please refer to the Supplementary Material for the specifics of the collection process. In every frame of every video of our video collection, we applied the method of [12] to detect faces and extract from each detected face a set of 68 landmarks, according to the MULTI-PIE markup scheme [29]. Afterwards, we applied the following steps:

False detections removal: This was implemented by tracking each detected face in the first frame throughout the processed video. A face is kept if its bounding box (BB) stays within a reasonable margin, chosen experimentally to be half the width of the BB, compared to its location in the previous frame. We pruned videos in which we lost track of the face for K consecutive frames (chosen experimentally to be 5) before reaching the desired number of tracked frames F (chosen experimentally to be 2000). This step helped to remove false detections arising due to a failure in the face detector or out-of-context detections, e.g. a facial photo in the background of a video, faces that pop in/out of the camera viewing angle, etc. This step resulted in pruning 1000 videos (16.7% of the initial dataset).

Temporal smoothing: Extracted landmarks were temporally smoothed using cubic splines. This was performed to alleviate the effects of the potential jitters in the extracted landmarks between consecutive frames and to fill in the possible gaps (frames with lost tracking) that persisted for less than K frames.

3D facial reconstruction from videos: For every video, we followed the process described in Sec. III-A and estimated the facial shape parameters ($\mathbf{i}$, $\mathbf{e}_f$ for $f = 1, ..., F$). The final output of pseudo-ground truth creation is the sequence of expression vectors $\{\mathbf{e}_f\}$. However, we also utilise the identity vector $\mathbf{i}$ in the next step as one means of error pruning.

Error pruning: With such a large number of videos, there will be some cases of videos where 3D reconstruction has failed. This is an unavoidable byproduct of the fact that the adopted landmark localization, even though very robust, might not be sufficiently accurate for cases of extremely challenging facial videos. Our approach compensates for that by two stages of pruning problematic videos:

a) Automatic pruning: We are based on the fact that under the adopted 3D face modelling (1), the coordinates of the estimated identity vector $\mathbf{i}$ of each video are assumed independent, identically distributed random variables that follow a normal distribution. Therefore, we classify as outliers and automatically prune the videos that correspond to an estimated value of $\|\mathbf{i}\|$ above an appropriate threshold. More details are given in the Supplementary Material. This resulted in automatically pruning 300 more videos (5% of the initial dataset).

b) Manual pruning: There might be a few problematic videos that "survived" the automatic pruning. For that reason, we inspected the reconstructions of all remaining videos and manually flag and prune videos where it is evident that the 3D face reconstruction has failed. In this step we manually pruned 250 videos (around 4% of the initial dataset).

To conclude, our constructed training set consists of videos of our collection that survived the aforementioned steps of video pruning. It consists of 5,000 videos (83.33% of the initial dataset) with 1,500 different identities and around 9M frames. For exemplar visualisations, please refer to the Supplementary Material.

C. DeepExp3D Network

The constructed training dataset of videos is rich in the facial expressions that are viewed from different angles and under various illumination conditions throughout the video, as well as in identities (1500 in total). This substantially facilitates the process of training a Convolutional Neural Network (CNN) $\mathscr{N} : I \rightarrow \mathbf{e}$, aiming at regressing the 3DMM facial expression coefficients (referred to as $\mathbf{e}$ in equation (1)) from a given RGB image I. The network $\mathscr{N}(I)$ learns during the training phase how to map from the image space to the facial expression space irrespective of the subject's identity shown in image I. This is achievable by the virtue of the utilized facial 3DMM which represents the reconstructed face as a summation of identity and expression parts on top of the model mean face ($\bar{\mathbf{x}}$), see equation (1). We extract vectors $\mathbf{e}$ from our dataset as a result of the fitting approach explained in section III-A and use them as pseudo annotations for training $\mathscr{N}$ in a supervised manner. However, to avoid teaching $\mathscr{N}$ the exact behaviour of our linear model-based fitting approach for estimating the facial expression parameters, we fine-tune our DCNN ($\mathscr{N}$) on the 4DFAB dataset [8]. The 4DFAB dataset is a large-scale database of dynamic high-resolution 3D faces (more than 1.8M 3D face) with subjects displaying both spontaneous and posed facial expressions. The ResNet [30] network structure was selected and trained after replacing the output softmax layer by a linear regression layer of $n_e = 28$ neurons. Before starting the training, dataset frames were aligned to a template of size 224×224 having the 68 points mark-up [56] projected from the mean 3D face $\bar{\mathbf{x}}$ into the image space. In total, the trained DeepExp3D is a mapping: $\mathscr{R}^{224\times224} \rightarrow \mathscr{R}^{28}$. Note

