Recursive Joint Attention for Audio-Visual Fusion in Regression-Based Emotion Recognition

Gnana Praveen R Eric Granger Patrick Cardinal

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

48th IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), June 2023

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

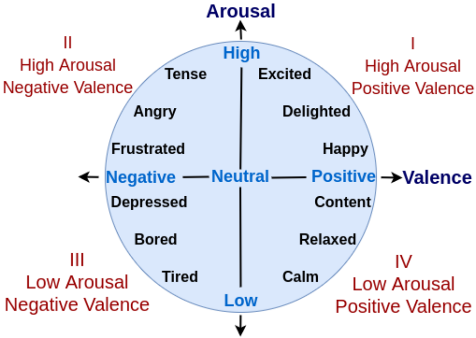
Outline

1. Dimensional Emotion Recognition
2. Motivation
3. Proposed Approach
4. Results and Discussion
5. Conclusion

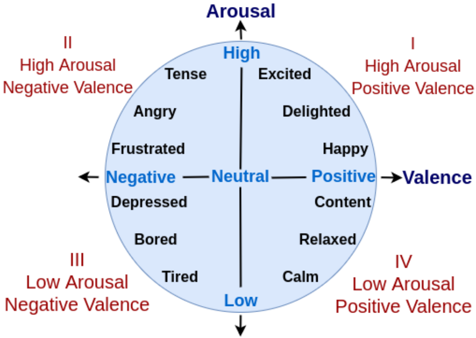
Outline

1. Dimensional Emotion Recognition
2. Motivation
3. Proposed Approach
4. Results and Discussion
5. Conclusion

Circumplex Model of Affect

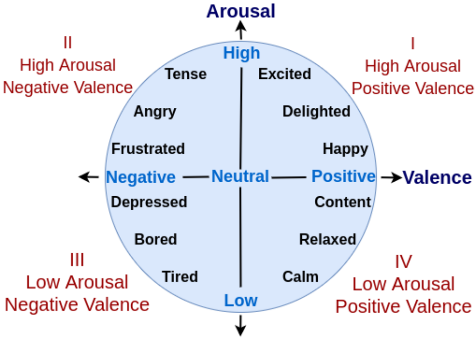


Circumplex Model of Affect



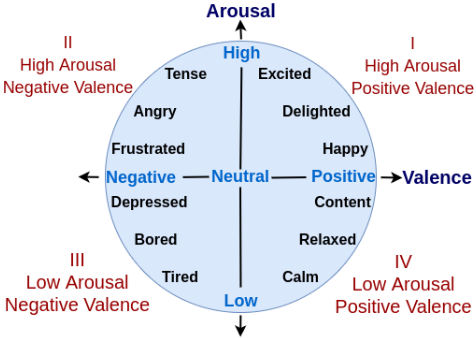
- Problem: estimating regression values in the valence-arousal space based on videos

Circumplex Model of Affect



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

Circumplex Model of Affect



- Problem: estimating regression values in the valence-arousal space based on videos
- **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 Emotion Recognition

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

A-V Fusion for Emotion Recognition

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

A-V Fusion for Emotion Recognition

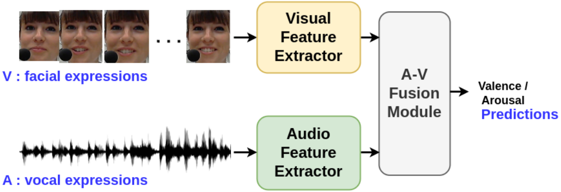
- Audio (A) and Visual (V) are the widely used contact-free modalities for emotion recognition
- A and V channels provide complementary information
- Fusion of A and V channels is expected to outperform uni-modal approaches

A-V Fusion for Emotion Recognition

- Audio (A) and Visual (V) are the widely used contact-free modalities for emotion recognition
- A and V channels provide complementary information
- Fusion of A and V channels is expected to outperform uni-modal approaches
- In this paper we focus on this scenario:

A-V Fusion for Emotion Recognition

- Audio (A) and Visual (V) are the widely used contact-free modalities for emotion recognition
- A and V channels provide complementary information
- Fusion of A and V channels is expected to outperform uni-modal approaches
- In this paper we focus on this scenario:



Challenges for A-V Fusion

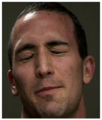
- How to extract efficient multi-modal feature representation of A-V modalities from videos?

