Exploiting multi-CNN features in CNN-RNN based Dimensional Emotion Recognition on the OMG in-the-wild Dataset

Dimitrios Kollias and Stefanos Zafeiriou

Abstract—This paper presents a novel CNN-RNN based approach, which exploits multiple CNN features for dimensional emotion recognition in-the-wild, utilizing the One-Minute Gradual-Emotion (OMG-Emotion) dataset. Our approach includes first pre-training with the relevant and large in size, Aff-Wild and Aff-Wild2 emotion databases. Low-, mid- and high-level features are extracted from the trained CNN component and are exploited by RNN subnets in a multi-task framework. Their outputs constitute an intermediate level prediction; final estimates are obtained as the mean or median values of these predictions. Fusion of the networks is also examined for boosting the obtained performance, at Decision-, or at Model-level; in the latter case a RNN was used for the fusion. Our approach, although using only the visual modality, outperformed state-of-the-art methods that utilized audio and visual modalities. Some of our developments have been submitted to the OMG-Emotion Challenge, ranking second among the technologies which used only visual information for valence estimation; ranking third overall. Through extensive experimentation, we further show that arousal estimation is greatly improved when low-level features are combined with high-level ones.

Index Terms—Deep convolutional and recurrent neural architectures; CNN plus Multi RNN; low-, mid-, high-level features; multi-CNN feature extraction and aggregation; multi-task learning; facial image analysis; valence; arousal; emotion recognition in-the-wild; AffWildNet; AffWild and AffWild2 emotion databases; OMG-Emotion database and Challenge.

1 Introduction

Automatic analysis of facial behaviour is the cornerstone of many application areas, including Human-Computer and Robot Interaction, Pervasive Computing, Ambient Intelligence and Virtual Reality. The research area of facial behaviour analysis includes the problems of: i) the recognition of the so-called six universal expressions (i.e., Anger, Disgust, Fear, Happy, Sad, Surprise), plus Neutral, influenced by the seminal work of Ekman [12], ii) the recognition of spontaneous expressions including mental states (pain intensity [17] and compound expressions [10]), iii) the detection of the facial Action Units (AU) and estimation of their intensity, according to the Facial Action Coding System [11] which provides a standardised taxonomy of facial muscles' movements, iv) the detection of micro-expressions, and v) the estimation of facial affect in a continuous dimensional space (e.g., valence and arousal). Related research can assist in flagging complex behavioral patterns such as deception, depression, autism, spectrum disorders and schizophrenia [1], [20], [27], [44], [56], [57].

The main focus of this paper is on dimensional emotion models, which are appropriate to represent not only extreme, but also subtle emotions appearing in everyday human-computer interactions. According to the dimensional approach [55] [60], affective behavior is described by a number of latent continuous dimensions. The most commonly used dimensions include valence (indicating how positive or negative an emotional state is) and arousal (measuring the power of emotion activation). Valence and arousal relate readily to specific functions of regions of the brain [16], [42], [58]; the parietal region of the right hemisphere appears to play a special role in the mediation of arousal, whereas the frontal regions appear to play a special role in emotional valence. A third dimension, tension, is also introduced but often excluded due to difficulties in consistently identifying what the dimension describes: tension, control, or potency (dominance). Fig. 1 shows the 2-D Valence-Arousal Space, introduced in [51]. Estimation of valence and arousal continuous values related to affect constitutes the problem examined in the following.

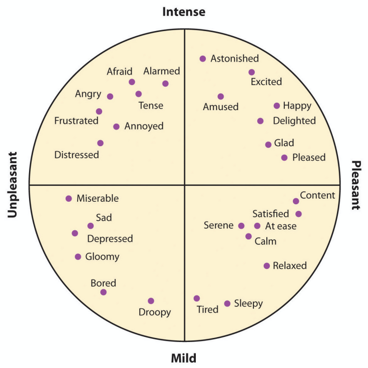

Fig. 1: The 2-D Valence-Arousal Space, as seen in [36]

In order to facilitate research on the above problems, many databases have been generated and annotated, most

- D. Kollias and S.Zafeiriou are with the Department of Computing, Imperial College London, United Kingdom
D. Kollias is also with the School of Computing and Mathematical Sciences, University of Greenwich
E-mail: dimitrios.kollias15@imperial.ac.uk
E-mail: s.zafeiriou@imperial.ac.uk

of which are in well-controlled conditions. In the beginning, data and annotations were scarce, hence research relied on extracting highly engineered handcrafted features and designing ad-hoc learning strategies [40], [41], [48], [52]. Naturally, as the amount of data and annotations grew, research has started to capitalise on data-intensive technologies, such as deep learning [5], [6], [19], [21]–[23], [26], [37].

It is now widely accepted, in both the computer vision and machine learning communities, that progress in a particular application domain is significantly catalysed when a large number of datasets are collected in unconstrained conditions (also referred as ”in-the-wild” data). Hence, facial analysis could not only focus on spontaneous behaviors, but also on behaviors captured in unconstrained conditions. In-the-wild dimensional databases have been generated, such as the audiovisual OMG-Emotion Dataset [3], Aff-Wild [61] [28], Aff-Wild2 [33] [29] [31] [25] and SEWA [53] ones, as well as AffectNet [43] which includes only static images.

Regarding the pipeline of facial behavior analysis, the standard paradigm has been to: i) detect and/or track the face in an image sequence, ii) detect and/or track facial landmarks, iii) extract handcrafted features[1], either around the landmarks, or on the face region as a whole, and iv) use the features and the landmarks for classification/regression using affective labels. Recently this paradigm has shifted from utilizing handcrafted features to utilizing features learned by deep Convolutional Neural Networks (CNNs) and/or Recurrent Neural Networks (RNNs). This shift was motivated by the striking performance achieved when utilizing deep neural networks (DNNs) in a variety of emotion recognition tasks [9], [14], [32], [45], [62].

In this paper, we address the issue of estimating valence and arousal utilizing the One-Minute-Gradual Emotion Dataset (OMG-Emotion Dataset), based on visual information only. We present novel deep neural architectures that provide best performance in valence and arousal estimation, as well as the submissions we made to the OMG-Emotion Challenge, which were ranked very high, especially for valence estimation.

The first main contribution of this paper is the development of CNN plus multi-RNN architectures for valence and arousal estimation in a multi-task optimization formulation. In this formulation, low-, mid- and high- level features are extracted from different layers of the CNN part and passed as input to the RNN part. The intuition for this is that these features include rich information which can be advantageous for the studied task.

These architectures are of two different types; in the first, the features extracted from, say, K CNN layers are concatenated and passed as input to a single RNN, whereas in the other, they are passed to K RNNs. In the experimental section, it is shown that the latter type outperformed all other developed architectures and even state-of-the-art networks that used not only the visual, but also the audio modality. Our work deviates from others, such as [6], [8], [37], that either: i) use standard CNN-RNN networks in which the output of the CNN is passed to the RNN, or ii) apply ensemble methodologies, using features extracted from many CNN networks (but not using features from multiple layers of the same network) and fusing them.

Both facial images and landmarks (after applying a Procrustes Analysis) are provided as inputs to these architectures. Additionally, ensemble formulations are proposed, using different levels of fusion (Model- or Decision-level) on the proposed architectures; these formulations are shown to further boost the obtained performance. In model-level fusion, our proposal is to perform fusion through a RNN instead of the typical fully connected layer.

Another contribution of this work is the approach to fit the developed architectures to the OMG-Emotion dataset characteristics and in particular to the dataset's annotation at utterance level. To deal with this, we split each utterance into sequences, which were individually processed by the above architectures. The mean or median of the predicted valence-arousal values were computed per sequence. Then, the means/medians were averaged at utterance level to provide the final valence and arousal estimates. This procedure deviates from related works that uniformly (or randomly) sample a constant number of frames from each utterance, assign to each of them the annotation value of the utterance and compute the prediction per frame [8].

An additional contribution of this work is the pre-training of the proposed architectures on the large-scale emotionally rich Aff-Wild database and on its larger extension, the Aff-Wild2. Other works [50], [59], [63] used networks that were not pre-trained on same task (valence-arousal estimation) but on other tasks (face recognition, object detection). The pre-training on these specific databases provided our developed architectures with the ability to effectively capture the dynamics of the OMG-Emotion in-the-wild dataset and thus provided a better performance.

The main findings of our approach have been: i) low-level features when combined with high-level ones in our CNN plus multi-RNN architectures, helped in boosting the networks' performance in arousal estimation; ii) CNN plus multi-RNN architectures outperformed standard CNN plus RNN ones showing that features extracted from previous layers contain useful and rich information for valence-arousal prediction; iii) better results were obtained when the features extracted from previous layers were processed by independent RNNs instead of being concatenated and fed to a single RNN; iv) better results were obtained when using a RNN instead of a fully connected layer for model-level fusion; v) when using the visual modality, network performance for valence estimation is much higher than the corresponding for arousal estimation.

