HAL open science

Leveraging recent advances in deep learning for audio-Visual emotion recognition

Liam Schoneveld, Alice Othmani, Hazem Abdelkawy

To cite this version:

Liam Schoneveld, Alice Othmani, Hazem Abdelkawy. Leveraging recent advances in deep learning for audio-Visual emotion recognition. Pattern Recognition Letters, 2021, 146, pp.1-7. 10.1016/j.patrec.2021.03.007 . hal-04032955

HAL Id: hal-04032955

https://hal.u-pec.fr/hal-04032955v1

Submitted on 24 Apr 2023

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Distributed under a Creative Commons Attribution - NonCommercial 4.0 International License

Version of Record: https://www.sciencedirect.com/science/article/pii/S0167865521000878
Manuscript_e01ec380f194891be9246ea8c05204c3

Pattern Recognition Letters
journal homepage: www.elsevier.com

Leveraging Recent Advances in Deep Learning for Audio-Visual Emotion Recognition

Liam Schoneveld[a], Alice Othmani[b,**], Hazem Abdelkawy[a,b]

[a] *Powder AI Research*
[b] *Université Paris-Est, LISSI, UPEC, 94400 Vitry sur Seine, France*

ABSTRACT

Emotional expressions are the behaviors that communicate our emotional state or attitude to others. They are expressed through verbal and non-verbal communication. Complex human behavior can be understood by studying physical features from multiple modalities; mainly facial, vocal and physical gestures. Recently, spontaneous multi-modal emotion recognition has been extensively studied for human behavior analysis. In this paper, we propose a new deep learning-based approach for audio-visual emotion recognition. Our approach leverages recent advances in deep learning like knowledge distillation and high-performing deep architectures. The deep feature representations of the audio and visual modalities are fused based on a model-level fusion strategy. A recurrent neural network is then used to capture the temporal dynamics. Our proposed approach substantially outperforms state-of-the-art approaches in predicting valence on the RECOLA dataset. Moreover, our proposed visual facial expression feature extraction network outperforms state-of-the-art results on the AffectNet and Google Facial Expression Comparison datasets.

© 2020 Elsevier Ltd. All rights reserved.

1. Introduction

Darwin concluded through his observations and descriptions of human emotional expressions that emotions adapt to evolution, are biologically innate, and universal across all human and even non-human primates (Matsumoto [2001]). Formal, systematic research studies have since been realized on the universality of emotions. This work demonstrated: (i) the universality of six basic emotions (anger, disgust, fear, happiness, sadness and surprise) and (ii) the cultural differences in spontaneous emotional expressions (Ekman et al. [1987]).

A human's emotion resulting from an interaction with stimuli is referred to as an *affect*. In psychology, an affect refers to the mental counterparts of internal bodily representations associated with emotions. In fact, humans express affect through facial, vocal or gestural behaviors. The notion of affect is subjective, and in the literature it is represented by two alternative views: **the categorical view** where affects are represented as discrete states with a wide variety of affective displays and **the dimensional view**, where we suppose that affects might not be culturally universal and alternatively, should be represented in a continuous arousal-valence space.

Recently, a trend in the scientific community has emerged towards developing new technologies for processing, interpreting or simulating human emotions through Affective Computing or through Artificial Emotional Intelligence. Consequently, a broad range of applications have been developed in Human-Computer Interaction, health informatics and assistive technologies. Initial research on affect recognition focused mainly on unimodal approaches, with speech emotion recognition (SER) and facial expression recognition (FER) (Rouast et al. [2019]) treated as separate problems. More recently however, work in affective computing has paid more attention to multimodal emotion recognition by developing approaches to multimodal data fusion.

Research on affect recognition has seen considerable progress as the focus has shifted from the study of laboratory-controlled databases to databases covering real-world scenarios. In traditional emotion recognition databases, subjects posed a particular basic emotion in laboratory-controlled conditions. In more recent databases, videos are obtained from real-life scenarios with *in-the-wild* environmental conditions and less constrained settings, which exhibit characteristics like illumination variation, noise, occlusion, non-frontal head poses,

**Corresponding author:
e-mail: alice.othmani@u-pec.fr (Alice Othmani)

© 2021 published by Elsevier. This manuscript is made available under the CC BY NC user license
https://creativecommons.org/licenses/by-nc/4.0/

and so on. Today, automatic emotion recognition of the six basic emotions in acted visual and/or audio expressions can be performed with high accuracy. However, in-the-wild emotion recognition is a more challenging problem due to the fact that spontaneously occurring behavior varies more widely in its audio profile, visual aspects, and timing.

In the present era, deep learning-based approaches are revolutionizing many areas of technology. Automatic emotion recognition likewise can benefit from the effectiveness of deep learning. In this paper, we propose a new approach for audio-visual emotion recognition (AVER). This approach is based on pre-training separate audio and visual deep convolutional neural network (CNN) recognition modules. A fusion module is then trained on the specific audio-visual dataset of interest. The fusion module is trained with the combination of generic emotion recognition features extracted by our pre-trained audio and visual components. The remainder of the paper is organized as follows. Section 2 reviews the literature on AVER and presents the contributions of our paper. Section 3 describes the proposed approach. Section 4 presents the experiments then reports and discusses the results. Finally, Section 5 concludes the paper and suggests future work.

2. Related Works and paper contributions

2.1. Related Work

Multimodal fusion for emotion recognition concerns the family of machine learning approaches that integrate information from multiple modalities in order to predict an outcome measure. Such is usually either a class with a discrete value (e.g., happy vs. sad), or a continuous value (e.g., the level of arousal/valence). Several literature review papers survey existing approaches for multimodal emotion recognition (Rouast et al. [2019]; BaltruÅąaitis et al. [2018]; Zeng et al. [2008]; Poria et al. [2017]). There are three key aspects to any multimodal fusion approach: (i) which features to extract, (ii) how to fuse the features, and (iii) how to capture the temporal dynamics.

