Multimodal Ambivalence/Hesitancy Recognition in Videos for Personalized Digital Health Interventions

Manuela González-González[3], Soufiane Belharbi[1], Muhammad Osama Zeeshan[1], Masoumeh Sharafi[1], Muhammad Haseeb Aslam[1], Lorenzo Sia1[1], Nicolas Richet[1], Marco Pedersoli[1], Alessandro Lameiras Koerich[2], Simon L Bacon[3] & Eric Granger[1]

[1]LIVIA, Dept. of Systems Engineering, ETS Montreal, Canada [2]LIVIA, Dept. of Software and IT Engineering, ETS Montreal, Canada
[3]MBMC, Dept. of Health, Kinesiology & Applied Physiology, Concordia University, Montreal, Canada

ACII 2026 PUEBLA, MEXICO

Ambivalence/Hesitancy Recognition & Behaviour Change Interventions

Health-related behaviour change refers to processes that support individuals in adopting and maintaining healthy behaviours to:

- Prevent the development or worsening of chronic diseases
- Reduce early mortality
- Improve mental and physical health and well-being

During **in-person interventions**, therapists/clinicians are able to identify when individuals are **ambivalent and hesitant** towards changing a behaviour, and are able to help them overcome it.

Ambivalence/Hesitancy (A/H):

- **Ambivalence**: The **simultaneous** presence of **competing positive and negative** feelings, ideas, thoughts, or emotions towards one same object or goal
- **Hesitancy**: A **conflicted** state **between willingness and resistance to act**. A state in which a person has not entirely made up their mind about how to act
- A/H play a primary role for individuals delaying, avoiding, or abandoning health interventions
- A/H are considered the same in the literature
- Conflicting/subtle emotions: between or within modalities
- A/H manifest as affective inconsistency across modalities or within a modality, such as language, facial, vocal expressions, and body language

Online interventions:

- Cost-effective and scalable compared to in-person interventions
- Personalized digital health (eHealth) interventions
- Requires automatic expression recognition in videos from multiple modalities
- Requires machine learning models to efficiently recognize A/H to guide health behaviour interventions
- Currently, only one A/H dataset is available (`BAH` [1])

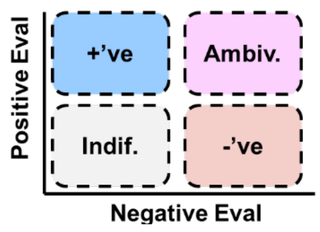

Figure 1. Ambivalence: competing positive and negative feelings.

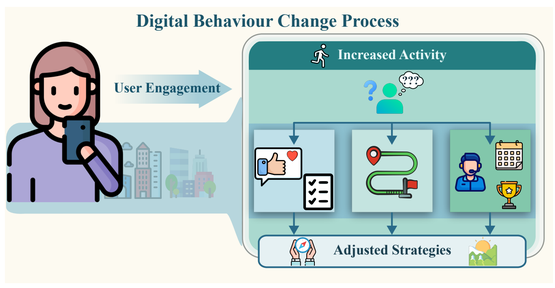

Figure 2. Conceptual illustration of the theoretical pathway by which ambivalence/hesitation (A/H) recognition can influence a digital behaviour change intervention: a user interacts with a smartphone app targeting physical activity, while the app detects A/H and adaptively adjusts cues and strategies to deliver more personalized support.

Highlights of this Work

We explore the application of deep learning models for the newly introduced task, A/H recognition. Our experiments cover different learning scenarios, in particular:

- Supervised learning
- Personalization using domain adaptation (UDA, SFDA, TTA)
- Zero-shot inference using recent Multimodal Large Language Models (MLLMs)

Experiments of A/H recognition are done at the frame and video level using the recent, unique, and video-based dataset for A/H recognition, the Behavioural Ambivalence/Hesitancy (`BAH`) dataset [1].

Behavioural Ambivalence/Hesitancy (`BAH`) Dataset [1]

- License: open access, custom, research, non-commercial and not-for-profit use
- Task: A/H recognition in videos
- 300 participants across Canada
- Online recorded videos: answers to 7 predefined questions
- 1,427 videos (10.6 hours, where 1.8 hours contain A/H)
- 916,618 frames
- Annotation: video/frame level, cues, inconsistencies

Figure 3. Download `BAH` dataset: github.com/LIVIAETS/bah-dataset.

