Master's Thesis

Neurosymbolic AI for Social Cognition

Jarne Demoen
Vrije Universiteit Brussel

4/11/2025

1 Introduction

Humans effortlessly integrate facial expressions, body language, and situational cues to infer emotions and intentions. A raised eyebrow conveys doubt, while a smile can convey happiness or trust. Machines, however, have a harder time interpreting body language, context, and multi-person group structure to derive the overall meaning of a situation. Modern techniques such as machine learning and deep learning can recognize visual patterns with high accuracy, but they do so in a purely subsymbolic manner: neural networks map inputs to outputs without providing insight into how decisions are made. As a result, even if a system predicts the correct emotion of a social scene, it remains unclear *why* the model reached that conclusion. This lack of transparency motivates the need for learning frameworks that support both perception and reasoning.

This thesis explores how neurosymbolic AI can help machines take their next step towards genuine social cognition. Social cognition is the human ability to interpret emotions, intentions, and the subtle dynamics that shape everyday social interactions.

Machines are increasingly expected to navigate human environments. For example, they support people at home, assisting clinicians, mediating online interactions, or working together with humans in shared physical spaces. In all of these settings, an accurate recognition of emotions is essential for safe, trustworthy, and socially aware AI systems. Yet today's emotion-recognition models remain largely pattern-driven. They could detect a smile, but they often miss whether it is a joyful smile at a wedding, a polite smile in a meeting, or a strained smile in a stressful situation. This motivates the need for approaches that integrate perceptual cues with structured, human-like knowledge.

Neurosymbolic AI provides a promising path forward by combining two complementary paradigms:

- **Symbolic AI**, which represents knowledge through a human-readable structure, such as logical rules, for example, "people are usually happy at weddings" or relationships between events.
- **Subsymbolic AI** (deep learning), which excels at uncovering patterns in raw data, such as recognizing facial expression in images.

- **Neurosymbolic AI** which unifies both approaches by using neural networks for perception and symbolic reasoning modules to interpret structured relationships.

In this work, symbolic knowledge refers to contextual cues that describe the social setting, such as the environment type and information derived from multiple people's facial expressions. Rather than solely relying on pixel-level features, the system incorporates a structured representation of what is actually happening in the scene and who is expressing which emotion. This allows us to investigate how relationships between context and group emotional configuration may influence the final emotion prediction.

This thesis addresses the following research questions:

1. Can explicit symbolic knowledge about context and group composition improve the accuracy of emotion recognition in social scenes?
2. Can such knowledge reduce the amount of training data required compared to a purely subsymbolic baseline?
3. Does a neurosymbolic setup make it easier to interpret and explain the model's decisions, for example, by inspecting which rules were used?

To study these questions, the experiments use the FindingEmo dataset, which contains naturalistic, multi-person social scenes annotated for valence, arousal, and discrete emotion categories. Its emphasis on contextual and group-based emotional understanding makes it well-suited for evaluating neurosymbolic approaches.

2 Literature Review

Understanding emotion in social situations requires both a description of what emotions are and an understanding of how they appear in real-world images. Research in affective science and computer vision has therefore developed several frameworks, ranging from psychological models of emotion to datasets and computational tools for learning from them.

2.1 Psychological Models of Emotion

One of the most important theories for representing human affect is Russell's Circumplex Model of Affect, which proposes that emotions can be mapped to a two-dimensional space defined by Valence (how positive the emotion feels) and Arousal (how intense the emotion feels). Empirical studies showed that when people judge the similarity of emotion-related terms or report their own feelings, these feelings naturally arrange themselves around a circular structure in this two-dimensional space. Emotions with similar valence and arousal values occupy nearby positions, while opposite emotions lie across the circle as shown in Figure 1 . This model provides a continuous, psychologically grounded representation that is widely used in affect annotation.