Extracted features: several handcrafted features have been designed for AVER. These low-level descriptors concern mainly geometric features like facial landmarks. Meanwhile, commonly-used audio signal features include spectral, cepstral, prosodic, and voice quality features. Recently, deep neural network-based features have become more popular for AVER. These deep learning-based approaches fall into two main categories. In the first, several handcrafted features are extracted from the video and audio signals and then fed to the deep neural network (Ringeval et al. [2015]; He et al. [2015]; Othmani et al. [2019]; Rejaibi et al. [2019]; Muzammel et al. [2020]). In the second category, raw visual and audio signals are fed to the deep network (Tzarakis et al. [2017]; Tzirakis et al. [2018]; Basnet et al. [2019]). Deep convolutional neural networks (CNNs) have been observed as outperforming other AVER methods (Rouast et al. [2019]).

Multimodal features fusion: An important consideration in multimodal emotion recognition concerns the way in which the audio and visual features are fused together. Four types of strategy are reported in the literature: feature-level fusion, decision-level fusion, hybrid fusion and model-level fusion (Zhang et al. [2017]; Poria et al. [2017]).

Feature-level fusion also called *early-fusion* concerns approaches where features are immediately integrated after extraction via simple concatenation into a single high-dimensional feature vector. Such is the most common strategy for multimodal emotion recognition. *Decision-level fusion* or *late fusion* concerns approaches that perform fusion after an independent prediction is made by a separate model for each modality. In the audio-visual case, this typically means taking the predictions from an audio-only model, and the prediction from a visual-only model, and applying an algebraic combination rule of the multiple predicted class labels such as 'min', 'sum', and so on. *Score-level fusion* is a subfamily of the decision-level family that employs an equally weighted summation of the individual unimodal predictors. *Hybrid fusion* combines outputs from early fusion and from individual classification scores of each modality. *Model-level fusion* aims to learn a joint representation of the multiple input modalities by first concatenating the input feature representations, and then passing these through a model that computes a learned, internal representation prior to making its prediction. In this family of approaches, multiple kernel learning (Chen et al. [2014]), and graphical models (BaltruÅąaitis et al. [2018, 2013]) have been studied, in addition to neural network-based approaches.

Modelling temporal dynamics: audio-visual data represents a dynamic set of signals across both spatial and temporal dimensions. Rouast et al. [2019] identify three distinct methods by which deep learning is typically used to model these signals: *Spatial feature representations:* concerns learning features from individual images or very short image sequences, or from short periods of audio. *Temporal feature representations:* where sequences of audio or image inputs serve as the model's input. It has been demonstrated that deep neural networks and especially recurrent neural networks are capable of capturing the temporal dynamics of such sequences (Kim et al. [2017]). *Joint feature representations:* in these approaches, the features from unimodal approaches are combined. Once features are extracted from multiple modalities at multiple time points, they are fused using one of strategies of modality fusion (Ringeval et al. [2015]).

2.2. Contributions of this work

In this work, a higher-performing deep neural network-based approach for AVER is presented. The proposed model is a fusion of two deep neural networks: (i) a deep CNN model, trained with knowledge distillation, for FER and (ii) a modified and fine-tuned VGGish model for SER. A model-level fusion based approach is employed to fuse the audio and the visual feature representations. To model the temporal dynamics, the spatial and temporal representations are processed using recurrent neural networks. The contributions of this work can be summarized as follows:

- A new high-performing deep neural network-based approach for AudioVisual Emotion Recognition (AVER)
- Learning two independent feature extractors – one for audio and one for face images – that are specialised

for emotion recognition, and that could be employed for any downstream audiovisual emotion recognition task or dataset

- Applying knowledge distillation (specifically, *self-distillation*), alongside additional unlabeled data for FER
- Learning the spatio-temporal dynamics via a recurrent neural network for AVER.

3. Proposed multimodal deep CNN architecture

Our proposed multimodal deep CNN architecture is made up of three components:

3.1. Visual facial expression embedding network

The first component of our multimodal architecture is a deep convolutional neural network (CNN) for facial expression recognition. The input to this network is a single RGB face image, detected and cropped using Multi-Task Cascaded Convolutional Networks (MTCNN) (Zhang et al. [2016]). The output of this network is a compact vector of dimension D_{face}.

We refer to this network as our 'facial expression embedding network', and we train it using *knowledge distillation* (Hinton et al. [2014]). Knowledge distillation is a two step process whereby a *teacher* network is trained on the task of interest, and then a (typically smaller) *student* network is trained on predictions made by the teacher. Specifically in this work, we leverage the benefits of *self-distillation*, whereby the student network has the same size as (or at least, is not smaller than) the teacher network. Self-distillation was recently leveraged to achieve state-of-the-art results on the well-known Imagenet classification dataset (Xie et al. [2020]). It has also been shown theoretically that self-distillation can improve a model's performance via a regularization effect (Mobahi et al. [2020]). We use self-distillation to improve the performance of our facial expression embedding network. The training procedure for this network thus consists of two phases:

1. Training a teacher model: our teacher model is a fine-tuned FaceNet (Schroff et al. [2015]), trained simultaneously on two different visual facial expression recognition datasets (Section 3.1.1).
2. Training a student model: a second CNN is additionally trained to mimic the outputs of this fine-tuned FaceNet (Section 3.1.2).

3.1.1. The teacher network

The starting point for our teacher model is a pre-trained FaceNet (Schroff et al. [2015]). This model is then trained to learn specialised features for facial emotion recognition using two datasets:

- **AffectNet** (Mollahosseini et al. [2019]), which consists of around 440,000 in-the-wild face crop images, each of which is human-annotated into one of eight facial expression categories (Neutral, Happy, Sad, Surprise, Fear, Disgust, Anger and Contempt).
- **Google Facial Expression Comparison (FEC)** (Agarwala and Vemulapalli [2019]), which consists of around 700,000 triplets of unique face crop images. Annotations denoting the most similar pair of face expressions in each triplet are provided. The goal is to train a model that places the similar pair closer together in a learned embedding space.

