Multimodal Learning for Detecting Stress under Missing Modalities

Julie Mordacq[1,2] Leo Milecki[1,3] Maria Vakalopoulou[1,3] Steve Oudot[1,2,*] Vicky Kalogeiton[2,4,*]

[1]Inria Saclay [2]Ecole Polytechnique [3]CentraleSupelec, Paris-Saclay [4]LIX, CNRS, IP Paris *equal supervision

Motivation

Detect **stress** induced by specific triggers in **multimodal** data under **missing** modalities [1].

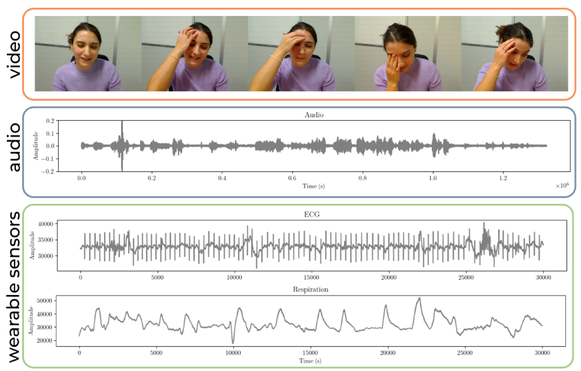


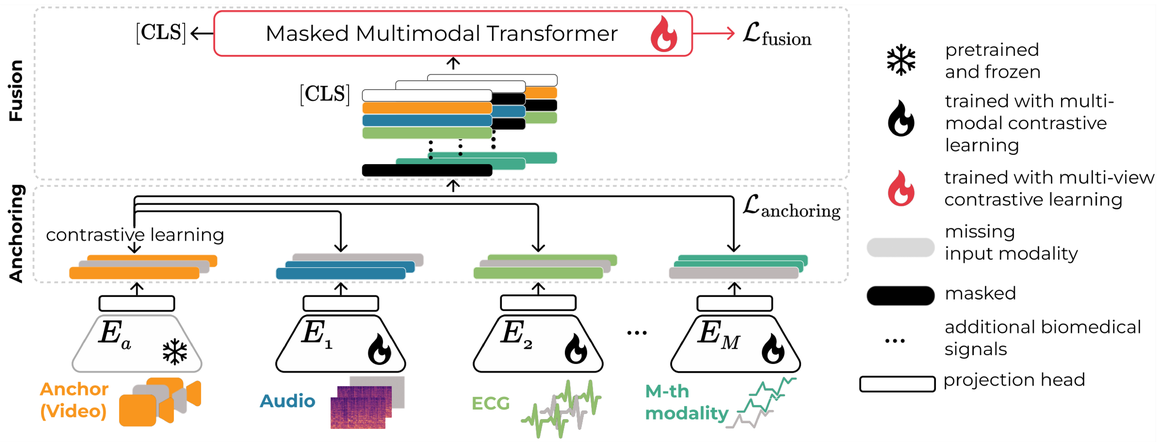
Figure: Recording physiological responses to stress-inducing stimuli

Dealing with missing modalies. Methods tackle missing modalities present limitations including (i) not being easily generalizable to **more than two** modalities [2] (ii) requiring **all** modalities during training.

Contributions

- Aligning all modalities in **the space of the strongest modality** to learn a **joint embedding space**
- a **Fusion strategy to handle missing modalities** during training and inference
- Set **new SOTA** on `StressID` [3] in various modality settings

Our method



Anchoring

Unimodal representation. Each modality is encoded using a specific encoder and projected into a fixed size representation: $f_m^i \in \mathbb{R}^d$.

Anchoring the unimodal representations. Modality-specific representations are aligned to the one of the **video** using the infoNCE loss:

$$\mathcal{L}_{\mathcal{A},\mathcal{M}_m} = -\sum_{i=1}^{B} \log \frac{\exp(\cos(f_a^i, f_m^i)/\tau)}{\sum_{k=1}^{B} \exp(\cos(f_a^i, f_m^k)/\tau)}$$

Given M modalities, we define the **anchoring loss**:

$$\mathcal{L}_{\text{anchoring}} = \sum_{m=1,\mathcal{M}_m \neq \mathcal{A}}^{M} (\mathcal{L}_{\mathcal{A},\mathcal{M}_m} + \mathcal{L}_{\mathcal{M}_m,\mathcal{A}})$$

Masked Multimodal Transformer

Obtaining a global representation. The unimodal representations are stacked and prepended a `[CLS]` token.

Masking missing modalities. The strategy to deal with missing modalities is applied to the scaled dot-product [4]:

$$O_i = \sum_j z_{ij} \frac{\exp(Q_i^T K_j/\sqrt{d_k})}{\sum_{\{j',z_{ij'}=1\}} \exp(Q_i^T K_{j'}/\sqrt{d_k})} V_j$$

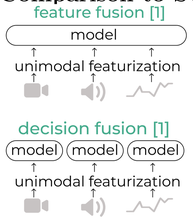
Modality Dropout. Create two simultaneous views of a batch and within one of the view, hide up to $M-1$ modalities.

$$\mathcal{L}_{\text{fusion}} = -\sum_{i=1}^{B} \log \frac{\exp(\cos(\text{CLS}^i, \text{CLS}^{i'})/\tau)}{\sum_{k=1}^{B} \exp(\cos(\text{CLS}^i, \text{CLS}^{k'})/\tau)}$$

Results

`StressID` [3] is designed for stress identification. It includes sensors, audio and video recordings. It presents *711* recordings and a missing modality ratio of $\eta = 12.5\%$.

Comparison to SOTA.



	ACC	F1
Video	62(4)‡	67(3)‡
Biomedical signals	58(4)‡	66(5)‡
Audio	62(4)‡	67(4)‡
Feature Fusion [3]	61(3)‡	66(4)‡
Decision Fusion [3]	65(5)‡	72(5)‡
Ours	69.5(3.7)	75.9(4.3)

Table: **Comparison to SOTA**

Robustness to Missing Modalites.

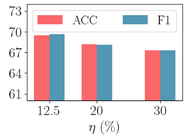


Figure: **Evaluation on different modality ratios,** η

	ACC	Δ	F1	Δ
	69.5(2.9)		69.6(3.1)	
no video	61.2(4.6)	8.3	63.0(4.2)	6.6
no audio	68.3(2.9)	1.2	68.4(3.0)	0.9

Table: **Evaluation on two modality scenarios.** For each scenario, we remove one modality from the test set.

References

[1] J. Mordacq, L. Milecki, et al. ADAPT: Multimodal learning for detecting physiological changes under missing modalities. In *MIDL*, 2024.

[2] M. Ma, J. Ren, et al. Smil: Multimodal learning with severely missing modality. In *AAAI*, 2021.

[3] H. Chaptoukaev, V. Strizhkova, et al. Stressid: a multimodal dataset for stress identification. *NeurIPS*, 2023.

[4] Leo Milecki, Vicky Kalogeiton, et al. Contrastive masked transformers for forecasting renal transplant function. In *MICCAI*, 2022.