Continuous Emotion Recognition via Deep Convolutional Autoencoder and Support Vector Regressor

Sevegni Odilon Clement Allognon*, Alceu de S. Britto Jr.† and Alessandro L. Koerich*

*École de Technologie Supérieure, Université du Québec, Montréal, QC, Canada
Email: sevegni-odilon.allognon.1@ens.etsmtl.ca, alessandro.koerich@etsmtl.ca
†Pontifical Catholic University of Paraná, Curitiba, PR, Brazil
Email: alceu@ppgia.pucpr.br

***Abstract*—Automatic facial expression recognition (FER) is an important research area in the emotion recognition and computer vision. Applications can be found in several domains such as medical treatment, driver fatigue surveillance, sociable robotics, and several other human-computer interaction systems. Therefore, it is crucial that the machine should be able to recognize the emotional state of the user with high accuracy. In recent years, deep neural networks have been used with great success in recognizing emotions. In this paper, we present a new model for continuous emotion recognition based on FER by using an unsupervised learning approach based on transfer learning and autoencoders. The proposed approach also includes preprocessing and post-processing techniques which contribute favorably to improving the performance of predicting the concordance correlation coefficient for arousal and valence dimensions. Experimental results for predicting spontaneous and natural emotions on the RECOLA 2016 dataset have shown that the proposed approach based on visual information can achieve concordance correlation coefficient of 0.516 and 0.264 for valence and arousal, respectively.**

***Index Terms*—Deep learning, Unsupervised learning, Representation learning, Facial expression recognition**

I. Introduction

The visual recognition of emotional states usually involves analyzing a person's facial expression, body language, or speech signals. Facial expressions contain abundant and valuable information about the emotion and thought of human beings. Facial expressions naturally transmit emotions even if a subject wants to mask his/her emotions. Several researchers suggest that there are emotional strokes produced by the brain and shown involuntarily by our corps through the face [1]. Emotions are an important process for human-to-human communication and social contact. Thus, emotions need to be considered to achieve better human-machine interaction.

In psychology research [1], [2], there are three emotion theories to model the emotion state: discrete theory, appraisal theory and dimensional theory. The discrete theory claims that there exists a small number of discrete emotions (i.e., anger, disgust, happiness, neutral, sadness, fear, and surprise) that are inherent in our brain and recognized universally [3]. Such a theory has been largely adopted in research on emotion recognition. However, it has some drawbacks as it does not take into consideration people who exhibit non-basic, subtle and complex emotions like depression. It results that these basic discrete classes may not reflect the complexity of the emotional state expressed by humans. As a result, the appraisal theory has been introduced. This is a theory where emotions are generated through continuous, recursive subjective evaluation of both our own internal state and the state of the outside world [3]. Nonetheless, the appraisal theory is still an open research problem on how to use it for automatic measurement of emotional state. Finally, for the dimensional theory, the emotional state considers a point in a continuous space. This third theory can model the subtle, complicated and continuous emotional state. It models emotions using two independent dimensions, i.e. arousal (relaxed vs. aroused) and valence (pleasant vs. unpleasant) as shown in Fig. 1. The valence dimension indicates how positive or negative an emotion is, and it ranges from unpleasant to pleasant. The arousal dimension indicates how excited or apathetic an emotion is, ranging from sleepiness or boredom to frantic excitement [4].

The typical computational approach for emotion recognition is to take every single data as a single unit (e.g., a frame of a video sequence) independently. It can be made as a standard regression problem for every frame using the so-called static (frame-based) regressors. Many researches have been scrutinized by predicting emotion in continuous dimensional space from the recognition of discrete emotion categories. However, emotion recognition is a challenging task because human emotions lack temporal boundaries. Moreover, each individual expresses and perceive emotions in different ways. In addition, one utterance may contain more than one emotion.

Several deep learning architectures such as convolutional neural networks (CNNs), autoencoder (AE), memory enhanced neural network models such as long short-term memory models (LSTM), have recently been used successfully for emotion recognition. Traditionally facial expression recognition (FER) consists of feature extraction utilizing handcrafted representations such as local binary pattern (LBP) [5], [35]–[39],

978-1-7281-6926-2/20/$31.00 ©2020 IEEE

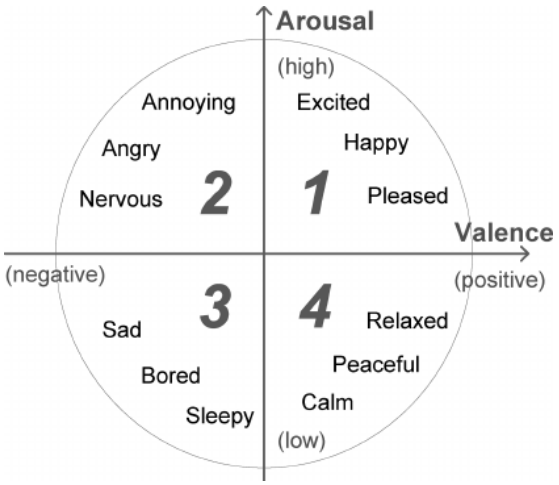

Fig. 1. Valence-Arousal 2D dimension plane [45].

histogram of oriented gradients (HOG) [7], [40], scale invariant feature transform (SIFT) [6], Gabor wavelet coefficients [28]–[33], [37], Haar features [31], [34], principal component analysis [41], [42], 3D shape parameters [43] and then predict the emotion from these extracted features. For instance, Shan et al. [5] formulated a boosted-LBP feature and combined it with a support vector machine (SVM) classifier. Berretti et al. [6] computed the SIFT descriptor on 3D facial landmarks of depth images and used SVM for the classification. Albiol et al. [7] proposed a HOG descriptor-based elastic bunch graph matching (EBGM) algorithm that is more robust to changes in illumination, rotation, small displacements and that achieved a higher accuracy compared to the classical Gabor–EBGM ones.

