Emotion Recognition Using Fusion of Audio and Video Features

Juan D. S. Ortega, Patrick Cardinal and Alessandro L. Koerich

Abstract— **In this paper we propose a fusion approach to continuous emotion recognition that combines visual and auditory modalities in their representation spaces to predict the arousal and valence levels. The proposed approach employs a pre-trained convolution neural network and transfer learning to extract features from video frames that capture the emotional content. For the auditory content, a minimalistic set of parameters such as prosodic, excitation, vocal tract, and spectral descriptors are used as features. The fusion of these two modalities is carried out at a feature level, before training a single support vector regressor (SVR) or at a prediction level, after training one SVR for each modality. The proposed approach also includes preprocessing and post-processing techniques which contribute favorably to improving the concordance correlation coefficient (CCC). Experimental results for predicting spontaneous and natural emotions on the RECOLA dataset have shown that the proposed approach takes advantage of the complementary information of visual and auditory modalities and provides CCC of 0.749 and 0.565 for arousal and valence, respectively. The proposed approach outperforms the baseline system and several traditional approaches based on auditory and visual handcrafted features.**

I. Introduction

Understand the emotional state of people is a phenomenon that has attracted the attention and also fascinate researchers from different branches of science for a while. Psychologists, psychiatrists, neuroscientists, and computer scientists constantly try to untangle the combination of variables that best describes an emotional state. The circumplex model, proposed by Russell [12] generates a circular representation of the emotional space. This model suggests that there are two independent neurophysiological systems and their combination generates a representation of an emotional state. This model is represented by coordinated axis of two dimensions, where the vertical axis represents the alertness/activeness level (arousal) in the emotion and the horizontal axis represents the pleasure/displeasure level (valence). Both dimensions can be normalized in the range between -1 and +1. A vast mixture of emotions can be described with the linear combination of arousal and valence.

In the last years, several works attempt to predict valence and arousal using machine learning algorithms [5], [11], [19], [1]. Eyben *et al.* [5] proposed a fully automatic audiovisual recognition approach based on Long Short-Term Memory (LSTM) modeling of word-level audio and video features. Evaluations carried out on the SEMAINE dataset have shown how acoustic, linguistic, and visual features contribute to the recognition of different affective dimensions. Ringeval *et al.* [11] studied the relevance of machine learning algorithms to integrate contextual information to automatically predict emotion from several raters in continuous time domains. Evaluations carried out on the RECOLA dataset have achieved concordance correlation coefficient (CCC) of 0.804 and 0.528 for arousal and valence reespectively. Weber *et al.* [19] improved the performance of multimodal prediction with low-level features by adding high-level geometry-based features as well as by fusing the unimodal predictions trained on each training subject before performing the multimodal fusion. The high-level features have improved the performance of the multimodal prediction of arousal and the subject's fusion generalized poorly in multimodal prediction, particularly on valence. Ding *et al.* [3] presented the fusion of facial expression recognition and audio emotion recognition at score level. For facial emotion recognition, a pre-trained deep Convolutional Neural Network (CNN) fine-tuned on FER2013 dataset for feature extraction and kernel SVM, logistic regression and partial least squares are used as classifiers. For audio emotion recognition, a deep LSTM Recurrent Neural Network (LSTM-RNN) is trained directly on the Emotion Recognition in the Wild 2016 challenge dataset. Experimental results have shown that the proposed approach achieved state-of-the-art performance with an overall accuracy of 53.9% on the test dataset. Yan *et al.* [21] presented a transductive deep transfer learning architecture based on VGGface16 CNN which is used to jointly learn optimal nonlinear discriminative features from non-frontal facial expressions. Cross-dataset experiments on BU-3DEF and Multi-PIE datasets have shown that the proposed approach outperforms the state-of-the-art cross-database facial expression recognition methods. Yan *et al.* [20] propose a multi-cue fusion emotion recognition framework by modeling human emotions from three complementary cues: facial texture, facial landmark action and audio signal, and then fusing them together. Models are fused at both feature level and decision level and the experimental results on AFEW and CHEAVD datasets have shown the effectiveness of the proposed approach. Tzirakis *et al.* [17] proposed an emotion recognition system using auditory and visual modalities. They utilized a pre-trained CNN to extract features from the speech and a deep residual network to extract features from the visual modality.