TABLE II

COMPARISON OF REQUIRED RUN-TIME TO PRODUCE ESTIMATIONS OF FACIAL EXPRESSION FROM A SINGLE IMAGE.

Method	ITW [5]	SfM3DMM [36]	3DDFA [68]	ExpNet [7]	**Ours**
Time (sec)	6.4	3.0	0.6	0.088	**0.02**

finally that 70% of the dataset was used for training and the rest were halved for testing and validation. While training, the network minimises the ℓ_2 norm error between the output and the ground-truth facial expression parameters.

D. Back-end Emotion Classifier

To classify the generated facial expression vectors $\mathbf{e} \in \mathscr{R}^{28}$ produced by the DeepExp3D network $\mathscr{N}$, Error Correcting Output Codes (ECOC) method [14] was utilised to solve this 7-class (neutral + six basic emotions) classification problem. ECOC strategy combines multiple binary learners to solve the multi-class classification problem. Our binary learner of choice is Support-Vector Machines (SVM). 10-fold cross validation with the one-versus-all [50] coding scheme were implemented to train/test the emotion classifier. The SVM hyper-parameters were optimized using the Bayesian optimization approach [58]. 68-landmarks were extracted from the images of all the employed emotion datasets in table I and used to register them to the mean face template, as done in section III-C.

IV. EXPERIMENTAL RESULTS

In this section, we present extensive qualitative and quantitative evaluations and comparisons of our pipeline, as well as its intermediate steps.

Implementation and runtimes. Our method uses the ResNet [30] CNN structure with 50 layers, implemented in TensorFlow [17]. For both training and testing, we use a machine with Nvidia Tesla V100 GPU and Intel(R) Xeon(R) CPU E5-1660 v4@3.20GHz. Our overall FER framework achieves 20ms of total processing time per image (i.e. 50 fps when applied on videos). Using the same machine, we also ran methods that solve the same (FER) or closely-related problems (3D shape estimation with disentanglement of identity and expression) using single-image input, see Table II. We observe that our method is at least 4 to 320 times faster than the other tested methods. This is mainly due to the particularly compact and descriptive representation of facial expressions that is achieved in our framework.

A. Facial Expression Recognition

To evaluate our FER method, 5 publicly available datasets for emotion recognition were used, namely: Radboud [37], KDEF [47], RAF-DB [41], CFEE [15], CK+ [46]. All five datasets have basic emotion [16] annotations (happy, sad, fearful, angry, surprised, disgusted), as well as the neutral expression. Presentation of results per dataset follows:

First of all, the **Radboud dataset** [37] has 67 subjects imaged from 5 different angles each at the same time. To test the performance of our network on recognising an emotion

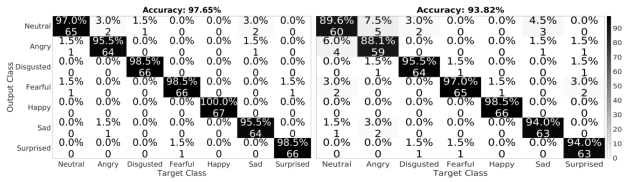

Fig. 2. Confusion matrices generated by our emotion classifier when running 10-fold cross validation on RadFD[37] on only frontal images (**left**) and one of the semi-profile images (**right**) of each subject.