A number of studies in the literature have focused on predicting emotion from facial expressions by using deep neural networks (DNN). Zhao et al. [8] combined deep belief networks (DBN) and multi-layer perceptron for FER. Mostafa et al. [9] used recurrent neural networks to study emotion recognition from facial features. A large majority of these scientific studies had been carried out using handcrafted features. Despite the fact that these approaches reported good accuracy for the prediction, the handcrafted feature has its inherent drawbacks; either unintended features that do not benefit classification may be included or important features that have a great influence on the classification may get omitted. This is because these features are "crafted" by human experts, and the experts may not be able to consider all possible cases and include them in feature vectors.

With the recent success achieved in deep learning, a trend in machine learning has emerged towards deriving a representation directly from the raw input signal. Such a trend is motivated by the fact that CNNs learn representation and discriminant functions through iterative weight updated by backpropagation and error optimization. Therefore, CNNs could include critical and unforeseen features that humans hardly come up with and hence contribute to improving the performance. CNNs have been employed in many works but oftentimes, they require a high number of convolutional layers to learn a good representation due to the high complexity of facial expression images. The disadvantage of increasing network depth is the complexity of the network as the training time, which can grow significantly with each additional layer. Furthermore, increasing network complexity requires more training data and it makes it more difficult to find the best network configuration as well as the best initialization parameters.

In this paper, we introduce unsupervised feature learning to predict the emotional state in an end-to-end approach. We aim to learn good representations in order to build a compact continuous emotion recognition model with a reduced number of parameters that produce a good prediction. We propose a convolutional autoencoder (CAE) architecture, which learns good representations from facial images while keeping a low dimensionality of the representation. The encoder is used to compress the data and the decoder is used to reproduce the original image. The representation learned by the CAE is used to train a support vector regressor (SVR) to predict the affective state of individuals. In the proposed architecture we did not take into account the temporal correlation between adjacent frames. The main contributions of this paper are: (i) learning a compact but meaningful representation of continuous affective states; (ii) the representation is learned directly from the raw images; (iii) the representation is learned from unlabeled raw data and it achieves concordance correlation coefficients (CCCs) that are comparable to the state-of-the-art in continuous emotion recognition.

The structure of this paper is as follows. Section II provides the most recent studies on emotion recognition from facial expressions. Section III introduces the proposed approach. In Section IV, we describe the dataset used in this study. We present our results in Section V. Conclusions and perspectives of future work are presented in the last section.

II. Related Work

Several studies have been proposed to model the FER problem using raw face images with DNNs. Tang [10] used L2-SVM objective function to train DNNs for classification. Lower layer weights are learned by backpropagating the gradients from the top layer linear SVM by differentiating the L2-SVM objective function with respect to the activation of the penultimate layer. Liu et al. [11] proposed a 3D-CNN and deformable action part constraints in order to locate facial action parts and learn part-based features for emotion recognition. In the same vein, Liu et al. [12] extracted image-level features with pre-trained CNN models. Yu and Zhang [13] proved that the random initialization of neural networks allowed to vary network parameters and also renders the classification ability of diverse networks. Because of that, the ensemble technique usually shows concrete performance improvement. Kahou et al. [14] proposed an approach that combines multiple DNNs for different data modalities such as facial images, audio, bag of mouth features with CNN, deep restricted Boltzmann machine and the output of such modalities are averaged to

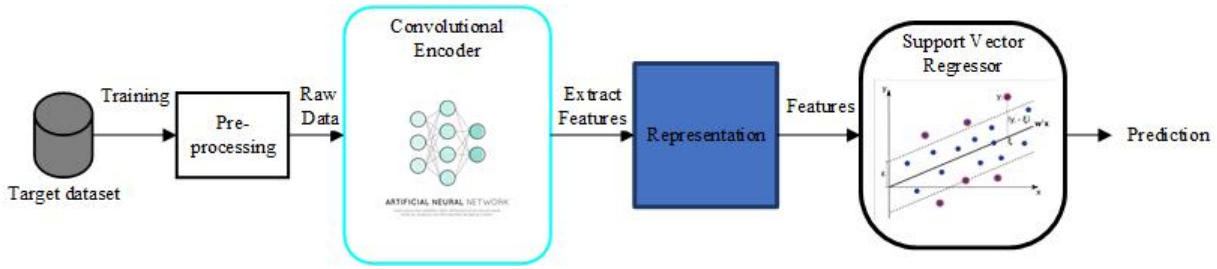

Fig. 2. An overview of the proposed architecture.

take a final decision. Liu et al. [27] presented a boosted DBN to combine feature learning/strengthen, feature selection and classifier construction in a unified framework. Features are fine-tuned and jointly selected to form a strong classifier that can learn highly complex features from facial images and more importantly, the discriminative capabilities of selected features are strengthened iteratively according to their relative importance to the strong classifier. Mollahosseini et al. [44] proposed a single component architecture made up of two convolutional layers each, followed by max pooling and four inception layers. The inception layers increase the depth and width of the network while keeping the computational budget constant.

