Cross Attentional Audio-Visual Fusion for Dimensional Emotion Recognition

Gnana Praveen R Eric Granger Patrick Cardinal

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

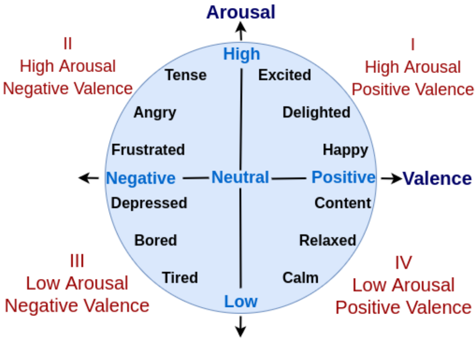
December 18 2021

LABORATOIRE
D'IMAGERIE, DE VISION
ET D'INTELLIGENCE
ARTFICIELLE

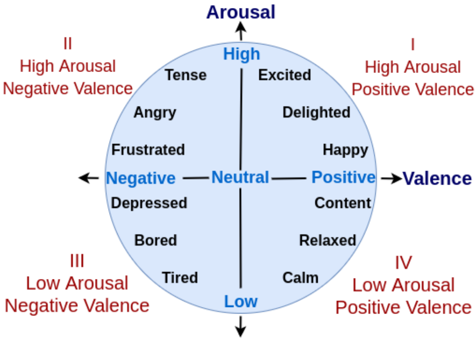
Outline

Outline

Dimensional Emotion Recognition

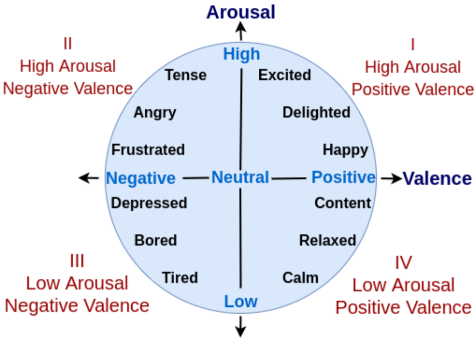


Dimensional Emotion Recognition



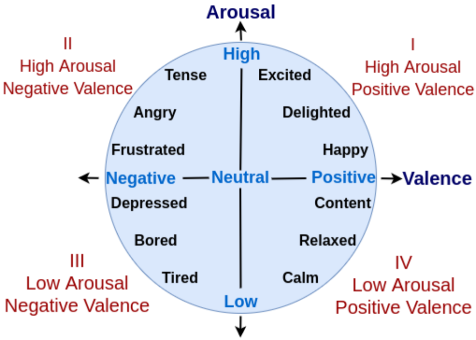
- Problem : Estimating regression values in the valence-arousal space

Dimensional Emotion Recognition



- Problem : Estimating regression values in the valence-arousal space
- Valence denotes the range of emotions from very sad (negative) to very happy (positive)

Dimensional Emotion Recognition



- Problem : Estimating regression values in the valence-arousal space
- Valence denotes the range of emotions from very sad (negative) to very happy (positive)
- Arousal reflects the energy or intensity of emotions from very passive to very active

A-V Fusion for Dimensional Emotion Recognition

- Audio (A) and Visual (V) are the widely used contact free modalities for emotion recognition

A-V Fusion for Dimensional Emotion Recognition

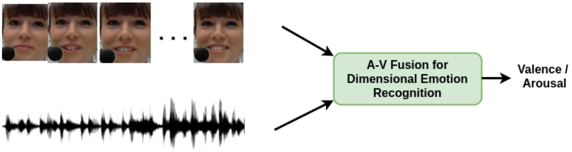
- Audio (A) and Visual (V) are the widely used contact free modalities for emotion recognition
- A and V channels provide complimentary relationship to obtain comprehensive information

A-V Fusion for Dimensional Emotion Recognition

- Audio (A) and Visual (V) are the widely used contact free modalities for emotion recognition
- A and V channels provide complimentary relationship to obtain comprehensive information
- Efficient fusion of A and V channels are expected to outperform uni-modal approaches

A-V Fusion for Dimensional Emotion Recognition

- Audio (A) and Visual (V) are the widely used contact free modalities for emotion recognition
- A and V channels provide complimentary relationship to obtain comprehensive information
- Efficient fusion of A and V channels are expected to outperform uni-modal approaches



Challenges for A-V Fusion

- How to extract efficient multi-modal feature representation of A-V modalities?