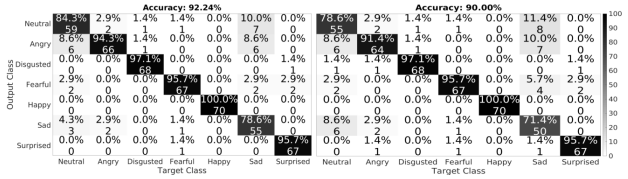

Fig. 3. Confusion matrices generated by our emotion classifier when running 10-fold cross validation on the KDEF dataset on only frontal images (**left**) and only one of the semi-profile images (**right**) of each subject.

from dissimilar view angles, we run two experiments. In the first, the frontal image of each subject showing a specific emotion was kept ($67 \times 7 = 469$ images in total), while in the second experiment one of the semi-profile faces (captured from 45/135 degrees) of each subject was selected randomly and used for 10-fold cross validation. Figure 2 reports the confusion matrices and accuracies obtained in both cases. The average MSE of expression parameters generated from semi-profile and frontal images is 0.008 over all the subjects. The comparable generated accuracies in figure 2, as well as the small MSE, demonstrate the ability of the DeepExp3D in producing view-angle independent expression estimations. As shown in table III, our proposed approach produces the highest **accuracy** (97.63%) on the RadFD [37] dataset compared to recent state-of-the-art methods.

The **KDEF dataset** [47] is similar in structure to RadDF [37] where each of the 70 subjects was pictured from five different angles at the same time (0°, 45°, 90°, 135°, 180°). Each subject was asked to elicit the same emotion twice, only one thereof was picked randomly. Only frontal images ($7 \times 70 = 490$ in total) were employed in the reported results in table III. We attain the best **accuracy** compared to other state-of-the-art methods (92.24%), revealing the power of our DeepExp3D in generating separable facial expressions according to their basic emotion label. Figure 3 demonstrates the confusion matrices generated by our emotion classifier on either frontal images (left) or semi-profile (45°) images (right). Both reported confusion matrices emphasize the high separability of happy and disgusted labels from the rest of the emotions (100% and 97.1%, respectively), while it seems that the sad and neutral expressions tend to group closely (78.6% and 84.3% for frontal, and 71.4% and 78.6% for semi-profile images, respectively).

The **CFEE dataset** [15] was collected from 230 subjects with two groups of labelled images, basic and compound. Images labelled with basic emotions (total of 1836) were passed to the DeepExp3D and then for training/testing our

TABLE III

COMPARISON OF FACIAL EXPRESSION RECOGNITION ACCURACIES ON 5 WIDELY-USED BENCHMARKS.

Dataset	Approach	Acc.(%)	Dataset	Approach	Acc.(%)	Dataset	Approach	Acc.(%)
RadFD	ExpNet [7]	75.00	RAF-DB	ExpNet [7]	55.20	KDEF	ExpNet [7]	71.00
	Ali et al. [1]	85.00		Li& Deng [40]	74.20		Zavarez et al. [67]	72.55
	Zavarez et al. [67]	85.97		Lin et al. [43]	75.73		Ali et al. [1]	78.00
	Jiang& Jia [33]	94.52		Fan et al. [18]	76.73		Ruiz-Garcia et al. [55]	86.73
	Mavani et al. [48]	95.71		Gosh et al. [22]	77.48		Yaddaden et al. [65]	90.62
	Wu& Lin [63]	95.78		Shen et al. [57]	78.60		**Ours**	**92.24±0.70**
	Sun et al. [59]	96.93		Vielzeuf et al. [61]	80.00	CK+	ExpNet [7]	61.17
	Yaddaden et al. [65]	97.57±1.33		Deng et al. [10]	81.83		Wang et al. [62]	86.3
	Ours	**97.65±1.00**		Ours	82.06±0.73		Jung et al. [34]	92.35
CFEE	ExpNet [7]	62.50		**Li et al.** [42]	**83.27**		**Ours**	**96.45±0.8**
	Ours	96.43±1.1						
	Du et al.[15]	**96.84±9.73**						

emotion classifier. Our obtained **average accuracy per class** is comparable to the state-of-the-art on this dataset by Du et al. [15], see table III.

The **RAF benchmark** [41] is the most challenging among all utilised FER datasets in this paper. This dataset was collected from the internet, no lab-controlled conditions. The authors of [41] sought the help of well-trained annotators for segregating the dataset into basic and compound emotion images. We use the basic emotion images, which are 13395 in total, and estimate their facial expressions using DeepExp3D. The train/test splits provided by the authors of [41] were used for training/testing our emotion classifier. Our produced **average accuracy per class** is comparable to the highest accuracy reported on this dataset (82.06 vs 83.27).