The rest of this paper is organized as follows. Section 2 reviews related work and existing state-of-the-art methods for facial expression recognition with emphasis on the dimensional model of affect. Section 3 gives a brief description of the databases used in our experiments, i.e., OMG-Emotion Dataset, Aff-Wild and Aff-Wild2 databases. Section 4 presents the pre-processing steps which were essential to obtain a common input representation for analysis. Section 5 presents the developed methods, i.e., the created novel deep neural architectures, including ensembles and fusion of networks, for valence-arousal estimation. Section 6 describes specific implementation details that we followed to

[1] Examples of handcrafted features include Histogram of Oriented Gradients (HoGs), Scale Invariant Feature Transform (SIFT), Local Binary Patterns (LBPs) and features from multiscale and multiorientation Gabor filterbanks

achieve the best results. Section 7 provides an evaluation of our approach by analysing the obtained results and presenting comparisons with other methods, in terms of achieved performance. Finally, Section 8 presents the conclusions.

2 Related Work

One of the first deep learning architectures for valence and arousal estimation was proposed in [19]. In this work, both frame-based CNN and CNN plus RNN architectures were proposed and compared. The CNN consisted of 3 convolutional layers; the first two layers were followed by max pooling layers and the third by a quadrant pooling layer. A fully connected layer was then used, followed by the output layer. The CNN plus RNN architectures consisted of the previously described CNN network (keeping its weights fixed) without the top regression layer, followed by a single RNN layer that gave the final estimates. This methodology achieved very high valence and arousal correlations in a part of the RECOLA database [54].

The authors in [6] explored and fused different handcrafted and deep learning features from all available modalities (acoustic, visual, and textual). They also considered the interlocutor influence (a person's influence on the interacting partner's behaviors) for the acoustic features.

In more detail, the authors extracted: i) from the acoustic modality, hand-crafted features, such as MFCCs, loundness, F0, jitter, shimmer and features learned from the SoundNet [2], ii) from the visual modality, features learned from VGG-FACE [49] and DenseNet that had been pre-trained on the FER+ [4] dataset (annotated in terms of the basic expressions), and iii) from the textual modality, word vectors that were used as features. All those features were fused and passed as input to a LSTM network that produced the estimates for valence, arousal and likability. This approach was the winning of AVEC 2017 Challenge that utilized the SEWA database.

The authors of [5] presented the FATAUVA-Net method, which is a deep learning framework in which a core layer, an attribute layer, an AU layer and a valence-arousal layer were trained sequentially. The core layer was a series of convolutional layers, followed by the attribute layer which extracted facial area's features (face, eye, eyebrow, mouth). These layers were used in supervised learning of AUs. Finally, AUs were employed as mid-level representations to estimate the intensity of valence and arousal. This methodology produced the highest results of the First Affect-in-the-wild Challenge [28] which was the first challenge on the estimation of valence and arousal in-the-wild, using the Aff-Wild database for recognition of affect.

Best results in the Aff-Wild database have been obtained by the authors of [24] [28]. In these works, the authors performed a large number of experiments, training CNN and CNN-RNN networks on the Aff-Wild for emotion recognition. The best performing network, AffWildNet, consists of the convolutional and pooling parts of the ResNet-50 network [15], followed by a fully connected layer, a 2-layer GRU [7] and the output layer that provided the final valence-arousal estimates. This network was further fine-tuned on the RECOLA and AFEW-VA databases, producing state-of-the-art performance.

Table 1 provides a summary of the performance of the above-described methods on the respective databases.

TABLE 1: State-of-the-art algorithms for valence-arousal estimation, their performances and utilized databases

Work	Databases Used	Methods	Results

[19]	part of
	RECOLA
	as used in the
	AVEC Challenge
		CNN-RNN visual only:
		(conv + max-pool) x2
		+ conv + quadrant-pool
		+ RNN
			Valence:
			RMSE = 0.107
			PCC = 0.554
			CCC = 0.507

[6]	SEWA	(1) audio:
		handcrafted + SoundNet features
		(2) visual:
		VGG-FACE + DenseNet features
		(3) text:
		word vectors - features
		fusion of (1), (2), (3) + LSTM
			Valence - Arousal:
			RMSE = 0.081 - 0.086
			PCC = 0.758 - 0.702
			CCC = 0.756 - 0.672

[5]	Aff-Wild	1) core layer: series of conv. layers
		2) attribute layer: facial features
		3) AU layer
		4) Valence & Arousal layer
			Valence - Arousal
			MSE = 0.123 - 0.095
			CCC = 0.396 - 0.282

[24]
[28]
	Aff-Wild;
	whole RECOLA;
	AFEW-VA;
		AffWildNet:
		ResNet-50 + FC + GRU
			CCC:
			Valence - Arousal
			Aff-Wild : 0.570 - 0.430
			RECOLA : 0.526 - 0.273
			AFEW-VA: 0.515 - 0.556

3 The utilized in-the-wild Dimensional Emotion Databases

In this Section we provide a short description of the Aff-Wild and its extension Aff-Wild2 database, which have been used to pre-train the developed deep neural network architectures, as well as the OMG-Emotion database, analysis of which is the target of this paper.

3.1 Aff-Wild Database

The Aff-Wild database [28] [61] has been the first large scale captured in-the-wild database that has been annotated by 8 lay experts with regards to valence and arousal. It consists of 298 videos and displays reactions of 200 subjects, with a total video duration of more than 30 hours. The total number of frames in this database is 1,224,100. Regarding subjects' gender, 130 are male and 70 female. The Aff-Wild database served as benchmark for the Aff-Wild Challenge, organized in conjunction with CVPR 2017. The aim for this database was to collect spontaneous facial behaviors in arbitrary recording conditions. To this end, the videos were collected using Youtube. The main keyword that was used to retrieve the videos was reaction.

3.2 Aff-Wild2 Database

The Aff-Wild database has recently been augmented with new Youtube videos having a total length of 13 hours and 5 minutes, thus forming the Aff-Wild2 database [29], [31], [33]. This database has been the basis for the ABAW Competition [25]. The aim has been to extend the spontaneous facial behaviors in arbitrary recording conditions met in Aff-Wild, whilst significantly increasing the number of different subjects in it. All the additional videos have been annotated by four experts and contain a wide range in subjects': age (from babies and young children, to elderly people); ethnicity (subjects are caucasian, hispanic or latino, asian, black, or african american); profession (e.g. actors, athletes, politicians, journalists); head pose; illumination conditions; occlusions; emotions. In total, Aff-Wild2 consists of 558 videos with 2,786,201 frames. 11 out of those videos display

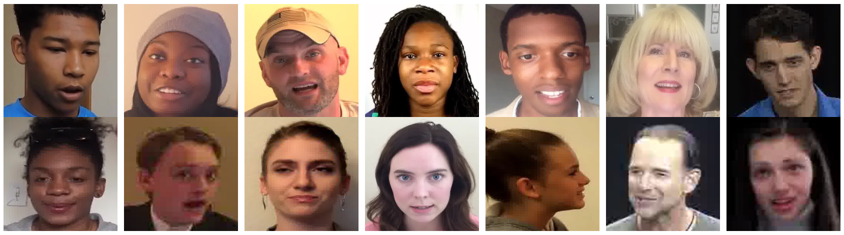

Fig. 2: Sample images from the OMG-Emotion dataset showing people displaying various in-the-wild emotions

two subjects, all of which have been annotated. The total number of subjects is 458, with 279 of them being male and 179 female.

3.3 OMG-Emotion Dataset

The One-Minute Gradual-Emotional Behavior dataset (OMG-Emotion dataset) [3] contains in-the-wild videos from Youtube where emotion expressions emerge and develop over time based on monologued scenarios. Figure 2 shows some frames from this dataset, with various people displaying different emotions under many occasions/circumstances. This dataset is annotated in terms of valence and arousal and also contains a large number of different identities.

The OMG-Emotion dataset served as a benchmark for the One-Minute Gradual-Emotion Recognition (OMG-Emotion) Challenge [2], held jointly with the Special Session on Neural Models for Behavior Recognition at the WCCI/IJCNN 2018. In particular, the dataset is split into training, validation and test sets in a subject independent manner, meaning that each subject appears strictly in only one of these sets. The training set consists of 231 videos composed of 2442 utterances, the validation set consists of 60 videos composed of 617 utterances and the test set consists of 204 videos composed of 2229 utterances. Each utterance has an average length of 8 seconds and each video has an average length of around 1 minute.

