Decoding the Unspoken: Details on Methodology for FSDE Design Expo 2025

M. Soroush Izadan

FSDE Design Expo 2025

1 Introduction

Effective human-robot interaction (HRI) requires robots to perceive and interpret subtle human social cues in real-time. We present a unified recursive Bayesian framework for estimating key social factors: Facial Approachability (A_F), Head Pose-Based Engagement (E_F), and Motion-Based Approachability (A_M). These estimates, represented as Gaussian distributions (μ, σ^2), allow robots to dynamically adapt their behavior based on a probabilistic understanding of the human state, incorporating temporal dynamics and principled uncertainty management.

2 Core Bayesian Estimation Framework

All estimators employ a recursive Bayesian approach, modeling the target social factor ($X \in \{A_F, E_F, A_M\}$) as a Gaussian random variable. The state at time t is inferred from relevant observations $O^{(t)}$ and the previous state $X^{(t-1)}$.

The core update cycle involves two steps:

1. Prediction (Time Update): The prior belief from the previous time step is propagated forward, incorporating process noise ($\sigma^2_{process}$) to account for potential state drift over time Δt:

$$p(X^{(t)}|X^{(t-1)}) = \mathcal{N}(\mu^{(t-1)}_{prior}, (\sigma^{(t-1)}_{prior})^2 + \sigma^2_{process} \cdot \Delta t) = \mathcal{N}(\mu^{(t)}_{predict}, (\sigma^{(t)}_{predict})^2) \tag{1}$$

2. Update (Measurement Update): A likelihood function $p(O^{(t)}|X^{(t)})$ is computed based on the current observations. This likelihood, modeled as $\mathcal{N}(\mu^{(t)}_{likelihood}, (\sigma^{(t)}_{likelihood})^2)$, represents how well different hypothetical states $X^{(t)}$ explain the observations. The predicted prior is then combined with the likelihood using standard Gaussian update equations to yield the posterior belief:

$$(\sigma^{(t)}_{posterior})^2 = \left(\frac{1}{(\sigma^{(t)}_{predict})^2} + \frac{1}{(\sigma^{(t)}_{likelihood})^2} \right)^{-1} \tag{2}$$

$$\mu^{(t)}_{posterior} = (\sigma^{(t)}_{posterior})^2 \left(\frac{\mu^{(t)}_{predict}}{(\sigma^{(t)}_{predict})^2} + \frac{\mu^{(t)}_{likelihood}}{(\sigma^{(t)}_{likelihood})^2} \right) \tag{3}$$

The posterior mean $\mu_{posterior}^{(t)}$ is typically clamped to $[0, 1]$ and the variance $(\sigma_{posterior}^{(t)})^2$ is lower-bounded.

3 Facial Approachability Estimation (A_F)

3.1 Goal

Estimate the perceived approachability based on facial expressions and gaze direction, drawing from social psychology findings.

3.2 Input Cues ($O^{(t)}$)

- Detected facial emotion probabilities (e.g., neutral, happy, sad, etc.).
- Gaze direction (e.g., direct vs. averted).

3.3 Core Logic & Likelihood

- Emotions are mapped to base approachability means (μ_E), informed by psychological ratings (e.g., happy $\rightarrow$ high μ_E, anger $\rightarrow$ low μ_E).
- Gaze direction modulates the means for certain emotions (e.g., direct gaze slightly increases μ_E for happy).
- The overall $\mu_{likelihood}$ is a weighted average of the emotion-gaze modulated means, weighted by temporally smoothed emotion probabilities.
- The $\sigma_{likelihood}^2$ incorporates base uncertainties for each emotion and the variance across contributing emotions.
- Temporal smoothing of emotion probabilities using an EWMA-like model enhances robustness to noise.
- Uncertainty quantification considers classification entropy, potentially grouped by valence.

4 Head Pose-Based Engagement Estimation (E_F)

4.1 Goal

Estimate cognitive engagement based on head orientation dynamics, interpreting focus vs. scanning behavior.

4.2 Input Cues ($O^{(t)}$)

- Head pose (yaw, pitch) classified into Regions of Interest (ROIs) with hysteresis.
- History of confirmed ROI dwell times.
- History of confirmed ROI changes.

4.3 Core Logic & Likelihood

Engagement (E_F) is estimated from two primary head pose features:

- **ROI Dwell Time Entropy (H_{norm}):** Calculated over a time window. Low entropy (focus on few ROIs) maps to high $\mu_{likelihood}$ via an inverse sigmoid function. High entropy (scatter across many ROIs) maps to low $\mu_{likelihood}$.
- **EWMA of ROI Change Rate ($\mathcal{E}_{\Delta ROI}$):** Captures recent head movement activity. Low rate (stability) maps to high $\mu_{likelihood}$ via an inverse sigmoid. High rate (active scanning) maps to low $\mu_{likelihood}$.
- The two likelihoods are combined using inverse variance weighting. Disagreement between the entropy and rate cues increases $\sigma^2_{likelihood}$, based on comparing normalized cue outputs.

5 Motion-Based Approachability Estimation (A_M)

5.1 Goal

Estimate approachability based on locomotion patterns, interpreting speed and movement variability.

5.2 Input Cues ($O^{(t)}$)

Filtered velocity estimates ($\mathbf{v}_t = [v_x, v_y]^T$) from Kalman-filtered motion tracking data, used to derive:

- Current Speed (S_t).
- EWMA of Speed Change ($\mathcal{E}^{(t)}_{\Delta S}$).
- EWMA of Heading Change ($\mathcal{E}^{(t)}_{\Delta H}$), using circular differences.

5.3 Core Logic & Likelihood

Approachability (A_M) is inferred from fusing speed and variability cues:

- **Speed Cue (S_t):** Low speed maps to high $\mu_{likelihood}$ (inverse sigmoid). High speed maps to low $\mu_{likelihood}$. Crucially, the *influence* of this cue is weighted (w_S) based on the speed's distance from a neutral midpoint, reducing its impact for ambiguous mid-range speeds.
- **Variation Cues ($\mathcal{E}_{\Delta S}, \mathcal{E}_{\Delta H}$):** High change rates map to high $\mu_{likelihood}$ (standard sigmoid). Low change rates map towards a neutral $\mu_{likelihood}$ (e.g., ≈ 0.5), indicating stability is less informative than active variation for *increasing* approachability in this model.
- Likelihoods are combined using modulated inverse variance weighting (incorporating w_S). Disagreement between normalized cue outputs increases $\sigma^2_{likelihood}$.

The use of Kalman-filtered velocity provides robustness against noisy raw motion data.

6 System Implications

This suite of estimators provides a multi-faceted, probabilistic assessment of human social state. By combining A_F, E_F, and A_M, potentially through higher-level fusion or rule-based systems, robots can gain a richer understanding for guiding navigation, interaction initiation, dialogue management, and overall socially appropriate behavior in dynamic human environments. The continuous nature and explicit uncertainty representation are key for robust real-world deployment.