In this paper we propose an approach based on transfer learning and multimodal fusion for the problem of continuous emotion recognition in video. Facial features are

Juan David Silva Ortega, Patrick Cardinal and Alessandro Lameiras Koerich are with Department of Software and IT Engineering, École de Technologie Supérieure, University of Québec, H3C 1K9, Montréal, QC, Canada. juan-david.silva-ortega.1@ens.etsmtl.ca, patrick.cardinal@etsmtl.ca, alessandro.koerich@etsmtl.ca

978-1-7281-4569-3/19/$31.00 ©2019 IEEE

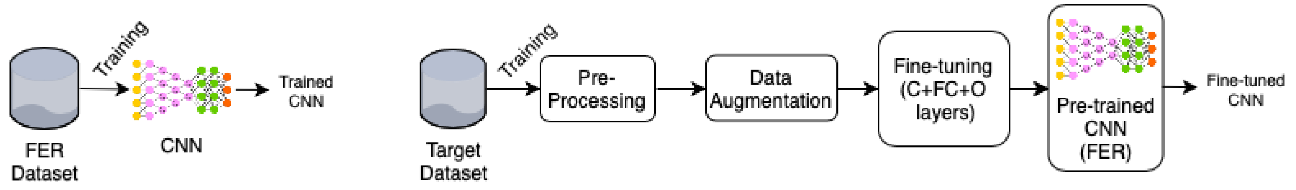


Fig. 1: (a) Pre-training a CNN in a source dataset; (b) Fine-tuning the CNN on the target dataset.

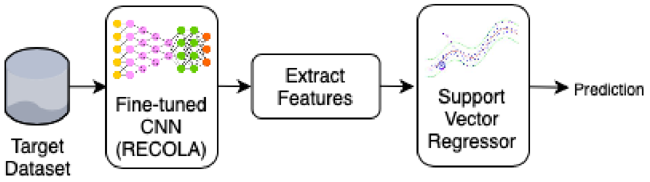


Fig. 2: Using the fine-tuned CNN to extract features and train an SVR.

extracted from video frames by a pre-trained CNN and such features are fused with handcrafted audio features extracted from the subjects' voice. We evaluate two fusion schemes, early and late as well as the several pre- and post-processing techniques to improve the prediction of arousal and valence levels. The main contributions of this paper are: (i) a compact representation learned from raw data that is comparable to the state-of-the-art; (ii) a multimodal fusion approach of a learned representation and handcrafted features that yields to a compact representation and that achieved CCCs that are comparable to the state-of-the-art as well.

This paper is organized as follows. Section II presents the proposed approach for dealing with raw video, the proposed fusion schemes and the pre- and post-processing techniques. Section III presents the experimental results for predicting valence and arousal on RECOLA dataset. Finally, in the last section we present our conclusions and perspectives of future work.

II. Proposed Approach

Fig. 1 shows a general structure of the proposed approach to predict arousal and valence levels. We start with transfer learning, a technique that allows us to import information from another model that was trained on a similar task. The idea is to use a source dataset (FER) which provides a large number of face images with emotional content to train a CNN. Transfer learning allows us to use such a pre-trained CNN in our target task instead of training a CNN initialized with random parameters. The target dataset (RECOLA) is then used to fine-tune some parameters (layers) of such a CNN. The target dataset provides video data, audio, and physiological signals as well as handcrafted features extracted from these modalities. Video frames that allow us to build a regressor to predict levels of arousal and valence of a subject. Because of the nature of the datasets and the strong relation between the tasks (emotional prediction or classification), we can say empirically that even if we train two network models, one for each dataset, the feature space should be correlated.

