Calculating people's attraction based on their facial expressions, body movement, voice tone and words spoken

Mitrea Ioan David

October 29, 2025

1. **A Multimodal Emotion Recognition System: Integrating Facial Expressions, Body Movement, Speech, and Spoken Language** [1]

   Why I chose it: It demonstrates the effectiveness of a multimodal architecture for emotion recognition, using the same modalities I target (facial expressions, body movement, speech, and language).

   Description: Proposes a multimodal emotion recognition system that combines four data sources—facial expressions, body movements, voice, and spoken language. The author argues this approach substantially increases accuracy. The paper presents the system architecture and reports evaluation statistics on the test datasets used.

2. **Learning situated emotions** [2]

   Why I chose it: It argues that emotions are grounded in prior experiences; the emotions we feel now depend on what we have felt before.

   Description: People do not have only a universal version of an emotion (e.g., fear or anger). Instead, they learn situation-specific variants of these emotions that fit particular contexts (e.g., fear of physical injury versus fear of social rejection). Over time, these situational forms become part of a person's "real self."

3. **Facial expression and body gesture emotion recognition: A systematic review on the use of visual data in affective computing** [3]

   Why I chose it: It surveys the full range of approaches used to classify emotions from visual data, providing a comprehensive overview of algorithms and techniques.

   Description: Reviews common methods for recognizing emotions from facial expressions and body gestures, summarizing the state of the art and methodological trends in affective computing with visual data.

4. **Recurrent Neural Networks (RNNs): A gentle Introduction and Overview** [4]

   Why I chose it: It clarifies RNNs and their applicability to sequential data. I apply this principle because current emotions are influenced by previously experienced emotions.

   Description: Offers a detailed introduction to RNNs and LSTMs, explaining core architectures, what an RNN/LSTM cell is, and how backpropagation through time works. It also discusses challenges of simple RNNs, such as vanishing and exploding gradients.

5. **Automatic Speech Recognition: A survey of deep learning techniques and approaches** [5]

   Why I chose it: It helps map the landscape of speech recognition techniques and their evolution, informing the choice of a recognition model for my work.

   Description: Highlights and compares deep learning–based methods for converting human speech to text, covering datasets, metrics, toolkits, and recent models.

6. **A test of multimodal communication in humans using 881 judgements of men and women's physical, vocal, and olfactory attractiveness** [6]

   Why I chose it: Reinforces the multimodal perspective by evaluating physical appearance, vocal, and olfactory attractiveness simultaneously.

   Description: Tests the redundancy hypothesis—that different modalities may not add unique information but provide overlapping cues available through other senses—within human judgments of attractiveness.

7. **Body sway predicts romantic interest in speed dating** [7]

   Why I chose it: Matches our first-meeting setup and uses body movement to predict mutual liking. It also tests background music, which we can adopt.

   Description: They recorded partners' movements during real speed dates and measured how much one person's movement followed the other. This directional coupling predicted romantic interest beyond attractiveness, and groovy background music increased willingness to meet again.

8. **Virtual speed dating: Utilizing online-meeting platforms to study initial attraction and relationship formation** [8]

   Why I chose it: It proposes some other inputs that people could give at the end of the interaction: attractiveness, warmth, status, similarity, and romantic interest.

Description: Participants first completed a baseline survey, then joined virtual speed-dating events on Zoom where they met multiple partners for 4-minute video dates, rated each on traits like attractiveness, warmth, similarity, and romantic interest, were matched when both expressed mutual interest, attended recorded virtual first dates with their matches, and completed weekly follow-up surveys for four weeks to track relationship development.

9. **Attraction in every sense: How looks, voice, movement and scent draw us to future lovers and friends** [9]

   Why I chose it: Reinforces our choosing of multimodal approach importance in calculating attraction between lovers and friends and also the fact that olfactory senses don't matter that much.

   Description: Participants were exposed to recordings of individuals through four sensory modalities—faces (visual), voices (auditory), body movements (visual motion), and body odours (olfactory)—as well as combined audio-video (multimodal) presentations; they then rated each stimulus for attractiveness, allowing researchers to compare how attraction judgments aligned across senses and identify correlations and redundancies between modalities.

10. **Romantic Attraction (Chapter 4 from The Science of Romantic Relationships)** [10]

    Why I chose it: Attraction accompanies with it a positive emotion so that ensures our research has basis (correlating positive expressions with attraction).

    Description: This chapter explores interpersonal attraction, the subjective appeal of another person, which is often accompanied by a positive emotional reaction and an affiliative motivation for greater closeness to that person. This chapter organizes the many specific traits that enhance attraction in terms of characteristics that offer domain-general rewards and characteristics that advance specific evolutionary goals. The chapter then reviews the characteristics that are most consistently desirable, including physical attractiveness, social status, warmth/kindness, intelligence, proximity, familiarity, similarity, and reciprocity by reviewing relevant research findings, as well as exceptions and boundary conditions.

11. **Methods to improve reliability of video-recorded behavioral data** [11]

    Why I chose it: Provides a way to make the people feel more natural when they first meet, giving more accuracy to our input data.

    Description: Video-recorded real patient–clinician consultations to test ways to make recordings more natural and reliable. They used small,

unobtrusive cameras, familiarized participants beforehand, standardized setup and coder training, and analyzed how these methods improved the consistency and authenticity of recorded interactions.

12. **Multimodal Machine Learning: A Survey and Taxonomy** [12]

    Why I chose it: It proposes multiple ways (Dynamic Time Warping, graphical models, etc.) to align the input sources (voice and video) and compares alignment approaches useful for multimodal fusion.