Our teacher model's architecture (Fig. 1) is almost identical to the model proposed in Agarwala and Vemulapalli [2019], the only difference is that we add an additional output head for the AffectNet loss. A pre-trained FaceNet[1] is taken up until the Inception 4e block. This is followed by a 1x1 convolution and a series of five untrained DenseNet (Huang et al. [2017]) blocks. After this, another 1x1 convolution followed by global average pooling reduces this representation to a single D_{face} dimensional vector. After pooling, two independent linear transformations serve as output heads. These heads take the D_{face}-dimensional facial expression representation vector as input and make separate predictions for the AffectNet and FEC tasks. A 32-dimensional embedding is used for the FEC triplets task, while an 8-dimensional output head produces class logits for AffectNet (which has 8 classes). The teacher network training procedure is detailed in Algorithm 1, while implementation details are provided in the supplementary materials.

To improve the regularization effects of self-distillation through model ensembling, we in fact train two teacher networks, and concatenate their outputs to serve as distillation targets (see Section 3.1.2 for details). The only difference between the two teachers are the random seeds used for initialization, and penultimate layer dimensionalities: we use $D_{\text{face}} = 128$ for one teacher network, and $D_{\text{face}} = 256$ for the other.

3.1.2. Student network

Our student network is a DenseNet201 pre-trained on Imagenet.[2] The student network training procedure is essentially the same as described in Algorithm 1, except that we additionally sample batches of unlabeled data from an internal dataset, which we refer to as *PowderFaces*. The PowderFaces dataset was created by downloading approximately 20,000 short, publicly-available videos from various online sources such as YouTube. To increase the frequency of faces in the dataset, specific search terms and topics were used when searching for videos, such as 'podcast', 'interview', or 'monologue'. MTCNN face detection was then applied to the extracted frames from those videos, producing approximately 1 million individual face crops. The sampled batches of face crops from the Google FEC, AffectNet, and PowderFaces datasets are passed through our two teacher networks. Each of the two teacher networks produces predictions for the Google FEC task (32-dimensional) and AffectNet class logits

[1] In this, we used a FaceNet pre-trained on the VGGFace2 dataset, as we found performance to be slightly improved. The pre-trained FaceNet model architecture and weights were obtained from https://github.com/timesler/facenet-pytorch.

[2] We use the implementation and pre-trained Imagenet weights provided in the `torchvision` Python package.

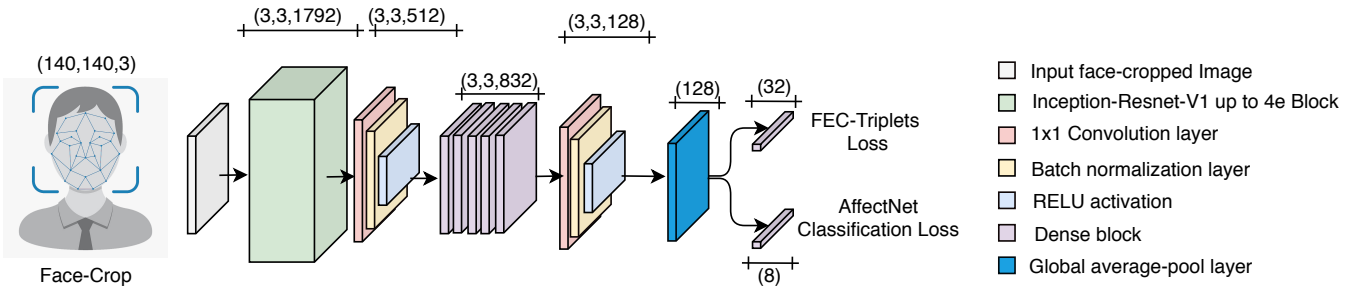

Fig. 1. Our facial expression recognition neural network architecture, before distillation (i.e. the *teacher network*). Faces are detected and cropped using MTCNN. The resulting 140x140 RGB images are then fed to an Inception Resnet V1 until the Inception 4e block. This is followed by a 1x1 convolution layer (1x1 Conv), batch normalization (BN) and a ReLU activation. Then, five DenseNet blocks are applied. Finally, another set of 1x1 Conv, BN and ReLU is applied. The output is then averaged over the spatial dimensions, resulting in a vector of size D_{face} (in the figure, $D_{\text{face}} = 128$). Two separate linear layers then give us the final model outputs – a vector for the Google FEC triplets task, and class logits for AffectNet. The model is trained to minimize both the AffectNet and Google FEC losses simultaneously. The numbers over each block represent the tensor's output shape after applying that block.