Challenges for A-V Fusion

- How to extract efficient multi-modal feature representation of A-V modalities?
- How to effectively leverage the complimentary relationship of A-V modalities ?

Challenges for A-V Fusion

- How to extract efficient multi-modal feature representation of A-V modalities?
- How to effectively leverage the complimentary relationship of A-V modalities ?
- How to handle wide range of variations in facial expressions due to pose, illumination, identity-bias, etc. ?

Challenges for A-V Fusion

- How to extract efficient multi-modal feature representation of A-V modalities?
- How to effectively leverage the complimentary relationship of A-V modalities ?
- How to handle wide range of variations in facial expressions due to pose, illumination, identity-bias, etc. ?
- How to handle wide range of variations in vocal expressions due to speaker identity-bias, background noise, etc ?

Outline

A-V Fusion Approaches for Dimensional Emotion Recognition

- [Tzirakis et al., 2017] extracted A and V features from Resnet-50 and 1D CNN respectively, which is concatenated and fed to LSTM

A-V Fusion Approaches for Dimensional Emotion Recognition

- [Tzirakis et al., 2017] extracted A and V features from Resnet-50 and 1D CNN respectively, which is concatenated and fed to LSTM
- [Schoneveld et al., 2021] explored knowledge distillation using student-teacher model for V modality and 2D CNN for A modality using spectrograms, which is further concatenated and fed to LSTM

A-V Fusion Approaches for Dimensional Emotion Recognition

- [Tzirakis et al., 2017] extracted A and V features from Resnet-50 and 1D CNN respectively, which is concatenated and fed to LSTM
- [Schoneveld et al., 2021] explored knowledge distillation using student-teacher model for V modality and 2D CNN for A modality using spectrograms, which is further concatenated and fed to LSTM
- [Tzirakis et al., 2021] investigated various fusion strategies along with attention mechanisms including self-attention.

A-V Fusion Approaches for Dimensional Emotion Recognition

- [Tzirakis et al., 2017] extracted A and V features from Resnet-50 and 1D CNN respectively, which is concatenated and fed to LSTM
- [Schoneveld et al., 2021] explored knowledge distillation using student-teacher model for V modality and 2D CNN for A modality using spectrograms, which is further concatenated and fed to LSTM
- [Tzirakis et al., 2021] investigated various fusion strategies along with attention mechanisms including self-attention.
- [Parthasarathy and Sundaram, 2021] explored transformers with cross modal attention, where cross attention is integrated with self attention

Limitations of SOA Approaches

- Most of the existing approaches focus on modeling the intra-modal relationships

Limitations of SOA Approaches

- Most of the existing approaches focus on modeling the intra-modal relationships
- The inter-modal relationships are not effectively explored to capture the complimentarity of A-V modalities

Limitations of SOA Approaches

- Most of the existing approaches focus on modeling the intra-modal relationships
- The inter-modal relationships are not effectively explored to capture the complimentarity of A-V modalities
- Though attention models have been explored with transformers, they fail to capture the complimentary relationship of A-V modalities

Limitations of SOA Approaches

- Most of the existing approaches focus on modeling the intra-modal relationships
- The inter-modal relationships are not effectively explored to capture the complimentarity of A-V modalities
- Though attention models have been explored with transformers, they fail to capture the complimentary relationship of A-V modalities
- The semantic relevance among A-V features are not effectively captured in the existing approaches

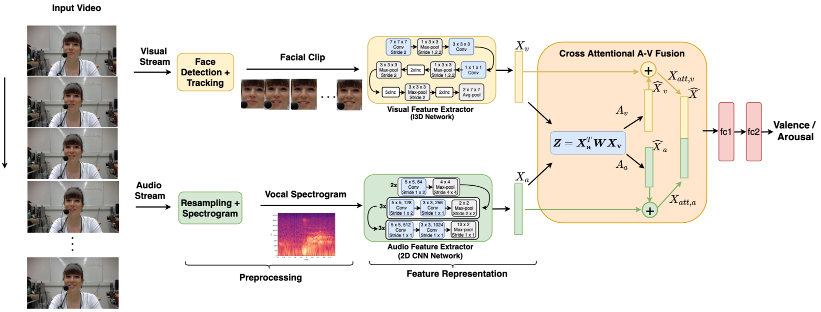
Outline

Overall Framework

- The training mechanism has three major modules : V Network, A Network and Cross-Attentional A-V Fusion