So far, several approaches for FER have used CNNs with different architectures and different image preprocessing techniques. Mostly, they used supervised approaches that require the labeling of a large number of face images, which is expensive and time-consuming. This is a limitation because nowadays a lot of unlabeled data are created continuously. It is imperative that automatic FER deals with this case and take advantage of it. The proposed approach differs from the previous ones in a way that it has the ability to handle large datasets with the unsupervised approach, to learn the inherent relevant features without using explicitly provided labels and then predict emotional state with high accuracy.

III. Proposed Approach

In this section, we describe the overall architecture of the proposed approach, which is made up of three parts, as shown in Fig. 2. A key component of the proposed approach is the convolution operation and the autoencoder. Traditionally, most studies on FER are based on handcrafted features. However, after the recent success of CNNs in several classification tasks, many works on FER are now based on supervised approaches for representation learning using CNNs. In contrast to previous works in FER, the proposed approach starts with the supervised learning of a CNN on a source dataset for a classification task. The learned weights of the convolutional layers are reused in a CAE, which is trained on a target dataset in an unsupervised fashion. The meaningful representation learned by the CAE on the target dataset is used to train a regression model to predict continuous emotions.

In the first stage, we begin with training a CNN on a source dataset for a classification task. The idea is to use the transfer learning technique that allows us to import information from such a trained model to jump start the development process of the unsupervised approach on a new or similar task. The key concept is to use FER 2013 dataset, which has been used in ICMLW2013[1] [10] to recognize discrete emotions in pictures. This dataset provides a large number of facial images with emotional content to train a CNN. Once the CNN model is trained on the FER 2013 dataset, we use the convolutional layers (CLs) of such a pre-trained CNN to initialize the CLs of the CAE, as shown in Fig. 3.

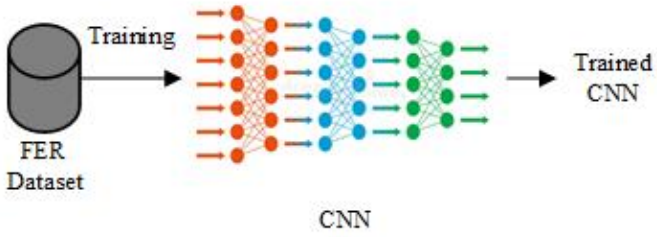

Fig. 3. Pre-training a CNN in a source dataset for a classification task

TABLE I
Pre-training the CNN architecture

Layers Type	Filter Dimension	Kernel Size	Activation
Conv2D	64	3×3	ReLU
BatchNormalization	-	-	-
Conv2D	64	3×3	tanh
Max Pool	-	2×2	-
BatchNormalization	-	-	-
Conv2D	128	2×2	ReLU
Max Pool	-	2×2	
Flatten	-	-	-
Fully Connected	100	-	tanh
Dropout(0.5)	-	-	-
Fully Connected	50	-	ReLU
Fully Connected	10	-	tanh
Fully Connected	7	-	softmax

The architecture of the proposed CNN is described in Table I. The architecture proposed by Sun et al. [15] which achieved 67.8% of accuracy on the test set of the FER 2013 dataset. In order to enhance the learned representation and

[1]30th Intl Conf on Machine Learning - Worksh on Representational Learning

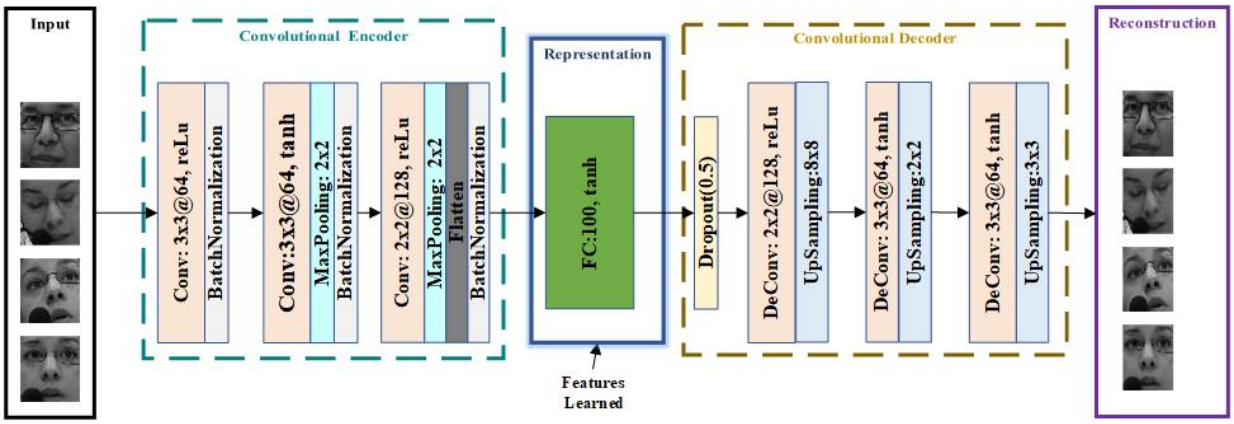

Fig. 4. Proposed Convolutional Autoencoder (CAE)