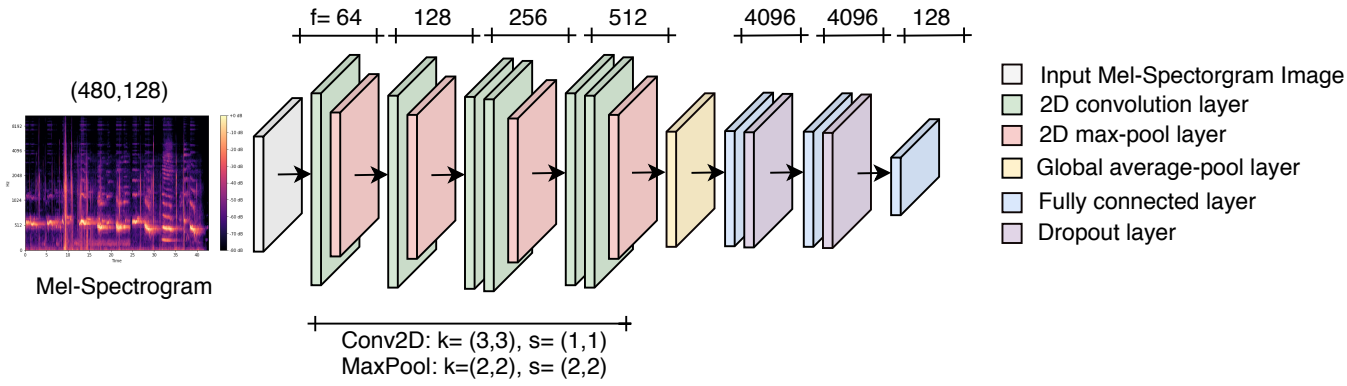

Fig. 2. Modified VGGish backbone feature extractor for Speech Emotion Recognition. The Mel-Spectogram is computed from the audio signal and then fed to the modified VGGish backbone network consisting of 6 convolutional layers followed by 3 fully connected layers of size (4096, 4096 and 128) to output an embedding vector of size 128. The size of the feature maps (f) of each convolutional and fully-connected layer are shown above each block of operations. The kernel size (k) and stride (s) are specified below the convolution blocks.

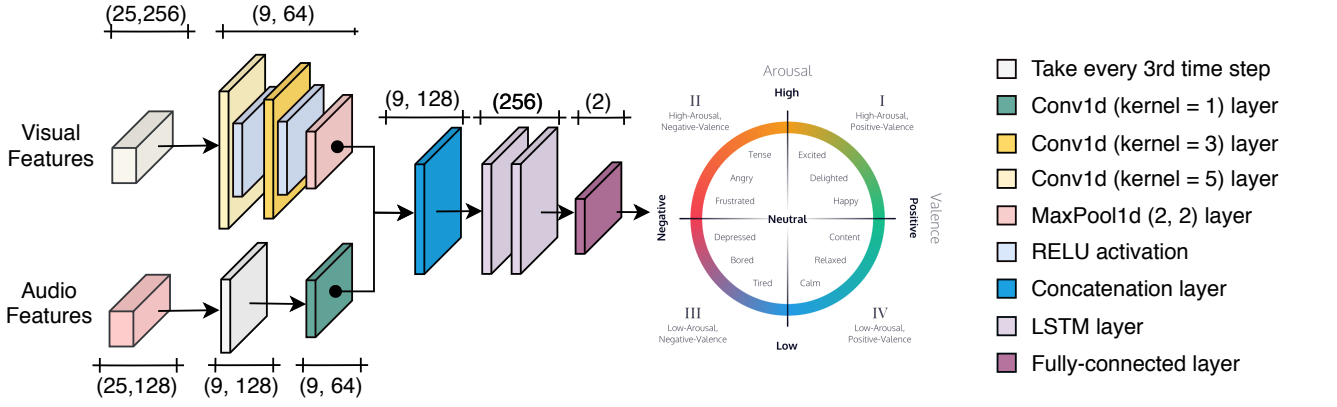

Fig. 3. Our audio-visual fusion network architecture. The audio and the visual embedding vectors are each fed to a small, independent convolutional network. This results in one tensor of size (9, 64) for each modality. We concatenate these two to give a tensor of size (9, 128), which is fed to a two-layer LSTM network with dimensionality of 256. Taking the final time step's output of this LSTM gives a single vector of size 256, which is pass through a single fully-connected layer with two outputs, which after a tanh activation gives our predictions between -1 and 1 for arousal and valance.

(8-dimensional). These four vectors (i.e., two vectors from two teacher networks) are individually L2-normalized The four normalized vectors are then concatenated, producing one long vector of dimension 80. A knowledge distillation loss (we specifically use 'Relational Knowledge Distillation' (Park et al. [2019]) for our loss function) is then calculated by comparing the output of a third output head in the student network to this 80-dimensional target vector. This knowledge distillation loss is then added to the standard AffectNet and Google FEC losses, which are calculated as per the teacher network training procedure. Implementation details for our student network are provided in the supplementary materials.

3.2. Audio embedding network for emotion recognition

This section details our proposed deep learning-based approach for recognizing emotions from audio segments. The approach is based on fine-tuning the VGGish model (Hershey et al. [2017]) on the RECOLA dataset (Ringeval et al. [2013]).

3.2.1. Audio pre-processing

For each input audio file, a set of Mel-spectrogram representations (R) are created. The input audio files are down-sampled at a 16KHz sampling rate. Then, the short-time Fourier transform (STFT) is performed to create windows with length (l) of 40 milliseconds and a hop length of 40 milliseconds. To create the latter windows, a set of 128 Mel filters (M_f) are applied with a Mel frequency range of 125-7500 Hz. Finally, for each audio file, a tensor of shape $[R, l, M_f]$ is generated to create a compatible pre-processed data input for the VGGish backbone network.

