A Joint Cross-Attention Model for Audio-Visual Fusion in Dimensional Emotion Recognition

R Gnana Praveen[1], Wheidima Carneiro de Melo[1], Nasib Ullah, Haseeb Aslam[1], Osama Zeeshan[1], Théo Denorme[1], Marco Pedersoli[1], Alessandro L. Koerich[1], Simon Bacon[2], Patrick Cardinal[1], and Eric Granger[1]

[1] LIVIA, École de technologie supérieure, Montreal, Canada
[2]Dept. of Health, Kinesiology & Applied Physiology, Concordia University, Montreal, Canada

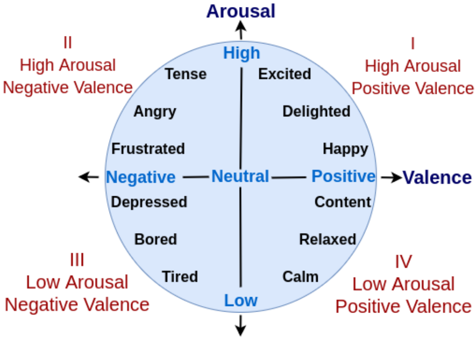
FRSQ Double Chair in AI and Digital Health

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

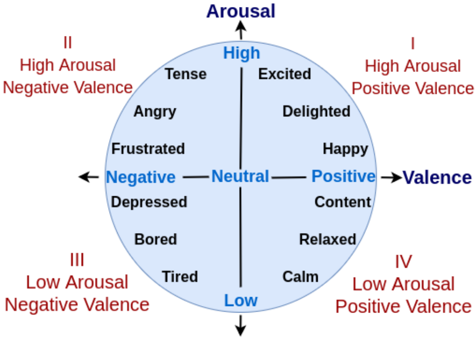
Outline

Outline

Dimensional Emotion Recognition

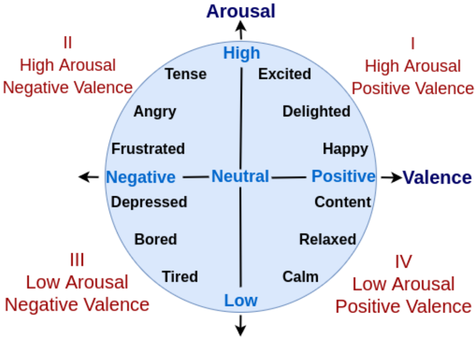


Dimensional Emotion Recognition



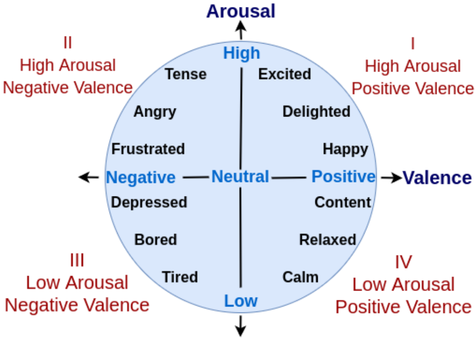
- Problem: Estimating regression values in the valence-arousal space

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

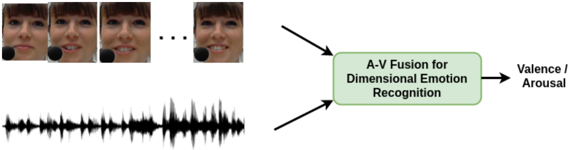
- Audio (A) and Visual (V) are the widely used contact free modalities for emotion recognition
- A and V channels provide complimentary relationship to obtain comprehensive information

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

- In general, there is a variation of expressions across subjects and annotation bias.

Challenges for A-V Fusion

- In general, there is a variation of expressions across subjects and annotation bias.
- How to extract efficient multi-modal feature representation of A-V modalities?