For annotating the collected data, the Amazon Mechanical Turk tool was used, resulting, on average, in five independent annotations per utterance. Each annotator was given the full contextual information of the video up to that point when annotating the dataset. That means that each annotator could take into consideration not only the visual and audio information but also the context of each video, i.e. what was spoken in the current and previous utterances through the context clips provided by the annotation tool. In this manner, each annotation is based on multimodal information.

Each utterance was given a specific valence and arousal value, based on the gold standard of the five annotations. Valence annotations range in $[-1, 1]$, whereas arousal ones range in $[0, 1]$. In Fig. 3, on top row are the 2-D histograms of the annotations in the OMG-Emotion training, validation and test sets, respectively, and in the bottom row are the corresponding datasets' annotations' distributions.

Additionally, this dataset contains categorical annotations for each utterance; transcripts of what was spoken in each of the videos are also provided.

4 Pre-processing: Face Detection & Alignment, Image Resizing & Normalization

Data pre-processing consists of all processing steps that are required for starting the extraction of meaningful features from the data. The usual steps are face detection, face alignment, image resizing and image normalization. The first step is to extract face bounding boxes from all video frames. In order to do so, we used the Deformable Part

[2] https://www2.informatik.uni-hamburg.de/wtm/omgchallenges/omg_emotion2018_session.html

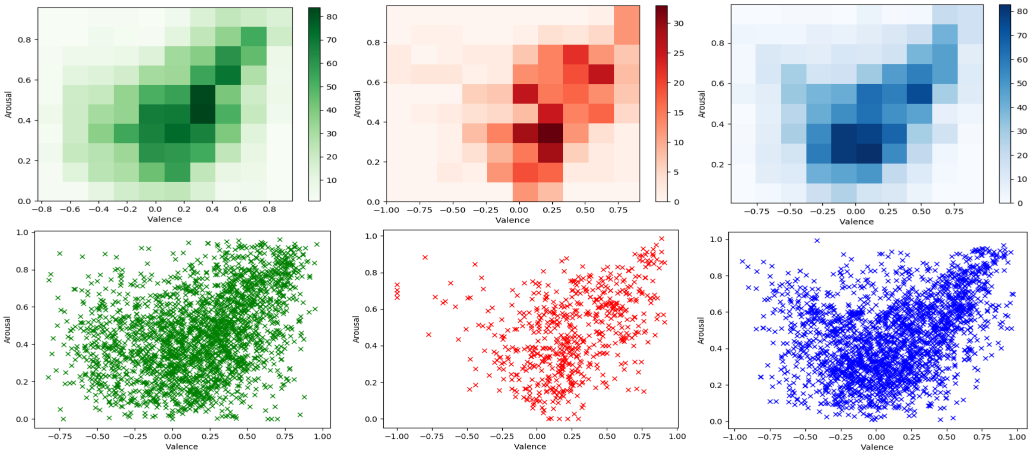

Fig. 3: Histograms (on the top row) and Distributions (on the bottom row) in the 2-D Valence-Arousal space of utterance-level annotations of the training (in green), validation (in red) and test (in blue) OMG-Emotion sets

Model (DPM) detector ffld2 [39] that has proven to be highly efficient and accurate for face detection in-the-wild.

For face alignment, we extracted facial landmarks and implemented the Generalized Procrustes Analysis [13]. In our implementations, we first used the facial landmark detector inside the dlib library [18] to locate 68 facial landmarks in all frames. We used as reference and rigid points, 5 anchor points that corresponded to the location of the left eye, right eye, nose and mouth in a prototypical frontal face. For every frame, we used its 5 facial landmarks corresponding to the location of the same facial components; we performed Procrustes transformation, which eliminates in-plane rotation, isotropic scaling and translation, on the coordinates of these 5 landmarks and the coordinates of the 5 landmarks of the frontal face; we imposed this transformation to the whole new frame to perform the alignment. All cropped and aligned images were then resized to $96 \times 96 \times 3$ pixel resolution and their intensity values were normalized to the range $[-1, 1]$. Those images, along with the 68 facial landmarks, were then used as inputs for training our networks, as described in the following Section.

5 The Developed Architectures

This section presents the proposed framework for dimensional emotion recognition, by describing the CNN, the standard CNN plus RNN, the proposed CNN plus Multi RNN architectures and then an ensemble methodology for fusion.

In all architectures presented in this Section, we compared a uni-task learning approach, independently for valence and arousal, to multi-task learning approach. The latter provided better performance in estimation of both affective dimensions. This result is in agreement with [47] which claims that there exist inter-correlations between the valence and arousal emotion dimensions. This relation between emotion dimensions in isolation, (i.e., without including features), has been well-supported by related research in psychology [35] [34]. That is the reason why in the following, we focus on the multi-task case.

5.1 CNN architectures

We experimented with three state-of-the-art networks: VGG-Face, ResNet-50 and DenseNet-121. These networks were first pre-trained either on the Aff-Wild or the Aff-Wild2 database and then trained on the OMG-Emotion training set. To design the structure of these networks, we took into account the procedure used to annotate the OMG-Emotion dataset. According to this, each utterance was labeled with a single pair of valence and arousal values. We split each utterance into smaller parts-sequences, each consisting of the same number of consecutive frames. Then, we assigned to each of those parts-sequences of frames, the label of the corresponding utterance.

Training of the CNN networks was performed as shown in Fig.4. In more detail, each CNN was provided with an input sequence and was trained to predict, for each frame in the sequence, the respective valence-arousal pair of values. The 68 facial landmarks (per each frame of the input sequence) were also provided as additional inputs to the CNN networks. The final valence (arousal) prediction was computed as the mean, or median (both approaches were considered) of the per-frame valence (arousal) values in that sequence.

In Fig.4, the CNN structure can be any of the VGG-FACE, ResNet-50 and DenseNet-121 ones. In the VGG-FACE CNN case, the landmarks were concatenated with the outputs of the last pooling layer of the network and were given as input to the first fully connected layer, that consisted of 4096 units. In this way, both outputs and landmarks were mapped to the same feature space, before performing the prediction. In the ResNet-50 (and DenseNet-121) case, the landmarks were concatenated with the averaged pooled features of the ResNet-50 (DenseNet-121) network and were given as input to a fully connected layer consisting of 1500 units. This layer was followed by the output layer which provided the final estimates for valence-arousal pair.

5.2 Standard CNN plus RNN architectures

In order to consider the contextual information in the data and more specifically the temporal dependencies of facial expressions in each utterance, we designed standard CNN plus RNN architectures. In the following we present the different CNN-RNN architectures that we have developed and used in the experimental study. In these architectures, the output of the CNN's last pooling layer is being fed to a fully connected layer, whose output constitutes the input of the RNN layers. These architectures were pre-trained on either the Aff-Wild, or the Aff-Wild2 databases. We then used two different strategies for training these architectures: i) keeping the CNN weights fixed and training the remaining architecture (i.e., the fully connected layers and the RNNs), or ii) training the whole architecture in an end-to-end manner (by jointly training the CNN and RNN parts). The latter approach provided the best results.

5.2.1 *AffWildNet*

At first, we considered the AffWildNet [28] as the best performing network on the Aff-Wild database and re-trained it on the OMG-Emotion database. As shown in Table 2, the AffWildNet is a CNN-RNN network consisting of the convolutional and pooling parts of ResNet-50 followed by a fully connected layer of 1500 units, followed by a 2-layer GRU, with each layer having 128 units. In this architecture, the landmarks are concatenated with the averaged pooled features of the ResNet-50 and being fed as input to the fully connected layer consisting of 1500 units. Similarly to the CNN case described in the previous Subsection, the CNN-RNN network receives an input sequence of frames, then predicts, for each frame, the valence-arousal values and finally computes the mean, or median, of these values, which is the final estimate. This architecture is the same as in Fig.4, if one replaces the CNN network with the CNN-RNN (i.e., the AffWildNet).

5.2.2 *DenseNet-RNN*

We also used a DenseNet-RNN structure that is quite similar to that of the AffWildNet described in the previous Subsection. The only difference is that it uses the DenseNet-121 network's convolutional and pooling layers.