On the **Extended Cohn-Kanade (CK+)** [46] dataset, our method manages to generate the highest accuracy (96.45%) compared to other methods. This dataset has 327 sequences of frontal images originating from 210 subjects. Similar to [7], we keep the peak frame of each sequence and associate it with the label of this sequence.

Overall, quite similarly in all experimented benchmarks, the trained emotion classifier recognises the happy, fearful, surprised and disgusted emotions better than the rest (neutral, sad, angry). This can be mainly referred to two essential factors: 1) intensity of the related action units when deconstructing each emotion according to the Emotional Facial Action Coding System (EFACS) [20], 2) ability of employed expression basis (U_{exp}) in capturing the relevant action units. The trained DeepExp3D tends to capture well mouth- jaw- and cheeks-related motions (action units 6, 12, 14, 15, 16, 20, 26 [20]), e.g. lip corner puller/depressor, lower lip depressor, lip stretcher, jaw drop, etc., which are essential in characterising the happy, surprised, disgusted and fearful emotions. On the other hand, subtle details around the eyes, like inner brow raiser, brow lowerer, upper lid raiser, lid tightener, etc., which are judgmental for discerning emotions like sadness and anger, appear to be more challenging for the DeepExp3D. This can be explained by the fact that action units 6, 12, 14, 15, 16, 20, 26 are better represented in the original FaceWarehouse model utilised to annotate our collected dataset of videos presented in section III-B, as well as the 4DFAB dataset [8] used for the fine-tuning stage. Please see the supplementary materials for more results and visualisations. Additionally, the extracted landmarks might degrade the results largely if they fail to annotate the 68 targeted locations on the face with good accuracy.

B. Emotional Stress Analysis

In this section, we investigate the ability of our proposed framework in detecting stress conditions using only facial videos. Stress is widely conceived as a complex emotional state which can be identified by biosignals [23]. However, their recording may not always be convenient and practical for daily monitoring, thus research community pursuits stress identification only using facial cues, which constitutes a quite challenging task. Related literature is limited regarding the combination of biosignals with deep learning frameworks [32], [27], or only visual cues [24], [26]. In this work, we evaluate the performance of our method against other state-of-the-art methods in stress identification. Towards that end, we utilize the dataset (SRD'15) used in [24] which has 24 subjects (with age 47.3±9.3 years) and 288 videos in total. Each subject performed 11 experimental tasks (either neutral or stressful). The whole experiment was divided into 4 phases: 1) social exposure, 2) Emotional recall, 3) stressful images/mental task, 4) stressful videos.

The frames of each recorded video were labeled as either 'stressful' or 'non-stressful', according to the task under investigation. Next, our method was used to perform facial expression recognition from each frame and a 5-fold cross validation was carried out, while making sure frames coming from the same subject do not exist in both training and testing folds at the same time. The experiments were repeated 10 times and the average stress recognition accuracy of each phase is reported in table IV. Note that the first phase (social exposure) was not taken into account as it contains a task with speech which affects 'per se' head motility compared to a neutral non speech task as explained in [25]. For comparisons, we have also followed the same protocol and applied the method of [24] which uses head motility and the method of [27] which uses heart activity signals (IBI), with their results also shown in table IV. We observe that the proposed method achieves high accuracy and outperforms the other tested methods. This is an especially promising result for stress analysis, as our method uses only non-invasive and frame-based visual features.

TABLE IV

STRESS DETECTION ACCURACY COMPARISON ON DATASET USED IN [24]

Phase	**Head motility** [24]	**DWNet1D** [27]	**Ours**
Emotional Recall	82.99 %	83.50 %	**86.70 %**
Stressful images	85.42 %	**92.60 %**	88.42 %
Stressful videos	85.83 %	85.90 %	**88.83 %**
Average	84.75 %	87.33 %	**87.98 %**

Fig. 4. Estimated expressions (second row) from selected test images (first row) of our dataset with the Ground Truth (GT) expressions (bottom row). Both estimated and GT expressions are visualized with the 3D GT identity.

