Continuous Emotion Recognition with Spatiotemporal Convolutional Neural Networks

Thomas Teixeira¹, Éric Granger, Alessandro Lameiras Koerich
École de Technologie Supérieure, Université du Québec, H3C 1K3, Montreal, QC, Canada

Abstract

The attention in affect computing and emotion recognition has increased in the last decade. Facial expressions are one of the most powerful ways for depicting specific patterns in human behavior and describing human emotional state. Nevertheless, even for humans, identifying facial expressions is difficult, and automatic video-based systems for facial expression recognition (FER) have often suffered from variations in expressions among individuals, and from a lack of diverse and cross-culture training datasets. However, with video sequences captured in-the-wild and more complex emotion representation such as dimensional models, deep FER systems have the ability to learn more discriminative feature representations. In this paper, we present a survey of the state-of-the-art approaches based on convolutional neural networks (CNNs) for long video sequences recorded with in-the-wild settings, by considering the continuous emotion space of valence and arousal. Since few studies have used 3D-CNN for FER systems and dimensional representation of emotions, we propose an inflated 3D-CNN architecture, allowing for weight inflation of pre-trained 2D-CNN model, in order to operate the essential transfer learning for our video-based application. As a baseline, we also considered a 2D-CNN architecture cascaded network with a long short term memory network, therefore we could finally conclude with a model comparison over two approaches for spatiotemporal representation of facial features and performing the regression of valence/arousal values for emotion prediction. The experimental results on RAF-DB and SEWA-DB datasets have shown that these fine-tuned architectures allow to effectively encode the spatiotemporal information from raw pixel images, and achieved far better results than the current state-of-the-art.

Keywords: Facial Expression Recognition, Convolutional Neural Networks, Recurrent Neural Networks, Inflated 3D-CNN, Dimensional Representation of Emotion.

Email addresses: thomas.teixeira.1@ens.etsmtl.ca (Thomas Teixeira), eric.granger@etsmtl.ca (Éric Granger), alessandro.koerich@etsmtl.ca (Alessandro Lameiras Koerich)

¹Corresponding author
1. Introduction

Facial expressions are the results of peculiar positions and movements of facial muscles over time. According to previous studies, face images and videos as consecutive set of frames are a huge source of information for representing individual emotional state (Li & Deng, 2018). Depending on some identity bias of subjects such as gender, age, culture and ethnicity but also depending on the quality of face expression recordings (illumination, head pose, context), the detection of spontaneous facial expressions in-the-wild is very challenging. Although this field of research has shown a growing interest for a couple of decades, previous works have basically stated the fundamentals of human affect theory (Ekman & Friesen, 1971; Ekman, 1994) by firstly lying on discrete models, splitting emotions into few basic categories such as anger, disgust, fear, happiness, sadness, surprise, etc., which could be cross-culture recognizable. Nowadays, however, the capacity of generalization of these models have been questioned, whereas describing emotions into multi-dimensional spaces has shown more representativeness and accuracy in human psychology (Jack et al., 2012).

There are three levels for describing emotion: pleasantness, attention and level of activation. Specifically, emotion recognition datasets annotate emotional state with two values named valence and arousal, where valence represents the level of pleasantness, and arousal represents the level of activation, each of these values lying into $[-1; 1]$ range. With these values, we are able to project individual emotional state into a 2D space called the circumplex model, as shown in Figure 1. Arousal level is represented in the vertical axis and valence level on the horizontal one.

A number of existing facial expression recognition (FER) systems used handcrafted features at first such as local binary patterns (LBP) (Zavaschi et al., 2013; Shan et al., 2009), LBP-TOP (Wang et al., 2014), and scale-invariant feature transform (SIFT) descriptors (Lowe, 1999), but the advances in deep learning have introduced very competitive methods for for learning representation from images of objects. Traditionally, the design of automatic FER systems has been established on learning from annotated data. Previous studies, firstly exploited FER datasets with subjects associated to a unique category of emotion and a single frame. The state-of-the-art results achieved on FER2013 (Goodfellow et al., 2013), TFD (Susskind et al., 2010), and SFEW (Dhall et al., 2011) datasets were notably obtained using Convolutional Neural Networks (CNNs) (Georgescu et al., 2019; Kim et al., 2015; Liu et al., 2017; Zhang et al., 2015; Guo et al., 2016; Kim et al., 2016; Pramerdorfer & Kampel, 2016). Besides the previous FER systems were based on static images, it has been shown that temporal information and dynamic facial components have a crucial importance for describing face expressions (Li & Deng, 2018). By definition, people express emotions in a dynamic process, and learning spatiotemporal structures has become the current trend in recent studies by building deep FER networks on video data (Fayolle & Droit-Volet, 2014).

A classical approach for dealing with image sequences is to aggregate per frame deep features on a whole clip before final emotion prediction, such as in Ding et al. (2016). In addition to the feature aggregation, Bargal et al. (2016) concatenated mean, variance,
minimum and maximum over sequence feature vector thus bringing statistical information. However since feature aggregation cannot exploit inter-correlations between frames and is not able to depict temporal dependencies, this approach has strong limitations. To circumvent this issue, recurrent neural networks (RNN) or 3D-CNN architectures can integrate data series as input, provided that data are sequentially ordered and transitions have a substantial potential of information. While improved type of RNNs, long short term memory network (LSTM) can deal with sequential data with variable length in both directions, 3D-CNN exploits textures variations from sequence of images by extending convolutional kernels to a third dimension. Hence 3D-CNNs are well suited for computer vision applications. Tran et al. (2015) built a 3D-CNN network for action recognition task on UCF101 dataset, set of videos classified over 101 action categories. They finally demonstrated that 3D-CNNs overwhelmed 2D-CNNs on different video analysis benchmarks, and could bring efficient and compact features. Then 3D-CNNs transferred to FER applications, and recently led several FER studies based on this architecture (Abbasnejad et al. 2017; Fan et al. 2016; Nguyen et al. 2017; Liu et al. 2018; Barros & Wermter 2016; Zhao et al. 2018; Ouyang et al. 2017).

A popular approach for dealing with temporal sequences of frames is a cascaded network, in which architectures for representation learning and discrimination are stacked on top of each other, thus various levels of features are leaned by each block and processed by the following until the final prediction. Particularly combining CNN and LSTM
architectures has shown effectiveness to obtain spatiotemporal representations (Ji et al., 2013; Tran et al., 2015). For instance, Ouyang et al. (2017) used a VGG-16 CNN to extract 16-frame sequence features and fed an LSTM to get six basic emotion categories predictions. They pre-processed video frames with a multi-task cascade CNN (MTCNN) to detect faces and describe each video by a single 16-frame window. Similarly, Vielzeuf et al. (2017) used a VGG-16 CNN and an LSTM as part of a network ensemble with a 3D-CNN and an LSTM and an audio network. Particularly, they used a method called multi-instance learning (MIL) method to create bag-of-windows for each video with an specific overlapping ratio. Each sequence was described by a single label and contribute to the overall prediction of the matching video clip.

Since deep neural networks (DNNs) are highly data-dependent, there are strong limitations for designing FER systems based on DNNs, even more since FER datasets are often small and task oriented (Li & Deng, 2018). Considering this fact, training a deep model on FER datasets usually leads to overfitting. In other words, end-to-end training is not practicable if one may learn representation and a discriminant with deep architectures on images with few pre-processing. In this way, some previous work showed that additional task-oriented data for pre-training networks or fine-tuning on well-known pre-trained models could greatly help on building better performing FER models (Campos et al., 2015; Xu et al., 2014). Pre-training deep neural networks is then essential for not leading architectures to overfitting. In this way, several state-of-the-art models have been developed and shared for research purposes such as VGG-Face CNN (Parkhi et al., 2015). VGG-Face is a CNN based on the VGG-16 architecture created by Simonyan & Zisserman (2015), with the purpose of circumventing the lack of data by building an architecture for face identification and verification and make it available for the research community. This network was trained on about three million images of 2,600 different subjects, which makes this architecture specially adapted both for face and emotion recognition tasks. Recent works that have performed well in FER challenges such as EmotiW (Dhall et al., 2018) or AVEC (Ringeval et al., 2015) are based on VGG-Face architecture. Wan et al. (2017) combined linear discriminant analysis and weighted principal component analysis with VGG-Face for feature extraction and dimension reduction in a face recognition task. Knyazev et al. (2017), as part of the EmotiW challenge, fine-tuned a VGG-Face on the FER2013 dataset (Goodfellow et al., 2013), and aggregated frame features for finally classifying emotions on video sequences with a linear support vector machine. Finally, Ding et al. (2017) proposed peculiar fine-tuning techniques with VGG-Face. They constrained their own network to act like VGG-Face network by transferring the distribution of outputs from late layers rather than transferring the weights. Making different architectures act the same.