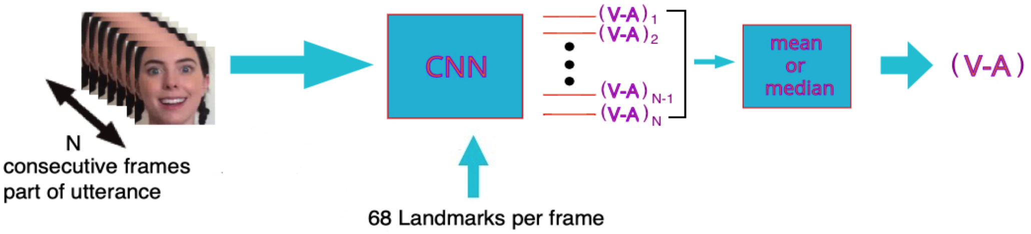

Fig. 4: The developed CNN structure. It gives only one valence-arousal (V-A) estimate per input sequence of consecutive frames. The CNN component can be any of the VGG-FACE, ResNet-50 and DenseNet-121 networks. The 68 landmarks are concatenated with the extracted features from the last pooling layer of the CNN component and are passed to the fully connected layer that precedes the output layer.

TABLE 2: The AffWildNet architecture

block 1	ResNet-50 conv & pooling parts	
block 2	fully connected 1 dropout	1500
block 3	GRU layer 1 dropout	128
block 4	GRU layer 2	128
block 5	fully connected 2	2

5.3 CNN plus Multi-RNN networks

In general, features extracted from the low CNN layers contain rich, complete and time varying information, whilst high-level features are highly specific and characteristic of the specific problem studied. Taking this into account, we have developed and used CNN plus Multi-RNN networks; these networks extract low-, mid- and high- level features from different layers of the CNN and pass them through RNNs. These networks are split into two different types through different methodologies: the first, referred as CNN-1RNN, concatenates the extracted features from 3 CNN layers and passes them to a single RNN, whereas the other, referred as CNN-3RNN, processes them independently through 3 RNN subnets.

It should be mentioned that we also tested other networks: CNN-2RNN (extracting features from 2 CNN layers and pass them independently to 2 RNNs); CNN-2RNN-1FC (similarly as before, with the outputs from the 2 RNNs being concatenated and passed to a fully connected layer; in this way they are both mapped to the same feature space, before performing the final prediction); CNN2-to-1RNN (extracting features from 2 CNN layers, concatenating them and passing them as input to a single RNN); CNN-3RNN-1FC (the outputs from the 3 RNNs being concatenated and passed to a fully connected layer, before performing the final prediction). These architectures provided performance that was around 4-5% lower than the performance of the CNN-1RNN and CNN-3RNN networks, presented next in this Section.

5.3.1 CNN-3RNN networks

The CNN-3RNN networks include the convolutional and pooling layers of VGG-FACE, followed by a fully connected layer of 4096 units. The 68 facial landmarks are concatenated with the features extracted from the last pooling layer of VGG-FACE and are passed to this fully connected layer. Then, low-, mid- and high-level features are extracted and each one is processed by a 2-layer GRU network that predicts the valence and arousal values. Each GRU layer comprises 128 units. Similarly to the architectures described in Subsection 5.3, the CNN-3RNN networks are provided with an input sequence of frames (and the corresponding landmarks of each frame), predicting, for each frame, the valence-arousal values; their mean, or median constitute the final estimates.

Fig.5, presents an example of CNN-3RNN networks, named CNN-3RNN-2nd-pool_last-pool_fc. In this network: i) the features extracted from the fully connected layer are passed as input to a RNN network, denoted RNN_1 in Fig.5; ii) the features extracted from the last pooling layer (before being concatenated with the landmarks) are passed as input to a second RNN network, denoted RNN_2 in Fig.5; iii) the features extracted from the second pooling layer (following the fourth convolutional layer) are passed as input to another RNN network, denoted RNN_3 in Fig.5. Fig.6 depicts the exact structure of the afore-mentioned RNN_i, $i \in \{1, 2, 3\}$, networks. All networks have the same structure; a 2-layer GRU network, with each layer having 128 units. Next, the outputs of the 3 RNNs are concatenated and passed to the output layer that performs the valence-arousal prediction. As shown in the experimental Section 7, this network based on the features extracted from these specific layers provided the best results in these type of networks.

5.3.2 CNN-1RNN networks

The CNN-1RNN types of networks consist of the convolutional and pooling layers of VGG-FACE, followed by a fully connected layer of 4096 units. The 68 facial landmarks are concatenated with the features extracted from the last pooling layer of VGG-FACE and are passed to this fully connected layer. Then, low-, mid- and high-level features are extracted, concatenated and passed to a 2-layer GRU network that predicts the valence and arousal values. Each GRU layer comprises 128 units. Similarly to the other architectures described above, the CNN-1RNN networks are provided with an input sequence of frames (and the corresponding landmarks of each frame), predicting, for each frame, the valence-arousal values; their mean, or median, are the final estimates.

Fig.7 presents one example of CNN-1RNN networks, which we call CNN-1RNN-2nd-pool_last-pool_fc. In this

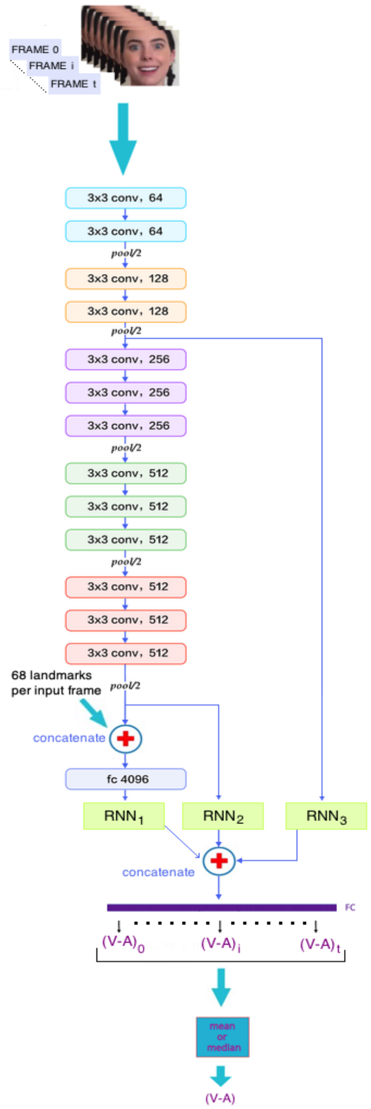

Fig. 5: The CNN-3RNN-2nd-pool_last-pool_fc. It provided a valence-arousal (V-A) estimate per input sequence of consecutive frames. The ’68 landmarks’ are concatenated with the features of the last ’pool’ layer and passed as input to the ’fc’ layer. This architecture provided the best results.

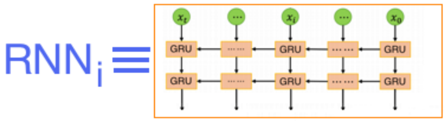

Fig. 6: The structure of each RNN network in the CNN-3RNN architecture displayed in Fig. 5.

network, the features extracted from: i) the second pooling layer (following the fourth convolutional), ii) the last pooling layer (following the 13th convolutional and before being concatenated with the landmarks) and iii) the fully connected layer, are concatenated and passed to the RNN. As shown in the experimental Section 7, this network based on the features extracted from these specific layers provided the best results in these type of networks.

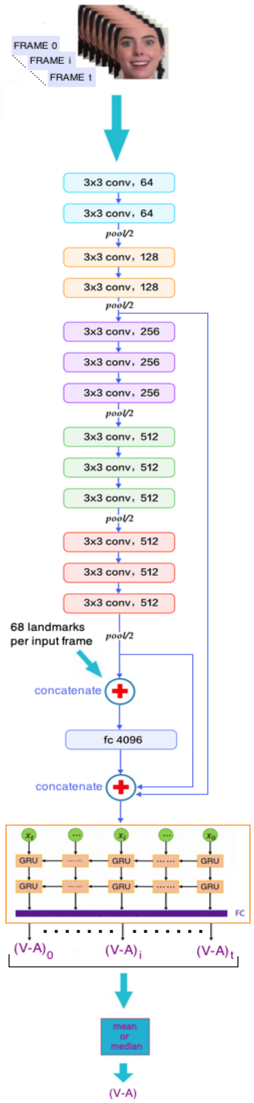

Fig. 7: The CNN-1RNN-2nd-pool_last-pool_fc architecture. It provides a valence-arousal (V-A) estimate per input sequence of consecutive frames. The ’68 landmarks’ are concatenated with the features of the last ’pool’ layer and passed as input to the ’fc’ layer.

5.4 Ensemble Methodology

In this Subsection we describe an ensemble approach which fuses the developed networks at: i) Model-level and ii) Decision-level. Model-level fusion is based on concatenating the high level features extracted by different networks, whilst Decision-level fusion is based on weighted averaging the predictions provided by different networks. On the one hand, Model-level fusion takes advantage of the mutual information in the data. On the other hand, the averaging procedure in Decision-level fusion reduces variance in the ensemble regressor (thus achieving higher robustness), while preserving the relative importance of each individual model.