However, before fine-tuning the CNN on the target dataset, we propose to use certain preprocessing techniques to improve the quality of the training data, such as: (i) frame suppression to discard video frames where a face is not detected; (ii) frame quality selection to filter detected face are far from being frontal face images; (iii) delay compensation to realign labels and frames to compensate the reaction lag of annotators [9]. Finally, as the amount of training data is critical to train a CNN, given its high number of trainable parameters, a data augmentation strategy is used to address this issue.

The proposed CNN shown in Fig. 3, is based on the architecture presented by Sun *et al.* [14], which achieved 67.8% of accuracy on the test set of FER dataset. We have introduced some modifications on such an architecture in an attempt to improving the learned representation. We have changed the number of convolutional layers and fully-connect layers, to reduce the number of parameters of the network and have a best trade-off between the complexity and the amount of data available for training. The proposed architectures are shown in Table I.

Before fine-tuning these architectures for our target task, which is the regression problem of predicting continuous emotions, we have replaced the output layer by a single neuron to obtain a real value at the output. For the fine-tuning process, we started by freezing (making them non-trainable) all convolutional layers and fine-tuning only the fully connected layers. Then, we have progressively unfrozen the convolutional layers to evaluate if the network could improve the representation learned on the FER dataset to the target dataset (RECOLA). Besides fine-tuning the CNNs, we have also trained CNNs end-to-end, which means that they were initialized randomly. Finally, as illustrated in Fig. 2, we have used such CNNs to generate feature vectors from video frames by taking the representation provided by the fully connected layers and use them as the input of another learning algorithm, or to fuse such feature vectors with feature vectors of another modality as a complementary representation of the emotional state.

For the multimodal fusion, we propose to use the extended Geneva Minimalistic Acoustic Parameter Set (eGeMAPS), which are low level descriptors that cover spectral, cepstral, prosodic and voice quality information of the voice record. Such features have been used in the RECOLA baseline [18] together with other modalities. Table II shows the average

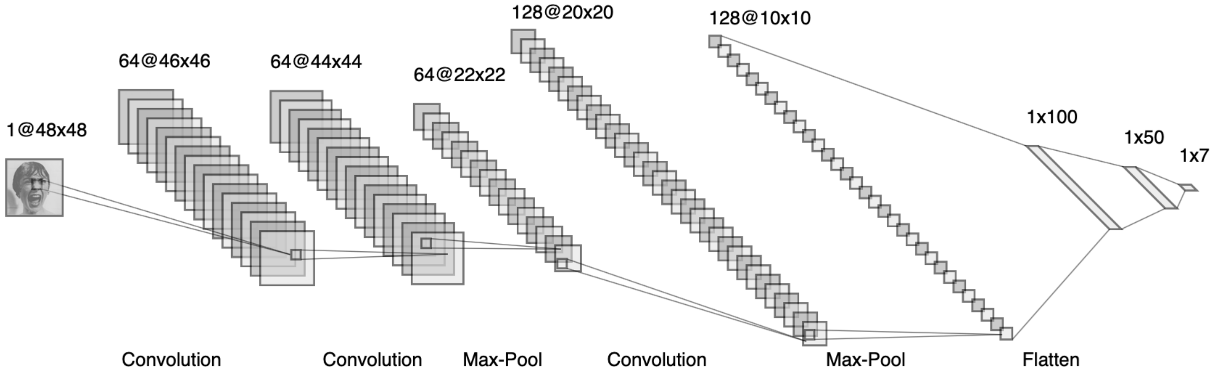


Fig. 3: The CNN architecture for the FER dataset with three convolutional layers (CL) and two full connected (FC) layers.

TABLE I: Three CNN architectures based on [14]: (A) With less CL and FC layers; (B) With a reduced number of CL and FC layers; (C) With more CLs.