reduce the number of parameters of the network to achieved a best trade-off between complexity and the amount of data available for training, we reduced the number of CLs and fully connect layers (FC). The CNN model has three CLs and four FCs as shown in Table I. The first CL has 64 filters of size 3×3, ReLU activation followed by a batch normalization (BN) layer. The second CL takes as input the response-normalized feature maps of the first CL and filters it with 64 kernels of size 3×3, *tanh* activation followed by maxpooling operation to reduce the dimensionality and avoid overfitting. The third CL has 128 kernels of size 2×2 and ReLU activation. The first, second, third and fourth FCs have 100, 50, 10, and 7 neurons respectively. The CNN is trained up to 500 epochs using a categorical cross-entropy loss function with Adam optimizer.

In the next step, we have the unsupervised approach for representation learning. A CAE takes an face image as input and tries to reconstruct it back using a reduced number of units from the latent space representation. A CAE is made up of an encoder and a decoder part. The image is passed through the encoder, which is a sequence of CLs that produces at the encoded layer a low-dimensional representation of the face image called latent space representation. The decoder takes this latent space representation from the encoded layer and try to reconstruct the original image. The decoder part is made up of a sequence of transposed convolution (or deconvolution) layers which increase the resolution of the units of the latent space. The architecture of the proposed CAE is shown in Fig. 4. The convolutional encoder has three CLs and one fully connected layer which were "transferred" from the pre-trained CNN. The encoded layer is in fact a FC layer with a certain number of neurons. The decoder part has three transposed CLs. The first transposed CL filters the output of the encoded layer with 128 kernels of size 2×2, ReLU activation and an upsampling layer. The second CL takes as input the output of the first transposed CL and filters it with 64 kernels of size 3×3, *tanh* activation followed by an upsampling layer. Finally, the third CL has 64 kernels of size 3×3, ReLU activation connected to the outputs of the second CL.

In the final step, we have the supervised approach for the regression task that uses as input features the representation learned at the encoded layer of the autoencoder. An SVR is trained with the features generated by the CAE with the aim of predicting continuous emotions. We follow the same strategy that has been used in AVEC 2016[2] [21] to predict arousal and the valence for each video frame. We use grid search to find the best combination of the complexity parameter C and epsilon ϵ of the SVR that maximizes a performance measure.

A. Post-Processing

As the proposed architecture does not take into consideration the temporal correlation between adjacent frames, this may affect the predictions. To circumvent this problem, we post-process the predictions with a median filter, scaling, centering, and delay compensation which together allow us to improve the performance.

Median filtering smooths our predictions by reducing the high-frequency components by filtering the 1D output array with a window of size between 0.04 and 20 seconds. The scaling factor β_{tr} is the ratio of the gold standard (GS_{tr}) and the prediction (Pr_{tr}) as shown in (1) over the training set. The prediction on the development set is multiplied by the factor β_{tr} with the purpose of rescaling the predictions as shown in (2).

$$\beta_{tr} = \frac{GS_{tr}}{Pr_{tr}} \tag{1}$$

$$Pr'_{dev} = \beta_{tr} \cdot Pr_{dev} \tag{2}$$

The centering technique entails just subtracting the predictions (y) by the mean of gold standard predictions $(\bar{y}_{GS})$ as shown in (3), where y' is the corrected value.

$$y' = y - \bar{y}_{GS} \tag{3}$$

In the annotation process of facial expression, the annotator needs to sense the stimulus, perceive the emotional message and make a decision in real time. Then, we note a reaction lag

[2]Audio-Visual + Emotion Recognition Challenge

between the annotation and the underlying emotional content. Therefore, the ratings make by this annotator considering each dimension may not be reliable and match with the reality. [23] argues that the delay varies between different raters and can range anywhere between 2-10 seconds. To deal with it we use the delay compensation of annotation. The delay compensation is basically achieved by shifting the input features relative to the ground truth labels during the training and testing.

IV. Datasets

In this section we present the source dataset used for training the CNN for the classification task and the target dataset used for representation learning and the regression task. Furthermore, we also present the preprocessing techniques that have been used to detect and align face images within video frames of the target dataset.

A. FER 2013 Dataset

The FER 2013 dataset has been created by Carrier & Courville and it is publicly available[3]. We used the FER 2013 dataset to train a CNN for the task of classifying facial expressions, as the starting point in the pre-training stage for the proposed model shown in Fig. I. The FER 2013 dataset is made up of grayscale images of 48×48 pixels that comprise six acted emotions (disgust, anger, fear, joy, saddens, surprise) and neutral and it is split into 28,709 images for training, 3,589 for validation and 3,589 for test.

