Master 2 Internship Proposal

Engagement Detection for Human-Robot Interaction

Location:	ISIR, Sorbonne University, Paris, France
Team:	ACIDE
Supervisors:	Hamed Rahimi (`hamed.rahimi@sorbonne-universite.fr`)
Duration:	5-6 months
Keywords:	Human-Robot Interaction, Computer Vision, Foundation Models, World Models, Deep Learning.

Context

Engagement detection [1, 2] is a critical prerequisite for socially appropriate human–robot interaction (HRI), enabling robots to decide *when* to approach a person, *whether* to initiate conversation, and *how* to do so without violating social norms or personal space. In everyday human interaction, engagement is communicated implicitly through posture, gaze, motion patterns, ongoing activity, and spatial arrangements, all of which inform whether an interaction is welcome. In HRI, these signals are closely tied to concepts of proxemics, privacy, and conversational readiness [3]. Prior work has shown that robots which respect engagement cues and personal space are perceived as more intelligent, trustworthy, and less intrusive, particularly in shared and public environments [4].

Problem Statement

Most existing engagement detection systems rely on short-term changes in observable cues such as motion in the visual field, head pose, eye gaze, or the presence of speech to infer engagement states [3]. While effective in constrained settings, these approaches often treat engagement as a reactive signal derived from low-level sensory changes, rather than as a latent social and intentional state. As a result, such systems struggle to distinguish between visually similar but semantically different situations (e.g., a person glancing toward a robot versus actively seeking interaction), and they generalize poorly across contexts, activities, and multi-person scenes. Moreover, cue-based approaches provide limited mechanisms for explicitly encoding privacy, "do-not-disturb" intent, or task-focused human behavior, increasing the risk of socially inappropriate or intrusive robot actions.

Objectives & Scientific Approach

The objective of this work is to reconceptualize engagement detection as an intention-aware world-modeling problem that integrates perception, prediction, and social reasoning. Rather than relying solely on instantaneous audio-visual changes, we aim to leverage **video foundation models** [5] and **self-supervised predictive learning** to infer latent human states such as attention, intention, availability, and interaction readiness over time. Inspired by recent advances in world modeling and joint predictive architectures, this approach seeks to learn structured representations of human activity, spatial context, and social dynamics. This will allow a robot to anticipate whether an approach or conversation would be appropriate. By modeling engagement as a temporally grounded and context-dependent phenomenon, the robot can make proactive yet conservative decisions, initiating interaction only when the likelihood of acceptance is high.

Work Plan

The successful candidate will work on the following tasks:

- **Literature Review:** Analysis of state-of-the-art methods in engagement detection, proxemics, and video foundation models (e.g., VideoMAE, multimodal LLMs).
- **Methodology Design:** Developing a pipeline that utilizes video foundation models to extract temporal features regarding human activity and social context.
- **Implementation:** Training/Fine-tuning a predictive model to infer ”interaction readiness” based on historical context rather than instantaneous cues.
- **Evaluation:** Validating the approach on public HRI datasets or real-world scenarios, comparing against baseline cue-based methods.

Candidate Profile

- **Education:** Master 2 student in Computer Science, Robotics, AI, or related fields.
- **Technical Skills:**
  - Strong programming skills in Python.
  - Experience with Deep Learning frameworks (PyTorch or TensorFlow).
  - Knowledge of Computer Vision (Transformers, Foundation Models) is a strong plus.
- **Soft Skills:** Autonomy, scientific curiosity, and good writing/communication skills in English.

How to Apply

Please send a CV and recent grade transcripts to **Hamed Rahimi** with the subject line *”[M2 Application] Engagement Detection Internship”*.

References

[1] C. Oertel, G. Castellano, M. Chetouani, J. Nasir, M. Obaid, C. Pelachaud, and C. Peters. Engagement in Human-Agent Interaction: An Overview. *Front. Robot. AI*, 7:92, 2020. doi: 10.3389/frobt.2020.00092.

[2] S. Lee, J. Lee, and D. Kwon. Implementation of engagement detection for human–robot interaction. *Sensors*, 24(10):3154, 2024.

[3] M. Daza, C. Dongo, E. Bustamante, J. Berrocal, and J. M. Murillo. An approach of social navigation based on proxemics. *Applied Sciences*, 11(4):1860, 2021.

[4] C. Sirithunge, H. Ranasinghe, and D. Alahakoon. An evaluation of human conversational preferences in human–robot interaction. *Computational Intelligence and Neuroscience*, 2021.

[5] N. Madan, A. Møgelmose, R. Modi, Y. S. Rawat, and T. B. Moeslund. Foundation models for video understanding: A survey. *arXiv preprint arXiv:2405.03770*, 2024.