Layers	**Architecture**		
	A	B	C
Zero-Padding(2D)	1×1		1×1
Convolution	64×3×3	64×3×3	256×3×3
Zero-Padding(2D)			1×1
Max Pooling	2×2		
Zero-Padding(2D)	1×1		
Convolution	128×3×3	64×3×3	256×3×3
Max Pooling	2×2	2×2	
Zero-Padding(2D)	1×1		1×1
Convolution	256×3×3	128×3×3	256×3×3
Max Pooling	2×2	2×2	2×2
Zero-Padding(2D)	1×1		1×1
Convolution	256×3×3		256×3×3
Zero-Padding(2D)			1×1
Convolution			256×3×3
Zero-Padding(2D)			1×1
Convolution			256×3×3
Max Pooling	2×2		2×2
Fully Connected	1024	100	1024
Fully Connected		50	1024
Fully Connected	7	7	7
# of Parameters	3.3M	1.4M	41.7M

contribution to the final prediction of video appearance (V_{App}) and geometric (V_{Geo}), electrocardiogram (ECG), heart rate and heart rate variability (HR), electrodermal activity (EDA), skin conductance level (SCL) and resistance (SCR). The average audio contribution is 0.320 which is the highest among all modalities. Therefore, we choose such a modality to build our multimodal approach.

TABLE II: RECOLA baseline: average valence-arousal contribution of each modality in the final prediction.

Audio	**Video**		**ECG**	**HR**	**EDA**	**SCL**	**SCR**
	App	**Geo**					
0.320	0.085	0.170	0.03	0.115	0.035	0.075	0.170

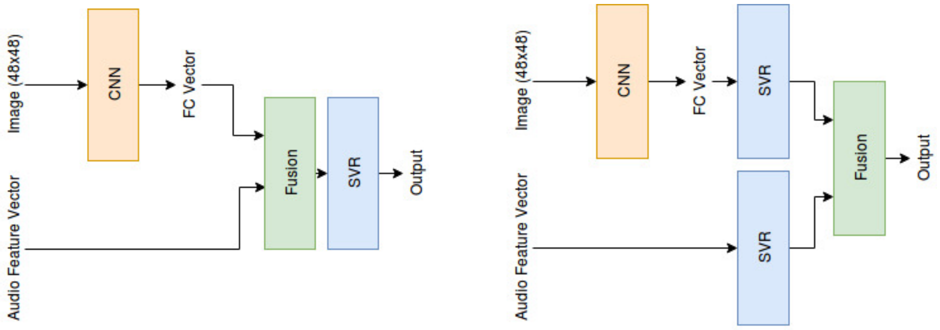
Considering that we use CNNs as feature extractors, we must define how to use such features to make predictions. Valstar *et al.* [18] show how the features of the RECOLA baseline perform well with an SVR algorithm. Therefore, we also use an SVR and compare its performance with that of the CNN. Also, tuning an SVR in the fusion approach is less complex than tuning a CNN that takes into account multiple modalities, because an SRV requires adjusting only two hyperparameters instead of a large number of weights. The grid search strategy can find a suitable combination of C and ϵ given a range of values for both hyperparameters. An SVR algorithm can either predicts the arousal and valence levels by using single or multiple data modalities. Therefore, different fusion techniques can be used for such an aim.

A. Fusion

We have evaluated the early and the late schemes for fusing video and audio features. Early fusion, as shown in Fig. 4a, stacks all representations as a single multimodal vector and train a single SVR. Therefore, the assumption is that the model will learn from the richest source of information [13], [22], [15]. Late fusion, as illustrated in Fig. 4b uses unimodal vectors to train different SVRs and the predictions provided by each regressor (one by modality) and further fused by averaging them into one single prediction. Late fusion is more difficult to tune because we must train and find the hyper-parameters of at least two models. On the other hand, late fusion is more flexible than early fusion because each model has its own representation space and its own hyper-parameters.

B. Post-Processing

The propose approach does not takes into account the correlation that exists between consecutive frames. Therefore, the predictions of our model may have a noisy behavior. To mitigate such a problem, post-processing techniques such as median filtering, scaling factor, and centering [16] allow the enhancement of the predictions of the model. Median filtering smooths our predictions by filtering the 1D output array with a window ranging between 0.4 and 8 seconds. The scaling factor β can be calculated by Eq. 1 as the ratio of the gold standard (GS) and the prediction (Pr) over the training set. The prediction on the development set (Eq. 2) is simply multiplied by this factor with the aim of rescaling the output.