Challenges for A-V Fusion

- In general, there is a variation of expressions across subjects and annotation bias.
- 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

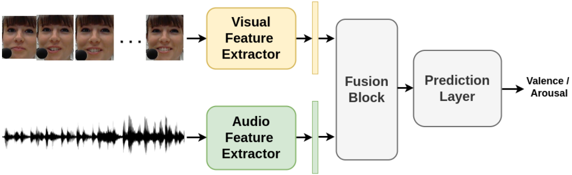
- In general, there is a variation of expressions across subjects and annotation bias.
- How to extract efficient multi-modal feature representation of A-V modalities?
- How to effectively leverage the complimentary relationship of A-V modalities?
- How to handle the wide range of variations in facial expressions due to pose, illumination, identity-bias, etc.?

Challenges for A-V Fusion

- In general, there is a variation of expressions across subjects and annotation bias.
- 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 the wide range of variations in facial expressions due to pose, illumination, identity-bias, etc.?
- How to handle the wide range of variations in vocal expressions due to speaker identity-bias, background noise, etc.?

Outline

Overview of A-V Fusion Approaches



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
- These 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
- These relationships are not effectively explored to capture the complimentarity of A-V modalities
- Though attention models have been explored with transformers, they do not effectively capture the complimentary relationship of A-V modalities

Limitations of SOA Approaches

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

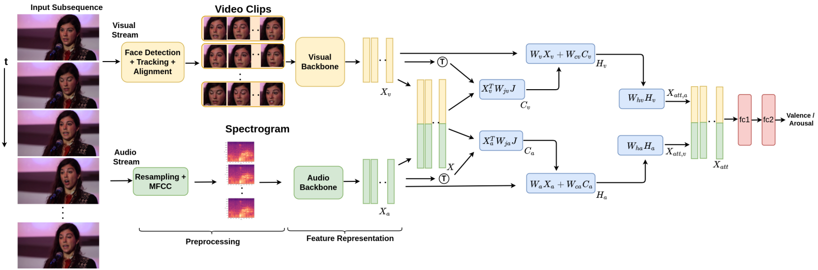
Outline

Overall Framework

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

Overall Framework

- The training mechanism has three major modules: V Network, A Network, and Joint 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

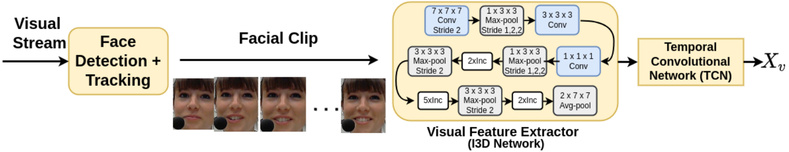
- I3D is widely used for the task of action recognition. Inspired by the performance of I3D, we use I3D for feature extraction
- We inflate Inception v-1 architecture from 2D-CNN pretrained model on ImageNet

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 inflate Inception v-1 architecture from 2D-CNN pretrained model on ImageNet
- The features are further fed to Temporal Convolutional Network (TCN) to capture long-term temporal relationship.

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 inflate Inception v-1 architecture from 2D-CNN pretrained model on ImageNet
- The features are further fed to Temporal Convolutional Network (TCN) to capture long-term temporal relationship.



Audio Network

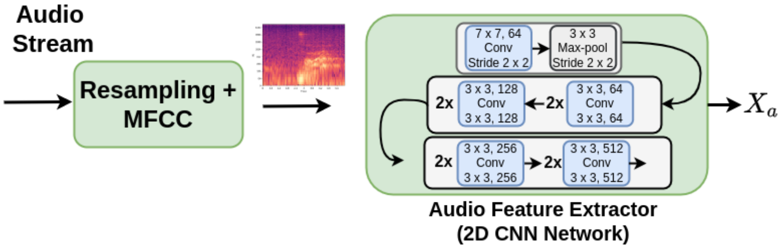
- Spectrograms are obtained from the speech signal and fed to the 2D CNN network (Resnet18)

Audio Network

- Spectrograms are obtained from the speech signal and fed to the 2D CNN network (Resnet18)
- The spectrograms are fed to the network, which is trained from scratch