B. RECOLA Dataset

To train and evaluate the proposed CAE architecture and the SVR, we use the RE-mote COLlaborative and Affective (RECOLA) dataset introduced by Ringeval et al. [18] to study socio-affective behaviors from multimodal data in the context of remote collaborative work for the development of computer-mediated communication tools [19]. However in this study we use the same subset of the RECOLA dataset that was used in the Audio/Visual Emotion Challenge and Worksh (AVEC) 2015 and 2016 challenges [17], [21] as we do not have access to the full dataset. This subset contains four modalities that are audio, video, electrocardiogram (ECG) and electro-dermal activity (EDA). The dataset is split equally in three partitions, training (9 subjects), validation (9 subjects) and test (9 subjects) by stratifying (i.e., balancing) the gender and the age of the subjects. The labels (valence and arousal) are re-sampled at a constant frame rate of 40 ms. In addition, we do not have the labels for the test set. We ensure that no validation data is used for unsupervised feature learning. The CAE has been trained with all unlabeled video data.

C. Face Detection and Alignment

On FER several obstacles appear in our path to achieve a suitable prediction. One of them being the fact that humans as unpredictable entities are in a constant movement even in a face to face conversation and because of this, sometimes, the subject does not look directly into the camera. Other issues arise like a delay on the annotated labels that is imposed by the annotator and the absence of bounding box coordinates for several frames. Therefore, we evaluate different strategies to avoid these problems and end up with good quality facial images from developing our approach.

We started with the dropping frames issue. Over the RECOLA dataset, a certain number of frames do not have the bounding box coordinates to extract the face. Even with our own-implemented face-detector we cannot extract the section over the image that contains the subject's face for all the frames. Because of this, we tried to preserve all the dataset by using the entire image without the bounding box. However, the frame quality selection to filter detected face is far from the frontal image and the delay compensation is then used to realign labels and frames in order to compensate the late reaction of annotators. By doing so, we reduced the dataset size, which is actually not a good option for our approach, because the unsupervised algorithms perform well with a large amount of data. Thus, instead of dropping the blank frames or frames where faces are not well detected, we decided to replace them by other frames that were well detected. It is a kind of data augmentation strategy whereby the frames without the bounding box have been slightly changed to detect the face of participants.

V. Experiments and Results

This section presents the metrics used and the experiments undertaken to evaluate the proposed approach. The experimental results are analyzed and compared to previous works.

A. Evaluation Metrics

Concordance correlation coefficient (CCC) [16] is calculated as the evaluation metric for the AVEC challenges [17], [21]. It combines Pearson's correlation coefficient (PCC) with the square difference between the mean of the two compared time-series as denoted in (4).

$$\rho_c = \frac{2\rho\sigma_x\sigma_y}{\sigma_x^2 + \sigma_y^2 + (\mu_x - \mu_y)^2} \tag{4}$$

where ρ is the PCC between two time-series (e. g., prediction and gold standard), σ_x^2 and σ_y^2 are the variance of each time-series and μ_x and μ are the mean value of each time-series. As a result, predictions that are well correlated with the gold standard but shifted in value are penalised in proportion to the deviation.

CCC will help to evaluate emotion recognition in terms of continuous time and continuous valued dimensional affect into two dimensions: arousal and valence [17]. The problem of dimensional emotion recognition can thus be posed as a regression problem through these two dimensions.

B. Experimental Setup

For raw signal, we cropped faces of the subject's video to have the images with the size 48×48. The image size 48×48 is used to reduce the computation complexity and

[3]https://www.kaggle.com/c/challenges-in-representation-learning-facial-expression-recognition-challenge/

because the pre-trained model used the FER 2013 dataset, which consists of 48×48 pixel grayscale images of faces. To train the proposed model, we initialized the network with the pre-trained weights from the FER dataset. We used the Adam optimization method and the mean square error (MSE) as loss function with a fixed learning rate of 10^{-5} throughout all experiments. We evaluated different batch sizes, learning rates and epochs in order to determine the best setup for training the CNN model.

The CAE uses the same CL of the pre-trained CNN model. On the other hand, all FC layers are replaced by a encoded layer, which is in fact represented as a FC layer. We evaluated different dimensions for the encoded layer of the CAE, as shown in Table III. For regularization of the network, we also used dropout with p=0.25 for the encoded layer. This step is important as our models have a large number of parameters and not regularizing the network makes it prone to overfitting on the training data.

We have carried out different experiments by freezing the CLs and fine-tune (training) just the encoded layer as shown in Table II. This means that the weights of the CNN learned in the classification task do not change. Alternately, we unfreeze progressively the CLs from the deeper to the first CL and retrain the network with the RECOLA dataset in an unsupervised fashion.

TABLE II
CCC SCORES FOR AROUSAL AND VALENCE DIMENSIONS BY FREEZING DIFFERENT NUMBER OF CLS.

	CCC		
Dimension	0 CL frozen	2 CL frozen	1 CL frozen
Valence	0.397	0.399	**0.516**
Arousal	0.027	0.035	**0.264**

We noticed that unfreezing only the deepest CL gives the best result. By the way, when all CLs are frozen, the model did not train properly, meaning that the value of the loss function did not decrease significantly. In the end, a chain of post-processing methods is applied, namely, median filtering (size of window was between 0.04s and 20s) [20], centering, scaling and delay compensation as described in Section III-A. Any of these post-processing techniques were kept when we have observed an improvement in the CCC.