Recent works used 3D convolutional kernels for spatiotemporal description of visual information. The first 3D-CNN models were developed for action recognition task (Ji et al., 2013; Tran et al. 2015; Liu et al., 2018). Pre-trained 3D-CNNs on action recognition datasets were then made available and transferred to affect computing research (Fan et al., 2016; Nguyen et al., 2017). Ouyang et al. (2017) combined VGG-Face and LSTM among other CNN-RNN and 3D-CNN networks for finally obtaining ensemble network,
multi-modality fusion (video+audio) and output seven categorical emotion predictions. From our knowledge, no 3D-CNN model has ever been evaluated for FER systems with dimensional models, and only few models have been used on the categorical task. This is mainly due to the lack of available datasets for FER on video format, adequately exploiting the temporal dimension. Indeed more image datasets are available and pre-training on 2D-CNNs for FER systems are still the best choice. To circumvent this issue, Carreira & Zisserman (2017) developed i3D network, which is able to learn 3D feature representations based on 2D datasets. They used an inflated inception network to extend learned weights in 2D to a third dimension. In this way, they developed various pre-trained networks based on the same architecture with combination of ImageNet and Kinetics datasets either on images or optical flow inputs.

Most of the previous studies on FER are based on the categorical representation of emotion but some studies have also dealt with continuous representations of emotion, which have been proved to be effective on both image and video datasets. Discrete models are a very simple representation of emotion and they cannot be cross-culture generalizable. For instance, smiles can be either attributed to happiness or to fearness or to disgust depending on the context. On the other hand, dimensional models can distinguish emotions upon a better basis which are arousal and valence levels (Cardinal et al., 2015). These two values, widely used in the psychology field, can assign a wider range of emotional states. As firstly suggested by Simonyan & Zisserman (2014) and Kim et al. (2018), building feed-forward networks combining different high-level features, namely color features, texture features (LBP) and shape features (SIFT descriptors). Researchers have shown that low and high-level features complement each other and their combination could shrink the affective gap, which is defined as the concordance between signal properties or features and the desired output values. Other works focused on emotion recognition at group level by studying not only facial expressions but also body posture or context (Mou et al., 2015), or by exploring various physiological signals such as electrocardiogram and respiration volume (Ben Henia & Lachiri, 2017). Kollias & Zafeiriou (2018) compared and used exhaustive variations of CNN-RNN models for valence-arousal predictions on Aff-Wild dataset (Kollias et al., 2019). Past studies have particularly worked on full-length short video clips in order to predict a unique categorical label (Dhall et al., 2018; Ringeval et al., 2015). However with current datasets and dimensional models, almost every frames are annotated and several peaks of emotions can be distinguished (Kossaifi et al., 2019). Therefore, a unique label cannot be attributed to a single video clip. The straightforward approach is to split videos into several mini-clips, and averaging the predictions on consecutive frames of sequence into a unique value. Nevertheless, the duration of emotion is not standardized and it is almost totally dependent on random events such as environmental context or subject identity. In this way, windowing video clips is challenging since detecting the most significant sequence for a single unity of emotion is not straightforward. Therefore, fixing arbitrary sequence lengths could bring important biases in emotion prediction and can lead to a loss of information.

In this paper, we propose and compare different approaches for multivideo sequence
analysis for continuous emotion prediction in-the-wild based on the following assumptions: (i) splitting videos into several mini-clips can help improve predictions for continuous values of emotion, not only by increasing the amount of training data, but also by isolating facial expression units; (ii) state-of-the-art architectures and proven models such as CNN-RNN and 3D-CNNs are promising to model spatiotemporal relations; (iii) optimizing fine-tuned architectures help model convergence when a low amount of training data is available. The main contributions of this paper can be summarized as: (i) double transfer learning over VGG and ResNet architectures with ImageNet and RAF-DB datasets; (ii) a comparative study for multi-sequence learning on a regression problem (valence and arousal prediction) with 2D-CNN and 3D-CNN; (iii) experiments were conducted over various pre-trained architectures, various windows settings such as sequence length, overlapping ratio, fusion of annotations; (iv) the implementation of various initialization and fine-tuning techniques for building and training 3D-CNN models.

This paper is organized as follows. Section 2 presents an overview of the proposed approach for continuous emotion recognition including preprocessing steps, model architectures, pre-training and fine-tuning procedures and post-processing steps. Section 3 presents the datasets used in the emotion recognition task, the experimental results and the performance achieved by the proposed architectures. Finally, the last section presents the conclusion and perspective of future work.

![Figure 2: Overview of pre-trained and fine-tuning of 2D architectures.](image)

2. Proposed Approach

We propose a two-step approach for continuous emotion prediction. In the first step, to circumvent the lack of sequences of continuous labeled videos, we rely on three source image datasets: ImageNet, VGG-Face and RAF-DB. ImageNet and VGG-Face, which contains generic object images and face images, respectively, are used for pre-training three 2D-CNN architectures: VGG-11, VGG-16 and ResNet50. RAF-DB is closer to
the target dataset since it contains face images annotated with discrete emotions, and it is used for fine-tuning the 2D-CNN architectures previously trained on ImageNet and VGG-Face, as shown in Figure 2. Such 2D-CNNs will be used as baseline models with the target dataset.

In the second step, we adapt such baseline models for spatiotemporal continuous emotion recognition and fine-tune them on a target dataset. We use two strategies to model sequential information of videos, as shown in Figure 3: (i) a cascade approach where an LSTM is added after the last convolutional layer of the 2D-CNNs to form a 2D-CNN-LSTM; (ii) inflating the 2D convolutional layers of the 2D-CNNs to a third dimension to build a i3D-CNN. The second step also includes pre-processing of the videos frames, as well as post-processing of the predictions.

Figure 3: Overview of the two used deep architectures for depicting spatiotemporal features on video sequences: a cascaded network 2D-CNN-LSTM and an i3D-CNN network.

In the next sections we describe the process for pre-training and fine-tuning the 2D-CNNs, the pre-processing steps used to locate face images within video frames and to build the sequences of frames to feed the spatiotemporal models, the architecture of the 3D models, and post-processing of the emotion predictions.

2.1. Pre-Training and Fine-Tuning of 2D-CNNs

Training CNNs on small datasets systematically leads to overfitting. To circumvent this issue, CNNs can be pre-trained or fine-tuned on datasets similar or not to the target task [Campos et al. 2015; Xu et al. 2014]. Well-known CNN architectures such as AlexNet [Krizhevsky et al. 2017], VGG [Simonyan & Zisserman 2015], and GoogleNet form an important set of baselines for a large number of tasks, particularly pre-training such networks on ImageNet dataset constitutes a powerful tool for representation learning. However, recent FER studies have shown that VGG-Face architectures, which are trained on a very large dataset of face images are overwhelming architectures trained on ImageNet for FER applications [Kaya et al. 2017]. Furthermore, Li & Deng (2018) have shown that multi-stage fine-tuning can provide an even better performance. We can particularly mention FER2013 [Goodfellow et al. 2013], TFD [Susskind et al. 2010] or more recently RAF-DB [Li et al. 2017a; Li & Deng 2019] datasets as good sources of additional data for FER tasks. Besides, Tannugi et al. (2019) and Li & Deng (2020) pursued interesting work on cross-dataset generalization task by switching in turn source and target FER datasets and evaluating performance of FER models. Li & Deng (2020) have shown that
datasets are strongly biased and they have developed accordingly, novel architecture that can learn domain-invariant and discriminative features.