Algorithm 1: Visual model: training the teacher network. Given feature extractor network f_Θ, Google FEC output head g_ϕ, AffectNet output head h_θ, number of training steps N, AffectNet loss weight α.

for *iteration in range(N)* **do**
- $(\mathbf{X}_{\text{FEC}}, \mathbf{y}_{\text{FEC}}) \leftarrow$ batch of Google FEC triplets and labels
- $(\mathbf{X}_{\text{Aff}}, \mathbf{y}_{\text{Aff}}) \leftarrow$ batch of AffectNet images and class labels
- $\mathbf{e}_{\text{FEC}} \leftarrow f_\Theta(\mathbf{X}_{\text{FEC}})$ ▹ Face embeddings for FEC images
- $\mathbf{e}_{\text{Aff}} \leftarrow f_\Theta(\mathbf{X}_{\text{Aff}})$ ▹ Face embeddings for AffectNet images
- $\mathbf{v}_{\text{FEC}} \leftarrow g_\phi(\mathbf{e}_{\text{FEC}})$ ▹ Predict vectors for triplet loss
- $\mathbf{p}_{\text{Aff}} \leftarrow h_\theta(\mathbf{e}_{\text{Aff}})$ ▹ Predict class probabilities for AffectNet
- $L_{\text{FEC}} = \texttt{triplet_loss}(\mathbf{v}_{\text{FEC}}, \mathbf{y}_{\text{FEC}})$
- $L_{\text{Aff}} = \texttt{cross_entropy_loss}(\mathbf{p}_{\text{Aff}}, \mathbf{y}_{\text{Aff}})$
- $L = L_{\text{FEC}} + \alpha * L_{\text{Aff}}$ ▹ Total loss for training step
- Obtain all gradients $\Delta_{\text{all}} = (\frac{\partial L}{\partial \Theta}, \frac{\partial L}{\partial \phi}, \frac{\partial L}{\partial \theta})$
- $(\Theta, \phi, \theta) \leftarrow \texttt{SGD}(\Delta_{\text{all}})$ ▹ Update feature extractor and output heads' parameters simultaneously

end

3.2.2. VGGish backbone network

Our deep model for audio-based emotion recognition is based on a modified version of the VGGish model (Hershey et al. [2017]). Our starting point is the original VGGish model, pre-trained on the Audio Set dataset (Gemmeke et al. [2017]). The VGGish backbone consists of 6 convolutional layers that output 64, 128, 256, and 512 feature maps (f) respectively. For each convolution layer, a kernel (k) with size 3x3, and stride (s) of 1x1 is used. A max pooling layer with a kernel (k) of size 2x2, and stride (s) 2x2 is then applied. We take this VGGish backbone, but replace its last convolution and max pooling layers with a global average pooling layer. The resulting model produces an output vector of dimension 256. After this, we add three randomly-initialized, fully-connected layers, with output dimensionalities of 4096, 4096, and 128. These layer sizes were chosen to mimic the fully-connected penultimate layers of the original VGGish architecture. The aim of these layers is to extract a standard embedding vector with size of 128 that reflects the emotional characteristics of the input audio segment.

We take this expanded VGGish backbone architecture and fine-tune it end-to-end on the RECOLA dataset. Two separate VGGish networks are fine-tuned: one to predict arousal and the other to predict valence. We pass inputs of size $[480, 128]$ to the VGGish model, which are the mel-spectrogram representations of 30 seconds of audio from one of the videos in the RECOLA dataset. The target used for fine tuning is then the average ground truth arousal or valence for the target values corresponding to the input 30 seconds of audio. We predict this target by passing the 128-dimensional audio representation through a fully-connected layer f_ϕ with a `tanh` activation. The training procedure for fine tuning our audio feature extraction model is detailed in Algorithm 2.

Algorithm 2: VGGish fine-tuning algorithm for predicting arousal. Given the VGGish feature extractor network f_Θ, arousal prediction head f_ϕ, number of training steps N.

for *iteration in range(N)* **do**
- $(\mathbf{X}, \mathbf{y}) \leftarrow$ batch of RECOLA spectrograms and targets
- $\mathbf{e} \leftarrow f_\Theta(\mathbf{X})$ ▹ Calculate VGGish embeddings for batch
- $\mathbf{p} \leftarrow f_\theta(\mathbf{e})$ ▹ Predict arousal for all elements in batch
- $Loss = -\texttt{concordance_correlation_coeff}(\mathbf{p}, \mathbf{y})$
- Obtain all gradients $\Delta_{\text{all}} = (\frac{\partial Loss}{\partial \Theta}, \frac{\partial Loss}{\partial \theta})$
- $(\Theta, \theta) \leftarrow \texttt{Adam}(\Delta_{\text{all}})$ ▹ Update VGGish model, output head

end

3.3. Audio-visual fusion model

A model-level fusion based approach is considered as shown in Fig. 3. Visual features are extracted by taking face crops from the video sequence of interest using MTCNN and passing them to our student network (Section 3.1.2). Audio features are extracted using our fine-tuned VGGish backbone (the temporal granularity of these features is increased by removing the

global average pooling over the temporal dimension from our fine-tuned VGGish model). To have the same reduced size, audio and visual features are passed through small, independent pre-transform networks consisting of 1D convolutions, where the convolution filters slide over the temporal dimension. The pre-transform networks are designed to give an output tensor of shape [9, 64]. These are concatenated into a single audio-visual features tensor, which is passed to a two-layer LSTM network with hidden and output dimensionality of 256. The final output of the LSTM network is then taken, and passed to a simple linear transform with two outputs. These outputs are passed through a `tanh` activation to produce the final arousal and valence predictions. Our loss function is the negative CCC. Further implementation details for our fusion network are provided in the supplementary materials.

4. Experimental results

4.1. Datasets

The performances of the proposed approach have been evaluated using the REmote COLlaborative and affective interactions (RECOLA) corpus (Ringeval et al. [2013]). In RECOLA, participants' spontaneous interactions were collected while being engaged in a remote discussion that aimed to manipulate their moods. Then, six annotators measured the emotional state present in all sequences continuously on the valence and arousal dimensions. 27 audio-visual recordings of 5 minutes of interaction, which includes 9 videos for training and 9 for validation, are publicly available. In order to perform a fair comparison, the test set annotations (for the last 9 videos) of the AVEC challenge are not given. We use the training videos to train our models, validate on the validation videos, and submitted our results on the test set to the RECOLA dataset managers for evaluation. We also evaluate our visual feature extraction network on held-out evaluation data from the AffectNet and Google FEC datasets, which were introduced in Section 3.1.1.