Overall Framework

- The training mechanism has three major modules : V Network, A Network and Cross-Attentional A-V Fusion



Visual Network

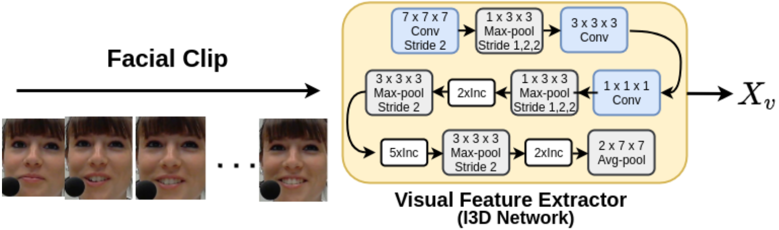
- I3D is widely used for the task of action recognition. Inspired by the performance of I3D, we use I3D for feature extraction

Visual Network

- I3D is widely used for the task of action recognition. Inspired by the performance of I3D, we use I3D for feature extraction
- We have inflated inception v-1 architecture from 2D pretrained model on ImageNet

Visual Network

- I3D is widely used for the task of action recognition. Inspired by the performance of I3D, we use I3D for feature extraction
- We have inflated inception v-1 architecture from 2D pretrained model on ImageNet



Audio Network

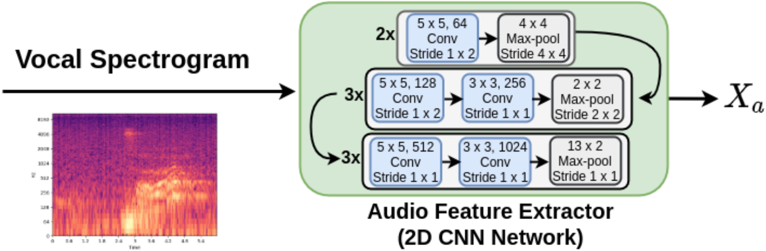
- Spectrograms are obtained from the speech signal and fed to the 2D CNN network

Audio Network

- Spectrograms are obtained from the speech signal and fed to the 2D CNN network
- The spectrograms are fed to the network, which is trained from scratch

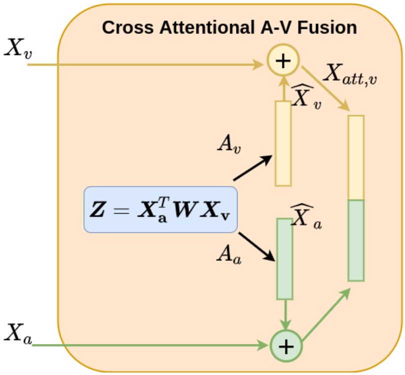
Audio Network

- Spectrograms are obtained from the speech signal and fed to the 2D CNN network
- The spectrograms are fed to the network, which is trained from scratch

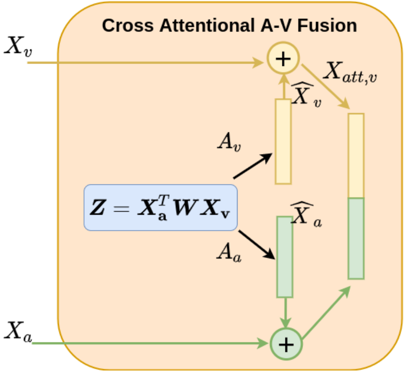


Cross Attentional AV Fusion

Cross Attentional AV Fusion



Cross Attentional AV Fusion



- The V features ($\boldsymbol{X_v}$) and A features ($\boldsymbol{X_a}$) features are fed to the cross attentional module

Cross Attentional Fusion

- Cross attentional fusion was found to be efficient in capturing the semantic relevance across the modalities

Cross Attentional Fusion

- Cross attentional fusion was found to be efficient in capturing the semantic relevance across the modalities
- It estimates the cross correlation across the A-V features to capture the complimentary relationship

Cross Attentional Fusion

- Cross attentional fusion was found to be efficient in capturing the semantic relevance across the modalities
- It estimates the cross correlation across the A-V features to capture the complimentary relationship
- The cross correlation helps A-V features to interact between each other and gives a measure of semantic relevance across the modalities