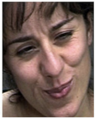
Challenges for A-V Fusion

- How to extract efficient multi-modal feature representation of A-V modalities from videos?
- How to effectively leverage the complementary information of A-V modalities?

Challenges for A-V Fusion

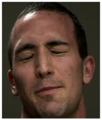
- How to extract efficient multi-modal feature representation of A-V modalities from videos?
- How to effectively leverage the complementary information of A-V modalities?
- How to handle a wide range of variations in V: facial expressions due to pose, identity bias, occlusion, etc.?

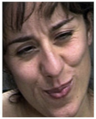





Challenges for A-V Fusion

- How to extract efficient multi-modal feature representation of A-V modalities from videos?
- How to effectively leverage the complementary information of A-V modalities?
- How to handle a wide range of variations in V: facial expressions due to pose, identity bias, occlusion, etc.?






- How to handle a wide range of variations in A: vocal expressions due to speaker identity bias, background noise, etc?

Outline

1. Dimensional Emotion Recognition
2. Motivation
3. Proposed Approach
4. Results and Discussion
5. Conclusion

A-V Fusion Approaches for Dimensional Emotion Recognition

- [Tzirakis et al., 2017] extract 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] extract A and V features from Resnet-50 and 1D CNN respectively, which is concatenated and fed to LSTM
- [Tzirakis et al., 2021] investigate various fusion strategies along with attention mechanisms including self-attention.

A-V Fusion Approaches for Dimensional Emotion Recognition

- [Tzirakis et al., 2017] extract A and V features from Resnet-50 and 1D CNN respectively, which is concatenated and fed to LSTM
- [Tzirakis et al., 2021] investigate various fusion strategies along with attention mechanisms including self-attention.
- [Parthasarathy and Sundaram, 2021] explore transformers with cross-modal attention, where cross-attention is integrated with self-attention

A-V Fusion Approaches for Dimensional Emotion Recognition

- [Tzirakis et al., 2017] extract A and V features from Resnet-50 and 1D CNN respectively, which is concatenated and fed to LSTM
- [Tzirakis et al., 2021] investigate various fusion strategies along with attention mechanisms including self-attention.
- [Parthasarathy and Sundaram, 2021] explore transformers with cross-modal attention, where cross-attention is integrated with self-attention
- [Praveen et al., 2023] explore joint cross-attentional fusion to jointly leverage the intra and inter-modal relationships across A and V modalities

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 explored to capture the complementarity 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 explored to capture the complementarity of A-V modalities
- Though attention models have been explored with transformers, they fail to capture the complementary relationship of A-V modalities

Outline

1 Dimensional Emotion Recognition

2 Motivation

3 Proposed Approach

4 Results and Discussion

5 Conclusion

Overall Framework

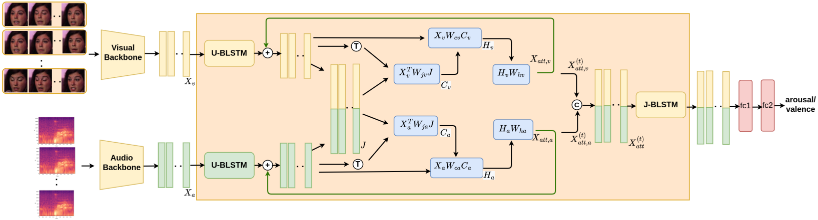
- The backbones of A and V models are trained and obtained separately to produce deep features

Overall Framework

- The backbones of A and V models are trained and obtained separately to produce deep features
- These A and V features are used to train the proposed A-V fusion model