C. Evaluation of our Framework's Intermediate Steps

Even though the final output of our proposed pipeline is the emotion class, we have also conducted detailed experiments to evaluate the intermediate steps of our framework. First of all, we qualitatively and quantitatively evaluate the accuracy of estimating the **3D expression parameters**, which is the intermediate pipeline step taken as output of DeepExp3D. We test its performance on the test split of our **FaceVid** dataset both qualitatively and quantitatively. First of all, Figure 4 presents qualitative results of the proposed method as compared to the Ground Truth (GT) reconstructions. We show both the estimated and GT expressions on top of the mean face and the GT identity parameters. We observe that the estimations provided by our method are visually very close to the GT. Furthermore, we compare our DeepExp3D with: 1) a baseline approach following a linear shape model fitting proposed in [31], and 2) ITW, a state-of-the-art 3D reconstruction method from in-the-wild images [5]. We provide both methods with the same facial expression model (FaceWarehouse [6]) and compute the average of the Mean Squared Error (MSE) between the estimations and the ground truth over the test split. Our method achieves by far the lowest (best) MSE with **0.007**, while ITW [5] and the baseline [31] obtain 0.026 and 0.098, respectively. More results and visualisations are available in the supplementary material.

Furthermore, we evaluate our approach on **3D face reconstruction from videos** (Sec. III-A), which is used to construct the pseudo-ground truth annotations of our training dataset (**FaceVid**). This evaluation is presented in the Supplementary Material, from which it can be concluded that our proposed approach outperforms the compared state-of-the-art 3D reconstruction methods and achieves satisfactory accuracy for usage in pseudo-ground truth creation.

V. CONCLUSION

In this paper, we have proposed a framework for the automatic recognition of human emotions from monocular images. Our framework utilises a well-trained deep CNN (termed as **DeepExp3D**), of our own implementation, capable of estimating the 3D facial expression parameters from a single image, even in challenging scenarios. We have extensively evaluated the performance of our trained **DeepExp3D** and compared it with state-of-the-art methods for 3D reconstruction from in-the-wild images. Our **DeepExp3D** demonstrates a superior performance in regressing the facial expression coefficients when trained on the same facial expression model as the competitors. We have also extensively tested the potential of the trained DeepExp3D in recognising facial expressions when combined with an mSVM classifier on 5 widely-used benchmarks (taken under either controlled or in-the-wild conditions). Our reported emotion recognition results reveal the competitive performance of our proposed framework when compared with recent state-of-the-art approaches on the same datasets. We report the highest accuracy on the KDEF [47] (92.24%), RadFD [37] (97.63%), CK+ [46] (96.45%) datasets, and the second best on CFEE [15] and RAF-DB [41] (with 0.41% and 1.21% difference from the best, respectively).

REFERENCES

[1] A. M. Ali, H. Zhuang, and A. K. Ibrahim. An approach for facial expression classification. *IJBM*, 9(2):96–112, 2017.

[2] P. Barros, G. I. Parisi, C. Weber, and S. Wermter. Emotion-modulated attention improves expression recognition: A deep learning model. *Neurocomputing*, 253:104–114, 2017.

[3] V. Blanz and T. Vetter. A morphable model for the synthesis of 3d faces. In *Proceedings of the 26th annual conference on Computer graphics and interactive techniques*, pages 187–194. ACM Press/Addison-Wesley Publishing Co., 1999.

[4] J. Booth, A. Roussos, A. Ponniah, D. Dunaway, and S. Zafeiriou. Large scale 3d morphable models. *IJCV*, 2018.

[5] J. Booth, A. Roussos, E. Ververas, E. Antonakos, S. Ploumpis, Y. Panagakis, and S. Zafeiriou. 3d reconstruction of "in-the-wild" faces in images and videos. *T-PAMI*, 2018.

[6] C. Cao, Y. Weng, S. Zhou, Y. Tong, and K. Zhou. Facewarehouse: A 3d facial expression database for visual computing. *IEEE Transactions on Visualization and Computer Graphics*, 20(3):413–425, 2014.