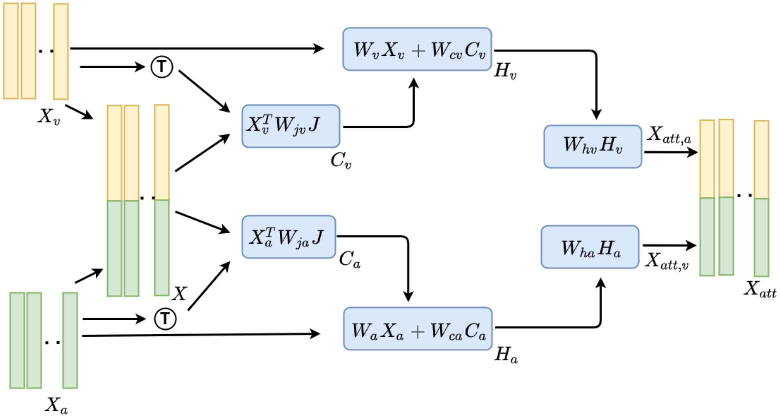
Audio Network

- Spectrograms are obtained from the speech signal and fed to the 2D CNN network (Resnet18)
- The spectrograms are fed to the network, which is trained from scratch



Joint Cross Attentional AV Fusion

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



Joint Cross Attentional Fusion

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

Joint 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

Joint 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

Joint 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
- Using joint representation in the cross attention module helps to jointly model the intra and inter modal relationships.

Joint Cross Attentional AV Fusion

Joint Cross Correlation matrix

$$\boldsymbol{C}_{\mathbf{a}} = \tanh\left(\frac{\boldsymbol{X}_{\mathbf{a}}^{\top}\boldsymbol{W}_{\mathbf{ja}}\boldsymbol{J}}{\sqrt{d}}\right) \quad \text{and} \quad \boldsymbol{C}_{\mathbf{v}} = \tanh\left(\frac{\boldsymbol{X}_{\mathbf{v}}^{\top}\boldsymbol{W}_{\mathbf{jv}}\boldsymbol{J}}{\sqrt{d}}\right)$$

where $\boldsymbol{W}_{\mathbf{ja}}, \boldsymbol{W}_{\mathbf{jv}}$: *learnable parameters*

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

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

$\boldsymbol{J}$: combined A-V deep features of given video sequence

$\boldsymbol{d}$: feature dimension of concatenated features

Joint Cross Attentional AV Fusion

Joint Cross Attention Weights

$$\boldsymbol{H}_{\mathbf{a}} = ReLu(\boldsymbol{W}_{\mathbf{a}}\boldsymbol{X}_{\mathbf{a}} + \boldsymbol{W}_{\mathbf{ca}}\boldsymbol{C}_{\mathbf{a}}^{\top})$$
$$\boldsymbol{H}_{\mathbf{v}} = ReLu(\boldsymbol{W}_{\mathbf{v}}\boldsymbol{X}_{\mathbf{v}} + \boldsymbol{W}_{\mathbf{cv}}\boldsymbol{C}_{\mathbf{v}}^{\top})$$

where $\boldsymbol{W}_{\mathbf{a}}, \boldsymbol{W}_{\mathbf{v}}, \boldsymbol{W}_{\mathbf{ca}}, \boldsymbol{W}_{\mathbf{cv}}$: learnable parameters

T : Transpose Operation

Joint Cross Attentional AV Fusion

Joint Cross Attention Weights

$$\boldsymbol{H}_{\mathbf{a}} = ReLu(\boldsymbol{W}_{\mathbf{a}}\boldsymbol{X}_{\mathbf{a}} + \boldsymbol{W}_{\mathbf{ca}}\boldsymbol{C}_{\mathbf{a}}^{\top})$$
$$\boldsymbol{H}_{\mathbf{v}} = ReLu(\boldsymbol{W}_{\mathbf{v}}\boldsymbol{X}_{\mathbf{v}} + \boldsymbol{W}_{\mathbf{cv}}\boldsymbol{C}_{\mathbf{v}}^{\top})$$

where $\boldsymbol{W}_{\mathbf{a}}, \boldsymbol{W}_{\mathbf{v}}, \boldsymbol{W}_{\mathbf{ca}}, \boldsymbol{W}_{\mathbf{cv}}$: learnable parameters
T : Transpose Operation