Overall Framework

- The backbones of A and V models are trained and obtained separately to produce deep features
- These A and V features are used to train the proposed A-V fusion model



Recursive Joint Cross Attentional A-V Fusion

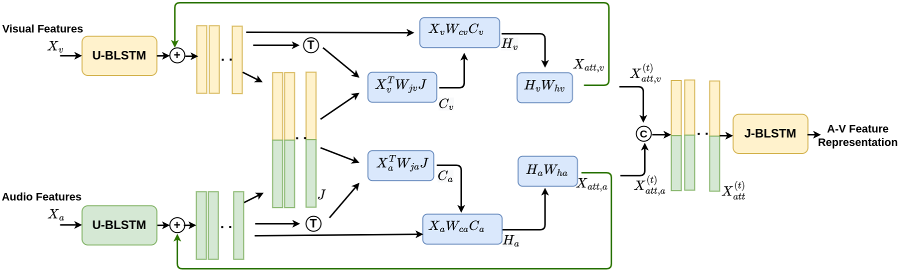


Figure: Block diagram of the proposed recursive joint attention model with Bi-directional LSTMs.

Recursive Joint Cross Attentional A-V 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

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

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

Recursive Joint Cross Attentional A-V 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

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

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

Joint Cross Attention Weights

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

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

Recursive Joint Cross Attentional A-V Fusion

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

Recursive Joint Cross Attentional A-V Fusion

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

Recursive Attended features

- The recursive joint cross-attention iteratively refines the A-V features, producing more robust A-V feature representations

$$\boldsymbol{X}^{(t)}_{\mathbf{att,a}} = \boldsymbol{W}^{(t)}_{\mathbf{ha}}\boldsymbol{H}^{(t)}_{\mathbf{a}} + \boldsymbol{X}^{(t-1)}_{\mathbf{a}}$$
$$\boldsymbol{X}^{(t)}_{\mathbf{att,v}} = \boldsymbol{W}^{(t)}_{\mathbf{hv}}\boldsymbol{H}^{(t)}_{\mathbf{v}} + \boldsymbol{X}^{(t-1)}_{\mathbf{v}}$$

where $\boldsymbol{W}^{(t)}_{\mathbf{ha}}, {\boldsymbol{W}_{\mathbf{hv}}}^{(t)}$: learnable parameters

Outline

1. Dimensional Emotion Recognition
2. Motivation
3. Proposed Approach
4. Results and Discussion
5. Conclusion

Experimental Setup

- Datasets: Affwild2 and Fatigue (private) datasets

Experimental Setup

- Datasets: Affwild2 and Fatigue (private) datasets
  - Partitions of ABAW3 challenge [Kollias, 2022] of Affwild2: training, validation, and testing partitions have 341, 71, and 152 videos respectively

Experimental Setup

- Datasets: Affwild2 and Fatigue (private) datasets
  - Partitions of ABAW3 challenge [Kollias, 2022] of Affwild2: training, validation, and testing partitions have 341, 71, and 152 videos respectively
  - Fatigue has 27 videos captured from 18 participants, suffering from degenerative diseases inducing fatigue

Experimental Setup

- Datasets: Affwild2 and Fatigue (private) datasets
  - Partitions of ABAW3 challenge [Kollias, 2022] of Affwild2: training, validation, and testing partitions have 341, 71, and 152 videos respectively
  - Fatigue has 27 videos captured from 18 participants, suffering from degenerative diseases inducing fatigue
- Performance Measure:
  - Concordance Correlation Coefficient (CCC)

Ablation Study [Affwild2]

- Backbones: R3D and Resnet18 for V and A (spectrograms) modalities respectively to obtain the deep features

Table: **Performance of our approach with components of BLSTM and recursive attention on Affwild2 data**

Method	Valence	Arousal
JA Fusion w/o recursion		
Fusion w/o U-BLSTM	0.670	0.590
Fusion w/o J-BLSTM	0.691	0.646
Fusion w/ U-BLSTM and J-BLSTM	0.715	0.688
JA Fusion w/ recursion		
JA Fusion with $t = 2$	**0.721**	**0.694**
JA Fusion with $t = 3$	0.706	0.652
JA Fusion with $t = 4$	0.685	0.601

