CROSS ATTENTIONAL AUDIO-VISUAL FUSION FOR DIMENSIONAL EMOTION RECOGNITION

R. Gnana Praveen, **Eric Granger**, **Patrick Cardinal**

Laboratoire d'imagerie, de vision et d'intelligence artificielle (LIVIA),
École de technologie supérieure, Montreal, Canada.

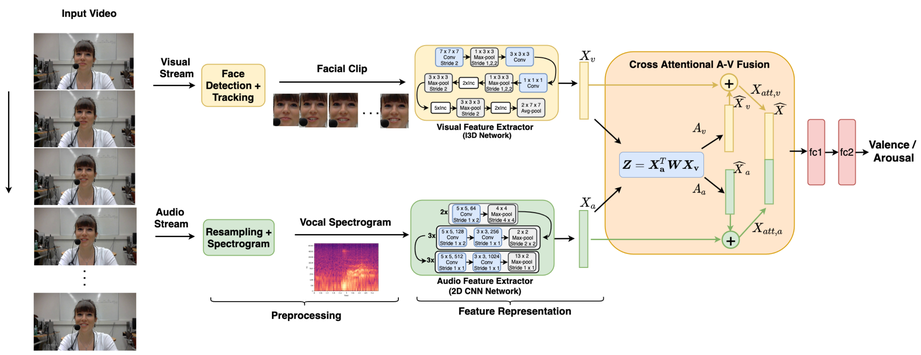
Motivation

- Dimensional emotion recognition deals with the problem of estimating emotion levels on a finer granular level using valence and arousal
- Although human emotions can be expressed through various modalities, we focus on audio (A) and visual (V)
- The objective is to develop ML models for efficient fusion of A and V for dimensional emotion recognition

Related Work

- Most of the existing approaches have modeled intra-modal relationship using LSTM based fusion (Tzirakis et al., JSTSP 2017), (Schoneveld et al., PR Letters 2021)
- Though attention models have been explored, they fail to explicitly capture the complementary relationship (Tzirakis et al., IF 2021), (Partha et al., SLT 2021)
- In this paper, we propose to leverage cross attention based on cross correlation to model the inter-relationship and complementarity of A-V modalities

Proposed Approach



- The proposed approach has three major blocks **Audio Feature Extractor**, **Visual Feature Extractor** and **Cross Attentional Module**
- The V features ($\boldsymbol{X}_v$) are obtained using I3D network inflated from inception architecture pretrained on ImageNet
- The A features ($\boldsymbol{X}_a$) are obtained from 2D CNN using spectrograms
- The obtained A ($\boldsymbol{X}_a$) and V ($\boldsymbol{X}_v$) features are used to compute cross correlation matrix ($\boldsymbol{Z}$)

$$\boldsymbol{Z} = \boldsymbol{X}_a^T \boldsymbol{W} \boldsymbol{X}_v \quad \text{where } \boldsymbol{W} \in \mathbb{R}^{K \times K} \text{ : learnable parameter and K is feature dimension} \tag{1}$$

- The cross attention weights of A ($\boldsymbol{A}_a$) and V ($\boldsymbol{A}_v$) modalities are obtained using column-wise softmax of $\boldsymbol{Z}$ and $\boldsymbol{Z}^T$ respectively
- The attention maps of A ($\widehat{\mathbf{X}_\mathbf{a}}$) and V ($\widehat{\mathbf{X}_\mathbf{v}}$) modalities are obtained as

$$\widehat{\mathbf{X}_\mathbf{a}} = \mathbf{X}_\mathbf{a}\mathbf{A}_\mathbf{a} \quad \text{and} \quad \widehat{\mathbf{X}_\mathbf{v}} = \mathbf{X}_\mathbf{v}\mathbf{A}_\mathbf{v} \tag{2}$$

- The attended features of A ($\boldsymbol{X}_{\mathbf{att,a}}$) and V ($\boldsymbol{X}_{\mathbf{att,a}}$) modalities are

$$\boldsymbol{X}_{\mathbf{att,a}} = \tanh(\boldsymbol{X}_\mathbf{a} + \widehat{\mathbf{X}_\mathbf{a}}) \tag{3}$$

$$\boldsymbol{X}_{\mathbf{att,v}} = \tanh(\boldsymbol{X}_\mathbf{v} + \widehat{\mathbf{X}_\mathbf{v}}) \tag{4}$$