[7] F.-J. Chang, A. T. Tran, T. Hassner, I. Masi, R. Nevatia, and G. Medioni. Expnet: Landmark-free, deep, 3d facial expressions. In *FG*, 2018.

[8] S. Cheng, I. Kotsia, M. Pantic, and S. Zafeiriou. 4dfab: A large scale 4D database for facial expression analysis and biometric applications. In *CVPR*.

[9] S. Cheng, I. Marras, S. Zafeiriou, and M. Pantic. Statistical non-rigid icp algorithm and its application to 3d face alignment. *IVC*, 2017.

[10] J. Deng, G. Pang, Z. Zhang, Z. Pang, H. Yang, and G. Yang. cgan based facial expression recognition for human-robot interaction. *IEEE Access*, 7:9848–9859, 2019.

[11] J. Deng, A. Roussos, G. Chrysos, E. Ververas, I. Kotsia, J. Shen, and S. Zafeiriou. The menpo benchmark for multi-pose 2d and 3d facial landmark localisation and tracking. *IJCV*, 2018.

[12] J. Deng, Y. Zhou, S. Cheng, and S. Zaferiou. Cascade multi-view hourglass model for robust 3d face alignment. In *FG*, 2018.

[13] A. Dhall, A. Kaur, R. Goecke, and T. Gedeon. Emotiw 2018: Audio-video, student engagement and group-level affect prediction. In *ICMI*, 2018.

[14] T. G. Dietterich and G. Bakiri. Solving multiclass learning problems via error-correcting output codes. *JAIR*, 1994.

[15] S. Du, Y. Tao, and A. M. Martinez. Compound facial expressions of emotion. *Proceedings of the National Academy of Sciences*, 111(15):E1454–E1462, 2014.

[16] P. Ekman. Facial expression and emotion. *American psychologist*, 48(4):384, 1993.

[17] M. A. et al. TensorFlow: Large-scale machine learning on heterogeneous systems, 2015. Software available from tensorflow.org.

[18] Y. Fan, J. C. Lam, and V. O. Li. Multi-region ensemble convolutional neural network for facial expression recognition. In *ICANN*, 2018.

[19] B. Fasel and J. Luettin. Automatic facial expression analysis: a survey. *Pattern recognition*, 36(1):259–275, 2003.

[20] W. V. Friesen, P. Ekman, et al. Emfacs-7: Emotional facial action coding system. *Unpublished manuscript, University of California at San Francisco*, 2(36):1, 1983.

[21] B. Gecer, S. Ploumpis, I. Kotsia, and S. Zafeiriou. Ganfit: Generative adversarial network fitting for high fidelity 3d face reconstruction. *arXiv preprint arXiv:1902.05978*, 2019.

[22] S. Ghosh, A. Dhall, and N. Sebe. Automatic group affect analysis in images via visual attribute and feature networks. In *ICIP*, 2018.

[23] G. Giannakakis, D. Grigoriadis, K. Giannakaki, O. Simantiraki, A. Roniotis, and M. Tsiknakis. Review on psychological stress detection using biosignals. *IEEE Transactions on Affective Computing*, 2019.

[24] G. Giannakakis, D. Manousos, V. Chaniotakis, and M. Tsiknakis. Evaluation of head pose features for stress detection and classification. In *2018 IEEE EMBS International Conference on Biomedical & Health Informatics (BHI)*, pages 406–409. IEEE, 2018.

[25] G. Giannakakis, D. Manousos, P. Simos, and M. Tsiknakis. Head movements in context of speech during stress induction. In *FG*, 2018.

[26] G. Giannakakis, M. Pediaditis, D. Manousos, E. Kazantzaki, F. Chiarugi, P. G. Simos, K. Marias, and M. Tsiknakis. Stress and anxiety detection using facial cues from videos. *Biomedical Signal Processing and Control*, 31:89–101, 2017.

[27] G. Giannakakis, E. Trivizakis, M. Tsiknakis, and K. Marias. A novel multi-kernel 1d convolutional neural network for stress recognition from ecg.