Comparison with state-of-the-art approaches

Table: **CCC performance of the proposed and state-of-the-art methods for A-V fusion on Affwild2 data**

Method	**Type of Fusion**	**Valence**	**Arousal**
Validation Set			
Kuhnke et al. [Kuhnke et al., 2020]	Feature Concatenation	0.493	0.613
Zhang et al. [Zhang et al., 2021]	Leader Follower Attention	0.469	0.649
Rajasekhar et al [Rajasekhar et al., 2021]	Cross Attention	0.541	0.517
Rajasekhar et al. [Praveen et al., 2022]	Joint Cross Attention	0.670	0.590
Ours	Recursive JA + BLSTM	**0.721**	**0.694**
Test Set			
Meng et al. [Meng et al., 2022]	LSTM + Transformers	0.606	0.596
Vincent et al. [Karas et al., 2022]	LSTM + Transformers	0.418	0.407
Rajasekhar et al [Praveen et al., 2022]	Joint Cross Attention	0.451	0.389
Ours	Recursive JA + BLSTM	**0.467**	**0.405**

Results with Fatigue (private) Data

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

Method	**Fatigue Level**
Audio only (2D-CNN: Resnet18)	0.312
Visual only (3D-CNN: R3D)	0.415
Feature Concatenation	0.378
Cross Attention [Rajasekhar et al., 2021]	0.421
Recursive JA + BLSTM (Ours)	**0.447**

Outline

1. Dimensional Emotion Recognition
2. Motivation
3. Proposed Approach
4. Results and Discussion
5. Conclusion

Conclusion

- A recursive joint cross attentional A-V fusion model is proposed for dimensional emotion recognition to effectively capture the intra- and inter-modal relationships across A and V modalities.

Conclusion

- A recursive joint cross attentional A-V fusion model is proposed for dimensional emotion recognition to effectively capture the intra- and inter-modal relationships across A and V modalities.
- Joint cross-attention is employed in a recursive fashion, while still leveraging the intra-modal relationships using BLSTMs.

Conclusion

- A recursive joint cross attentional A-V fusion model is proposed for dimensional emotion recognition to effectively capture the intra- and inter-modal relationships across A and V modalities.
- Joint cross-attention is employed in a recursive fashion, while still leveraging the intra-modal relationships using BLSTMs.
- Extensive set of experiments conducted on Affwild2 and Fatigue (private) datasets shows that the proposed approach outperforms SOTA

Thank you for your attention!

References I

Karas, V., Tellamekala, M. K., Mallol-Ragolta, A., Valstar, M., and Schuller, B. W. (2022).
Time-continuous audiovisual fusion with recurrence vs attention for in-the-wild affect recognition.
In *CVPRW*.

Kollias, D. (2022).
Abaw: Valence-arousal estimation, expression recognition, action unit detection & multi-task learning challenges.
In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2328–2336.

References II

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

Meng, L., Liu, Y., Liu, X., Huang, Z., Jiang, W., Zhang, T., Liu, C., and Jin, Q. (2022).
Valence and arousal estimation based on multimodal temporal-aware features for videos in the wild.
In *CVPRW*, pages 2344–2351.

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

References III

Praveen, R. G., Cardinal, P., and Granger, E. (2023).
Audio-visual fusion for emotion recognition in the valence-arousal space using joint cross-attention.
IEEE Transactions on Biometrics, Behavior, and Identity Science.

Praveen, R. G., de Melo, W. C., Ullah, N., Aslam, H., Zeeshan, O., Denorme, T., Pedersoli, M., Koerich, A. L., Bacon, S., Cardinal, P., and Granger, E. (2022).
A joint cross-attention model for audio-visual fusion in dimensional emotion recognition.
In *CVPRW.*

References IV

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

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

References V

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 J. of Selected Topics in Signal Processing, 11(8):1301–1309.

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