Cross Attentional Fusion

- Cross attentional fusion was found to be efficient in capturing the semantic relevance across the modalities
- It estimates the cross correlation across the A-V features to capture the complimentary relationship
- The cross correlation helps A-V features to interact between each other and gives a measure of semantic relevance across the modalities
- Cross correlation based cross attention was successfully applied in few shot classification [Hou et al., 2019] and weakly supervised action localization [Lee et al., 2021]

Cross Attentional AV Fusion

Cross Correlation matrix

$$\boldsymbol{Z} = \boldsymbol{X_a}^T \boldsymbol{W} \boldsymbol{X_v}$$

where $\boldsymbol{W}$: learnable parameter

$\boldsymbol{X_v}$: deep features of V modality of given video sequence

$\boldsymbol{X_a}$: deep features of A modality of given video sequence

Cross Attentional AV Fusion

Cross Correlation matrix

$$\boldsymbol{Z} = \boldsymbol{X_a^T W X_v}$$

where $\boldsymbol{W}$: learnable parameter
$\boldsymbol{X_v}$: deep features of V modality of given video sequence
$\boldsymbol{X_a}$: deep features of A modality of given video sequence

Cross Attention Weights

$$\boldsymbol{A_a}_{i,j} = \frac{e^{\boldsymbol{Z}_{i,j}/T}}{\sum_{k=1}^{K} e^{\boldsymbol{Z}_{k,j}/T}} \quad \text{and} \quad \boldsymbol{A_v}_{i,j} = \frac{e^{\boldsymbol{Z}^T{}_{i,j}/T}}{\sum_{k=1}^{K} e^{\boldsymbol{Z}^T{}_{i,k}/T}}$$

where $\boldsymbol{Z}_{i,j}$: ith row and jth column of $\boldsymbol{Z}$
T : softmax temperature

Cross Attentional AV Fusion

Attention Maps

$$\widehat{\boldsymbol{X_a}} = \boldsymbol{X_a}\boldsymbol{A_a} \quad \text{and} \quad \widehat{\boldsymbol{X_v}} = \boldsymbol{X_v}\boldsymbol{A_v}$$

Cross Attentional AV Fusion

Attention Maps

$$\widehat{\boldsymbol{X_a}} = \boldsymbol{X_a A_a} \quad \text{and} \quad \widehat{\boldsymbol{X_v}} = \boldsymbol{X_v A_v}$$

Final Attended features

$$\boldsymbol{X_{att,a}} = \tanh(\boldsymbol{X_a} + \widehat{\boldsymbol{X_a}})$$

$$\boldsymbol{X_{att,v}} = \tanh(\boldsymbol{X_v} + \widehat{\boldsymbol{X_v}})$$

Cross Attentional AV Fusion

Attention Maps

$$\widehat{\boldsymbol{X}_a} = \boldsymbol{X}_a \boldsymbol{A}_a \quad \text{and} \quad \widehat{\boldsymbol{X}_v} = \boldsymbol{X}_v \boldsymbol{A}_v$$

Final Attended features

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

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

- The final attended features are further concatenated and fed to fully connected layers for valence / arousal prediction

Outline

1. Dimensional Emotion Recognition
2. Motivation for Cross Attentional A-V Fusion
3. Proposed Approach
4. Results and Discussion
5. Conclusion

Experimental Setup

- We have evaluated our proposed approach on RECOLA and Fatigue (private) datasets
- RECOLA dataset has been used for various challenges such as AVEC 2015 and AVEC 2016, where 9 subjects are used for training and 9 for validation
- Fatigue is a private dataset, which has 27 videos captured from 18 participants, suffering from degenerative diseases inducing fatigue
- Concordance Correlation Coefficient (CCC) is used to measure the performance of the proposed approach

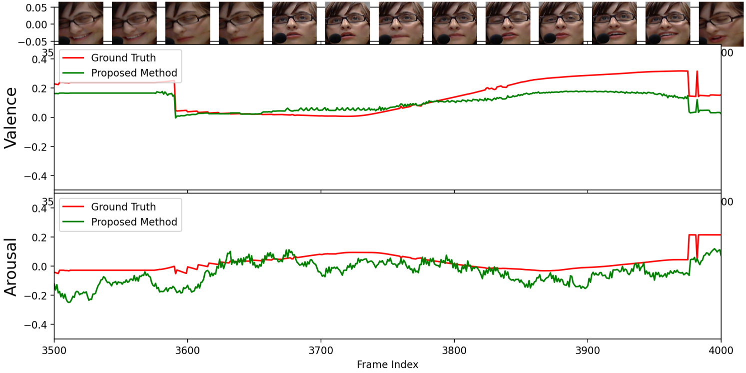
Ablation Study

- For V modality, we have explored 2D CNN and I3D model whereas for A modality, we have used the same deep network for all the experiments