(a) (b)

Fig. 4: (a) Early fusion: video and audio features are concatenated and used to train an SVR; (b) Late fusion: video and audio features are used to train two SVRs and the predictions of both are fused to generate an output.

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

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

Finally, for centering the prediction distribution and align the ranges between it, we subtract the mean of the gold standard $\bar{y}_{GS}$ by the prediction y:

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

where y' is the corrected prediction value.

III. Experimental Results

The target dataset used to evaluate the proposed approach is RECOLA, which is a multimodal dataset composed of 9.5 hours of audio, visual, and physiological recordings captured from 46 French-speaking participants. However, only the data of 18 speakers is publicly available from which nine speakers compose the training set and another nine speakers compose the developing set. We only used the audio and the visual modalities, because we aim a non-invasive approach for emotion detection.

Two metrics are used to evaluate the behavior of the different CNN architectures on the target dataset. The Mean Absolute Error (MAE), which offers equity over all the absolute differences between golden standard and predictions and is defined as:

$$\text{MAE} = \frac{1}{n}\sum_{j=1}^{n} |y_j - y_{GSj}| \tag{4}$$

where n represents the total number of samples, y_{GS} is the gold standard and y is the prediction.

The second metric is the Concordance Correlation Coefficient (CCC) which is defined in Eq. 5. It combines the Pearson correlation coefficient (ρ) with the squared difference between the mean of the predictions (μ_y) and the mean of the golden standard ($\mu_{y_{GS}}$). This metric measures the association between variables and penalizes the score even if the model predicts the emotion well, but it shifts the value.

$$\text{CCC} = \frac{2\rho\sigma_y\sigma_{y_{GS}}}{\sigma_y^2 + \sigma_{y_{GS}}^2 + (\mu_y - \mu_{y_{GS}})^2} \tag{5}$$

where σ_y^2 and $\sigma_{y_{GS}}^2$ are the variance of the prediction and the gold standard, respectively.

A. FER Dataset

The FER dataset is used to pre-train the CNN architectures proposed in Section II. This dataset is composed of gray scale images of 48×48 pixels that comprise seven acted emotions (disgust, anger, fear, joy, saddens, surprise, neutral). The FER dataset is split into 28,709 images for training, 3,589 for validation, and 3,589 for test. Table III shows the accuracy of the propose CNN architectures (Table I) on the FER test set. The architecture B, which has less trainable parameters than architectures A and C, achieved the best accuracy of 67.7%, which is comparable to the results reported in [3], [6]. Therefore, for all further steps we adopted architecture B as our baseline.

TABLE III: Accuracy for the three proposed architectures of Table I on FER dataset.

Architecture	Accuracy (%)	
	Train	Test
A	59.6	62.5
B	**90.3**	**67.7**
C	25.0	24.7

B. Transfer Learning

Given the CNN pre-trained on FER dataset, we replace the output layer by a single neuron with linear activation to implement a regressor. Next, we freeze all convolutional layers (CL) and fine-tune (training) just the fully connected (FC) layers using the target dataset (RECOLA). Progressively, we unfreeze one convolutional layer per training session from the deeper to the first convolution layer and retrain the network with RECOLA dataset. Table IV shows that the highest CCCs are achieved by keeping frozen all the convolutional filters learned on the FER dataset and fine-tuning only the fully connected layers. Unfreezing convolutional layers reduces the CCC for both arousal and valence dimensions.

TABLE IV: CCC for the arousal and valence dimensions by fine-tuning none, one, two and all convolutional layers (CL) of the CNN.

Dimension	**CCC**			
	Trainable CLs			
	None	**3**	**2, 3**	**1, 2, 3**
Arousal	**0.239**	0.215	0.208	0.203
Valence	**0.314**	0.301	0.297	0.295

C. Pre-Processing

The RECOLA dataset provides annotation for face localization within the video frames. However, for several video frames the annotation is not available. Even our face detector is not able to find the face in these video frames. Therefore, we can discard these frames since they may not contain a frontal face. However, if we drop all these frames, we end up discarding 22% and 12% of the video frames of the training and developing datasets, respectively. This allows us to improve the quality of the training data, nevertheless we reduce the amount of data for training the CNN. We introduced a second face detector that is able to retain 25% of the data discarded previously. Besides that, we augment the remaining data using low-level transformations to reduce overfitting while training the CNN.