5.4.1 Model-level Fusion

Let us consider the CNN-1RNNs and CNN-3RNNs described in the previous Subsection. We concatenate the outputs of all the RNNs in the above networks and provide them, as input, either: i) to another single RNN layer with 128 GRU units, or ii) to a fully connected layer with 128 units; the output layer follows. We denote the resulting networks as Model-level Fusion + RNN and Model-level Fusion + FC, respectively. Similarly to the previous Subsections, for each frame in the input sequence of frames, this model-level fusion network predicts the valence-arousal values and then computes their mean, or median, as final estimates.

5.4.2 Decision-level Fusion

Let us consider again the CNN-1RNNs and CNN-3RNNs described above. The final valence (arousal) estimate $O_v^{dec.-level}$ ($O_a^{dec.-level}$), is computed as a weighted average of the final valence (arousal) estimates, $o_v^n (o_a^n)$, of these networks; each weight is proportional to the corresponding network performance on the validation set:

$$O_i^{dec.-level} = \frac{1}{\sum_n t_i^n} \sum_n t_i^n \cdot o_i^n, \tag{1}$$

where $i \in \{v, a\}$ (v stands for valence, a stands for arousal), t_i^n is equal to the Concordance Correlation Coefficient (CCC), ρ_i, for valence or arousal, computed on the validation set, with n denoting the CNN-1RNNs or CNN-3RNNs; the CCC has been the evaluation criterion of the OMG-Emotion Challenge, taking values in $[-1, 1]$ and is defined as follows :

$$\rho_i = \frac{2 s_{i,xy}}{s_{i,x}^2 + s_{i,y}^2 + (\bar{x}_i - \bar{y}_i)^2}, \tag{2}$$

where $i \in \{v, a\}$, $s_{i,x}$ and $s_{i,y}$ are the variances of the valence/arousal labels and predicted values respectively, $\bar{x}_i$ and $\bar{y}_i$ are the corresponding mean values and $s_{i,xy}$ is the covariance value.

6 Network Training Details

In the following, we provide further information regarding the parameters used in the developed architectures (learning rate, dropout probability value, batch size, sequence length), the loss function that was formulated for our problem and the series of post-processing steps that were applied to the obtained estimates of valence and arousal.

6.1 Implementation Details

In all developed CNN, CNN plus RNN and CNN plus Multi-RNN architectures, dropout with 0.5 probability value was applied on the fully connected layers that were on top of the convolutional and pooling layers of CNN networks (VGG-FACE, ResNet-50 and DenseNet-121). Additionally, dropout with 0.8 probability value was applied after the first GRU layer of the RNNs.

For training our CNN networks, different sequence sizes were used, ranging from 40 to 100, with the size of 80 frames providing the best results. In the CNN plus RNN and CNN plus Multi-RNN cases, we used a batch size of 4 and sequence length of 80 consecutive frames. When training the CNN architectures, the learning rate was chosen to be 10^{-4}. When training end-to-end the CNN plus RNN architectures, the learning rate was either 10^{-4} or 10^{-5}; when training them, keeping their respective CNN parts fixed, it was 10^{-3}. All networks were trained using Tensorflow on a Quadro GV100 Volta GPU and the training time was about a day.

6.2 Objective Function

Since the evaluation criterion of the OMG-Emotion Challenge was the CCC, our loss function was based on that criterion and was defined as:

$$\mathcal{L}_{total} = 1 - \frac{\rho_a + \rho_v}{2}, \tag{3}$$

where ρ_a and ρ_v are the CCC for the arousal and valence.

6.3 Post-Processing

Finally, for all investigated methods, a chain of post-processing steps was applied. These steps included: i) median filtering of the - per frame - predictions within a sequence and ii) smoothing of the - per utterance - predictions (especially to those that consisted of too few frames). Any of these post-processing steps was kept when an improvement was observed on the CCC over the validation set, and applied then, with the same configuration to the test partition.

7 Experimental Results

In all conducted experiments, best results were obtained when the final estimates were the median of the, per frame, valence and arousal estimates within a sequence. In all developments, we trained the DNNs with the training set, evaluated them on the respective validation set and selected the best networks according to the validation performance. There were no significant differences between training the DNN multiple times and then averaging the predictions, or using a 10-fold cross validation.

We examined to include a level of encoding for matching the size of landmarks with the size of the CNN features before fusing them. We first passed the 68 landmarks to a fully connected layer of 512, 1024, or 2048 units and then fused this output with the features extracted from the CNN. However, we did not notice any significant difference in performance, although the developed architectures were more complex and bigger in terms of learnable parameters.

7.1 CNN-RNN Component Analysis

Table 3 shows the performance of the developed CNN, standard CNN plus RNN, CNN plus Multi-RNN and ensemble architectures, pre-trained on the Aff-Wild2 database, with and without the post-processing steps described in Subsection 6.3 (for all networks: p-value $\leqslant 10^{-20} \ll 0.05$). The VGG-FACE has achieved the best performance compared to the ResNet-50 and DesNet-121 networks. This is expected as the VGG-FACE network has been pre-trained with a large dataset for face recognition (many human faces have been, therefore, used in its construction), thus better filters are already established in comparison to the ResNet-50 and DesNet-121 that have been pre-trained on objects. Additionally, after further pre-training on Aff-Wild2, a better tuning of these filters is attained in the VGG-FACE case.

Additionally, AffWildNet and DenseNet-RNN networks achieved a better performance than all CNN networks. The former networks are standard CNN plus RNNs in which the RNN is used in order to model the contextual information in the data, taking into account temporal variations and thus a better performance is expected.

One can also note that both CNN-1RNN-2nd-pool_last-pool_fc and CNN-3RNN-2nd-pool_last-pool_fc exhibit a much improved performance (between 6% and 10% on average) when compared to CNN plus RNN architectures. This validates our sence that low-level CNN features together with high-level ones provide useful information for our task. Additionally, CNN-3RNN-2nd-pool_last-pool_fc outperformed CNN-1RNN-2nd-pool_last-pool_fc showing that it is better to exploit the low- and high-level features' time variations via RNNs, independently, and then concatenate them, rather than concatenate them first and process them through the use of a single RNN.

Table 3 validates that using the ensemble methodology is better than using a single network. This is because different networks produce quite different features; fusing them exploits all these representations that include rich information. It can also be observed that Model-level fusion method has a superior performance compared to that of the Decision-level one, since the features from different networks that are concatenated, contain richer information about the raw data than the final decision. In particular, in Model-level fusion, we concatenate these features and pass them through an RNN and the whole ensemble is trained end-to-end and optimized so that the concatenation of features can provide the best overall result. Moreover, in Model-level fusion, a better performance is achieved when a RNN, instead of a fully connected layer, is used for the fusion.

One can also notice that the post-processing steps helped to achieve a better performance, mainly in valence estimation. The median filter size that we used was 81 for valence (similar to the sequence length), whereas only 3 for the arousal. The arousal window size was small, but, when it was increased, the performance decreased. Our final observation is that the performance of the networks in arousal was worse than their performance in valence. This is expected because we only used the visual modality for training our networks; for arousal the audio cues appear to include more discriminating capabilities than facial features in terms of correlation coefficient; this conclusion confirms previous findings [46].

In the following, we compare the performance of the best performing networks of Table 3 with post-processing to that of networks trained from scratch, or being pre-trained with the Aff-Wild or the Aff-Wild2 database. Table 4 presents the results of this comparison. The Aff-Wild2 database, due to its big size and emotion diversity, boosted the performance of all networks pre-trained with it, in comparison to the performance of the networks trained directly with the OMG-Emotion set. This was also the case when we pre-trained the networks with the Aff-Wild database. Overall, networks pre-trained with the Aff-Wild2 achieved a better performance in comparison to networks pre-trained with the Aff-Wild database.

Between CNN-1RNN and CNN-3RNN types of architectures, a better performance was acquired when using the latter one. Next, we present an ablation study on extracting different CNN low-, mid- and high-level features in CNN-3RNN networks. Table 5 compares their performance (in all cases: p-value $\leqslant 10^{-25} \ll 0.05$). The first four rows of Table 5 show the performance of networks where a combination of low-, mid- and high-level features are extracted, whereas the next rows show the performance of networks where only low-, or only mid-, or only high-level features are extracted. Let us note that worst performances among all these types of networks were obtained when features were extracted from mid- CNN levels (convolutional layers 6-9). Generally, best performances were obtained when features were extracted from high- and from low-levels. The optimal combination (that provided the best performance) was through the use of CNN-3RNN-2nd-pool_last-pool_fc. One more observation is that low-level features (convolutional layers 3-5), especially when combined with high-level, significantly affected the performance in predicting both valence and arousal.