    Description: Provides an overview of techniques used in multimodal models, discussing strengths and weaknesses of different alignment and fusion methods when combining sets of input features. The survey is useful for choosing alignment strategies and understanding trade-offs when integrating modalities.

13. **Tensor Fusion Network for Multimodal Sentiment Analysis** [13]

    Why I chose it: It proposes a novel fusion method (Tensor Fusion Network) that explicitly models unimodal, bimodal and trimodal interactions — a promising candidate to include in our experiments.

    Description: Introduces the Tensor Fusion Network which combines unimodal representations into a 3-dimensional tensor formed by the 3-way Cartesian product of modality embeddings; this structure captures intra- and cross-modal interactions and can improve multimodal sentiment/emotion prediction.

14. **SWRR: Feature Map Classifier Based on Sliding Window Attention and High-Response Feature Reuse for Multimodal Emotion Recognition** [14]

    Why I chose it: Proposes an alternative to strict frame-to-frame alignment by using sliding-window attention and feature-reuse, an approach we can test to handle temporal misalignment between audio and video.

    Description: Proposes SWRR, which processes audio and visual features in overlapping sliding windows (1–2 s windows, 0.5–1 s hop) rather than relying on exact timestamp matching. Attention highlights salient moments in each window and a feature-reuse module carries strong cues between overlapping windows for smoother, more robust predictions.

15. **Improving Techniques for Convolutional Neural Networks Performance** [15]

    Why I chose it: Proposes sliding-window attention to handle temporal misalignment when modalities lack frame-level synchronization; it's practical for our experiments testing alternative temporal fusion strategies and enhancing real-time applicability.

Description: Convolutional Neural Networks (CNNs) have been extensively used in several application domains. Researchers have been exploring methods to enhance the accuracy of applications in accuracy-critical domains by either increasing the depth or width of the network. The presence of structures results in a significant increase in both computational and storage costs, hence causing a delay in response time. Convolutional Neural Networks have significantly contributed to the rapid development of several applications, including image classification, object detection, and semantic segmentation. However, in some applications that need zero tolerance for mistakes, such as automated systems, there are still certain issues that need to be addressed to achieve better performance. Then, despite the progress made so far, there are still limitations and challenges that must be overcome. Simultaneously, there is a need for reduced reaction time. Convolutional Neural Networks (CNNs) are now faced with significant obstacles of a formidable nature. This paper investigates different methods that can be used to improve convolutional neural network performance.

References

[1] K. Kraack. A multimodal emotion recognition system: Integrating facial expressions, body movement, speech, and spoken language. *arXiv preprint*, 2023.

[2] L. A. M. Lebois, C. D. Wilson-Mendenhall, W. K. Simmons, L. F. Barrett, and L. W. Barsalou. Learning situated emotions. *Neuropsychologia*, 145:106637, 2018.

[3] A. Sherstinsky. Facial expression and body gesture emotion recognition: A systematic review on the use of visual data in affective computing. *IEEE Transactions on Affective Computing*, 13(4):1648–1670, 2022.

[4] R. M. Schmidt. Recurrent neural networks (rnns): A gentle introduction and overview. *arXiv preprint arXiv:1912.05911*, 2019.

[5] H. Ahlawat, N. Aggarwal, and D. Gupta. Automatic speech recognition: A survey of deep learning techniques and approaches. *ACM Computing Surveys*, 55(7):1–34, 2022.

[6] Megan Nicole Williams and Coren Lee Apicella. A test of multimodal communication in humans using 881 judgements of men and women's physical, vocal, and olfactory attractiveness. *Heliyon*, 9(6):e16895, 2023.

[7] A. Chang, H. E. Kragness, W. Tsou, D. J. Bosnyak, A. Thiede, and L. J. Trainor. Body sway predicts romantic interest in speed dating. *Social Cognitive and Affective Neuroscience*, 16(1–2):185–192, 2021.

[8] J. E. French, L. J. Bolton, and A. L. Meltzer. Virtual speed dating: Utilizing online-meeting platforms to study initial attraction and relationship formation. *Personal Relationships*, 31(2):420–444, 2024.

[9] A. Schirmer, M. Franz, L. Krismann, V. Nöring, M. Große, M. Mahmut, and I. Croy. Attraction in every sense: How looks, voice, movement and scent draw us to future lovers and friends. *British Journal of Psychology*, 116(3):684–701, 2025.

[10] Theresa DiDonato and Brett Jakubiak. Romantic attraction. In *The Science of Romantic Relationships*, pages 123–157. Cambridge University Press, 2023.

[11] K. K. Haidet, J. Tate, D. Divirgilio-Thomas, A. Kolanowski, and M. B. Happ. Methods to improve reliability of video-recorded behavioral data. *Research in Nursing & Health*, 32(4):465–474, 2009.

[12] Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. Multimodal machine learning: A survey and taxonomy. 2017.

[13] Amir Zadeh, Minghai Chen, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. Tensor fusion network for multimodal sentiment analysis. 2017.

[14] Z. Zhao, T. Gao, H. Wang, and B. W. Schuller. Swrr: Feature map classifier based on sliding window attention and high-response feature reuse for multimodal emotion recognition. pages 2433–2437, 2023.

[15] Anonymous. Improving techniques for convolutional neural networks performance. *European Journal of Electrical Engineering and Computer Science*, 8(1):1–16, Jan 2024.