[28] I. J. Goodfellow, D. Erhan, P. L. Carrier, A. Courville, M. Mirza, B. Hamner, W. Cukierski, Y. Tang, D. Thaler, D.-H. Lee, et al. Challenges in representation learning: A report on three machine learning contests. In *International Conference on Neural Information Processing*, pages 117–124. Springer, 2013.

[29] R. Gross, I. Matthews, J. Cohn, T. Kanade, and S. Baker. Multi-pie. *Image and Vision Computing*, 28(5):807–813, 2010.

[30] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[31] P. Huber, P. Kopp, W. Christmas, M. Rätsch, and J. Kittler. Real-time 3d face fitting and texture fusion on in-the-wild videos. *IEEE Signal Processing Letters*, 24(4):437–441, 2017.

[32] B. Hwang, J. You, T. Vaessen, I. Myin-Germeys, C. Park, and B.-T. Zhang. Deep ecgnet: An optimal deep learning framework for monitoring mental stress using ultra short-term ecg signals. *TELEMEDICINE and e-HEALTH*, 24(10):753–772, 2018.

[33] B. Jiang and K. Jia. Robust facial expression recognition algorithm based on local metric learning. *Journal of Electronic Imaging*, 25(1):013022, 2016.

[34] H. Jung, S. Lee, J. Yim, S. Park, and J. Kim. Joint fine-tuning in deep neural networks for facial expression recognition. In *ICCV*, pages 2983–2991, 2015.

[35] M. R. Koujan, A. Akram, P. McCool, J. Westerfeld, D. Wilson, K. Dhaliwal, S. McLaughlin, and A. Perperidis. Multi-class classification of pulmonary endomicroscopic images. In *ISBI*, 2018.

[36] M. R. Koujan and A. Roussos. Combining dense nonrigid structure from motion and 3d morphable models for monocular 4d face reconstruction. In *CVMP*, 2018.

[37] O. Langner, R. Dotsch, G. Bijlstra, D. H. Wigboldus, S. T. Hawk, and A. Van Knippenberg. Presentation and validation of the radboud faces database. *Cognition and emotion*, 24(8):1377–1388, 2010.

[38] O. Leonovych, M. R. Koujan, A. Akram, J. Westerfeld, D. Wilson, K. Dhaliwal, S. McLaughlin, and A. Perperidis. Texture descriptors for classifying sparse, irregularly sampled optical endomicroscopy images. In *Annual Conference on Medical Image Understanding and Analysis*, pages 165–176. Springer, 2018.

[39] S. Li and W. Deng. Deep facial expression recognition: A survey. *arXiv preprint arXiv:1804.08348*, 2018.

[40] S. Li and W. Deng. Reliable crowdsourcing and deep locality-preserving learning for unconstrained facial expression recognition. *IEEE Transactions on Image Processing*, 28(1):356–370, 2019.

[41] S. Li, W. Deng, and J. Du. Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild. In *CVPR*, 2017.

[42] Y. Li, J. Zeng, S. Shan, and X. Chen. Patch-gated cnn for occlusion-aware facial expression recognition. In *ICPR*, 2018.

[43] F. Lin, R. Hong, W. Zhou, and H. Li. Facial expression recognition with data augmentation and compact feature learning. In *ICIP*, 2018.

[44] M. Liu, S. Li, S. Shan, R. Wang, and X. Chen. Deeply learning deformable facial action parts model for dynamic expression analysis. In *ACCV*, pages 143–157. Springer, 2014.

[45] M. Liu, R. Wang, S. Li, S. Shan, Z. Huang, and X. Chen. Combining multiple kernel methods on riemannian manifold for emotion recognition in the wild. In *ICMI*, 2014.

[46] P. Lucey, J. F. Cohn, T. Kanade, J. Saragih, Z. Ambadar, and I. Matthews. 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*, pages 94–101. IEEE, 2010.

[47] D. Lundqvist, A. Flykt, and A. Öhman. The karolinska directed emotional faces (kdef). *CD ROM from Department of Clinical Neuroscience, Psychology section, Karolinska Institutet*, 91:630, 1998.