4.2. Evaluation metric

The Concordance Coefficient Correlation (CCC) (Lawrence and Lin [1989]) is used to evaluate the performance of the proposed approach on RECOLA, as it is standard metric for emotion recognition on the RECOLA dataset. The CCC (Equation 1) measures the agreement between a vector of predicted ($Pred$) and true ($True$) values for a continuous variable:

$$CCC(True, Pred) = \frac{2 * Corr(True, Pred) * \sigma_{True} * \sigma_{Pred}}{\sigma^2_{True} + \sigma^2_{Pred} + (\mu_{True} - \mu_{Pred})^2} \quad (1)$$

Where $\mu_{\mathbf{x}}$ is the mean of $\mathbf{x}$, $\sigma_{\mathbf{x}}$ is the standard deviation of $\mathbf{x}$, and $Corr(\mathbf{x}, \mathbf{y})$ returns Pearson's correlation coefficient between $\mathbf{x}$ and $\mathbf{y}$.

4.3. Visual facial expression embedding network performance

We evaluate our visual facial expression embedding network on the standard evaluation subsets of the two datasets it was trained on:

Table 1. Performances of the proposed visual facial expression embedding network on the AffectNet validation set comparing to existing state-of-the-art methods

Methods	Accuracy
Georgescu et al. [2019]	59.6%
Siqueira et al. [2020]	59.3%
Ours (Teacher model)	61.3%
Ours (Student, no distillation)	58.8%
Ours (Distilled student, no PowderFaces)	61.1%
Ours (Distilled student)	**61.6%**

Table 2. Triplet prediction performances of the proposed visual facial expression embedding network on the Google FEC test set comparing to existing state-of-the-art methods

Methods	Accuracy
Agarwala and Vemulapalli [2019]	81.8%
Ours (Teacher model)	84.5%
Ours (Student, no distillation)	85.0%
Ours (Distilled student, no PowderFaces)	86.4%
Ours (Distilled student)	**86.5%**

1. **AffectNet:** for AffectNet, which requires classifying faces into eight discrete facial expression classes, we train a logistic regression model on the features extracted by our student network for the entire AffectNet training set.[3] This method achieves state-of-the-art results on the AffectNet validation set, with an accuracy of 61.6% (Table 1).
2. **Google FEC:** following Agarwala and Vemulapalli [2019], we evaluate using triplet accuracy on the Google FEC test set. Using this metric, we find our approach substantially improves on state-of-the-art on the FEC test set, with an accuracy of 86.5% (Table 2).

To experimentally verify the importance of the different components of our approach, we perform an ablation study. Training the student model architecture without the distillation loss component reveals the importance of distillation. Without distillation, our model's performance dropped substantially: to 58.8% on AffectNet and 85% on Google FEC.

Similarly, to determine the importance of the unlabeled PowderFaces dataset, we again train the student model with distillation, but without the additional distillation targets provided by using this unlabeled data. The results suggest that the additional unlabeled data may not be so important to our results: accuracy on AffectNet dropped only slightly to 61.1%, and performance on Google FEC reduced by only 0.1% to 86.4%. We discuss these ablation results further in Section 4.6.

4.4. Performance on RECOLA in the visual-only and audio-only context

To test performance on RECOLA of our visual and audio feature extractors separately, we retrain the same fusion architecture, but disable either the audio or visual feature inputs.

[3] When training this logistic regression, we re-weight the classes in the AffectNet training set to have equal representation, as per the validation set.

Table 3. RECOLA dataset results (in terms of CCC) for predicting arousal and valence on train, development and test sets.

	Valence			Arousal		
CCC	Train	Dev	Test	Train	Dev	Test
Visual only	.6	.55	.66	.49	.57	.57
Audio only	.55	.46	.52	.78	.80	.70
Audio-visual	.69	.63	.74	.78	.81	.72

Table 4. Performances of the proposed audio embedding network on the RECOLA dataset comparing to existing state-of-the-art methods. In parenthesis are the performances obtained in the development set. —— : no results reported in the original papers.

Methods	Arousal	Valence
Tzarakis et al. [2017]	.70 (.75)	.31 (.41)
Han et al. [2017]	.67 (.76)	.36 (.48)
He et al. [2015]	——(.80)	——(.40)
Ours	.70 (.80)	.52 (.46)

Visual-only: feeding the embeddings from our visual feature extractor into the visual-only version of our fusion model performs well on the RECOLA dataset (Table 3). Such reaches a CCC of 0.55 for predicting valence and 0.57 for predicting arousal on the validation set, while on the test set our CCC reaches 0.66 for valence and 0.57 for arousal. This result illustrates the robustness of our visual feature extractor: when predicting valence, our method achieves state-of-the-art performance when compared to other *multimodal* approaches, even though we use *only visual* features as input.

Audio-only: our results (Table 3) show that our modified VGGish backbone feature extractor for audio segments performs well, CCCs of 0.52 and 0.70 for valence and arousal, respectively on the RECOLA test set. The achieved results for arousal prediction match the existing state-of-the-art methods when only audio features are used (Table 4). This shows that our approach to transfer learning from the acoustic events domain, and the VGGish architecture, provide a robust means to extracting audio-based features for emotion recognition.

4.5. Fusion model performance and comparison with state-of-the-art approaches

Our multimodal fusion model achieves a CCC of 0.740 in valence prediction, and 0.719 in arousal prediction in RECOLA test set (Table 5). These results reflect the robustness of our learned visual and audio features extraction techniques, and the efficacy of the approach to fusing these modalities and accounting for temporal dynamics. Our proposed method, with a CCC of 0.740, substantially outperforms all existing methods in predicting valence, with the previous best performing approach of Tzarakis et al. [2017] achieving a CCC of 0.612. Simultaneously, our approach achieves strong results in predicting arousal with a CCC of 0.719, compared to the state-of-the-art performance of Ringeval et al. [2015], who obtained a CCC of 0.796.