C. Experimental Results

Table III shows the CCC scores achieved by the proposed approach when the SVR is trained on the representation learned by the CAE, for different dimensions of the encoder layer and delay compensation. We can see that when the encoder layer has a low dimension, the CCC score is very low as well. However, the CCC scores increase as the encoder layer dimension increases, and the best CCCs are achieved for a 900-dimensional encoded layer for both arousal and valence dimensions. This is due to the fact that by increasing the dimension of the encoder layer the CAE is able to learn more relevant representations. Nonetheless, Table III also shows that beyond 1,000 units, the CCC scores do not increase. This can be explained by the fact that the CAE does not find novel relevant features and this behavior is related to the size of the training set. By increasing the number of training samples, the CAE will probably continue to capture the more relevant features.

TABLE III
CCC SCORES FOR AN SVR TRAINED ON FEATURES TAKEN FROM THE ENCODED LAYER OF THE CAE. THE CCC SCORES FOR AROUSAL AND VALENCE DIMENSIONS CONSIDER A DELAY COMPENSATION OF 40 OR 30 FRAMES.

Dimension	**Encoded Layer Dimension**	**Delay Compensation**	**CCC**
Valence	100	40	0.197
Valence	500	40	0.324
Valence	700	40	0.361
Valence	900	40	**0.516**
Valence	1000	40	0.365
Valence	100	30	0.195
Valence	500	30	0.384
Valence	700	30	0.392
Valence	900	30	0.498
Valence	1000	30	0.395
Arousal	100	40	0.018
Arousal	500	40	0.071
Arousal	700	40	0.151
Arousal	900	40	**0.264**
Arousal	1000	40	0.119
Arousal	100	30	0.031
Arousal	500	30	0.092
Arousal	700	30	0.162
Arousal	900	30	0.257
Arousal	1000	30	0.114

We compare the performance achieved by our approach against the current state-of-the-art for the RECOLA dataset. Table IV shows the results for the valence dimension that predicts how positive or negative the emotion is and for the arousal dimension that shows how the excitation is. Most of these results have been submitted to the AVEC2016 challenge which used a subset of RECOLA dataset encompassing only 27 participants. We observed that the results obtained by Tzirakis et al. [24] are slightly higher than the proposed approach because they employed the RECOLA dataset, which has data from 46 participants and they have also modeled the temporal correlation between adjacent frames using an LSTM. The prediction of the valence dimension achieved by our approach outperforms Han et al. [25] even though the prediction of the arousal dimension remains the same. In comparison with AVEC2016 Baseline [21], the prediction of the valence dimension (unpleasant to pleasant) of the proposed approach outperforms the appearance features and it is slightly higher than the geometric features. On the other hand, the CCC for the arousal dimension (the degree of excitement) achieved by the proposed approach is slightly lower than those achieved with appearance and geometric features.

VI. CONCLUSION

In this paper we have proposed a novel approach for continuous emotion recognition based on convolutional autoencoder (CAE) and support vector regressor (SVR). In the first step the

TABLE IV
PERFORMANCE COMPARISON BETWEEN THE PROPOSED APPROACH (CAE+SVR) AND OTHER STATE-OF-THE-ART APPROACHES.

Approach	Features	Valence	Arousal
Tzirakis et al. [24]	Raw signal	0.620	0.435
AVEC 2016 Baseline [21]	Appearance	0.486	0.343
AVEC 2016 Baseline [21]	Geometric	0.507	0.272
Han et al. [25]	Mixed	0.265	0.394
Ortega et al. [26]	Deep	0.433	0.252
Proposed Approach	Raw signal	**0.516**	**0.264**

a meaningful representation is learned from a close-related source dataset but for a classification task. Such a learned representation is used to initialize the CAE, which is trained on a target dataset in an unsupervised fashion. Finally, the representation learned by the CAE on the target dataset is used to train a regression model based on SVR.

The proposed approach produces a compact but meaningful representation that outperformed several state-of-the-art approaches on the AVEC 20016 dataset. As a future work we plan to evolve the proposed approach by capturing the temporal correlation between adjacent frames as such an information seems very useful in continuous emotion prediction.

REFERENCES

[1] D. Grandjean, D. Sander, and K. R. Scherer, " Conscious emotional experience emerges as a function of multilevel, appraisal-driven response synchronization," Consciousness and Cognition, vol. 17, pp.484-495, 2008.

[2] S. Marsella and J. Gratch, "Computationally modeling human emotion," Commun. ACM, vol. 57, no. 12, pp. 56–67, 2014.

[3] H. Gunes, B. Schuller, M. Pantic, and R. Cowie, "Emotion representation,analysis and synthesis in continuous space: A survey," in IEEE Intl Conf and Worksh on Automatic Face and Gesture Recognition, pp. 827–834, 2011.