Table V shows a summary of the best CCC values for arousal and valence achieved by fine-tuning the CNN. The results consider the impact of the delay compensation and the preprocessing (PP) on the CCC. We see that the preprocessing increases the CCC for both arousal and valence dimensions. Delays of 70 and 50 frames correspond to 2.8 and 2.0 seconds, respectively.

TABLE V: The best CCC scores provided by the CNN for arousal and valence, for fine-tuning the convolutional layers (CLs), with delay compensation, and pre-processing (PP).

Dimension	**Trainable CLs**	**Delay**	**PP**	**CCC**
Arousal	1, 2, 3	70	Yes	**0.252**
Arousal	1, 2, 3	70	No	0.239
Arousal	None	70	Yes	0.203
Valence	1, 2, 3	50	Yes	**0.358**
Valence	1, 2, 3	50	No	0.314
Valence	None	50	Yes	0.294

D. Multimodal Fusion

In the proposed approach we use the CNNs only as feature extractors. Instead of selecting only the CNNs that lead to the best CCCs (Table V), we have also selected the CNNs that use transfer learning as initialization (no trainable CLs). Besides that, we have also considered the output of both the first (FC50) and the second (FC100) fully connected layers as feature representations. The idea behind that is to increase the diversity. The features extracted with the CNNs are then used to train an SVR.

Table VI shows the best CCCs achieved by the SVRs trained on CNN features. Reducing the dimension of the feature vector from 100 to 50 value leads to a reduction in CCC for valence but not for arousal. Furthermore, the best CCC achieved for arousal is for the features generated by fine-tuned CLs while for valence, the best CCC was achieved for features generated by the CNN with CLs learned on FER dataset. The reduced feature vectors preserve the essential information and generates better support vectors and therefore better predictions than a larger vector that may contain noisy features.

TABLE VI: CCC for SVR trained on CNN features for arousal and valence dimensions.

Dimension	**Trainable CLs**	**Delay**	**FC**	**CCC**
Arousal	1, 2, 3	70	100	0.146
Arousal	1, 2, 3	70	50	**0.154**
Arousal	None	70	100	0.148
Arousal	None	70	50	0.059
Valence	1, 2, 3	50	100	0.429
Valence	1, 2, 3	50	50	0.414
Valence	None	50	100	**0.433**
Valence	None	50	50	0.414

TABLE VII: Fusion results for **arousal** and **valence** with delay compensation of 70 and 50 respectively.

Arousal			
Modality	**Feature**	**Fusion**	**CCC**
Video	CNN [0CL, FC50]	Early	**0.749**
Audio	Recola		
Video	CNN [3CLs, FC50]	Early	**0.749**
Audio	Recola		
Video	CNN[3CLs, FC100]	Late	0.701
Audio	Recola		
Video	CNN [3CLs, FC50]	Late	0.715
Audio	Recola		
Valence			
Video	CNN [3CLs, FC100]	Early	**0.565**
Audio	Recola		
Video	CNN [0CL, FC100]	Early	**0.551**
Audio	Recola		
Video	CNN [0CL, FC100]	Late	0.543
Recola	Audio		
Video	CNN [3CLs FC50]	Late	0.522
Recola	Audio		

CL: # Frozen Convolutional Layers.

Table VII shows the CCCs for early and late multimodal fusion of CNN and audio features. It is important to highlight that SVRs trained on eGeMAPS audio features [18] achieved CCCs of 0.681 and 0.329 for arousal and valence, respectively. The results of Table VII show how complementary is the information between video and audio modalities and also, how each modality contributes to the final prediction. Fig. 5 summarizes the changes of the CCC according to the inclusion of a new technique (transfer learning by using FER dataset, preprocessing, and post-processing).