Attended features

$$\boldsymbol{X}_{\mathbf{att,a}} = \boldsymbol{W}_{\mathbf{ha}}\boldsymbol{H}_{\mathbf{a}} + \boldsymbol{X}_{\mathbf{a}}$$
$$\boldsymbol{X}_{\mathbf{att,v}} = \boldsymbol{W}_{\mathbf{hv}}\boldsymbol{H}_{\mathbf{v}} + \boldsymbol{X}_{\mathbf{v}}$$

where $\boldsymbol{W}_{\mathbf{ha}}, \boldsymbol{W}_{\mathbf{hv}}$: learnable parameters

Joint Cross Attentional AV Fusion

Joint Cross Attention Weights

$$\boldsymbol{H}_{\mathbf{a}} = ReLu(\boldsymbol{W}_{\mathbf{a}}\boldsymbol{X}_{\mathbf{a}} + \boldsymbol{W}_{\mathbf{ca}}\boldsymbol{C}_{\mathbf{a}}^{\top})$$
$$\boldsymbol{H}_{\mathbf{v}} = ReLu(\boldsymbol{W}_{\mathbf{v}}\boldsymbol{X}_{\mathbf{v}} + \boldsymbol{W}_{\mathbf{cv}}\boldsymbol{C}_{\mathbf{v}}^{\top})$$

where $\boldsymbol{W}_{\mathbf{a}}, \boldsymbol{W}_{\mathbf{v}}, \boldsymbol{W}_{\mathbf{ca}}, \boldsymbol{W}_{\mathbf{cv}}$: learnable parameters
T : Transpose Operation

Attended features

$$\boldsymbol{X}_{\mathbf{att,a}} = \boldsymbol{W}_{\mathbf{ha}}\boldsymbol{H}_{\mathbf{a}} + \boldsymbol{X}_{\mathbf{a}}$$
$$\boldsymbol{X}_{\mathbf{att,v}} = \boldsymbol{W}_{\mathbf{hv}}\boldsymbol{H}_{\mathbf{v}} + \boldsymbol{X}_{\mathbf{v}}$$

where $\boldsymbol{W}_{\mathbf{ha}}, \boldsymbol{W}_{\mathbf{hv}}$: learnable parameters

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

Outline

Experimental Setup

- We have evaluated our proposed approach on Affwild2 dataset (ABAW3 challenge).
- The training, validation and testing partitions has 341, 71 and 152 videos respectively.
- Concordance Correlation Coefficient (CCC) is used to measure the performance of the proposed approach

Ablation Study

Performance of our approach with different components on the development set of the Affwild2 dataset. The Resnet18 [He et al., 2016] is used to extract A features in all experiments.

V Backbone	Fusion Module	Valence	Arousal
I3D	Feature Concatenation	0.531	0.468
R3D	Feature Concatenation	0.517	0.493
I3D	Cross-Attention [Praveen et al., 2021]	0.541	0.517
I3D	Leader-Follower [Schoneveld et al., 2021]	0.592	0.521
Resnet18-GRU	Joint Cross-Attention (Ours)	0.632	0.520
R3D	Joint Cross-Attention (Ours)	0.642	**0.592**
I3D	Joint Cross-Attention (Ours)	0.657	0.580
I3D-TCN	Joint Cross-Attention (Ours)	0.663	0.584
I3D-TCN + R3D	Joint Cross-Attention (Ours)	**0.670**	0.590

Comparison with state-of-the-art approaches

CCC of the proposed approach compared to state-of-the-art methods for A-V fusion on the Affwild2 development set.

