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

Mitrea Ioan David

November 19, 2025

1 Data Collection

The quality and reliability of affective computing research depend critically on rigorous and multimodal data collection pipelines capable of capturing complex human emotional expressions. Traditional laboratory acquisition frameworks prioritize controlled environments and high-fidelity measurements; however, they often suffer from limited ecological validity and reduced generalizability [1]. Recent research has therefore shifted toward hybrid approaches that integrate audiovisual, physiological, and behavioral modalities collected in more naturalistic settings, thereby enabling richer affective modeling [2]. Following this methodological direction, our study implements a multimodal extraction pipeline designed to capture fine-grained facial, acoustic, and bodily behavioral cues from video recordings. The complete pipeline is summarized below.

1.1 Person Video Segmentation

To ensure that downstream feature extraction is performed only during time intervals in which the target participant is visible, we implemented per person temporal segmentation based on facial recognition. For each video frame, we computed face embeddings using `face_recognition` and compared them to a reference embedding of the target participant. A cosine distance threshold determined whether the detected face corresponded to the participant of interest. This yielded a set of visibility intervals

$$S = \{(t_i^{\text{start}}, t_i^{\text{end}}) \mid i = 1, \ldots, N\},$$

representing all temporal segments in which the participant was visually present. These intervals functioned as the structural backbone for subsequent audio and behavioral feature extraction.

1.2 Audio Extraction Aligned With Speaker Visibility

Because conventional speaker diarization models can be unreliable in multi-speaker naturalistic environments, we leveraged the segments where each person is visible S to construct a per participant audio track. Using `moviepy`, audio was extracted only from the intervals in which the participant was detected onscreen. Silent padding ensured the temporal alignment of the resulting WAV file with the original video duration. This audio track facilitates the analysis of vocal affect — such as prosody, energy, and temporal speaking patterns — while minimizing cross speaker contamination.

1.3 Facial Action Unit Extraction

Facial behavior was captured by extracting facial Action Units (AUs) based on a two stage mapping process. First, emotion probabilities were obtained using the DeepFace facial expression recognizer. Second, these categorical emotion estimates were mapped to continuous AU vectors using a rule mapping derived from established facial expression literature. Each frame within the visibility intervals produced a 17 dimensional AU vector:

$$\mathbf{a}_t = [\mathrm{AU1}_t, \ldots, \mathrm{AU17}_t].$$

Frames containing no detectable face (e.g., motion blur, occlusion) were assigned the zero vector to preserve temporal coherence. This procedure enabled robust modeling of dynamic facial affect across the entire temporal sequence.

1.4 Body, Hand, and Upper Body Movement Features

To capture nonverbal expressive behaviors beyond the face, we employed MediaPipe's FaceMesh, Hands, and Pose modules to extract kinematic and gestural information. For each frame, we tracked 3D keypoints for the hands, face contour, and upper body. Time aligned features included:

- **Hand movement dynamics**: frame-to-frame velocity, acceleration, and movement smoothness.
- **Gesture frequency**: counts of meaningful movements exceeding a motion threshold.
- **Face-touch gestures**: detection of hand to face contact events using spatial proximity between hand landmarks and facial meshes.
- **Upper-body motion**: displacement and temporal derivatives of shoulder, torso, and arm keypoints.

By integrating these multimodal cues—facial expression, vocal behavior, and gesture/movement dynamics—we obtain a rich representation of affective behavior capable of supporting downstream analyses such as affect recognition, behavioral profiling, and interpersonal dynamics modeling. This multimodal pipeline aligns with contemporary trends in affective computing that emphasize ecological validity, multimodal integration, and fine-grained temporal annotation [1, 2].

2 Methodology

Our approach models short temporal sequences of multimodal behavioral signals to predict binary labels - False if the person is not attracted to the other person and True otherwise. The methodology consists of feature aggregation every second, sequence construction, preprocessing, and sequential LSTM-based modeling. This section provides a precise description of each step, including hyperparameters and architectural details.

2.1 Sequence Construction and Feature Representation

Each date's video recordings are preprocessed to produce **per-second feature vectors $\mathbf{x}_t \in \mathbb{R}^{28}$** comprising:

- 17 **Facial Action Units (AUs) for each person**: per frame categorical emotions are mapped via a rule scheme to continuous AU intensities, then averaged within each second. Missing frames (no face detected) are zero filled before averaging.
- 4 **Hand and upper-body movement features for each person**: left/right hand velocity, cumulative gesture frequency, and cumulative face-touch events computed from MediaPipe landmarks.
- 3 **Acoustic features for the date**: energy (dB), pitch (Hz), and speaking rate extracted from each participant's audio segments constrained by visual presence.
- 4 **Sentiment features for each person**: compound, positive, neutral, and negative sentiment scores aggregated from manually adjusted ScribeFlow transcripts for utterances overlapping each second. Zero filled when participant is silent.