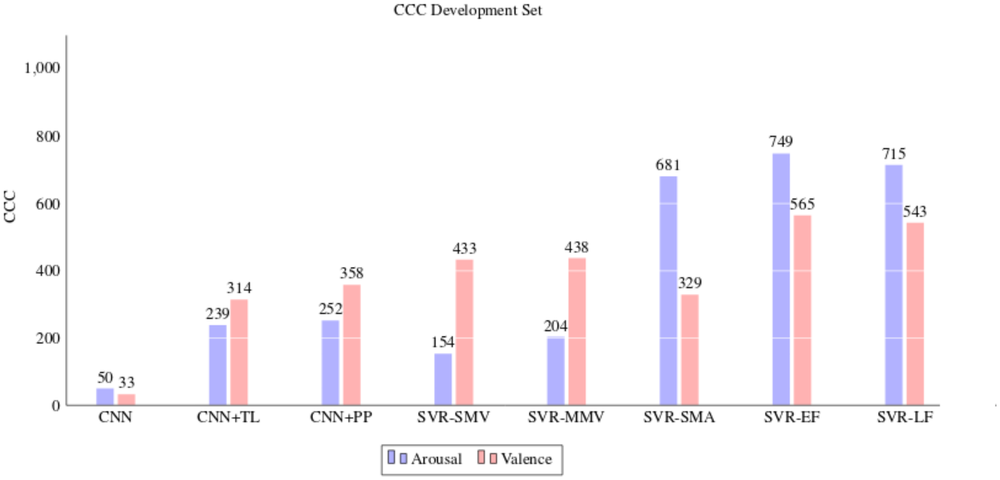


Fig. 5: Summary of the results by adding different elements: CNN alone; CNN with transfer learning and preprocessing (TL); CNN+TL plus post processing (CNN+PP); support vector regression with CNN features (SVR-SMV); support vector regression with audio features (SVR-SMA); multimodal early fusion (SVR-EF); multimodal late fusion (SVR-LF).

IV. Conclusion

In this paper we have presented a multimodal approach for continuous emotions recognition that combines visual and acoustic information to predict arousal and valence levels of speakers. The proposed approach is based on a pre-trained CNN to extract features which are further combined with handcrafted audio features. The proposed approach outperforms the baseline system for RECOLA, and several traditional approaches based on auditory and visual handcrafted features.

The experimental results have shown a positive impact of transfer learning in our model for the video modality, increasing the CCC in 0.05 and 0.06 for arousal and valence respectively. In the multimodal fusion, recent studies have shown that despite the fact that new and exotic fusion strategies are being developed, traditional fusion schemes are still able to generate competitive results [2], [4], [7], [8], [10].

References

[1] P. Cardinal, N. Dehak, A. L. Koerich, J. Alam, and P. Boucher. ETS System for AV+EC 2015. In *5th Intl Workshop on Audio/Visual Emotion Challenge*, pages 17–23, 2015.

[2] M. J. Cossetin, J. C. Nievola, and A. L. Koerich. Facial expression recognition using a pairwise feature selection and classification approach. In *Intl Joint Conf Neural Networks*, pages 5149–5155, 2016.

[3] W. Ding, M. Xu, D. Huang, W. Lin, M. Dong, X. Yu, and H. Li. Audio and face video emotion recognition in the wild using deep neural networks and small datasets. In *18th ACM Intl Conf on Multimodal Interaction*, pages 506–513, New York, USA, 2016.

[4] C. T. Duong, R. Lebret, and K. Aberer. Multimodal classification for analysing social media. *CoRR*, abs/1708.02099, 2017.

[5] F. Eyben, M. Kaiser, B. Schuller, G. Rigoll, and M. Wöllmer. LSTM-Modeling of continuous emotions in an audiovisual affect recognition framework. *Image and Vision Comp*, 31(2):153–163, 2012.

[6] I. J. Goodfellow, D. Erhan, P. L. Carrier, and A. Courville. Challenges in representation learning: A report on three machine learning contests. *Springer-Verlag*, 8228:117–124, 2013.