Globally, in this study we considered using three different data sources for double transfer learning (de Matos et al., 2019): VGG-Face, ImageNet, and RAF-DB. For the first two datasets, we already have three pre-trained architectures (VGG-11, VGG-16, ResNet50). On the other, we had to re-train such architectures on RAF-DB. We have evaluated several configurations for training and fine-tuning different CNN architectures with RAF-DB to find out how multi-stage fine-tuning can be well performed. In details, we fine-tuned CNN architectures by freezing weights from certain early layers and optimizing deeper ones. As architectures are divided into convolution blocks, we have frozen weights according to these blocks. The proposed architecture kept convolution blocks but classification layers (i.e. fully connected layers) were replaced by a stack of two fully connected layers with 512 and 128 units, respectively, and an output layer with seven units, since there are seven different emotion categories in RAF-DB: surprise, fear, disgust, happiness, sadness, anger, and neutral.

2.2. Pre-Processing

Face images are usually affected by background variations such as illumination, head pose, and face patterns linked to some identity bias. In this way, alignment and normalization are the two mostly preprocessing methods used in face recognition, which may aid learning discriminant and effective features. For instance, RAF-DB dataset contains aligned faces and subjects from SEWA-DB dataset are naturally facing a web camera while talking. Then, face alignment is not an important issue for this study. Furthermore, normalization only consists in scaling pixel values between 0 to 1 and to standardize input dimensions, faces have been resized to 100×80 pixels, which is the average dimension of faces founded in the target dataset. On the other hand, we detail other essential steps for face expression recognition in video sequences such as frame and face extraction, and window bagging.

2.2.1. Frame and Face Extraction

The videos of the target dataset (SEWA-DB), have been recorded at 50 frames per second (fps) rate. On the other hand, the valence and arousal annotations are available at each 10 ms, which corresponds to 10 fps. Therefore, it is necessary to replicate annotations for non-labeled frames when using 50 fps.

For locating and extracting faces from the frames of the SEWA-DB videos, we used a multi-task cascaded CNN (MTCNN) (Zhang et al., 2016), which has shown a great efficiency to elect the best bounding box candidates showing a complete face within the image. MTCNN employs three CNNs sequentially to decide which bounding box to keep according to particular criteria learned by deep learning. The face extractor network outputs bounding box coordinates and five facial landmarks: both eyes, nose and mouth extremities. Once faces are located, they are cropped using the corresponding bounding box. An overview of MTCNN architecture is shown in Figure 4. Only frames showing whole faces are kept, while other frames are discarded.
Figure 4: Exhaustive MTCNN architecture composed of a stack of CNNs (Zhang et al., 2016). A single meaningful bounding box and five facial landmarks are extracted from face images.

2.2.2. Sequence Learning

The target dataset contains long video sequences showing a variety of emotions along records of face expressions from a single subject. Duration of emotions is not clearly established and vary for each individual. Several studies have been previously carried out in order to get expression intensity variations by pointing peak and non-peak expressions along sequences. However, while whole video sequences represent multiple annotations at a specific sampling rate and not a single label, to represent a succession of diverse emotional states, we split the video sequences into several mini clips of fixed length with an specific overlapping ratio. This has two main advantages: (i) it increases the amount of data for training CNNs; (ii) it allows the investigation for which window settings learning of long sequences of valence and arousal annotations work best. We finally chose to evaluate two sequence lengths as number of consecutive frames in a single video (16 and 64) and three overlapping ratio for each sequence length (0.2, 0.5 and 0.8). For instance, a window of 16 consecutive frames with an overlap ratio of 0.5 contains the 8 frames of the previous window.

It was also important to check the integrity of contiguous video frames. Indeed some frames are deleted because no face was detected, hence damaging the continuity of the temporal information of emotion contained between each frame. The proposed strategy to divide videos into mini-clips may introduce important temporal gaps (two consecutive frames are faraway temporally). Therefore, we applied a tolerance (temporal difference between frames close enough) to select mini-clips that give sense to a unique emotion unit. Globally, the MTCNN can detect faces in the mini-clips and in average, 90% of the frames are kept, depending on the sequence length, overlapping ratio and frame sampling rate. Figure 5 presents the number of mini-clips available in training, validation and test sets according to such parameters. Finally, the last preprocessing step is to fuse annotations of multiple frames in one mini-clip to get a single emotion label for each window. For such aim we use either the average of labels or the extremum value of the labels to obtain a single valence and/or arousal value.
2.3. Spatiotemporal Models

We have developed two spatiotemporal models: (i) a cascaded network based on a VGG-16 network pre-trained on VGG-Face that can be fine-tuned or not on RAF-DB; (ii) an inflated network based on either VGG-11, VGG-16, or ResNet50 architectures pre-trained on different datasets (VGG-Face, RAF-DB, ImageNet).

2.3.1. Cascaded Networks

Long Short Term Memory networks (LSTMs) are a special kind of RNN, capable of learning order dependence as we may find in a sequence of frames from a video. The core of LSTMs is a cell state, which adds or removes information depending on the input, output and forget gates. The cell state remembers values over arbitrary time intervals and the gates regulate the flow of input and output information of the cell.

The architecture of the proposed cascade network combines the 2D convolutional layers of VGG-16 for representation learning with an LSTM to support sequence prediction, as shown in Figure 6. The LSTM has a single layer with 1,024 units, with random and uniform distribution initialization to extract temporal features from the face features learned by the 2D-CNN. In order to avoid overfitting, we added some dropout (20%) and recurrent dropout (20%) on LSTM units. Besides, there is also three fully connected layers stacked after the LSTM to improve the expressiveness and accuracy of the model.

The VGG-16-LSTM architecture may have two pre-training strategies: (i) VGG-16 pre-trained on VGG-Face dataset; (ii) VGG-16 pre-trained on VGG-Face and fine-tuned on RAF-DB. The former strategy adds extra information to the models, such as classification of basic emotions with RAF-DB, which could help to improve the performance on the regression task.
2.3.2. Inflated 3D-CNN (i3D-CNN)

The need to analyze a series of frames led us to the use of 3D-CNNs. 3D-CNNs produce activation maps that allow analyzing data where temporal is important. The main advantage of 3D-CNNs is to provide deep features from mini-clips that can strengthen the spatiotemporal relationship between frames. Contrary to 2D-CNNs, 3D-CNNs directly input batches of frame sequences rather than batches of frames for training. On the other hand, adding a third dimension increases the number of parameters of the model and that requires much larger training datasets than that required by 2D models. The main downside of this architecture is the lack of pre-trained models available for FER tasks. Regarding the amount of data available for the targeted task (face expression recognition on multidimensional representation of emotion), we cannot consider to train models in an end-to-end fashion. Then a solution was provided by Carreira & Zisserman (2017) with weight inflation of 2D-CNN pre-trained models. Inflating a 2D-CNN minimizes the need of large amounts of data for training properly a 3D-CNN as the inflation process reuses the weights of the 2D-CNNs. Figure 7 shows that the weight inflation principle consists of enlarging kernels of each convolution filter by one dimension. Regarding our target task, it means to extend the receptive field of each neuron to the time dimension (a.k.a. a sequence of frames). 2D convolutional kernels are then replicated as many times as necessary to fit the third dimension and form a 3D convolutional kernel. At first glance, pre-trained weights are just copied along time dimension and provide better approximation for initialization than randomness but do not constitute yet an adequate distribution for time dimension. With this in mind, the next issue is to find a method that fits best the transfer learning to time dimension with weight inflation by varying some parameters, such as: initialization, masking, multiplier and dilation.