[4] M. A. Nicolaou, H. Gunes, and M. Pantic, "Continuous prediction of spontaneous affect from multiple cues and modalities in valence-arousal space," IEEE Trans. Affective Comput., vol. 2, no. 2, pp. 92–105, Apr./Jun. 2011.

[5] C. Shan , S. Gong, and P. W. McOwan, "Facial expression recognition based on local binary patterns: A comprehensive study," Image and Vision Computing, vol. 27, pp. 803-816, 2009.

[6] S. Berretti, B. B. Amor, M. Daoudi , and A. D. Bimbo, " 3D facial expression recognition using SIFT descriptors of automatically detected keypoints," The Visual Computer, vol. 27, pp. 1021-1036, 2011.

[7] A. Albiol, D., Monzo, A., Martin, J., Sastre, and A. Albiol," Face recognition using HOG? EBGM," Pattern Recognition Letters, vol. 29, pp. 1537-1543, 2008.

[8] X. Zhao , X. Shi , and S. Zhang,"Facial Expression Recognition via Deep Learning," IETE Technical Review, vol. 32, pp.347-355,2015.

[9] A. Mostafa, M. I. Khalil, and H. Abbas,"Emotion Recognition by Facial Features using Recurrent Neural Networks,"in Intl Conf on Computer Engineering and Systems (ICCES), pp.417-422, 2018.

[10] Y. Tang ,"Deep learning using linear support vector machines," in ICML 2013 Challenges in Representation Learning Worksh, pp. 1-6, 2013.

[11] M. Liu, S. Li, S. Shan, R. Wang, and X. Chen," Deeply learning deformable facial action parts model for dynamic expression analysis," in Computer Vision–ACCV 2014, pp. 143–157. Springer, 2014.

[12] M. Liu, R. Wang, S. Li, S. Shan, Z. Huang, and X. Chen," Combining multiple kernel methods on riemannian manifold for emotion recognition in the wild," in 16th Intl Conf on Multimodal Interaction, pp. 494–501. ACM, 2014.

[13] Z. Yu, C. Zhang,"Image Based Static Facial Expression Recognition with Multiple Deep Network Learning,"Association for Computing Machinery, pp. 435-442, 2015.

[14] S. E. Kahou, C. Pal, X. Bouthillier, P. Froumenty, Ç . Gülçehre, R. Memisevic, P. Vincent, A. Courville, Y. Bengio, R. C. Ferrari, et al., "Combining modality specific deep neural networks for emotion recognition in video," in 15th ACM Intl Conf on multimodal interaction, pp. 543–550. ACM, 2013.

[15] B. Sun, C. Siming, L. Liandong, J. He, L. Yu, "Exploring Multimodal Visual Features for Continuous Affect Recognition," in 6th Intl Worksh on Audio/Visual Emotion Challenge, pp. 83-88, ACM, 2016.

[16] I. Lawrence K. Lin, "A Concordance Correlation Coefficient to Evaluate Reproducibility," Biometrics, vol. 45, pp. 255–268, JSTOR, 1989.

[17] F. Ringeval, B. Schuller, M. Valstar, S. Jaiswal, E. Marchi, D. Lalanne, R. Cowie, M. Pantic, "AV+EC 2015: The First Affect Recognition Challenge Bridging Across Audio, Video, and Physiological Data,"in 5th Intl Works on Audio/Visual Emotion Challenge, pp. 3-8, ACM, 2015.

[18] F. Ringeval, A. Sonderegger, J. Sauer, D. Lalanne, "Introducing the RECOLA multimodal corpus of remote collaborative and affective interactions,"in 10th IEEE Intl Conf and Works on Automatic Face and Gesture Recognition (FG), pp. 1-8, IEEE, 2013.

[19] F. Ringeval, A. Sonderegger, B. Noris, A. Billard, J. Sauer, D. Lalanne, "Humaine Association Conf on Affective Computing and Intelligent Interaction,"pp. 448-453, 2013.

[20] F. Ringeval, B. Schuller, M. Valstar, R. Cowie, M. Pantic, "in AVEC 2015: 5th Intl Audio/Visual Emotion Challenge and Works, pp. 1335-1336, 2015

[21] M. Valstar, J, Gratch, B. Schuller, F. Ringeval, D. Lalanne, M. Torres Torres, S. Scherer, G. Stratou, R. Cowie, M. Pantic, "AVEC 2016: Depression, Mood, and Emotion Recognition Works and Challenge," in 6th Intl Works on Audio/Visual Emotion Challenge, pp. 3-10, ACM, 2016.

[22] M. Kächele, P. Thiam, G. Palm, F. Schwenker and M. Schels, "Ensemble Methods for Continuous Affect Recognition: Multi-Modality, Temporality, and Challenges," in 5th Intl Works on Audio/Visual Emotion Challenge, pp. 9-16, ACM, 2015.

[23] S. Mariooryad and C. Busso, "Correcting timecontinuous emotional labels by modeling the reaction lag of evaluators," IEEE Transac on Affective Computing, vol. 6, no. 2, pp. 97–108, 2015.

[24] P. Tzirakis, G. Trigeorgis, M. A. Nicolaou, B. W. Schuller, and S. Zafeiriou, "End-to-end multimodal emotion recognition using deep neural networks," IEEE Journal of Selected Topics in Signal Processing, vol. 11, pp. 1301–1309, 2017.