[48] V. Mavani, S. Raman, and K. P. Miyapuram. Facial expression recognition using visual saliency and deep learning. In *ICCV*, pages 2783–2788, 2017.

[49] A. Mollahosseini, D. Chan, and M. H. Mahoor. Going deeper in facial expression recognition using deep neural networks. In *WACV*, 2016.

[50] N. Nilsson and L. Machines. Foundations of trainable pattern classifying systems. *McGraw-Hill, New York OBrien RM (2007) A caution regarding rules of thumb for variance ination factors. Qual Quant*, 41:673, 1965.

[51] N. Otberdout, A. Kacem, M. Daoudi, L. Ballihi, and S. Berretti. Deep covariance descriptors for facial expression recognition. *BMVC*, 2018.

[52] P. Paysan, R. Knothe, B. Amberg, S. Romdhani, and T. Vetter. A 3d face model for pose and illumination invariant face recognition. In *2009 Sixth IEEE International Conference on Advanced Video and Signal Based Surveillance*, pages 296–301. Ieee, 2009.

[53] X. Peng, X. Yu, K. Sohn, D. N. Metaxas, and M. Chandraker. Reconstruction-based disentanglement for pose-invariant face recognition. In *ICCV*, pages 1623–1632, 2017.

[54] E. Richardson, M. Sela, and R. Kimmel. 3d face reconstruction by learning from synthetic data. In *3DV*, 2016.

[55] A. Ruiz-Garcia, M. Elshaw, A. Altahhan, and V. Palade. Stacked deep convolutional auto-encoders for emotion recognition from facial expressions. In *IJCNN*, 2017.

[56] C. Sagonas, E. Antonakos, G. Tzimiropoulos, S. Zafeiriou, and M. Pantic. 300 faces in-the-wild challenge: Database and results. *Image and vision computing*, 47:3–18, 2016.

[57] F. Shen, J. Liu, and P. Wu. Double complete d-lbp with extreme learning machine auto-encoder and cascade forest for facial expression analysis. In *ICIP*, 2018.

[58] J. Snoek, H. Larochelle, and R. P. Adams. Practical bayesian optimization of machine learning algorithms. In *Advances in neural information processing systems*, pages 2951–2959, 2012.

[59] W. Sun, H. Zhao, and Z. Jin. An efficient unconstrained facial expression recognition algorithm based on stack binarized auto-encoders and binarized neural networks. *Neurocomputing*, 267:385–395, 2017.

[60] Y. Tang. Deep learning using linear support vector machines. *arXiv preprint arXiv:1306.0239*, 2013.

[61] V. Vielzeuf, C. Kervadec, S. Pateux, A. Lechervy, and F. Jurie. An occam's razor view on learning audiovisual emotion recognition with small training sets. In *Proceedings of the 2018 on ICMI*, 2018.

[62] Z. Wang, S. Wang, and Q. Ji. Capturing complex spatio-temporal relations among facial muscles for facial expression recognition. In *CVPR*, 2013.

[63] B.-F. Wu and C.-H. Lin. Adaptive feature mapping for customizing deep learning based facial expression recognition model. *IEEE access*, 6:12451–12461, 2018.

[64] S. Xiao, P. Ting, and R. Fu-Ji. Facial expression recognition using roi-knn deep convolutional neural networks. *Acta Automatica Sinica*, 42(6):883–891, 2016.

[65] Y. Yaddaden, M. Adda, A. Bouzouane, S. Gaboury, and B. Bouchard. User action and facial expression recognition for error detection system in an ambient assisted environment. *Expert Systems with Applications*, 112:173–189, 2018.

[66] S. Zafeiriou, G. G. Chrysos, A. Roussos, E. Ververas, J. Deng, and G. Trigeorgis. The 3d menpo facial landmark tracking challenge. In *ICCV*, pages 2503–2511, 2017.

[67] M. V. Zavarez, R. F. Berriel, and T. Oliveira-Santos. Cross-database facial expression recognition based on fine-tuned deep convolutional network. In *SIBGRAPI*, 2017.

[68] X. Zhu, Z. Lei, X. Liu, H. Shi, and S. Z. Li. Face alignment across large poses: A 3d solution. In *CVPR*, 2016.