These per second vectors are concatenated to form sequences of length $T = 15$ seconds:

$$\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_T] \in \mathbb{R}^{T \times 28}.$$

Sequences are generated with a sliding window (stride 1) to capture temporal dependencies, producing overlapping samples for model training. Missing modalities are handled deterministically via zero filling, allowing the model to learn from absent or occluded signals without introducing NaNs or imputed values.

2.2 Feature Normalization

All features are standardized using a **StandardScaler (z-score normalization)** fit on the training set. For efficiency, the training sequences are reshaped from (N_{train}, T, D) to $(N_{\text{train}} \cdot T, D)$ prior to fitting. Validation and test sequences are transformed using the same scaler.

2.3 Sequential LSTM Architecture

The model is a stacked LSTM followed by fully connected layers:

- **Input:** sequences of shape $(T = 15, D = 28)$.
- **LSTM layers:**
  1. LSTM(64), `return_sequences=True`, dropout=0.2
  2. BatchNormalization
  3. LSTM(32), `return_sequences=True`, dropout=0.2
  4. BatchNormalization
  5. LSTM(16), `return_sequences=False`, dropout=0.2
  6. BatchNormalization
- **Dense layers:** Dense(32) with ReLU + Dropout(0.3), Dense(16) with ReLU + Dropout(0.2)
- **Output:** Dense(1) with sigmoid activation, producing probability $\hat{y} \in [0, 1]$

Formally, the model learns a mapping:

$$f_\theta : \mathbb{R}^{T \times D} \to [0, 1], \qquad \hat{y} = f_\theta(\mathbf{X}),$$

where θ denotes all learnable parameters.

2.4 Training Objective and Optimization

The model is trained to minimize **binary cross-entropy**:

$$\mathcal{L}(\theta) = -\frac{1}{N} \sum_{i=1}^{N} w_{y_i} \left[y_i \log \hat{y}_i + (1 - y_i) \log(1 - \hat{y}_i) \right],$$

where w_{y_i} are optional class weights computed from the training data to address class imbalance. The optimizer is **Adam**, with default parameters (learning rate = 0.001).

2.5 Training Protocol

Training uses:

- **Batch size:** 16
- **Epochs:** up to 100 with EarlyStopping (monitor=`val_loss`, patience=20, restore_best_weights=True)
- **Learning rate reduction:** ReduceLROnPlateau (factor=0.5, patience=10, min_lr=1e-6)
- **Metrics:** accuracy, precision, recall

Per person sequences are split such that validation data contains the last 20% of each participant's sequences, preventing data leakage and ensuring evaluation reflects generalization to unseen individuals.

2.6 Summary

This methodology integrates multimodal behavioral signals into a sequential modeling framework. Temporal aggregation, sequence construction, normalization, and LSTM learning enable prediction of romantic interest while preserving interpretability and reproducibility. The approach adheres to best practices in affective computing and sequence-based behavioral modeling [2, 1].

3 Related Work

Our work lies at the intersection of multimodal affective computing and the psychology of romantic attraction. In contrast to many classical studies that examine isolated channels (e.g., only facial attractiveness or only vocal cues), we adopt an integrated perspective combining facial expressions, body movement, voice, and spoken language over time. This section reviews three main strands of prior research: (i) multimodal affective computing and behavioral signal processing, (ii) psychological and behavioral studies of attraction, and (iii) multimodal and multichannel approaches to human attractiveness.

3.1 Multimodal Affective Computing and Sequence Modeling

Multimodal affective computing has made substantial progress in integrating heterogeneous signals such as facial expression, body posture, speech, and language for emotion recognition and related tasks. Survey work highlights the importance of combining modalities at different levels (early, late, and hybrid fusion) and representing their temporal dynamics to capture complex affective states [3, 4]. Architectures such as the Tensor Fusion Network explicitly model unimodal, bimodal, and trimodal interactions between vision, audio, and language to improve sentiment analysis [5], while more recent approaches introduce attention-based mechanisms and feature reuse strategies to enhance multimodal emotion recognition from audiovisual data [6].