[25] J. Han, Z. Zhang, N. Cummins, F. Ringeval, B. Schuller, "Strength Modelling for Real-World Automatic Continuous Affect Recognition from Audiovisual Signals,"Elsevier, Image and Vision Computing, vol. 65, pp. 76-86, 2017.

[26] J. D. S. Ortega, P. Cardinal, A. L. Koerich, "Emotion Recognition Using Fusion of Audio and Video Features," in IEEE Intl Conf on Systems, Man, and Cybernetics (SMC), pp. 3827-3832, 2019.

[27] P. Liu, S. Han, Z. Meng, and Y. Tong, "Facial expression recognition via a boosted deep belief network," in IEEE Conf on Computer Vision and Pattern Recognition, pp. 1805–1812, 2014.

[28] M. S. Bartlett, G. Littlewort, M. G. Frank, C. Lainscsek, I. Fasel, and J. R. Movellan, " Recognizing facial expression: Machine learning and application to spontaneous behavior," in IEEE Conf on Computer Vision and Pattern Recognition, vol. 2, pp. 568–573, 2005.

[29] T. H. H. Zavaschi, L. E. S. Oliveira, A. L. Koerich, "Facial expression recognition using ensemble of classifiers," in IEEE Intl Conf on Acoustics, Speech and Signal Processing, pp. 1489–1492, 2011.

[30] Y. Tian, T. Kanade, and J. F. Cohn, " Evaluation of Gabor-waveletbased facial action unit recognition in image sequences of increasing complexity," In FG, pp. 229–234, May 2002.

[31] J. Whitehill, M. S. Bartlett, G. Littlewort, I. Fasel, and J. R. Movellan, "Towards practical smile detection," IEEE TPAMI, vol. 31, pp. 2106–2111, Nov. 2009.

[32] Y. Zhang and Q. Ji, " Active and dynamic information fusion for facial expression understanding from image sequences," IEEE TPAMI, vol. 27, pp.699–714, May 2005.

[33] Z. Zhang, M. Lyons, M. Schuster, and S. Akamatsu, "Comparison between geometry-based and Gabor-wavelets-based facial expression recognition using multi-layer perceptron," In FG, pp. 454–459, 1998.

[34] P. Yang, Q. Liu, and D. N. Metaxas, "Boosting coded dynamic features for facial action units and facial expression recognition," in IEEE Conf on Computer Vision and Pattern Recognition, pp. 1–6, June 2007.

[35] G. Zhao and M. Pietiainen. Dynamic texture recognition using local , "Binary patterns with an application to facial expressions," IEEE TPAMI, vol. 29, pp. 915–928, June 2007.

[36] M. F. Valstar, M. Mehu, B. Jiang, M. Pantic, and K. Scherer, " Metaanalyis of the first facial expression recognition challenge," IEEE T-SMC-B, vol. 42, pp. 966–979, 2012.

[37] T. H. H. Zavaschi, A. S. Britto Jr, L. E. S. Oliveira, and A. L. Koerich, "Fusion of feature sets and classifiers for facial expression recognition," Expert Systems with Applications, vol. 40, no. 2, pp. 646-655, 2013.

[38] T. Senechal, V. Rapp, H. Salam, R. Seguier, K. Bailly, and L. Prevost, " Combining AAM coefficients with LGBP histograms in the multi-kernel SVM framework to detect facial action units," in IEEE Intl Conf and Works on Automatic Face and Gesture Recognition, pp. 860–865, 2011.

[39] S. Jain, C. Hu, and J. K. Aggarwal, " Facial expression recognition with temporal modeling of shapes," In ICCV Works, pp. 1642–1649, 2011.

[40] Y. Hu, Z. Zeng, L. Yin, X. Wei, X. Zhou, and T. S. Huang, " Multiview facial expression recognition," in IEEE Intl Conf and Worksh on Automatic Face and Gesture Recognition, pp. 1–6, 2008.

[41] M. Mansano, A. S. Britto Jr., L. E. S. Oliveira, and A. L. Koerich, "2D Principal Component Analysis for Face and Facial-Expression Recognition," IEEE Computing in Science & Engineering, vol. 13, no. 3, pp. 9-13, 2008.

[42] A. L. Koerich, L. E. S. Oliveira, and A. S. Britto Jr., "Face recognition using selected 2DPCA coefficients," in 17th Intl Conf on Systems, Signals and Image Processing (IWSSIP), 2010.

[43] A. Lorincz, L. A. Jeni, Z. Szabó, J. F. Cohn, and T. Kanade, " Emotional expression classification using time-series kernels," in IEEE Conf on Computer Vision and Pattern Recognition Workss, pp. 889–895, IEEE, 2013.

[44] A. Mollahosseini, D. Chan, and M. H. Mahoor, "Going deeper in facial expression recognition using deep neural networks," in IEEE Winter Conf on Applications of Computer Vision (WACV), pp. 1–10, 2016.

[45] Y. Yi-Hsuan, C. H. H, "Machine Recognition of Music Emotion: A Review," Association for Computing Machinery, vol. 3, 2012.