Initialization: When replicating kernels for weight inflation, it is possible to just copy the weights $n$ times ($n$ being the dimension of time axis) or to center the weights. Centering means copying one time the 2D kernel and setting all other weights on both sides either
Figure 7: Representation of inflation method for a single convolution filter. 2D convolutional kernels are replicated along a new dimension; the temporal dimension, to obtain 3D convolutional kernels. Basically, $n \times n$ kernels are made cubic to obtain $n \times n \times n$ kernels. This process is applied to every convolutional filter to transform 2D convolutional layers into 3D convolutional layers.

randomly (with uniform distribution) or with zeros. We assume that pre-trained 2D kernels have a good capacity to generalize on images, then giving sufficiently distant distribution for all but one 2D kernels from the copied 2D kernel could have an impact on model convergence.

**Masking:** Assuming copied 2D kernels are well pre-trained for very similar task and well perform on images, the objective here is to adequately train inflated weights on time dimension. Then we consider not modifying centered weights during training in order to disseminate information learned from pre-trained weights to inflated weights.

**Multiplier:** The distribution of CNN weights and the range of targeted values for regression are closely related. Since values of valence and arousal range from $-1$ to $1$ and standard values of the weights often take values between $10^{-3}$ and $10^{-1}$, then rising targeted values by a factor could allow to scale up the distribution space and improve convergence.

**Dilation:** As suggested by Yu & Koltun (2016), we implemented dilated convolutions on our models. The dilation was performed only on time dimension. We divided the architectures into four blocks with increased level of dilation starting from level of 1 for convolutional layers (means no dilation) then 2, 4 and 8 level for top convolutional layers. Dilated convolution consists of larger receptive fields than conventional ones. In other words, neuron connections of one convolutional layer are spread among neurons of previous layers. Notably, this kind of implementation has shown a good performance for segmentation and object recognition task.
Figure 8: The proposed i3D-CNN based on inflation of 2D convolutional kernels of a pre-trained VGG-16. Video clips are fed to the i3D-CNN, then spatiotemporal face features are extracted and two fully connected (FC) layers perform the regression of valence and arousal values. For each convolutional layer a $3 \times 3 \times 3$ kernel is used, and the number of filters are indicated. We also detailed the number of units in FC layers.

Figure 8 shows the architecture of the inflated 3D-CNN, which is based on the inflation of 2D convolutional kernels of a pre-trained VGG-16. Such a i3D-CNN is then fine-tuned on the target dataset to perform regression of valence and arousal values with a sequence of fully-connected layers.

2.4. Post-Processing

The post-processing aims to improve the quality of the prediction by using some statistical information of the training set to reduce variance among datasets (Ortega et al., 2019). Due to data imbalance in the training set, some values of valence and arousal are difficult to reach. Neutral emotions, which imply valence and arousal levels close to zero, are more frequent in the training set than extreme valence and arousal values. There are three post-processing steps: scale normalization, mean filtering, and time delay.

Scale normalization consists in normalizing the predictions according to the label distribution in the training set. Valence and arousal predictions $(y')$ are normalized by the mean $(\bar{y}_{l_{tr}})$ and standard deviation $(\sigma_{l_{tr}})$ of the labels of the training set as:

$$y_{sn} = \frac{y' - \bar{y}_{l_{tr}}}{\sigma_{l_{tr}}}$$  \hspace{1cm} (1)

Mean filtering consists in centering predictions around mean values, increasing the linear relationship and correspondence to the labels. Valence and arousal predictions $(y')$ are centered by subtracting the mean values of the labels $(\bar{y}_{l_{tr}})$ of the training set and adding the mean values of the predictions $(\bar{y}_{tr}')$ on the training set as:

$$y_{mf} = y' - \bar{y}_{l_{tr}} + \bar{y}_{tr}'$$ \hspace{1cm} (2)

Finally, time delay is used to compensate some offset between the labels and the predictions due to the reaction-lag of annotators. Valence and arousal predictions $(y'(f))$
at frame $f$ are shifted over $t$ frames (precedent or subsequent) in order to align predictions and labels temporally as:

$$y_{td} = y'(f + t) \quad (3)$$

where $t$ is an integer in $[-10, 10]$.

3. Experimental Results

In this section we first present a brief description of the two FER datasets used in the experiments: RAF-DB and SEWA-DB. Next, we present the performance measures and summarize our experimental setting and the results achieved by the proposed 2D-CNN and i3D-CNN models.

3.1. Facial Expression Datasets

Real World Affective Faces Database (RAF-DB) is a real world dataset that contains 29,672 images downloaded from the Internet (Li et al., 2017b). Each image has been labeled by around 40 annotators. The dataset has two types of annotation: seven classes of basic emotions and 12 classes of compound emotions. We have only used the seven basic emotions (face images and labels). Other metadata such as facial landmarks, bounding box and identity bias such as age, gender, race are also provided but they have not been used in any step of the proposed approach. RAF-DB was used to fine-tune the pre-trained 2D-CNNs.

SEWA-DB is a large and richly annotated dataset consisting of six groups of subjects (around 30 people per group), from six different cultural background (British, German, Hungarian, Greek, Serbian, and Chinese) and divided into pairs of subjects (Kossaifi et al., 2019). Each pair had to discuss their emotional state and sentiment toward four adverts previously watched. The dataset consists of 64 videos (around 1,525 minutes of audio visual data), videos are split into three folders (34 training, 14 validation, 16 test). Since the labels are not provided for test set due to its use in FER challenges, we used the validation set as the test set and split the training set into a new training set (28 videos) and validation set (6 videos). Annotations are given for valence, arousal and level of liking. We only used valence and arousal annotations since previous studies have indicated that the level of liking is not well related with facial expressions.

3.2. Performance Metrics

The standard performance metrics used in continuous emotion recognition are the mean absolute error (MAE), Pearson correlation coefficient (PCC) and concordance correlation coefficient (CCC). PCC assess distance between target values and predictions and CCC establishes the strength of a linear relationship between two variables. The range of possible values lies in the interval $[-1; 1]$, where $-1$ or $1$ mean strong relation
and 0 means no relation at all. The MAE for a set of labels $y$ and predictions $y'$ is given by:

$$\text{MAE} = \frac{1}{n} \sum_{i=1}^{n} |y'_i - y_i|$$  \hfill (4)

The PCC is given as:

$$\rho = \frac{\sum_{i=1}^{n} (y_i - \bar{y}) (y'_i - \bar{y}')}{\sqrt{\sum_{i=1}^{n} (y_i - \bar{y})^2} \sqrt{\sum_{i=1}^{n} (y'_i - \bar{y}')^2}}$$  \hfill (5)

where $n$ is the number of samples, $y_i$ is the $i$-th label, $y'_i$ is the $i$-th prediction, and $\bar{y}$ and $\bar{y}'$ are the mean of labels and mean of predictions, respectively.

The CCC combines the PCC with the squared difference between the mean of predictions $\bar{y}'$ and the mean of the labels $\bar{y}$. CCC shows the degree of correspondence between the label and prediction distributions based on the covariance and correspondence. The CCC between a set of labels $y$ and predictions $y'$ is given by:

$$\rho_c = \frac{2\rho s_y s_{y'}}{s_y^2 + s_{y'}^2 + (\bar{y} - \bar{y}')^2}$$  \hfill (6)

$s_y^2$ and $s_{y'}^2$ are the variance of $y$ and $y'$ respectively.

3.3. Training and Fine-Tuning the 2D-CNNs

Our first task is to specialize the three pre-trained CNN architectures (VGG-11, VGG-16 and ResNet50) for emotion recognition by fine-tuning them with RAF-DB dataset. These three architectures were pre-trained either on VGG-Face or ImageNet. For fine-tuning the pre-trained 2D-CNNs on RAF-DB, video frames have been resized to $100 \times 80 \times 3$, which is the mean shape of video frames of the target dataset (SEWA-DB). Learning rate has been fixed to $1 \times 10^{-5}$ and batches of size 16. Optimization has been performed with Adam optimizer. We have assigned different weights to each class, according to the number of samples, to deal with the data imbalance found in RAF-DB. This allows that classes with few samples can affect the weights of the model to the same extent as classes with many more samples. Moreover, we observed that low-level data augmentation such as like rotation, flipping, highlight variations, could help improving the performance. Although data augmentation cannot bring significant information for emotion recognition, it can prevent overfitting on single sample and improve model distributions.