Within this broader context, [7] proposes a multimodal emotion recognition system that integrates facial expressions, body movement, speech, and spoken language, demonstrating that jointly modeling these modalities yields more robust affect estimates than using any single channel alone. Methodologically, our data collection and feature-extraction pipeline is closely aligned with this line of work: we extract facial Action Units, hand and upper-body motion descriptors, acoustic prosodic features, and sentiment scores from transcripts, all temporally aligned at the per-second level. At the sequence-modeling level, we employ stacked recurrent neural networks to capture temporal dependencies, following the general recommendations and conceptual overview of recurrent architectures in [1].

Unlike most of these studies, however, our target variable is not an internal affective state such as valence or arousal, but an explicitly interpersonal outcome: a binary indicator of romantic interest toward another person. Thus, while our methodological foundation is typical of multimodal affective computing [2, 4, 7], our prediction task is shifted from intra-individual emotion classification to dyadic attraction estimation.

3.2 Psychological Perspectives on Romantic Attraction

Psychological research on romantic attraction has identified multiple contributing factors, including physical appearance, voice, movement, similarity, and context [8]. Recent work emphasizes that attraction is inherently multimodal and dynamic, emerging from the integration of cues across different sensory and behavioral channels during interaction. For example, [9] synthesize evidence showing that looks, voice, movement, and even scent jointly shape attraction to future lovers and friends, arguing that no single channel fully explains interpersonal appeal.

Experimental paradigms such as speed dating have been used to study the formation of initial romantic interest under controlled but ecologically valid conditions. [10] show that synchronized body sway and movement patterns during speed dates predict romantic interest, highlighting the importance of nonverbal motor coordination. [11] extend this paradigm to virtual environments, using online meeting platforms to investigate how initial attraction and relationship formation unfold when interaction is mediated by video conferencing. These studies demonstrate that attraction judgments are sensitive to subtle temporal patterns in behavior, not just static snapshots of appearance or isolated self-report measures.

Our work is conceptually aligned with this literature in that we also focus on initial romantic interest following brief interactions. However, instead of analyzing a single behavioral channel (e.g., body sway alone) or relying primarily on self-report and static ratings, we attempt to predict attraction by jointly modeling dynamic facial, bodily, vocal, and linguistic cues within short temporal windows. This aligns with the theoretical view that attraction is a situated, context-dependent process unfolding over time [2, 8].

3.3 Multimodal Attractiveness and Communication of Appeal

Beyond emotion recognition and laboratory paradigms, several studies explicitly examine how different modalities communicate attractiveness and desirability. [12] analyze 881 judgments of men's and women's physical, vocal, and olfactory attractiveness, showing that these channels each contribute to perceived attractiveness, with partially overlapping but distinct information. Similarly, [9] argue for an "attraction in every sense" perspective, in which looks, voice, movement, and scent collectively inform evaluations of potential romantic partners and friends. These findings reinforce the idea that multimodal integration is necessary to approximate attraction judgments.

Our study extends this multimodal perspective in two important ways. First, we move from static or cross-sectional attractiveness ratings to temporally resolved behavioral sequences, incorporating not just what a person looks or sounds like, but how their expressions, gestures, and vocal prosody evolve over 15-second windows. Second, whereas prior work often focuses on global attractiveness or desirability ratings aggregated across many observers, we aim to predict attraction at the dyadic level: whether a specific participant expresses romantic interest in a specific interaction partner. This shift from population level attractiveness to pair attraction introduces additional variability but brings the modeling task closer to the phenomenon of romantic choice.

3.4 Positioning and Expected Performance

Taken together, these strands of research suggest that (i) multimodal and temporal modeling can substantially improve the prediction of affective and social judgments [3, 5, 6, 7], and (ii) romantic attraction is shaped by an interplay of visual, vocal, and motor cues [8, 10, 9, 12]. Our approach leverages these insights by training a sequential neural model on synchronized facial, gestural, vocal, and sentiment features to classify binary romantic-interest outcomes.

In line with prior findings in speed-dating and attractiveness research, we expect that models relying solely on static or unimodal features (e.g., appearance only or voice only) would show limited performance when predicting individual attraction decisions. By contrast, our multimodal, sequence-based approach is expected to achieve performance that is at least comparable to other multimodal emotion and social judgment models [7, 6], and potentially superior to simpler baselines that ignore temporal dynamics or rely on a single channel. At the same time, consistent with the literature on dyadic processes [8], we do not anticipate perfect predictability: unique interpersonal "chemistry" and unobserved contextual factors likely impose an upper bound on achievable accuracy, even with rich multimodal data.

4 References

References

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

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

[4] 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.

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

[6] 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.

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

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

[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] 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.

[11] 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.

[12] 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.