4.6. Discussion

The results in Section 4.3 suggest that the unlabeled Powder-Faces dataset only provides a marginal benefit in performance, if any at all. This ran contrary to our expectations: we expected unlabeled data to help in this context, given a similar approach achieved state-of-the-art results on Imagenet (Xie et al. [2020]). In that work however, the unlabeled dataset consisted of 300 million images – 300 times more than the Imagenet dataset itself. In our case, our unlabeled dataset of one million images is only about twice the size of the number of faces in AffectNet and Google FEC combined. Thus, to truly conclude whether the benefits of unlabeled data as shown on Imagenet transfer well to facial expressions, many more unlabeled images are needed. We plan to address this point in future work.

On the other hand, it appears that self-distillation is indeed beneficial in the facial expressions domain. This is illustrated by the marked improvement in our accuracy on the Google FEC dataset when distillation is applied, compared to training without distillation. For AffectNet, the benefits of distillation are less pronounced, and it seems that the combination of using a pre-trained FaceNet model, plus training on Google FEC at the same time as AffectNet are all necessary components to achieving our results.

Finally, our results for valence on the RECOLA test set show a dramatic improvement over existing state-of-the-art approaches (Table 5). As mentioned, our visual-only model for RECOLA also achieved state-of-the-art for valence (Table 3). This provides further evidence for the effectiveness of our visual feature extraction approach.

5. Conclusion and Future work

This paper introduces a high-performing deep neural network-based approach for AVER that fuses a distilled visual feature extractor network with a modified VGGish backbone and a model-level fusion architecture. The proposed visual facial expression embedding network shows that end-to-end training on both AffectNet and FEC in tandem is a highly effective method for learning robust facial expression representations. We have also demonstrated that knowledge distillation can provide further improvements for facial expression recognition. The performance of our modified VGGish backbone feature extractor presents a promising new direction for predicting emotion from audio. Moreover, our deep neural network approach to multimodal fusion has been shown to be effective in AVER, outperforming the state-of-art methods in predicting valence on the RECOLA dataset. For future work, we plan to investigate the best strategy for continuous emotion encoding: classification with coarse categories, regression, label distribution learning or even ranking.

Acknowledgements

This work was supported by Powder, a deep tech startup. Powder is a video editing and sharing platform for gamers. `https://powder.gg/`

Table 5. RECOLA dataset results (in terms of CCC) for predicting arousal and valence. S.M.: Strength modeling of SVR + BLSM

Methods	Audio features	Visual features	Modality fusion	Arousal	Valence
Ringeval et al. [2015]	LLDs + BLSTM	LLDs	Feature-level	.761	.492
Ringeval et al. [2015]	LLDs + BLSTM	LLDs	Decision-level	.796	.501
He et al. [2015]	LLDs	LLDs	model-based (BLSTM)	.747	.609
Han et al. [2017]	LDDs+S.M.	Geom. + S.M.	modality and model-based	.685	.554
Tzarakis et al. [2017]	1D CNN	ResNet50	Model-level (2 LSTM)	.714	.612
Proposed	**Fine-tuned VGGish**	**Distilled CNN**	**Model-based (LSTM)**	**.719**	**.740**

References

Agarwala, A., Vemulapalli, R., 2019. A compact embedding for facial expression similarity, pp. 5676–5685.

BaltruÅąaitis, T., Chaitanya, A., Louis-Philippe, M., 2018. Multimodal machine learning: A survey and taxonomy. IEEE transactions on pattern analysis and machine intelligence 41, 423–443.

BaltruÅąaitis, T., Ntombikayise, B., Peter, R., 2013. Dimensional affect recognition using continuous conditional random fields. IEEE International Conference and Workshops on Automatic Face and Gesture Recognition , 1–8.

Basnet, R., Islam, M.T., Howlader, T., Rahmanb, S.M.M., Hatzinakos, D., 2019. Estimation of affective dimensions using cnn-based features of audiovisual data. Pattern Recognition Letters 128, 290–297.

Chen, J., Chen, Z., Chi, Z., Fu, H., 2014. Emotion recognition in the wild with feature fusion and multiple kernel learning, Proceedings of the 16th International Conference on Multimodal Interaction. pp. 508–513.

Ekman, P., Friesen, W.V., O'sullivan, M., Chan, A., Diacoyanni-Tarlatzis, I., Heider, K., Krause, R., LeCompte, W., Ayhan Pitcairn, T., Ricci-Bitti, P.E., Scherer, K., Tomita, M., Tzavaras, A., 1987. Universals and cultural differences in the judgments of facial expressions of emotion. Journal of personality and social psychology 53, 712.

Gemmeke, J.F., Ellis, D.P.W., Freedman, D., Jansen, A., Lawrence, W., Moore, R.C., Plakal, M., Ritter, M., 2017. Audio set: An ontology and human-labeled dataset for audio events.

Georgescu, M.I., Ionescu, R.T., Popescu, M., 2019. Local learning with deep and handcrafted features for facial expression recognition. IEEE Access 7, 64827–64836.

Han, J., Zhang, Z., Cummins, N., Ringeval, F., Schuller, B., 2017. Strength modelling for real-world automatic continuous affect recognition from audiovisual signals. Image and Vision Computing 65, 76–86.

He, L., Jiang, D., Yan, L., Pei, E., Wu, P., Sahli, H., 2015. Multimodal affective dimension prediction using deep bidirectional long short-term memory recurrent neural networks. Proceedings of the 5th International Workshop on Audio/Visual Emotion Challenge , 73–80.

Hershey, S., Chaudhuri, S., Ellis, D.P.W., Gemmeke, J.F., Jansen, A., Moore, C., Plakal, M., Platt, D., Saurous, R.A., Seybold, B., Slaney, M., Weiss, R., Wilson, K., 2017. Cnn architectures for large-scale audio classification. International Conference on Acoustics, Speech and Signal Processing (ICASSP) .