The performance achieved by the 2D-CNNs after fine-tuning on RAF-DB is presented in Table 1 where the suffix BN refers to batch normalization layers added to the original architectures after each convolution layer to improve model convergence and reduce overfitting. Furthermore, we indicate for each architecture the dataset used for pre-training as well as the convolution block (2.1 to 5.1) from which we start fine-tuning the architectures. In general, most of the fine-tuned models achieved an accuracy higher than the
baseline models (Jyoti et al., 2019). Jyoti et al. (2019) developed two CNNs to analyze action units detection efficiency on three datasets including RAF-DB. The first CNN was based on residual connections with densely connected blocks (RCNN) and the second architecture was a CNN consisting of four convolution layers and three fully connected layers. The baseline RCNN and CNN achieved 76.54% and 78.23% of accuracy on the test set of RAF-DB dataset, respectively. On the other hand, the proposed VGG-16 model pre-trained with VGG-Face achieved 79.90% of accuracy of RAF-DB dataset.

Other recent works, which employ attention networks achieved better performances (Wang et al., 2020; Li et al., 2019). In our work, models analyzed the whole aligned faces to detect emotions, but did not consider two common problems we encounter with face analysis in real-world scenarios: occlusions and pose variations. On the contrary, Wang et al. (2020) and Li et al. (2019) addressed these problems by using region-based attention networks. Attention modules are used to extract compact face representation based on several regions cropped from the face and adaptively adjusts the importance of facial parts. Therefore, these models learn to discriminate occluded and non-occluded faces while improving emotion detection in both cases.

Table 1: Results of the fine-tuning of VGG and ResNet50 models on RAF-DB. Convolution block denotes the initial level of fine-tuning of each model. Full denotes that all layers have been fine-tuned.

| Reference        | Model     | Dataset for Pre-Training | Convolution Block | Accuracy (%) |
|------------------|-----------|--------------------------|-------------------|--------------|
| Jyoti et al. (2019) | RCNN      | NA                       | NA                | 76.5         |
|                  | CNN       | NA                       | NA                | 78.2         |
| Li et al. (2019)  | ACNN      | ImageNet                 | NA                | 85.1         |
| Wang et al. (2020)| CNN+RAN   | MS-Celeb-1M              | NA                | 86.9         |
3.4. 2D-CNN-LSTM Architecture

After specializing the three pre-trained CNN architectures (VGG-11, VGG-16 and ResNet50) for emotion recognition by fine-tuning them on the RAF-DB dataset, we are able to develop cascaded networks based on such architectures for spatiotemporal continuous emotion recognition. For this task we developed two CNN-LSTM networks, one based on the VGG-16 architecture pre-trained on VGG-Face and fine-tuned on RAF-DB because such an architecture achieved the best results on RAF-DB test set, and a second one without fine-tuning on RAF-DB. Spatial features for each frame are sequentially outputted by the CNN and the LSTM extracts temporal information from a single mini-clip, respectively. Different configurations were evaluated by varying the sequence length, the overlapping ratio and the strategy to fuse the labels. The architectures were fine-tuned on the development set of SEWA-DB and the mean squared error (MSE) was used as cost function. Some other works have also considered CCC as cost function since it provides information about correspondence and correlation between predictions and annotations. However, we observed a better convergence while using the MSE.

Table 2 shows the results in terms of PCC and CCC considering different frames rates (fps), sequence lengths (SL), overlapping ratios (OR) and fusion modes (FM). In general, both extremum and mean fusion performed well and the best results for both valence and arousal have been achieved for sequences of 64 frames at 10 fps. The VGG-16 architecture benefits from fine-tuning on RAF-DB and achieved CCC values of 0.625 for valence and 0.557 for arousal on SEWA-DB. In addition to the correlation metrics, the proposed CNN-LSTM achieved a MAE of 0.05, which also indicates a good correspondence between predictions and annotations.

| Dataset for Configuration | Initialization | fps | Label (SL, OR, FM) | PCC  | CCC  |
|--------------------------|----------------|-----|-------------------|------|------|
| VGG-Face                 | 10  Valence    | 16  | 0.2, extremum     | 0.590| 0.560|
|                          | Arousal        | 16  | 0.2, mean         | 0.549| 0.542|
|                          | 50  Valence    | 64  | 0.2, mean         | 0.631| 0.625|
|                          | Arousal        | 64  | 0.2, extremum     | 0.558| 0.557|
| RAF-DB                   | 10  Valence    | 64  | 0.8, mean         | 0.582| 0.568|
|                          | Arousal        | 64  | 0.2, extremum     | 0.517| 0.517|

SL: Sequence length (in frames), OR: Overlapping ratio, FM: Fusion mode.

3.5. i3D-CNN Architecture

Another alternative for spatiotemporal modeling is to use the i3D-CNN. In this way, strong spatiotemporal correlations between frames are directly learned from video mini-clips by a single network. Thanks to weight inflation, we are able to use the pre-trained
2D-CNNs to build i3D-CNN architectures. The inflation method allows us to transpose learned information from various static tasks to dynamic ones, and therefore to perform the essential transfer learning for the detection of spatiotemporal features. With this in mind, we reused the 2D-CNN architectures shown in Table 1 and expand their convolutional layers to build i3D-CNNs considering two configurations, denote as C1 and C2 in Table 3. Due to the high number of trainable parameters, i3D-CNNs are particularly time-consuming to train and therefore we had to fix some basic hyperparameters instead of performing exploratory experiments to set such hyperparameters. Therefore, we evaluated only the best configuration found for the 2D-CNN-LSTM, as shown in Table 1, which uses batch sizes of 8, sequence length of 64 frames, overlapping ratio of 0.8, and frame rate of 10 fps. This is the main downside of our study, as the number of trainable parameters of i3D-CNNs is three times greater than the counterpart 2D-CNNs.

Table 3: Variable parameters of i3D-CNN architectures and possible values

| Parameters            | C1          | C2          |
|-----------------------|-------------|-------------|
| Inflation             | Centered    | Copied      |
| Weight Initialization | Random      | Zero        |
| Masking               | No          | Yes         |
| Dilation              | Bloc1: 1    | Bloc1: 1    |
|                       | Bloc2: 1    | Bloc2: 2    |
|                       | Bloc3: 1    | Bloc3: 4    |
|                       | Bloc4: 1    | Bloc4: 8    |
| Multiplier            | \times 1    | \times 100  |

Tables 4 and 5 show the best configurations of each architecture for valence and arousal, respectively. Globally, varying the configuration parameters has not shown any advantage of using particular values for inflation, masking, and dilation parameters. Table 6 shows the best performance obtained for each architecture for valence and arousal in terms of PCC and CCC values. Inflated 3D-CNNs for regression seem to be very sensitive to some configurations for training regarding the range of results achieved by different base models and datasets used in their initialization. In these conditions, it is difficult to state on the effect of a single parameter for inflation. VGG-16 with batch normalization and ResNet50 achieved the best results for both valence and arousal and have shown a good ability to predict these values compared to other base models. Surprisingly, the VGG-16 pre-trained on ImageNet achieved higher PCC and CCC for both valence and arousal than those base models pre-trained on VGG-Face and RAF-DB, which are source datasets closer to the target one. On the contrary, ResNet50 benefits from the initialization with VGG-Face. In summary, the best results range from 0.31 to 0.4 for PCC and from 0.25 to 0.32 for CCC. These performances still show a poor correlation between predictions and annotations but are comparable to the performance achieved by other studies of continuous emotion prediction that use SEWA-DB.
Table 4: i3D-CNN model configurations according to the best performances for predicting valence.