Method	Valence			Arousal		
	Audio	**Visual**	**Fusion**	**Audio**	**Visual**	**Fusion**
Kuhnke et al. [Kuhnke et al., 2020]	0.351	0.449	0.493	0.356	0.565	0.604
Zhang et al. [Zhang et al., 2021]	-	0.405	0.457	-	0.635	**0.645**
Rajasekhar et al. [Praveen et al., 2021]	0.351	0.417	0.552	0.356	0.539	0.531
Joint Cross-Attention (Ours)	0.351	0.417	0.663	0.356	0.539	0.584
Joint Cross-Attention (Ours)	0.351	-	**0.670**	0.356	-	0.590

Challenge Results on Test Set

CCC of the proposed approach compared to state-of-the-art methods for A-V fusion on Affwild2 test set.

Method	Modalities	Valence	Arousal	Mean
Situ-RUCAIM3 [Meng et al., 2022]	Audio, Visual	0.606	0.596	0.601
FlyingPigs [Zhang et al., 2022]	Audio, Visual, Text	0.520	0.602	0.561
PRL [Nguyen et al., 2022]	Visual	0.450	0.445	0.448
HSE-NN [Savchenko, 2022]	Visual	0.417	0.454	0.436
AU-NO [Karas et al., 2022]	Audio, Visual	0.418	0.407	0.413
Joint Cross-Attention (Ours)	Audio, Visual	**0.374**	**0.363**	**0.369**
Baseline [Kollias, 2022]	Visual	0.180	0.170	0.175

Outline

Conclusion

- A joint cross attentional A-V fusion model is proposed for dimensional emotion recognition

Conclusion

- A joint cross attentional A-V fusion model is proposed for dimensional emotion recognition
- Unlike prior approaches of A-V fusion, it jointly models the inter and intra modal relationships to leverage the complementary nature of A-V modalities

Conclusion

- A joint cross attentional A-V fusion model is proposed for dimensional emotion recognition
- Unlike prior approaches of A-V fusion, it jointly models the inter and intra modal relationships to leverage the complementary nature of A-V modalities
- Extensive set of experiments conducted on Affwild2 dataset shows the robustness of the proposed approach.

Thank you for your attention!

References I

He, K., Zhang, X., Ren, S., and Sun, J. (2016).
Deep residual learning for image recognition.
In *CVPR*.

Karas, V., Tellamekala, M. K., Mallol-Ragolta, A., Valstar, M., and Schuller, B. W. (2022).
Continuous-time audiovisual fusion with recurrence vs. attention for in-the-wild affect recognition.
arXiv:2203.13285.

Kollias, D. (2022).
Abaw: Valence-arousal estimation, expression recognition, action unit detection and multi-task learning challenges.
arXiv:2202.10659.

References II

Kuhnke, F., Rumberg, L., and Ostermann, J. (2020).
Two-stream aural-visual affect analysis in the wild.
In *FGW*.

Meng, L., Liu, Y., Liu, X., Huang, Z., Cheng, Y., Wang, M., Liu, C., and Jin, Q. (2022).
Multi-modal emotion estimation for in-the-wild videos.
arXiv:2203.13032.

Nguyen, H.-H., Huynh, V.-T., and Kim, S.-H. (2022).
An ensemble approach for facial expression analysis in video.
arXiv:2203.12891.

References III

Praveen, R. G., Granger, E., and Cardinal, P. (2021).
Cross attentional audio-visual fusion for dimensional emotion recognition.
In *FG*.

Savchenko, A. V. (2022).
Frame-level prediction of facial expressions, valence, arousal and action units for mobile devices.
arXiv:2203.13436.

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

References IV

Zhang, S., An, R., Ding, Y., and Guan, C. (2022).
Continuous emotion recognition using visual-audio-linguistic information: A technical report for abaw3.
arXiv:2203.13031.

Zhang, S., Ding, Y., Wei, Z., and Guan, C. (2021).
Continuous emotion recognition with audio-visual leader-follower attentive fusion.
In *ICCV Workshop*.