Hinton, G., Vinyals, O., Dean, J., 2014. Distilling the knowledge in a neural network. In Deep Learning Workshop NIPS .

Huang, G., Liu, Z., Van Der Maaten, L., Weinberger, K.Q., 2017. Densely connected convolutional networks, in: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 4700–4708.

Kim, D.H., Baddar, W.J., Jang, J., Ro, Y.M., 2017. Multi-objective based spatio-temporal feature representation learning robust to expression intensity variations for facial expression recognition. IEEE Transactions on Affective Computing 10, 223–236.

Lawrence, I., Lin, K., 1989. A concordance correlation coefficient to evaluate reproducibility. Biometrics , 255–268.

Matsumoto, D., 2001. The handbook of culture and psychology. chapter Culture and emotion. pp. 171–194.

Mobahi, H., Farajtabar, M., Bartlett, P.L., 2020. Self-distillation amplifies regularization in hilbert space. arXiv preprint arXiv:2002.05715 .

Mollahosseini, A., Hasani, B., Mahoor, M.H., 2019. Affectnet: A database for facial expression, valence, and arousal computing in the wild. IEEE Transactions on Affective Computing 10, 18–31.

Muzammel, M., Salam, H., Hoffmann, Y., Chetouani, M., Othmani, A., 2020. Audvowelconsnet: A phoneme-level based deep cnn architecture for clinical depression diagnosis. Machine Learning with Applications , 100005.

Othmani, A., Kadoch, D., Bentounes, K., Rejaibi, E., Alfred, R., Hadid, A., 2019. Towards robust deep neural networks for affect and depression recognition from speech. arXiv preprint arXiv:1911.00310 .

Park, W., Kim, D., Lu, Y., Cho, M., 2019. Relational knowledge distillation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3967–3976.

Poria, S., Cambria, E., Bajpai, R., Hussain, A., 2017. A review of affective computing: From unimodal analysis to multimodal fusion. Information Fusion 37, 98–125.

Rejaibi, E., Komaty, A., Meriaudeau, F., Agrebi, S., Othmani, A., 2019. Mfcc-based recurrent neural network for automatic clinical depression recognition and assessment from speech. arXiv preprint arXiv:1909.07208 .

Ringeval, F., Eyben, F., Kroubi, E., Thiran, J.P., Ebrahimi, T., Lalanne, D., Schuller, B., 2015. Prediction of asynchronous dimensional emotion ratings from audiovisual and physiological data. Pattern Recognition Letters 66, 22–30.

Ringeval, F., Sonderegger, A., Sauer, J., Lalanne, D., 2013. Introducing the recola multimodal corpus of remote collaborative and affective interactions. In 2013 10th IEEE international conference and workshops on automatic face and gesture recognition (FG) , 1–8.

Rouast, P.V., Marc, A., Raymond, C., 2019. Deep learning for human affect recognition: Insights and new developments. IEEE Transactions on Affective Computing. .

Schroff, F., Kalenichenko, D., Philbin, J., 2015. Facenet: A unified embedding for face recognition and clustering. Proceedings of the IEEE conference on computer vision and pattern recognition , 815–823.

Siqueira, H., Magg, S., Wermter, S., 2020. Efficient facial feature learning with wide ensemble-based convolutional neural networks. ArXiv abs/2001.06338.

Tzarakis, P., Trigeorgis, G., Nicolaou, M., BjÃűrn, S., Stefanos, Z., 2017. End-to-end multimodal emotion recognition using deep neural networks. IEEE Journal of Selected Topics in Signal Processing 11, 1301–1309.

Tzirakis, P., Jiehao, Z., Bjorn W., S., 2018. End-to-end speech emotion recognition using deep neural networks. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) , 5089–5093.

Xie, Q., Luong, M.T., Hovy, E., Le, Q.V., 2020. Self-training with noisy student improves imagenet classification, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 10687–10698.

Zeng, Z., Pantic, M., Roisman, G.I., Huang, T.S., 2008. A survey of affect recognition methods: Audio, visual, and spontaneous expressions. IEEE transactions on pattern analysis and machine intelligence 31, 39–58.

Zhang, K., Zhang, Z., Li, Z., Qiao, Y., 2016. Joint face detection and alignment using multitask cascaded convolutional networks. IEEE Signal Processing Letters 23, 1499–1503. doi:10.1109/LSP.2016.2603342.

Zhang, S., Zhang, S., Huang, T., Gao, W., Tian, Q., 2017. Learning affective features with a hybrid deep model for audioâĂŞvisual emotion recognition. IEEE Transactions on Circuits and Systems for Video Technology 28, 3030–3043.

Graphical Abstract

Leveraging Recent Advances in Deep Learning for Audio-Visual Emotion Recognition

Liam Schoneveld, Alice Othmani, Hazem Abdelkawy

Emotional expressions are the behaviors that communicate our emotional state or attitude to others. They are expressed through verbal and non-verbal communication. Complex human behavior can be understood by studying physical features from multiple modalities; mainly facial, vocal and physical gestures. Recently, spontaneous multi-modal emotion recognition has been extensively studied for human behavior analysis. In this paper, we propose a new deep learning-based approach for audio-visual emotion recognition. Our approach leverages recent advances in deep learning like knowledge distillation and high-performing deep architectures. The deep feature representations of the audio and visual modalities are fused based on a model-level fusion strategy. A recurrent neural network is then used to capture the temporal dynamics. Our proposed approach substantially outperforms state-of-the-art approaches in predicting valence on the RECOLA dataset. Moreover, our proposed visual facial expression feature extraction network outperforms state-of-the-art results on the AffectNet and Google Facial Expression Comparison datasets.