| Base Models | Dataset for Initialization | Models Initialization | Parameters |
|-------------|---------------------------|-----------------------|------------|
| VGG-11-BN   | RAF-DB                    | Centered              | I          | No   | Random | ×1 |
|             | ImageNet                  | Copied                | I          | No   | Random | ×100 |
| VGG-16      | VGG-Face                  | Centered              | I          | No   | Random | ×1 |
|             | RAF-DB                    | Copied                | I          | No   | Random | ×1 |
|             | ImageNet                  | Centered              | I          | No   | Random | ×1 |
| VGG-16-BN   | VGG-Face                  | Centered              | I          | No   | Random | ×1 |
|             | RAF-DB                    | Copied                | I          | No   | Random | ×1 |
|             | ImageNet                  | Copied                | I          | Yes  | Random | ×1 |
| ResNet50    | VGG-Face                  | Centered              | I          | Yes  | Random | ×100 |
|             | RAF-DB                    | Copied                | VIII       | No   | Random | ×1 |
|             | ImageNet                  | Centered              | VIII       | Yes  | Random | ×1 |

Table 5: i3D-CNN model configurations according to the best performances for predicting arousal.

| Base Models | Dataset for Initialization | Parameters |
|-------------|---------------------------|------------|
| VGG-11-BN   | RAF-DB                    | Centered   | I          | No   | Random | ×1 |
|             | ImageNet                  | Centered   | VIII       | No   | Random | ×1 |
| VGG-16      | VGG-Face                  | Centered   | I          | No   | Random | ×1 |
|             | RAF-DB                    | Copied     | I          | Yes  | Random | ×100 |
|             | ImageNet                  | Centered   | I          | No   | Random | ×1 |
| VGG-16-BN   | VGG-Face                  | Centered   | I          | Yes  | Random | ×1 |
|             | RAF-DB                    | Copied     | I          | No   | Random | ×1 |
|             | ImageNet                  | Centered   | I          | No   | Zero   | ×1 |
| ResNet50    | VGG-Face                  | Copied     | I          | Yes  | Random | ×100 |
|             | RAF-DB                    | Centered   | VIII       | Yes  | Random | ×1 |
|             | ImageNet                  | Centered   | I          | No   | Random | ×1 |
Table 6: Best performances of i3D-CNNs for predicting valence and arousal in terms of PCC and CCC values according to different models and their initialization.

| Base Models | Dataset for Valence | Dataset for Arousal |
|-------------|---------------------|---------------------|
|             | Valence PCC CCC     | Arousal PCC CCC     |
| VGG-11-BN   | RAF-DB 0.035 0.018 0.359 0.348 | ImageNet 0.040 0.025 0.342 0.203 |
| VGG-16      | RAF-DB 0.036 0.028 0.242 0.119 | ImageNet 0.209 0.190 0.391 0.189 |
| VGG-16-BN   | RAF-DB 0.123 0.101 0.284 0.165 | ImageNet 0.346 0.304 0.382 0.326 |
| ResNet50    | RAF-DB 0.113 0.063 0.262 0.207 | ImageNet 0.137 0.135 0.323 0.256 |

4. Discussion

The experiments carried out on SEWA-DB have shown that cascaded networks (2D-CNN-LSTM) achieved better results than i3D-CNNs. Notably, for the former, valence was better predicted than arousal in terms of CCC. On the contrary for the latter, arousal was better predicted than valence, also in terms of CCC. Previous works have raised the fact that intuitively face textures on video sequences are the main source of information for describing the level of positivity in emotions, hence valence values. In contrast, arousal is better predicted with voice frequencies, and audio signals. However, our work with inflated networks suggests that simultaneous learning spatiotemporal face features benefits the prediction of arousal values.

Regarding the complexity of the two proposed approaches for continuous emotion recognition, we had to make some trade-off that certainly impact the quality of our results for i3D-CNNs. But we have observed a high sensitivity in the training of this type of architecture according to various configurations. This implies that i3D-CNN architectures are very flexible and further improvement could lie in better initialization and tuning of the number and quality of parameters regarding the potential of this model. Furthermore, inflated weights provided good initialization for action recognition tasks which suggested that we could also take advantage of this method for emotion recognition. However, the main difference is that for action recognition, researchers had hundreds of various short videos for classification while we have relatively long videos of few subjects for regression of continuous values on SEWA-DB. This task is much more challenging due to the complexity of our networks. But though, the experimental results shown a great potential for further improvement if more data is available for fine-tuning the 3D models. Nevertheless, performances achieved with 2D-CNN-LSTM were very satisfying and demonstrated this type of architecture is still a good choice for this
Table 7: Comparison of the best results achieved by the proposed models and the state-of-the-art for SEWA-DB dataset.

| Reference               | Model                        | Valence        | Arousal        |
|-------------------------|------------------------------|----------------|----------------|
|                         |                              | PCC  | CCC  | PCC  | CCC  |
| Proposed                | 2D-CNN-LSTM$^1$              | 0.631| 0.625| 0.558| 0.557|
|                         | i3D-CNN$^2$                  | 0.346| 0.304| 0.382| 0.326|
|                         | SVR                          | 0.321| 0.312| 0.182| 0.202|
|                         | RF                           | 0.268| 0.207| 0.181| 0.123|
| Kossaifi et al. (2019)  | LSTM                         | 0.322| 0.281| 0.173| 0.115|
|                         | ResNet18 (RMSE)              | 0.290| 0.270| 0.130| 0.110|
|                         | ResNet18 (CCC)               | 0.350| 0.350| 0.350| 0.290|
| Chen et al. (2019)$^*$  | ST-GCN                       | NA   | 0.540| NA   | 0.581|
| Zhao et al. (2019)$^*$  | DenseNet-style CNN (DenseFace)| NA   | 0.580| NA   | 0.594|

$^1$VGG-16 fine-tuned with RAF-DB. $^2$Inflated VGG-16-BN, initialized with ImageNet. NA: Not Available.

5. Conclusion

In this paper, we have presented two CNN architectures for continuous emotion prediction in-the-wild. The first architecture is a cascade network based on a fine-tuned VGG-16 CNN and a LSTM. Such an architecture achieved state-of-the-art results on SEWA-DB dataset, producing CCC values of 0.625 and 0.557 for valence and arousal respectively. The second architecture is based on the concept of inflation, which transfer
knowledge from the 2D-CNN pre-trained models into 3D to model temporal features. The i3D-CNN achieved CCC values of 0.304 and 0.326 for valence and arousal respectively. These values are far below than those achieved by the 2D-CNN-LSTM. Due to the high number of parameters of i3D-CNNs (barely 3 times greater than 2D-CNNs), fine-tuning and hyperparameter tuning of such architectures require a huge computational efforts as well as huge datasets. Unfortunately, continuous emotion datasets are relatively small for such a task.

We have also shown that a double transfer learning strategy over VGG and ResNet architectures with ImageNet and RAF-DB datasets can improve the accuracy of the baseline models. It should be noticed that in this work, subjects were mostly facing the camera with relative clear view of the whole face. To some extent, this could imply some bias in the results when presenting diverse real world scenarios. Moreover, the i3D architecture complexity could at this time be a drag for live applications. Finally, to the best of our knowledge, it was the first time 3D-CNNs were used for regression application and the detection of valence-arousal values for emotion recognition.

There are some promising directions to expand the approaches proposed in this paper. One could take advantage of the development of more huge and complex cross-cultural datasets exploiting occlusion cases, pose variations or even scene break such as Aff-Wild dataset. In particular, with i3D-CNN, we believe deep learning algorithms possess the complexity and robustness to these specific cases and benefit from an adequate flexibility to analyze both the separability and combination of discriminant spatiotemporal features. Thus, having promising results for facial emotion recognition on video sequences. Here we have demonstrated a peculiar and flexible way of fine-tuning an inflated CNN and maybe this strategy could be transferred to other applications such as object and action recognition on video sequences.

References

Abbasnejad, I., Sridharan, S., Nguyen, D., Denman, S., Fookes, C., & Lucey, S. (2017). Using synthetic data to improve facial expression analysis with 3D convolutional networks. In 2017 IEEE Int’l Conf on Computer Vision Workshops (ICCVW) (pp. 1609–1618).