Data subsets	Train	Validation	Test	Total
Number participants	195	30	75	300
Number participants with A/H	144	27	75	246
Number videos	778	124	525	1427
Number videos with A/H	385	75	318	778
Number frames	501,970	79,538	335,110	916,618
Number frames with A/H	76,515	13,984	65,756	156,255
Total duration (hour)	5.80	0.92	3.87	10.60
Total duration with A/H (hour)	0.87	0.16	0.75	1.79

Table 1. `BAH` dataset split into train, validation, and test sets.

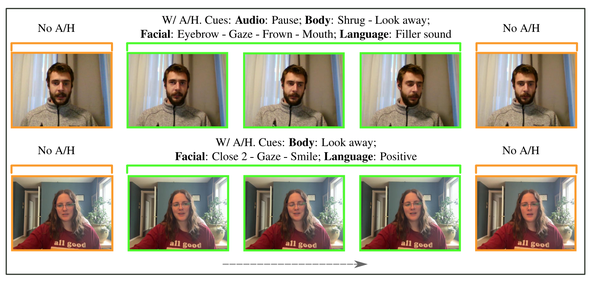

Figure 4. `BAH` examples of frames with (green) and without (orange) A/H with cues detailed in [1].

References

[1] M. González-González, S. Belharbi, M. O. Zeeshan, M. Sharafi, M. H Aslam, M. Pedersoli, A. L. Koerich, S. L. Bacon, and E. Granger. BAH dataset for ambivalence/hesitancy recognition in videos for digital behavioural change. In *ICLR*, 2026.

Experimental Results on `BAH` Dataset [1]

1. Supervised Learning

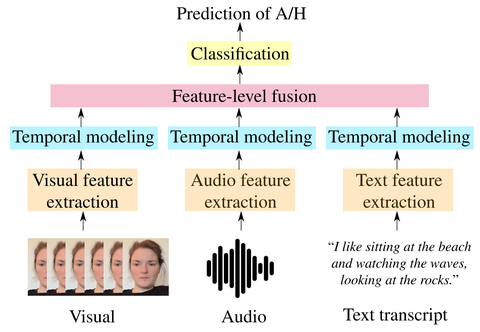

Figure 5. Multimodal model used for baseline evaluation.

	Modalities						
	V	A	T	V+A	V+T	A+T	V+A+T
`AVGF1`	0.5165	0.4658	0.5497	0.5205	0.5547	**0.5586**	0.5502
`AP`	0.2070	0.2238	0.2519	0.2225	0.2479	**0.2609**	0.2548

Table 2. **Impact of modalities**: Frame-level classification performance of multimodal models on the `BAH` dataset. V: Video, A: Audio, T: Text.

	Without context		With context (TCN)	
Backbone	`AVGF1`	`AP`	`AVGF1`	`AP`
APViT (TAFFC'22)	0.5051	0.1906	0.5019	0.2069
ResNet18 (CVPR'16)	0.5074	0.1940	0.5079	0.1993
ResNet34 (CVPR'16)	**0.5138**	0.1952	0.4998	0.1984
ResNet50 (CVPR'16)	0.4737	0.1942	0.4985	0.1915
ResNet101 (CVPR'16)	0.4929	**0.1967**	**0.5165**	**0.2070**
ResNet152 (CVPR'16)	0.4889	0.1843	0.5084	0.2058

Table 3. **Impact of architecture and context**: Performance of the visual modality on the test subset of `BAH` for frame-level classification.

Method	Fusion Approach	`AVGF1`	`AP`
LFAN (CVPRW'23)	Co-attention	0.5502	0.2548
CAN (CVPRW'23)	Concatenation	**0.5526**	**0.2631**
MT (CVPRW'24)	Transformer	0.5137	0.2134
JMT (CVPRW'24)	Cross-attention	0.5241	0.2139

Table 4. **Impact of feature fusion methods** on test set of the `BAH` dataset at frame-level classification.