- Finally, the A-V feature vector is obtained by concatenating the attended A and V features as

$$\widehat{\mathbf{X}} = [\boldsymbol{X}_{\mathbf{att,v}}; \boldsymbol{X}_{\mathbf{att,a}}] \tag{5}$$

- The feature ($\widehat{\mathbf{X}}$) is then fed to fully connected layers to predict valence or arousal

Experimental Datasets

The proposed approach has been evaluated on RECOLA and Fatigue datasets

Fig. 2: Samples from the RECOLA dataset

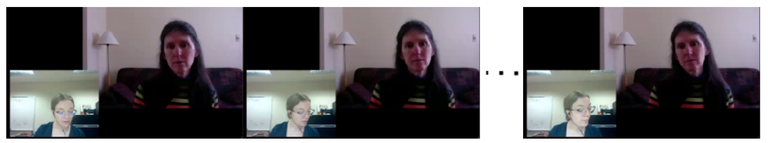


Fig. 3: Samples from the Fatigue dataset

Ablation Study

- For V modality, we explored 2D CNN and I3D networks
- For A modality, we have used the same 2D CNN for all the experiments
- We compared the proposed approach with self-attention, LSTM-based fusion and simple feature concatenation

Method: V + Fusion	Valence	Arousal
2D CNN + Feature Concatenation	0.538	0.680
2D CNN + LSTM	0.552	0.697
I3D + Feature Concatenation	0.579	0.732
I3D + Self Attention	0.623	0.787
I3D + Cross-Attention (ours)	0.685	0.835

Tab. 1: CCC performance of our proposed approach obtained with various fusion strategies on the RECOLA dataset.

Visual Results

Though the full frontal face of the subject is not available in the V modality, the proposed approach still tracks the groud truth by leveraging the A modality

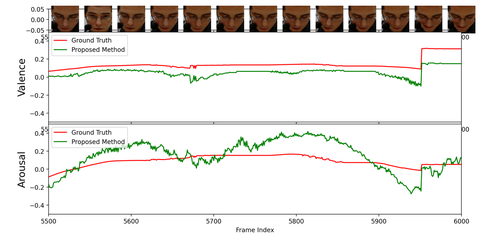


Fig. 4: Visualization of predictions of valence and arousal for subjects "dev 3"

Comparison to SOTA

[Tzirakis et al., 2017] and [Schoneval et al., 2021] explored deep models for A-V features and used LSTM based fusion. The proposed approach was found to outperform the state-of-the-art models by leveraging the inter-modal relationship of AV modalities. The experimental protocol of AVEC 2017 has been used for RECOLA dataset.

Method	Valence			Arousal		
	Audio	Visual	Fusion	Audio	Visual	Fusion
[He et al., 2015]	0.400	0.441	0.609	0.800	0.587	0.747
[Han et al., 2017]	0.480	0.592	0.554	0.760	0.350	0.685
[Tzirakis et al., 2017]	0.428	0.637	0.502	0.786	0.371	0.731
[Juan et al., 2019]	-	-	0.565	-	-	0.749
[Schoneval et al., 2021]	0.460	0.550	0.630	0.800	0.570	0.810
Proposed Approach	0.463	0.642	0.685	0.822	0.582	0.835
Proposed Approach (2-stage)	0.463	0.642	0.690	0.822	0.582	0.838

Tab. 2: CCC performance of the proposed approach and state-of-the-art models. All the results are presented on the RECOLA development set.

Results with Fatigue Data

The proposed approach has been further evaluated to estimate the fatigue level on a scale of 0 to 10. 27 videos are captured from 18 participants suffering from degenerate diseases inducing pain. 80% of data is used for training and 20 % for validation. We have experimented with simple feature concatenation and the proposed approach.

Method	Fatigue Level
Audio only (2D-CNN)	0.312
Visual only (I3D)	0.415
Feature Concatenation	0.378
Proposed Approach (Cross-Attention)	0.421

Tab. 3: CCC performance on fatigue dataset.