Bargal, S. A., Barsoum, E., Ferrer, C. C., & Zhang, C. (2016). Emotion recognition in the wild from videos using images. In Proc of the 18th ACM Int’l Conf on Multimodal Interaction ICMI ’16 (p. 433–436). doi:10.1145/2993148.2997627

Barros, P., & Wermter, S. (2016). Developing crossmodal expression recognition based on a deep neural model. Adaptive Behavior, 24. doi:10.1177/10597123166664017

Ben Henia, W. M., & Lachiri, Z. (2017). Emotion classification in arousal-valence dimension using discrete affective keywords tagging. In 2017 Int’l Conf on Engineering MIS (ICEMIS) (pp. 1–6).
Campos, V., Salvador, A., Giro-iNieto, X., & Jou, B. (2015). Diving deep into sentiment: Understanding fine-tuned CNNs for visual sentiment prediction. In 1st Int’l Workshop on Affect and Sentiment in Multimedia (p. 57–62).

Cardinal, P., Dehak, N., Koerich, A. L., Alam, J., & Boucher, P. (2015). ETS system for AV+EC 2015 challenge. In ACM Multimedia Conference 2015 (pp. 17–23). ACM. doi:10.1145/2808196.2811639

Carreira, J., & Zisserman, A. (2017). Quo vadis, action recognition? A new model and the kinetics dataset. In 2017 IEEE Conf on Computer Vision and Pattern Recognition, CVPR (pp. 4724–4733). doi:10.1109/CVPR.2017.502

Chen, H., Deng, Y., Cheng, S., Wang, Y., Jiang, D., & Sahli, H. (2019). Efficient spatial temporal convolutional features for audiovisual continuous affect recognition. In Proceedings of the 9th International on Audio/Visual Emotion Challenge and Workshop AVEC ’19 (p. 19–26). New York, NY, USA: Association for Computing Machinery. URL: https://doi.org/10.1145/3347320.3357690 doi:10.1145/3347320.3357690

Dhall, A., Goecke, R., Lucey, S., & Gedeon, T. (2011). Static facial expression analysis in tough conditions: Data, evaluation protocol and benchmark. In 2011 IEEE Int’l Conf on Computer Vision Workshops (ICCV Workshops) (pp. 2106–2112).

Dhall, A., Kaur, A., Goecke, R., & Gedeon, T. (2018). Emotiw 2018: Audio-video, student engagement and group-level affect prediction. In Int’l Conf on Multimodal Interaction (ICMI) (pp. 653–656). doi:10.1145/3242969.3264993

Ding, H., Zhou, S. K., & Chellappa, R. (2017). Facenet2expnet: Regularizing a deep face recognition net for expression recognition. In 12th IEEE Int’l Conf on Automatic Face & Gesture Recognition (pp. 118–126). doi:10.1109/FG.2017.23

Ding, W., Xu, M., Huang, D.-Y., Lin, W., Dong, M., Yu, X., & Li, H. (2016). Audio and face video emotion recognition in the wild using deep neural networks and small datasets. In Proc of the 18th ACM Int’l Conf on Multimodal Interaction ICMI ’16 (p. 506–513). doi:10.1145/2993148.2997637

Ekman, P. (1994). Strong evidence for universals in facial expressions: a reply to russell’s mistaken critique. Psychological bulletin, 115, 268–287.

Ekman, P., & Friesen, W. V. (1971). Constants across cultures in the face and emotion. Journal of personality and social psychology, 17, 124–129.

Fan, Y., Lu, X., Li, D., & Liu, Y. (2016). Video-based emotion recognition using CNN-RNN and C3D hybrid networks. In Proc of the 18th ACM Int’l Conf on Multimodal Interaction ICMI ’16 (p. 445–450). doi:10.1145/2993148.2997632
Fayolle, S. L., & Droit-Volet, S. (2014). Time perception and dynamics of facial expressions of emotions. *PLOS ONE, 9*, 1–9. doi:10.1371/journal.pone.0097944.

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. doi:10.1109/access.2019.2917266.

Goodfellow, I. J., Erhan, D., Carrier, P. L., Courville, A. C., Mirza, M., Hamner, B., Cukierski, W., Tang, Y., Thaler, D., Lee, D., Zhou, Y., Ramaiah, C., Feng, F., Li, R., Wang, X., Athanasakis, D., Shawe-Taylor, J., Milakov, M., Park, J., Ionescu, R. T., Popescu, M., Grozea, C., Bergstra, J., Xie, J., Romaszko, L., Xu, B., Zhang, C., & Bengio, Y. (2013). Challenges in representation learning: A report on three machine learning contests. In *20th Int’l Conf Neural Information Processing* (pp. 117–124). Springer volume 8228 of *Lecture Notes in Computer Science*. doi:10.1007/978-3-642-42051-1_16.

Guo, Y., Tao, D., Yu, J., Xiong, H., Li, Y., & D.Tao (2016). Deep neural networks with relativity learning for facial expression recognition. In *2016 IEEE Int’l Conf on Multimedia Expo Workshops (ICMEW)* (pp. 1–6).

Jack, R. E., Garrod, O. G., Yu, H., Caldara, R., & Schyns, P. G. (2012). Facial expressions of emotion are not culturally universal. *Proc of the National Academy of Sciences, 109*, 7241–7244.

Ji, S., Xu, W., Yang, M., & Yu, K. (2013). 3D convolutional neural networks for human action recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 35*, 221–231.

Jyoti, S., Sharma, G., & Dhall, A. (2019). Expression empowered residen network for facial action unit detection. In *14th IEEE Int’l Conf on Automatic Face & Gesture Recognition* (pp. 1–8). doi:10.1109/FG.2019.8756580.

Kaya, H., Gürpınar, F., & Salah, A. (2017). Video-based emotion recognition in the wild using deep transfer learning and score fusion. *Image and Vision Computing, .* doi:10.1016/j.imavis.2017.01.012.

Kim, B., Dong, S., Roh, J., Kim, G., & Lee, S. (2016). Fusing aligned and non-aligned face information for automatic affect recognition in the wild: A deep learning approach. In *2016 IEEE Conf on Computer Vision and Pattern Recognition Workshops (CVPRW)* (pp. 1499–1508).

Kim, B.-K., Lee, H., Roh, J., & Lee, S.-Y. (2015). Hierarchical committee of deep CNNs with exponentially-weighted decision fusion for static facial expression recognition. In *Proc of the 2015 ACM on Int’l Conf on Multimodal Interaction ICMI ’15* (p. 427–434). doi:10.1145/2818346.2830590.
Kim, H., Kim, Y., Kim, S. J., & Lee, I. (2018). Building emotional machines: Recognizing image emotions through deep neural networks. *IEEE Trans. Multim.*, 20, 2980–2992. doi:10.1109/TMM.2018.2827782

Knyazev, B., Shvetsov, R., Efremova, N., & Kuharenko, A. (2017). Convolutional neural networks pretrained on large face recognition datasets for emotion classification from video. *arXiv e-prints*, (p. arXiv:1711.04598). [arXiv:1711.04598](https://arxiv.org/abs/1711.04598).

Kollias, D., Tzirakis, P., Nicolaou, M. A., Papaioannou, Zhao, G., Schuller, B., Kotsia, I., & Zafeiriou, S. (2019). Deep affect prediction in-the-wild: Aff-wild database and challenge, deep architectures, and beyond. *International Journal of Computer Vision*, 127, 907–929. doi:10.1007/s11263-019-01158-4

Kollias, D., & Zafeiriou, S. (2018). A multi-component CNN-RNN approach for dimensional emotion recognition in-the-wild. *CoRR, abs/1805.01452*.

Kossaifi, J., Schuller, B. W., Star, K., Hajiyev, E., Pantic, M., Walecki, R., Panagakis, Y., Shen, J., M. Schmitt, Ringeval, F., & et al. (2019). Sewa db: A rich database for audio-visual emotion and sentiment research in the wild. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, (p. 1–1). doi:10.1109/tpami.2019.2944808

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2017). Imagenet classification with deep convolutional neural networks. *Commun. ACM*, 60, 84–90. doi:10.1145/3065386.