Backbone	Setting	`AVGF1`	`AP`
EmoCLIP (FG'24)	ZS	0.4525	0.5894
EmoCLIP (FG'24)	FT	0.5222	0.6275
Video-FocalNet (tiny) (ICCV'23)	FT	0.5294	0.6545
Video-FocalNet (small) (ICCV'23)	FT	0.5333	0.6558
Video-FocalNet (base) (ICCV'23)	FT	**0.5663**	**0.6732**

Table 5. **Video classification** performance on `BAH` test set. ZS: zero-shot. FT: finetuned.

2. Personalization using Domain Adaptation

	Methods	`AVGF1`	`AP`
Source-only		0.4894 ± 0.0999	0.3565 ± 0.1841
UDA	MMD (Ann.Stat.'13)	0.4931 ± 0.0943	0.3589 ± 0.1831
	ACPL (FG'24)	$\mathbf{0.5417 \pm 0.0728}$	$\mathbf{0.3739 \pm 0.1789}$
SFDA	SHOT (ICML'20)	0.4919 ± 0.1056	0.3520 ± 0.1656
	NRC (NeurIPS'21)	$\mathbf{0.5174 \pm 0.1041}$	$\mathbf{0.3688 \pm 0.1487}$
	Oracle	0.5864 ± 0.0751	0.4181 ± 0.1750
		`AVGF1`	`WAR`
TTA	TPT (NeurIPS'22)	0.3970 ± 0.0485	0.6568 ± 0.1076
	TDA (CVPR'24)	0.3990 ± 0.0419	0.6522 ± 0.1092
	DPE (NeurIPS'24)	0.3940 ± 0.0510	0.6676 ± 0.1135
	PromptAlign (NeurIPS'23)	0.3970 ± 0.0785	0.6720 ± 0.1125
	ReTA (ACM MM'25)	0.3980 ± 0.0833	0.6763 ± 0.1081
	T3AL (CVPR'24)	0.4070 ± 0.0400	0.6794 ± 0.1156
	TTA-CaP (CoRR'26)	$\mathbf{0.4112 \pm 0.0455}$	0.6929 ± 0.1128
	CLIP-AUTT (CoRR'26)	0.4111 ± 0.0446	$\mathbf{0.6983 \pm 0.1391}$

Table 6. **Performance of UDA, SFDA, and TTA settings**, with Source-only and Oracle as reference points. Results are reported as mean ± standard deviation over all target subjects.

3. Zero-shot Inference: MLLMs

		Video-Level		Frame-Level
Prompt	Setting	Video-LLaVA	AffectGPT	Video-LLaVA
Simple	ZS	0.2827	0.2933	0.4416
Definition 1 Only	ZS	0.3326	0.3772	0.4456
Definition 2 Only	ZS	0.3772	0.3772	0.1651
Transcript + Simple	ZS	NA	0.6016	NA
Transcript + Def 1	ZS	0.6341	0.6436	**0.4535**
Transcript + Def 1	FT	**0.7142**	**0.7229**	NA
Transcript + Def 2	ZS	0.3945	0.6462	0.1849

Table 7. **Zero-shot Inference with MLLMs**: Performance of video and frame-level predictions. See [1] for Def 1/2.

Conclusion & Future Directions

- Results over the A/H recognition task in videos using standard multimodal models and fusion techniques are **limited**
- Robust A/H recognition requires **specialized** multimodal, spatio-temporal models, and modality-fusion modules to detect conflicts within and across modalities

Recommendations for Future Work:

- The newly introduced A/H recognition task requires the model to detect conflicting affects across and within modalities
- To build more interpretable A/H recognition systems, we recommend a 2-level framework: The 1st level models affect per modality in an independent way, using off-the-shelf pretrained models such as sentiment analysis models. At the 2nd level, a dedicated fusion mechanism should be used to assess whether there is conflicting affect across/within modalities to make a decision. It should acquire an understanding of affect conflict to be able to detect it
- Statistics extracted from annotator cues in `BAH` could be leveraged to constrain the model to find conflicts across modalities
- A specialized temporal modeling approach could be used to detect within-modality conflict based on the output of the 1st level
- Statistics from the A/H segment durations could be considered (A/H occurs briefly)
- Context is important for A/H detection, especially for within-modality cases (hard to detect)
- Check out different methods for A/H recognition proposed in the challenge "Ambivalence/Hesitancy (AH) Video Recognition Challenge, 3rd" organized in 11th ABAW Workshop [Leaderboard][Slides][ABAW11 Leaderboards]