TABLE 3: CCC based evaluation, on the OMG test set, of valence & arousal predictions provided by our developed CNN, CNN plus RNN, CNN plus Multi-RNN and ensemble architectures. All networks are pre-trained on Aff-Wild2 with (without) post-processing. A higher CCC value indicates a better performance.

CCC	With (Without) Post-Processing		Mean
	Valence	Arousal	
VGG-Face	*0.378 (0.361)*	*0.203 (0.193)*	*0.291 (0.277)*
DenseNet-121	0.365 (0.350)	0.191 (0.184)	0.278 (0.267)
ResNet-50	0.359 (0.344)	0.195 (0.189)	0.277 (0.267)
AffWildNet	*0.409 (0.390)*	*0.224 (0.219)*	*0.317 (0.305)*
DenseNet-RNN	0.394 (0.378)	0.211 (0.209)	0.303 (0.294)
CNN-1RNN-2nd-pool_last-pool_fc	*0.449 (0.441)*	*0.303 (0.297)*	*0.376 (0.369)*
CNN-3RNN-2nd-pool_last-pool_fc	**0.472 (0.463)**	**0.329 (0.322)**	**0.401 (0.393)**
Decision-Level Fusion	0.501 (0.482)	0.332 (0.321)	0.417 (0.402)
Model-Level Fusion + FC	*0.518 (0.500)*	*0.348 (0.328)*	*0.433 (0.414)*
Model-Level Fusion + RNN	***0.535 (0.512)***	***0.365 (0.340)***	***0.450 (0.426)***

TABLE 4: CCC based evaluation, on the OMG test set, of valence & arousal predictions provided by various networks when: they are trained from scratch or are pre-trained with the Aff-Wild and Aff-Wild2 databases. A higher CCC value indicates a better performance.

Methods	Trained from Scratch		Pre-trained on Aff-Wild		Pre-trained on Aff-Wild2	
	Valence	Arousal	Valence	Arousal	Valence	Arousal
CNN-1RNN-2nd-pool_last-pool_fc	0.371	0.210	0.419	0.278	**0.449**	**0.303**
CNN-3RNN-2nd-pool_last-pool_fc	0.385	0.192	0.448	0.302	**0.472**	**0.329**
Model-level Fusion + RNN	0.431	0.265	0.511	0.342	**0.535**	**0.365**

Next, we present an ablation study on the use of landmarks as additional input to various networks. Table 6 compares the performance of the CNN-1RNN-2nd-pool_last-pool_fc, CNN-3RNN-2nd-pool_last-pool_fc and Model-level Fusion + RNN networks when the landmarks are and are not used as additional input. In all cases, using landmarks increases their performance by 1.2% - 1.9%.

Finally, to give more insight on the performance of the best CNN-3RNN (CNN-3RNN-2nd-pool_last-pool_fc), we analyzed its performance at different parts of the 2D Valence-Arousal Space. Table 7 presents the obtained valence and arousal performance in terms of Mean Squared Error (MSE) across 4 different regions of this Space. It can be

TABLE 5: Effect on CCC (on the OMG test set) of using features from different layers in the CNN-3RNN case. All networks are post-processed & pre-trained on Aff-Wild2. A higher CCC value indicates a better performance.

CNN-3RNN	CCC		Mean
	Valence	Arousal	
8th conv + last pool + fc	0.416	0.261	0.339
5th conv + last pool + fc	0.455	0.322	0.389
2nd pool + last pool + fc	**0.472**	**0.329**	**0.401**
3rd conv + 7th conv + fc	0.402	0.267	0.335
last conv + last pool + fc	0.440	0.248	0.344
6th conv + 7th conv + 8th conv	0.328	0.162	0.245
7th conv + 8th conv + 9th conv	0.334	0.172	0.253
3rd conv + 4th conv + 5th conv	0.345	0.185	0.265

TABLE 6: Effect on CCC (on the OMG test set) of (not) using landmarks as additional input to various networks. All networks are post-processed & pre-trained on Aff-Wild2. A higher CCC value indicates a better performance. V,A stand for Valence and Arousal

CCC	Without Landmarks			With Landmarks		
	V	A	Mean	V	A	Mean
CNN-1RNN-2nd-pool_last-pool_fc	0.429	0.291	0.360	**0.449**	**0.303**	**0.376**
CNN-3RNN-2nd-pool_last-pool_fc	0.454	0.310	0.382	**0.472**	**0.329**	**0.401**
Model-level Fusion + RNN	0.524	0.352	0.438	**0.535**	**0.365**	**0.450**

seen that better results have been obtained in the region with high arousal and positive valence; however the obtained MSE are not far away from the MSE across the whole 2D Valence-Arousal Space.

TABLE 7: Valence and Arousal MSE in areas of the 2D VA Space for the best CNN-3RNN. A lower MSE indicates a better performance. V,A stand for Valence and Arousal

2D VA-Space	V ∈ [0,1] A ∈ [0,0.5)	V ∈ [0,1] A ∈ [0.5,1]	V ∈ [-1,0) A ∈ [0,0.5)	V ∈ [-1,0) A ∈ [0.5,1]	V ∈ [-1,1] A ∈ [0,1]
CNN-3RNN-2nd-pool_last-pool_fc	MSE-V = 0.101 MSE-A = 0.031	MSE-V = 0.055 MSE-A = 0.021	MSE-V = 0.154 MSE-A = 0.061	MSE-V = 0.110 MSE-A = 0.040	MSE-V = 0.110 MSE-A = 0.041

7.2 Submissions to the OMG-Emotion Challenge

For the OMG-Emotion Challenge each team was allowed to have up to 3 submissions. We have submitted the CNN2-to-1RNN and CNN-3RNN-last-conv_last-pool_fc pre-trained on Aff-Wild models' predictions without post-processing (submission I) and with post-processing: either with median filtering (submission II) or with median filtering and smoothing (submission III; our best one). More details regarding our submissions can be found in [30].

7.3 Comparison with State-of-the-Art

Here we compare the performance of our best networks to the performances of state-of-the-art methods submitted to the OMG-Emotion Challenge. The authors of [50] developed the VNet and ANet models. VNet is a SphereFace [38] network, followed by a BLSTM, followed by a temporal pooling and the output layer. ANet is a VGG16 network with average pooling and accepts as input STFT maps extracted from the audio. In their fusion, the features extracted from VNet's temporal pooling and ANet's average pooling layers, are concatenated and passed to the output layer.

The authors of [59] developed two models. In the first model, denoted as openSMILE + LSTMs, features extracted from audio using openSMILE were passed through six 2-layer LSTMs, each predicting valence, arousal or both; the final prediction was their average. In the second model, denoted as VGG-FACE-BLSTM, the visual modality was used; frames from the utterances were passed through a fixed and pre-trained VGG-FACE followed by a 2-layer BLSTM that gave the final valence prediction.

The authors of [63] developed both single and ensemble networks, consisting of three models. In the first model, denoted as Single Multi-Modal, acoustic features were extracted using openSmile; visual features were extracted from a fixed and pre-trained VGG16 followed by 1-layer LSTM with attention mechanism; visual and acoustic features were passed into an SVM that performed the final predictions. The second model was similar to the first and extracted similar visual and acoustic features, but it also extracted acoustic features from SoundNet. All these features were passed to an SVM that performed the predictions. The late fusion of the two afore-mentioned models, is denoted as Ensemble I; the final predictions were a weighted sum of the models' predictions. The third model was an end-to-end trained VGG16 followed by 1-layer LSTM with attention mechanism that takes as input only visual data. The late fusion of the three developed models, is denoted as Ensemble II; again the final predictions were a weighted sum of the models' predictions.

Table 8 shows that our Model-level Fusion + RNN method outperforms all other methods -even those that have been trained using the audio modality as well- on both the valence and arousal estimation. Table 8 also shows that the CNN-3RNN-2nd-pool_last-pool_fc outperformed all state-of-the-art networks, regardless whether they additionally used the audio modality, except for: i) the Single Multi-Modal method that outperformed it on average by 0.015 (however this network used the audio modality as well; since the audio and speech contribute more to arousal estimation, this small difference is justified) and ii) Ensembles I and II, which are a fusion of many different networks that used the visual and audio modalities and thus again the difference in performance was expected.

TABLE 8: CCC based evaluation, on the OMG test set, of VA predictions provided by our best performing networks vs the state-of-the-art. V,A stand for valence and arousal. A higher CCC value indicates a better performance.