[7] Y. Li, J. Tao, B. Schuller, S. Shan, D. Jiang, and J. Jia. Mec 2017: Multimodal emot recog challenge. *1st Asian Conf Affect Comp Intell Interaction*, pages 1–5, 2017.

[8] P. P. Liang, A. Zadeh, and L.-P. Morency. Multimodal local-global ranking fusion for emot recognition. *20th ACM Intl Conf on Multimodal Interaction*, pages 472–476, 2018.

[9] S. Mariooryad and C. Busso. Analysis and compensation of the reaction lag of evaluators in continuous emotional annotations. In *2013 Humaine Ass Conf Affect Comp Intell Interac*, pages 85–90, 2013.

[10] J. D. S. Ortega, M. Senoussaoui, E. Granger, M. Pedersoli, P. Cardinal, and A. L. Koerich. Multimodal fusion with deep neural networks for audio-video emotion recognition. *arXiv preprint arXiv:1907.03196*, 2018.

[11] F. Ringeval, F. Eyben, E. Kroupi, A. Yuce, J.-p. P. Thiran, T. Ebrahimi, D. Lalanne, and B. Schuller. Prediction of asynchronous dimensional emotion ratings from audiovisual and physiological data. *Patt Recog Lett*, 66:22–30, 2015.

[12] J. A. Russell. A circumplex model of affect. *J of Personality and Social Psychology*, 39(6):1161–1178, 1989.

[13] C. G. M. Snoek, M. Worring, and A. W. M. Smeulders. Early versus late fusion in semantic video analysis. In *13th Annual ACM Intl Conf on Multimedia*, pages 399–402, New York, USA, 2005.

[14] B. Sun, S. Cao, L. Li, J. He, and L. Yu. Exploring multimodal visual features for continuous affect recognition. In *6th Intl Workshop on Audio/Visual Emotion Challenge*, pages 83–88, New York, USA, 2016.

[15] D. C. Tannugi, A. S. Britto Jr., and A. L. Koerich. Memory integrity of CNNs for cross-dataset facial expression recognition. *CoRR*, abs/1905.12082, 2019.

[16] G. Trigeorgis, F. Ringeval, R. Brueckner, E. Marchi, M. A. Nicolaou, and B. Schuller. Adieu features? end-to-end speech emotion recognition using a deep convolutional recurrent network. *IEEE Intl Conf Acoust Speech and Signal Proc*, pages 5200–5204, 2016.

[17] P. Tzirakis, G. Trigeorgis, M. A. Nicolaou, B. W. Schuller, and S. Zafeiriou. End-to-end multimodal emotion recognition using deep neural networks. *IEEE J Selec Top in Signal Proc*, 11(8):1301–1309, 2017.

[18] M. F. Valstar, J. Gratch, B. W. Schuller, F. Ringeval, D. Lalanne, M. Torres, S. Scherer, G. Stratou, R. Cowie, and M. Pantic. AVEC 2016 workshop and challenge. *CoRR*, abs/1605.01600, 2016.

[19] R. Weber, V. Barrielle, C. Soladié, and R. Séguier. High-Level Geometry-based Features of Video Modality for Emotion Prediction. In *6th Intl Workshop on Audio/Visual Emotion Challenge*, pages 51–58, New York, USA, 2016.

[20] J. Yan, W. Zheng, Z. Cui, C. Tang, T. Zhang, and Y. Zong. Multi-cue fusion for emotion recognition in the wild. *Neurocomputing*, 309:27–35, 2018.

[21] K. Yan, W. Zheng, T. Zhang, Y. Zong, and Z. Cui. Cross-database non-frontal facial expression recognition based on transductive deep transfer learning. pages 1–8, 2018.

[22] X. Zhu, L. Li, W. Zhang, T. Rao, M. Xu, Q. Huang, and D. Xu. Dependency exploitation: A unified cnn-rnn approach for visual emotion recognition. *26th Intl Joint Conf Artif Intell*, pages 3595–3601, 2017.