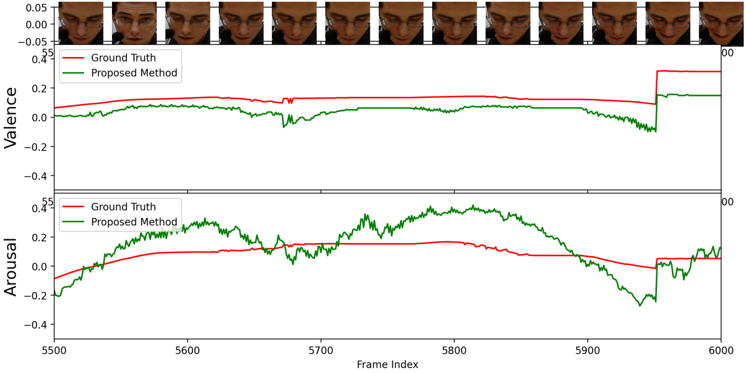
Table: **CCC performance of our proposed approach obtained with various components on the RECOLA dataset.**

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

Visualizations of valence and arousal predictions for subject "dev 1"



Visualizations of valence and arousal predictions for subject "dev 3"



Comparison with state-of-the-art approaches

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

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

Results with Fatigue (private) Data

Table: **CCC performance on Fatigue dataset.**

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

Outline

Conclusion

- We propose a cross attentional A-V fusion model for dimensional emotion recognition

Conclusion

- We propose a cross attentional A-V fusion model for dimensional emotion recognition
- Unlike prior approaches of A-V fusion, we focus on inter modal relationships to leverage the complementarity of A-V modalities

Conclusion

- We propose a cross attentional A-V fusion model for dimensional emotion recognition
- Unlike prior approaches of A-V fusion, we focus on inter modal relationships to leverage the complementarity of A-V modalities
- Contrary to prior cross attentional models, we explored cross attention based cross correlation in the context of regression

Conclusion

- We propose a cross attentional A-V fusion model for dimensional emotion recognition
- Unlike prior approaches of A-V fusion, we focus on inter modal relationships to leverage the complementarity of A-V modalities
- Contrary to prior cross attentional models, we explored cross attention based cross correlation in the context of regression
- Extensive set of experiments conducted on RECOLA and Fatigue (private) datasets shows that the proposed approach clearly outperforms SOTA

Thank you for your attention!

References I

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

He, L., Jiang, D., Yang, L., Pei, E., Wu, P., and Sahli, H. (2015).
Multimodal affective dimension prediction using deep bidirectional long short-term memory recurrent neural networks.
In *5th AVEC*.

References II

Hou, R., Chang, H., MA, B., Shan, S., and Chen, X. (2019).
Cross attention network for few-shot classification.
In *NIPS*.

Lee, J.-T., Jain, M., Park, H., and Yun, S. (2021).
Cross-attentional audio-visual fusion for weakly-supervised action localization.
In *ICLR*.

Ortega, J. D. S., Cardinal, P., and Koerich, A. L. (2019).
Emotion recognition using fusion of audio and video features.
In *IEEE International Conference on Systems, Man and Cybernetics (SMC)*, pages 3847–3852.

References III

Parthasarathy, S. and Sundaram, S. (2021).
Detecting expressions with multimodal transformers.
In *2021 IEEE Spoken Language Technology Workshop (SLT)*, pages 636–643.

Schoneveld, L., Othmani, A., and Abdelkawy, H. (2021).
Leveraging recent advances in deep learning for audio-visual emotion recognition.
Pattern Recognition Letters, 146:1–7.

Tzirakis, P., Chen, J., Zafeiriou, S., and Schuller, B. (2021).
End-to-end multimodal affect recognition in real-world environments.
Information Fusion, 68:46–53.

References IV

Tzirakis, P., Trigeorgis, G., Nicolaou, M. A., Schuller, B. W., and Zafeiriou, S. (2017).
End-to-end multimodal emotion recognition using deep neural networks.
IEEE Journal of Selected Topics in Signal Processing, 11(8):1301–1309.