Methods	Modality	CCC	
		Valence	Arousal
VNet [50]	V,A: visual	0.438	0.244
ANet + VNet [50]	V,A: audio + visual	0.442	0.236
openSMILE + LSTMs, VGG-FACE-BLSTM [59]	A: audio, V: visual	0.258	0.277
openSMILE + LSTMs, VGG-FACE-BLSTM + openSMILE + LSTMs [59]	A: audio, V: audio + visual	0.369	0.286
openSMILE + LSTMs [59]	V,A: audio	0.361	0.293
Single Multi-Modal [63]	V,A: audio + visual	0.484	0.345
Ensemble I [63]	V,A: audio + visual	0.496	0.356
Ensemble II [63]	V,A: audio + visual	0.499	0.361
CNN-3RNN-2nd-pool_last-pool_fc	V,A: visual	0.472	0.329
Model-level Fusion + RNN	V,A: visual	**0.535**	**0.365**

8 Conclusions

This paper presented the development of novel architectures for predicting valence-arousal, by utilizing the OMG-Emotion dataset. The proposed approach was based on visual information and achieved very good performance when tested on the OMG-Emotion test set. In the developed networks, features extracted from low-, mid- and high-CNN layers were either concatenated and fed to a single RNN, or processed by RNN subnets and then concatenated. Moreover an ensemble approach was proposed; the Model-level fusion through a RNN produced the best results. All developed networks were first pre-trained on the rich and large Aff-Wild or Aff-Wild2 databases.

Acknowledgments

The work of Dimitris Kollias was funded by a Teaching Fellowship of Imperial College London. Additionally, we would like to thank the reviewers for their valuable comments that helped us to improve this paper.

References

[1] Acharya, U.R., Oh, S.L., Hagiwara, Y., Tan, J.H., Adeli, H., Subha, D.P.: Automated eeg-based screening of depression using deep convolutional neural network. Computer methods and programs in biomedicine **161**, 103–113 (2018)

[2] Aytar, Y., Vondrick, C., Torralba, A.: Soundnet: Learning sound representations from unlabeled video. In: Advances in Neural Information Processing Systems, pp. 892–900 (2016)

[3] Barros, P., Churamani, N., Lakomkin, E., Siqueira, H., Sutherland, A., Wermter, S.: The omg-emotion behavior dataset. arXiv preprint arXiv:1803.05434 (2018)

[4] Barsoum, E., Zhang, C., Canton Ferrer, C., Zhang, Z.: Training deep networks for facial expression recognition with crowd-sourced label distribution. In: ACM International Conference on Multimodal Interaction (ICMI) (2016)

[5] Chang, W.Y., Hsu, S.H., Chien, J.H.: Fatauva-net : An integrated deep learning framework for facial attribute recognition, action unit (au) detection, and valence-arousal estimation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshop (2017)

[6] Chen, S., Jin, Q., Zhao, J., Wang, S.: Multimodal multi-task learning for dimensional and continuous emotion recognition. In: Proceedings of the 7th Annual Workshop on Audio/Visual Emotion Challenge, pp. 19–26. ACM (2017)

[7] Chung, J., Gulcehre, C., Cho, K., Bengio, Y.: Empirical evaluation of gated recurrent neural networks on sequence modeling. arXiv preprint arXiv:1412.3555 (2014)

[8] Deng, D., Zhou, Y., Pi, J., Shi, B.E.: Multimodal utterance-level affect analysis using visual, audio and text features. arXiv preprint arXiv:1805.00625 (2018)

[9] Ding, H., Zhou, S.K., Chellappa, R.: Facenet2expnet: Regularizing a deep face recognition net for expression recognition. In: 2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017), pp. 118–126. IEEE (2017)

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

[11] Ekman, P.: Facial action coding system (facs). A human face (2002)

[12] Ekman, P.: Darwin, deception, and facial expression. Annals of the New York Academy of Sciences **1000**(1), 205–221 (2003)

[13] Gower, J.C.: Generalized procrustes analysis. Psychometrika **40**(1), 33–51 (1975)

[14] Han, S., Meng, Z., Khan, A.S., Tong, Y.: Incremental boosting convolutional neural network for facial action unit recognition. In: Advances in neural information processing systems, pp. 109–117 (2016)

[15] He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778 (2016)

[16] Iordan, A., Dolcos, F.: Brain activity and network interactions linked to valence-related differences in the impact of emotional distraction. Cerebral cortex **27**(1), 731–749 (2017)

[17] Kaltwang, S., Rudovic, O., Pantic, M.: Continuous pain intensity estimation from facial expressions. In: International Symposium on Visual Computing, pp. 368–377. Springer (2012)

[18] Kazemi, V., Sullivan, J.: One millisecond face alignment with an ensemble of regression trees. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1867–1874 (2014)

[19] Khorrami, P., Le Paine, T., Brady, K., Dagli, C., Huang, T.S.: How deep neural networks can improve emotion recognition on video data. In: Image Processing (ICIP), 2016 IEEE International Conference on, pp. 619–623. IEEE (2016)

[20] Kim, J., Calhoun, V.D., Shim, E., Lee, J.H.: Deep neural network with weight sparsity control and pre-training extracts hierarchical features and enhances classification performance: Evidence from whole-brain resting-state functional connectivity patterns of schizophrenia. Neuroimage **124**, 127–146 (2016)

[21] Kollias, D., Cheng, S., Pantic, M., Zafeiriou, S.: Photorealistic facial synthesis in the dimensional affect space. In: European Conference on Computer Vision, pp. 475–491. Springer (2018)

[22] Kollias, D., Cheng, S., Ververas, E., Kotsia, I., Zafeiriou, S.: Deep neural network augmentation: Generating faces for affect analysis. International Journal of Computer Vision pp. 1–30 (2020)

[23] Kollias, D., Marandianos, G., Raouzaiou, A., Stafylopatis, A.G.: Interweaving deep learning and semantic techniques for emotion analysis in human-machine interaction. In: 2015 10th International Workshop on Semantic and Social Media Adaptation and Personalization (SMAP), pp. 1–6. IEEE (2015)

[24] Kollias, D., Nicolaou, M.A., Kotsia, I., Zhao, G., Zafeiriou, S.: Recognition of affect in the wild using deep neural networks. In: Computer Vision and Pattern Recognition Workshops (CVPRW), 2017 IEEE Conference on, pp. 1972–1979. IEEE (2017)

[25] Kollias, D., Schulc, A., Hajiyev, E., Zafeiriou, S.: Analysing affective behavior in the first abaw 2020 competition. arXiv preprint arXiv:2001.11409 (2020)

[26] Kollias, D., Sharmanska, V., Zafeiriou, S.: Face behavior\a la carte: Expressions, affect and action units in a single network. arXiv preprint arXiv:1910.11111 (2019)

[27] Kollias, D., Tagaris, A., Stafylopatis, A., Kollias, S., Tagaris, G.: Deep neural architectures for prediction in healthcare. Complex & Intelligent Systems **4**(2), 119–131 (2018)

[28] Kollias, D., Tzirakis, P., Nicolaou, M.A., Papaioannou, A., Zhao, G., Schuller, B., Kotsia, I., Zafeiriou, S.: Deep affect prediction in-the-wild: Aff-wild database and challenge, deep architectures, and beyond. International Journal of Computer Vision **127**(6-7), 907–929 (2019)

[29] Kollias, D., Zafeiriou, S.: Aff-wild2: Extending the aff-wild database for affect recognition. arXiv preprint arXiv:1811.07770 (2018)

[30] Kollias, D., Zafeiriou, S.: A multi-component cnn-rnn approach for dimensional emotion recognition in-the-wild. arXiv preprint arXiv:1805.01452 (2018)

[31] Kollias, D., Zafeiriou, S.: A multi-task learning & generation framework: Valence-arousal, action units & primary expressions. arXiv preprint arXiv:1811.07771 (2018)

[32] Kollias, D., Zafeiriou, S.: Training deep neural networks with different datasets in-the-wild: The emotion recognition paradigm. In: 2018 International Joint Conference on Neural Networks (IJCNN), pp. 1–8. IEEE (2018)

[33] Kollias, D., Zafeiriou, S.: Expression, affect, action unit recognition: Aff-wild2, multi-task learning and arcface. arXiv preprint arXiv:1910.04855 (2019)

[34] Kuppens, P., Tuerlinckx, F., Yik, M., Koval, P., Coosemans, J., Zeng, K.J., Russell, J.A.: The relation between valence and arousal in subjective experience varies with personality and culture. Journal of personality **85**(4), 530–542 (2017)

[35] Lane, R.D., Nadel, L.: Cognitive neuroscience of emotion. Oxford University Press (1999)

[36] Little, W., Vyain, S., Scaramuzzo, G., Cody-Rydzewski, S., Griffiths, H., Strayer, E., Keirns, N.: Introduction to sociology-1st canadian edition. BC Open Textbook project (2012)