Li, S., & Deng, W. (2018). Deep facial expression recognition: A survey. *arXiv preprint arXiv:1804.08348*.

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

Li, S., & Deng, W. (2020). A deeper look at facial expression dataset bias. *IEEE Transactions on Affective Computing*, (p. 1–1). doi:10.1109/taffc.2020.2973158

Li, S., Deng, W., & Du, J. (2017a). Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild. In *2017 IEEE Conf on Computer Vision and Pattern Recognition (CVPR)* (pp. 2584–2593).

Li, S., Deng, W., & Du, J. (2017b). Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild. In *2017 IEEE Conf on Computer Vision and Pattern Recognition (CVPR)* (pp. 2584–2593).

Li, Y., Zeng, J., Shan, S., & Chen, X. (2019). Occlusion aware facial expression recognition using CNN with attention mechanism. *IEEE Transactions on Image Processing*, 28, 2439–2450.
Liu, K., Liu, W., Gan, C., Tan, M., & Ma, H. (2018). T-C3D: Temporal convolutional 3D network for real-time action recognition. In *AAAI*.

Liu, X., Kumar, B. V. K. V., You, J., & Jia, P. (2017). Adaptive deep metric learning for identity-aware facial expression recognition. In *2017 IEEE Conf on Computer Vision and Pattern Recognition Workshops (CVPRW)* (pp. 522–531).

Lowe, D. G. (1999). Object recognition from local scale-invariant features. In *Proc of the 7th IEEE Int’l Conf on Computer Vision* (pp. 1150–1157 vol.2). Volume 2.

de Matos, J., Britto Jr., A. S., Oliveira, L. E. S., & Koerich, A. L. (2019). Double transfer learning for breast cancer histopathologic image classification. In *IEEE International Joint Conference on Neural Networks (IJCNN)* (pp. 1–6). doi: 10.1109/IJCNN.2019.8852092.

Mou, W., Celiktutan, O., & Gunes, H. (2015). Group-level arousal and valence recognition in static images: Face, body and context. In *2015 11th IEEE Int’l Conf and Workshops on Automatic Face and Gesture Recognition (FG)* (pp. 1–6). Volume 05.

Nguyen, D., Nguyen, K., Sridharan, S., Ghasemi, A., Dean, D., & Fookes, C. (2017). Deep spatio-temporal features for multimodal emotion recognition. In *2017 IEEE Winter Conf on Applications of Computer Vision (WACV)* (pp. 1215–1223).

Ortega, J. D. S., Cardinal, P., & Koerich, A. L. (2019). Emotion recognition using fusion of audio and video features. In *IEEE Intl Conf on Systems, Man, and Cybernetics (SMC)* (pp. 3827–3832). doi: 10.1109/SMC.2019.8914655.

Ouyang, X., Kawaai, S., Goh, E., Shen, S., Ding, W., Ming, H., & Huang, D.-Y. (2017). Audio-visual emotion recognition using deep transfer learning and multiple temporal models. In *19th ACM Int’l Conf on Multimodal Interaction* (pp. 577–582). doi: 10.1145/3136755.3143012.

Parkhi, O. M., Vedaldi, A., & Zisserman, A. (2015). Deep face recognition. In *Proc of the British Machine Vision Conf (BMVC)* (pp. 41.1–41.12). doi: 10.5244/C.29.41.

Pramerdorfer, C., & Kampel, M. (2016). Facial Expression Recognition using Convolutional Neural Networks: State of the Art. *arXiv e-prints*, (p. arXiv:1612.02903).

Ringeval, F., Schuller, B., Valstar, M., Jaiswal, S., Marchi, E., Lalanne, D., Cowie, R., & Pantic, M. (2015). Av+ec 2015: The first affect recognition challenge bridging across audio, video, and physiological data. In *Proc of the 5th Int’l Workshop on Audio/Visual Emotion Challenge AVEC ’15* (p. 3–8). doi: 10.1145/2808196.2811642.

Shan, C., Gong, S., & McOwan, P. W. (2009). Facial expression recognition based on local binary patterns: A comprehensive study. *Image and Vision Computing*, 27, 803–816.
Simonyan, K., & Zisserman, A. (2014). Two-stream convolutional networks for action recognition in videos. In *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems* (pp. 568–576).

Simonyan, K., & Zisserman, A. (2015). Very deep convolutional networks for large-scale image recognition. In *3rd Int’l Conf on Learning Representations, (ICLR)*.

Susskind, J. M., Anderson, A. K., & Hinton, G. E. (2010). The toronto face dataset. *Technical Report UTML TR 2010-001, U. Toronto,* .

Tannugi, D. L., Britto Jr., A. S., & Koerich, A. L. (2019). Memory integrity of cnns for cross-dataset facial expression recognition. In *IEEE Intl Conf on Systems, Man, and Cybernetics (SMC)* (pp. 3806–3811). doi:10.1109/SMC.2019.8914172

Tran, D., Bourdev, L. D., Fergus, R., Torresani, L., & Paluri, M. (2015). Learning spatiotemporal features with 3d convolutional networks. In *2015 IEEE Int’l Conf on Computer Vision, (ICCV)* (pp. 4489–4497). doi:10.1109/ICCV.2015.510

Vielzeuf, V., Pateux, S., & Jurie, F. (2017). Temporal multimodal fusion for video emotion classification in the wild. In *19th ACM Int’l Conf on Multimodal Interaction (ICMI)* (pp. 569–576). doi:10.1145/3136755.3143011

Wan, L., Liu, N., Huo, H., & Fang, T. (2017). Face recognition with convolutional neural networks and subspace learning. In *2017 2nd Int’l Conf on Image, Vision and Computing (ICIVC)* (pp. 228–233).

Wang, K., Peng, X., Yang, J., Meng, D., & Qiao, Y. (2020). Region attention networks for pose and occlusion robust facial expression recognition. *IEEE Trans. Image Process.*, 29, 4057–4069. doi:10.1109/TIP.2019.2956143

Wang, Y., See, J., Phan, R. C.-W., & Oh, Y.-H. (2014). Lbp with six intersection points: Reducing redundant information in lbp-top for micro-expression recognition. In *ACCV*.

Warr, P., Bindl, U., Parker, S., & Inceoglu, I. (2014). Four-quadrant investigation of job-related affects and behaviours. *European Journal of Work and Organizational Psychology*, 23, 342–363.

Xu, C., Cetintas, S., Lee, K.-C., & Li, L.-J. (2014). Visual sentiment prediction with deep convolutional neural networks. [arXiv:1411.5731](http://arxiv.org/abs/1411.5731)

Yu, F., & Koltun, V. (2016). Multi-scale context aggregation by dilated convolutions. In *4th Int’l Conf on Learning Representations (ICLR)*.

Zavaschi, T. H. H., Britto Jr., A. S., Oliveira, L. E. S., & Koerich, A. L. (2013). Fusion of feature sets and classifiers for facial expression recognition. *Expert Systems with Applications, 40*, 646–655. doi:10.1016/j.eswa.2012.07.074

27
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.

Zhang, Z., Luo, P., Loy, C. C., & Tang, X. (2015). Learning social relation traits from face images. In *2015 IEEE Int’l Conf on Computer Vision (ICCV)* (pp. 3631–3639). doi:10.1109/ICCV.2015.414.

Zhao, J., Li, R., Liang, J., Chen, S., & Jin, Q. (2019). Adversarial domain adaption for multi-cultural dimensional emotion recognition in dyadic interactions. In *Proceedings of the 9th International on Audio/Visual Emotion Challenge and Workshop AVEC ’19* (p. 37–45). New York, NY, USA: Association for Computing Machinery. URL: https://doi.org/10.1145/3347320.3357692 doi:10.1145/3347320.3357692.

Zhao, J., Mao, X., & Zhang, J. (2018). Learning deep facial expression features from image and optical flow sequences using 3D CNN. *The Visual Computer, 34*. doi:10.1007/s00371-018-1477-y.