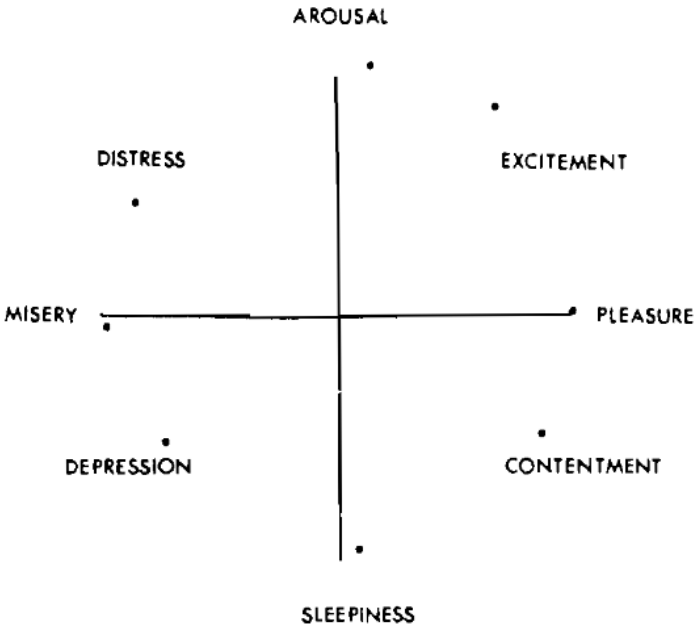

Figure 1: Eight affect concepts in a circular order

Alongside continuous representations, many emotion studies also rely on discrete taxonomies. Plutchik's Wheel of Emotions (Figure 2) organizes emotions into eight basic families (such as joy, anger, fear, or trust), each within varying levels of intensity. The wheel reflects psychological opposites (such as joy and sadness) and offers a structured approach for labeling discrete emotions.

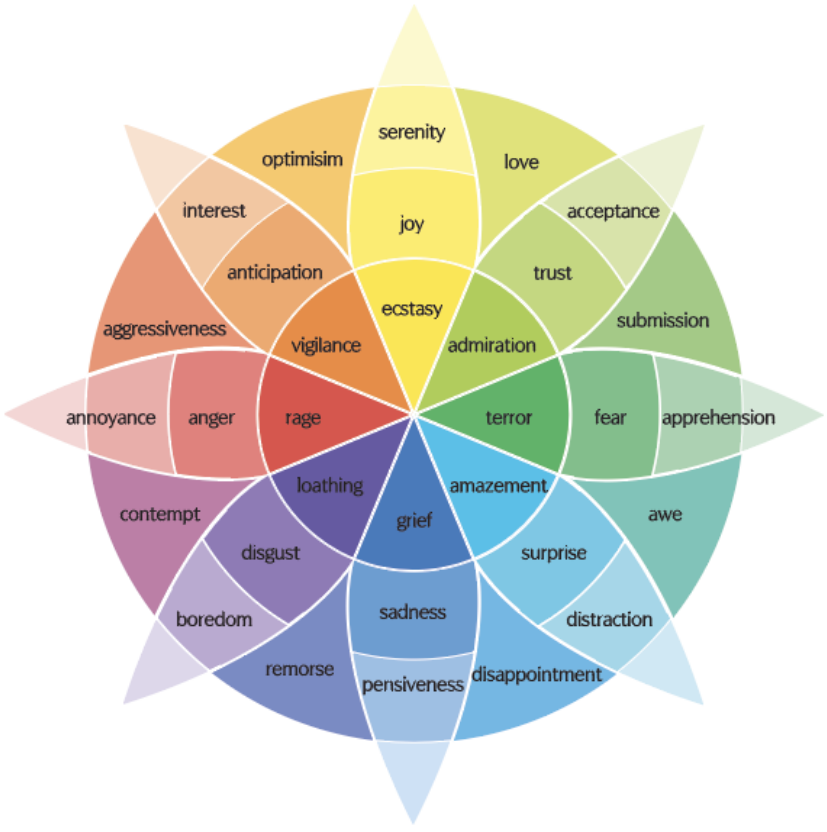

Figure 2: Plutchik's Wheel of Emotions (PWoE)

2.2 Emotion recognition in Multi-Person Social Scenes

Traditional computer-vision benchmarks for emotion recognition have focused on isolated faces with controlled pose, lighting, and expression. However, many real-world situations involve multiple individuals, rich social contexts, and interactions. All these cues together shape the overall emotional interpretation of a scene. Recognizing emotion in such settings requires combining facial cues with contextual information about the environment and the relationships between people.

The FindingEmo dataset was developed specifically to address this challenge. It contains images of natural social situations such as weddings, protests, sports events, family gatherings... These settings include multiple people and meaningful scenes. Each image is annotated with:

- continuous Valence and Arousal values, and
- discrete emotion labels derived from Plutchik's taxonomy (Figure 2).

This dual annotation scheme enables models to capture both broad affective tone and specific emotional categories. The dataset also highlights several challenges typical of real-world emotion understanding: class imbalance across emotion categories, greater reliability for Valence than for Arousal, and frequent confusion between semantically close emotions such as anger, fear, and disgust. Experiments with standard CNNs and vision transformers show that while facial cues and scene features both carry emotional information, simple fusion methods yield only limited improvement compared to what????. This illustrates the difficulty of reasoning across multiple cues.

This dataset, therefore, highlights the complexity of social emotion recognition: it is inherently multi-modal (faces + context), relational (multiple individuals in a single image), and approximate (human annotations include uncertainty). This motivates the interest in computational frameworks that can combine perception with structured reasoning.