BAH Dataset for Ambivalence/Hesitancy Recognition in Videos for Digital Behavioural Change

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

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

ICLR

Context: 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):
  - The **simultaneous** presence of **competing positive and negative** feelings, ideas, thoughts, or emotions towards one same object or goal
  - 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 are considered the same in the literature
  - Conflicting/subtle emotions

Online interventions:

- Personalized digital health (eHealth) interventions
- Automatic expression recognition for videos from multiple modalities
- Requires machine learning models to efficiently recognize A/H to guide health behaviour interventions
- However, no A/H datasets are available

This paper introduces the Behavioural A/H (BAH) dataset collected for multimodal recognition of A/H in videos.

BAH Capture and Annotation

Figure 1. Methodology for capturing the BAH dataset.

Annotation levels	Description	Annotated variables				
		Presence of AH	Level of certitude	Time stamps	Modality used	Cues
Level 1	Global annotation	Yes	Yes	No	No	No
Level 2	Frame level	Yes	Yes	Yes	No	No
Level 3	Modality focused	Yes	Yes	Yes	Yes	No
Level 4	Cue focused	Yes	Yes	Yes	Yes	Yes

Figure 2. Annotation levels of BAH dataset.

BAH Capture and Annotation

Question no.	Response	Prompt
1	Neutral	Tell us about an activity you commonly do after waking up.
2	Positive	Talk about an activity that brings you joy, for example, a hobby. Tell us why.
3	Negative	Talk about an activity you dislike doing, for example, a chore or something you find boring or annoying. Tell us why.
4	Ambivalent	Tell us about something you enjoy doing but wish you stopped doing (like a guilty pleasure) or something you don't do but wish you did.
5	Willing	Tell us about an activity you are almost always willing to do, for example with friends, at work, at home.
6	Resistant	Tell us about something people around you do, but that you would not be willing to do, for example, with friends, at work, at home.
7	Hesitant	Tell us about something you could have done already but haven't done yet, for example, something you are procrastinating or haven't made up your mind about.

Table 1. The 7 questions designed by our experts to create our videos for BAH dataset.

- 3 behavioural expert annotators
- Use of a codebook to guide annotation: conceptual clarity; cues (face, body, audio, language); and cross-modality inconsistencies
- Annotation: indicate the presence/absence of A/H
- Levels: videos/frames; timestamps: start/end A/H
- Additional information: annotation cues, inconsistencies

BAH Diversity

- BAH: Behavioural Ambivalence/Hesitancy (A/H) dataset
- Task: A/H recognition in videos
- 300 participants across Canada
- Online 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

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 2. BAH dataset split into train, validation, and test sets.

BAH Diversity

Figure 3. Examples of body language cues used by annotators to identify the occurrence of A/H: "looking away," and "changing posture."

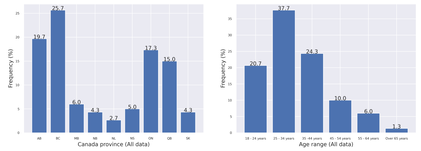

Figure 4. Canadian provinces and age range of participants.

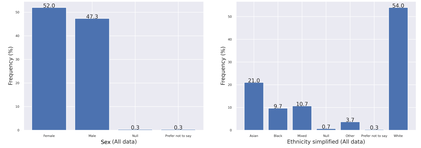

Figure 5. Sex and ethnicity of participants.

- See the paper for an analysis of the diversity and statistical properties.

Dataset	Affect	Modalities	Subject-based	Num. of participants	Num. of samples	Environment	Annotation
RAF-DB	Basic/compound emotions	Images	No	–	15,339 images	Wild	Image-level
AffectNet	Basic emotions	Images	No	–	450k images	Wild	Image label
Aff-wild2	Basic emotions, Valence/Arousal, Action Units	Video, audio	No	–	564 videos	Wild	Frame-level
MELD	Basic emotions	Video, audio	No	–	13000 utterances	Actors/TV-show	Frame-level
C-EXPR-DB	Compound emotions	Video, audio	No	–	400 videos	Wild	Frame-level
UNBC-McMaster	Pain estimation	Frames	Yes	25	200 videos	Lab	Frame-level
BioVid	Pain estimation	Frames, biomedical signals (GSR, ECG, and EMG at trapezius muscle)	Yes	90	18017 samples	Lab	Frame-level
RECOLA	Apparent Emotional Reaction Recognition	video, audio, physiology (electrocardiogram, and electrodermal activity)	Yes	46	46 videos	Lab	Frame-level
SEWA	Apparent Emotional Reaction Recognition	video, audio	Yes	398	1,990 videos	Wild	Frame-level
WEMAC	Discrete, dimensional emotions	Physiology (blood volume pulse, galvanic skin response, and skin temperature), audio	Yes	100	100 records	Lab	Self-reported
StressID	Stress	EDA, ECG, Respiration, Face video, Speech	Yes	65	587 videos	Lab	Frame-level
SchiNet	Estimation of Symptoms of Schizophrenia	video	Yes	91	91 videos	Wild	Video-level
MESC	Emotional Support Conversation	video, audio, text	Yes	–	1,019 dialogues	Wild	Utterance-level
IEMOCAP	Improvisations of scripted scenarios for basic emotions	video, audio, text	Yes	10 actors	–	Lab/Actors	Frame-level
BAH (ours)	Ambivalence/Hesitancy	Video, audio, transcript	Yes	300	1,427 videos	Wild	Video-level, Frame-level, A/H cues

Table 3. Common affective computing datasets for emotion modelling in health contexts.

BAH: Benchmarking

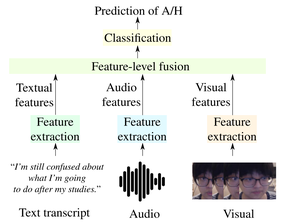

Figure 6. Multi-modal model for A/H recognition in videos at frame level.

- Importance of temporal/context modeling

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

Table 4. Visual modality performance on test set of BAH at frame-level classification.

- Importance of multi-modal learning

Modalities	AVGF1	AP
Visual	0.5165	0.2070
Audio	0.4658	0.2238
Text	0.5497	0.2519
Visual + Audio	0.5205	0.2225
Visual + Text	0.5547	0.2479
Audio + Text	**0.5586**	**0.2609**
Visual + Audio + Text	0.5502	0.2548

Table 5. Multimodal models performance on test set of BAH at frame-level classification.

- Importance of fusion

Method	Fusion Approach	AVGF1	AP
LFAN *(cvprw,2023)*	Co-attention	0.5502	0.2548
CAN *(cvprw,2023)*	Concatenation	**0.5526**	**0.2631**
MT *(cvprw,2024)*	Transformer	0.5137	0.2134
JMT *(cvprw,2024)*	Cross-attention	0.5241	0.2139

Table 6. Feature fusion performance on test set of BAH at frame-level classification.

- More results on zero-shot and domain adaptation (personalization) are in the supplementary material.
- **Conclusion**: A/H recognition is a new, challenging, and inherently multi-modal task. Existing multi-modal models yield limited performance. New methods, spatio-temporal modules, and fusion techniques are required for better recognition of conflicts between and within modalities.
- BAH access/code: github.com/LIVIAETS/bah-dataset

manuela.gonzalez@mail.concordia.ca, soufiane.belharbi@etsmtl.ca, Eric.Granger@etsmtl.ca