[37] Liu, C., Tang, T., Lv, K., Wang, M.: Multi-feature based emotion recognition for video clips. In: Proceedings of the 2018 on International Conference on Multimodal Interaction, pp. 630–634. ACM (2018)

[38] Liu, W., Wen, Y., Yu, Z., Li, M., Raj, B., Song, L.: Sphereface: Deep hypersphere embedding for face recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 212–220 (2017)

[39] Mathias, M., Benenson, R., Pedersoli, M., Van Gool, L.: Face detection without bells and whistles. In: European Conference on Computer Vision, pp. 720–735. Springer (2014)

[40] Meng, H., Bianchi-Berthouze, N.: Naturalistic affective expression classification by a multi-stage approach based on hidden markov models. In: International Conference on Affective Computing and Intelligent Interaction, pp. 378–387. Springer (2011)

[41] Meng, H., Huang, D., Wang, H., Yang, H., Ai-Shuraifi, M., Wang, Y.: Depression recognition based on dynamic facial and vocal expression features using partial least square regression. In: Proceedings of the 3rd ACM international workshop on Audio/visual emotion challenge, pp. 21–30. ACM (2013)

[42] Mickley Steinmetz, K.R., Kensinger, E.A.: The effects of valence and arousal on the neural activity leading to subsequent memory. Psychophysiology **46**(6), 1190–1199 (2009)

[43] Mollahosseini, A., Hasani, B., Mahoor, M.H.: Affectnet: A database for facial expression, valence, and arousal computing in the wild. arXiv preprint arXiv:1708.03985 (2017)

[44] Nasser, I.M., Al-Shawwa, M.O., Abu-Naser, S.S.: Artificial neural network for diagnose autism spectrum disorder (2019)

[45] Ng, H.W., Nguyen, V.D., Vonikakis, V., Winkler, S.: Deep learning for emotion recognition on small datasets using transfer learning. In: Proceedings of the 2015 ACM on International Conference on Multimodal Interaction, pp. 443–449. ACM (2015)

[46] Nicolaou, M.A., Gunes, H., Pantic, M.: Continuous prediction of spontaneous affect from multiple cues and modalities in valence-arousal space. IEEE Transactions on Affective Computing **2**(2), 92–105 (2011)

[47] Nicolaou, M.A., Zafeiriou, S., Pantic, M.: Correlated-spaces regression for learning continuous emotion dimensions. In: Proceedings of the 21st ACM international conference on Multimedia, pp. 773–776. ACM (2013)

[48] Nicolle, J., Rapp, V., Bailly, K., Prevost, L., Chetouani, M.: Robust continuous prediction of human emotions using multiscale dynamic cues. In: Proceedings of the 14th ACM international conference on Multimodal interaction, pp. 501–508. ACM (2012)

[49] Parkhi, O.M., Vedaldi, A., Zisserman, A.: Deep face recognition. In: BMVC, vol. 1, p. 6 (2015)

[50] Peng, S., Zhang, L., Ban, Y., Fang, M., Winkler, S.: A deep network for arousal-valence emotion prediction with acoustic-visual cues. arXiv preprint arXiv:1805.00638 (2018)

[51] Plutchik, R.: Emotion: A psychoevolutionary synthesis. Harpercollins College Division (1980)

[52] Ramirez, G.A., Baltrušaitis, T., Morency, L.P.: Modeling latent discriminative dynamic of multi-dimensional affective signals. In: International Conference on Affective Computing and Intelligent Interaction, pp. 396–406. Springer (2011)

[53] Ringeval, F., Schuller, B., Valstar, M., Gratch, J., Cowie, R., Scherer, S., Mozgai, S., Cummins, N., Schmitt, M., Pantic, M.: Avec 2017: Real-life depression, and affect recognition workshop and challenge. In: Proceedings of the 7th Annual Workshop on Audio/Visual Emotion Challenge, pp. 3–9. ACM (2017)

[54] Ringeval, F., Sonderegger, A., Sauer, J., Lalanne, D.: Introducing the recola multimodal corpus of remote collaborative and affective interactions. In: Automatic Face and Gesture Recognition (FG), 2013 10th IEEE International Conference and Workshops on, pp. 1–8. IEEE (2013)

[55] Russell, J.A.: Evidence of convergent validity on the dimensions of affect. Journal of personality and social psychology **36**(10), 1152 (1978)

[56] Tagaris, A., Kollias, D., Stafylopatis, A.: Assessment of parkinsons disease based on deep neural networks. In: International Conference on Engineering Applications of Neural Networks, pp. 391–403. Springer (2017)

[57] Tagaris, A., Kollias, D., Stafylopatis, A., Tagaris, G., Kollias, S.: Machine learning for neurodegenerative disorder diagnosissurvey of practices and launch of benchmark dataset. International Journal on Artificial Intelligence Tools **27**(03), 1850,011 (2018)

[58] Tom, N.L.S.B.L., et al.: Psychological and biological approaches to emotion. Psychology Press (1990)

[59] Triantafyllopoulos, A., Sagha, H., Eyben, F., Schuller, B.: audeering's approach to the one-minute-gradual emotion challenge. arXiv preprint arXiv:1805.01222 (2018)

[60] Whissel, C.: The dictionary of affect in language, emotion: Theory, research and experience: vol. 4, the measurement of emotions, r. Plutchik and H. Kellerman, Eds., New York: Academic (1989)

[61] Zafeiriou, S., Kollias, D., Nicolaou, M.A., Papaioannou, A., Zhao, G., Kotsia, I.: Aff-wild: Valence and arousal'in-the-wild'challenge. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, pp. 34–41 (2017)

[62] Zhao, X., Liang, X., Liu, L., Li, T., Han, Y., Vasconcelos, N., Yan, S.: Peak-piloted deep network for facial expression recognition. In: European conference on computer vision, pp. 425–442. Springer (2016)

[63] Zheng, Z., Cao, C., Chen, X., Xu, G.: Multimodal emotion recognition for one-minute-gradual emotion challenge. arXiv preprint arXiv:1805.01060 (2018)

Dimitrios Kollias , Fellow of the Higher Education Academy, holder of a Post-Graduate Certificate and member of the IEEE, is currently a Senior Lecturer in Computer Science with the School of Computing and Mathematical Sciences, University of Greenwich. He has been the recipient of the prestigious Teaching Fellowship of Imperial College London. He has obtained the Ph.D. from the Department of Computing, Imperial College London, where he was a member of the iBUG group. Prior to this, he received the Diploma/M.Sc. in Electrical and Computer Engineering from the ECE School of the National Technical University of Athens, Greece, and the M.Sc. in Advanced Computing from the Department of Computing of Imperial College London. He has published his research in the top journals and conferences on machine learning, perception and computer vision such as IJCV, CVPR, ECCV, BMVC, IJCNN, ECAI and SSCI. He is a reviewer in top journals and conferences, such as CVPR, ECCV, ICCV, AAAI, TNNL, TAC, Neurocomputing, Pattern Recognition and Neural Networks. He has been Competition Chair and Workshop Chair in IEEE FG 2020. He has won many grants and awards, such as from the City and Guilds College Association, the Imperial College Trust and the Complex & Intelligent Systems Journal. He has h-index 15 and i10-index 18. His research interests span the areas of machine and deep learning, deep neural networks, computer vision, affective computing and medical imaging.

Stefanos Zafeiriou is currently a Professor in Machine Learning and Computer Vision with the Department of Computing, Imperial College London. He also holds an EPSRC Fellowship. He received the Prestigious Junior Research Fellowships from Imperial College London in 2011 to start his own independent research group. He received the Presidents Medal for Excellence in Research Supervision for 2016. He received various awards during his doctoral and postdoctoral studies. He has been a Guest Editor of more than 6 journal special issues and co-organized more than 15 workshops/special sessions on specialized computer vision topics in top venues, such as CVPR/FG/ICCV/ECCV (including three very successfully challenges run in ICCV13, ICCV15 and CVPR'17 on facial landmark localisation/tracking). He has coauthored more than 70 journal papers mainly on novel statistical machine learning methodologies applied to computer vision problems, such as 2-D/3-D face analysis, deformable object fitting and tracking, shape from shading, and human behavior analysis, published in the most prestigious journals in his field of research, such as TPAMI, IJCV, TIP, TNNLS and many papers in top conferences, such as CVPR, ICCV, ECCV, ICML. His students are frequent recipients of very prestigious and highly competitive fellowships, such as the Google, Intel and Qualcomm ones. He has more than 12000 citations to his work, h-index 54, i10-index 159. He was the General Chair of BMVC 2017.