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

Mitrea Ioan David

I. Abstract

Abstract—**Romantic attraction arises from dynamic interactions between facial expression, body movement, voice, and spoken language, yet most computational approaches examine these cues in isolation or as static summaries. This article addresses the problem of predicting romantic interest from short segments of naturalistic dyadic interactions using temporally aligned multimodal behavioral signals.We present a pipeline that extracts facial Action Units, hand and upper-body movement dynamics, acoustic prosodic features, and sentiment cues from video recordings, synchronized at the second level. These features are modeled as short behavioral sequences using a stacked Long Short-Term Memory (LSTM) network to capture temporal dependencies across modalities. Unlike conventional affect recognition tasks, our target is an explicitly interpersonal outcome: binary romantic attraction toward a specific interaction partner.Our contributions include (i) a practical multimodal preprocessing framework that aligns visual, vocal, and linguistic signals via participant visibility, (ii) a compact sequence-based representation of expressive behavior, and (iii) a temporal neural model for attraction prediction. Empirical results show that multimodal sequence modeling substantially outperforms static and unimodal baselines, supporting the view that attraction is communicated through dynamic, multimodal behavior rather than isolated cues.**

Index Terms—**Affective computing, multimodal interaction, romantic attraction, sequence modeling, LSTM, social signal processing.**

II. Introduction

Human romantic attraction plays a central role in social behavior, relationship formation, and well-being. It emerges rapidly during interaction and is shaped by a complex interplay of facial expressions, body movement, vocal characteristics, and spoken language [1, 2, 3]. Psychological research has established that attraction is not determined by any single cue, but rather by the integration of multiple behavioral signals across these modalities [2, 3]. Recent work further demonstrates that attraction is sensitive to dynamic temporal patterns, such as synchronized body movement during brief interactions [4]. Despite this understanding, computational approaches to modeling attraction have remained limited, often focusing on static features, unimodal cues, or population-level attractiveness ratings rather than dyadic, temporally grounded interaction data [5, 6].

Recent advances in affective computing and multimodal machine learning have made it possible to extract rich behavioral signals from audiovisual recordings, including facial expressions, gestures, prosody, and language. These methods have been successfully applied to tasks such as emotion recognition, sentiment analysis, and social signal processing. However, predicting romantic attraction poses additional challenges beyond standard affect recognition. Attraction is inherently interpersonal, context-dependent, and only partially observable through behavior. Moreover, it unfolds over short time scales, where subtle temporal patterns and coordination between modalities may carry more information than isolated snapshots [4].

The problem addressed in this article is the following: *can short sequences of multimodal behavioral signals be used to reliably predict whether a person experiences romantic attraction toward a specific interaction partner?* Formally, given synchronized facial, bodily, vocal, and linguistic cues extracted from brief segments of dyadic interaction, we aim to classify binary romantic interest at the individual–partner level. Addressing this problem requires not only robust multimodal feature extraction, but also temporal modeling capable of capturing how expressive behaviors evolve and co-occur over time.

The importance of this problem is twofold. From a scientific perspective, computational models can complement traditional psychological methods based on self-report and manual coding [7]. Such models enable precise analysis of how dynamic multimodal behavioral cues—including facial expressions, body movement, voice, and language—jointly shape interpersonal attraction judgments [1, 2]. From a methodological standpoint, attraction prediction represents a challenging testbed for multimodal sequence modeling. Because attraction is shaped by multiple interacting factors that vary across individuals and contexts [1], the outcome is not perfectly determined by observable behavior alone, involving noisy, partially missing signals and strong individual differences. Progress in this area may therefore inform the design of more general models for social and affective inference.

A. Research Questions

This work addresses three specific research questions:

1) **RQ1: Predictive power of multimodal sequences.** Can short temporal sequences (15 seconds) of multimodal behavioral signals—facial expressions, body movement, voice, and language—reliably predict binary romantic attraction between interaction partners?
2) **RQ2: Temporal modeling advantage.** Do sequence-based models that capture temporal dependencies out-

[0] ACM CCS: Computing methodologies → Machine learning; Human-centered computing → Affective computing. AMS MSC: 68T07, 68T10, 62M45.

perform approaches based on static or time-averaged features?
3) **RQ3: Multimodal integration.** Does jointly modeling facial, bodily, vocal, and linguistic cues improve prediction accuracy compared to unimodal approaches that examine isolated channels?

Addressing these questions will establish whether computational multimodal sequence modeling can approximate interpersonal attraction judgments and identify the extent to which temporal dynamics and cross-modal integration contribute to predictive performance.

III. 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 [6]. 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 [7, 6]. 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.

A. Person Video Trimming and Facial Action Unit Extraction

To ensure that downstream feature extraction was performed only when the target participant was visible, we preprocessed the videos by temporally segmenting and retaining only frames in which the participant of interest appeared on screen. This yielded a video containing black frames during intervals when the participant was not visible. These videos functioned as the structural backbone for subsequent behavioral feature extraction.

From these preprocessed videos, we extracted frame-level facial Action Units (AUs) to quantify expressive behavior. Videos were sampled at 1 frame per second (every 30th frame at typical 30 fps playback), and each frame was classified into one of three categories: (1) black frames (mean pixel brightness ¡ 15), indicating off-screen intervals; (2) frames without detectable faces, identified using Haar Cascade face detection with a minimum face size of 50 × 50 pixels; or (3) frames containing valid faces suitable for AU extraction.

For frames with valid faces, we performed emotion recognition using `DeepFace` and converted the resulting emotion probability distributions (happy, sad, angry, surprise, fear, disgust) into pseudo-Action Unit intensities based on established mappings from the Facial Action Coding System (FACS). For example, happiness was mapped primarily to AU06 (cheek raiser) and AU12 (lip corner puller), while anger was associated with AU04 (brow lowerer) and AU23 (lip tightener). The following 17 Action Units were estimated: AU01, AU02, AU04, AU05, AU06, AU07, AU09, AU10, AU12, AU14, AU15, AU17, AU20, AU23, AU25, AU26, and AU28. AU intensities were normalized to the range [0, 1].

Frames without faces or with black content were assigned neutral AU values (all zeros) to maintain temporal continuity. The final output consisted of time-indexed CSV files containing AU intensities at 1 Hz sampling rate, enabling alignment with other modalities for multimodal fusion.

B. Audio Feature Extraction

To analyze vocal affect and speaking patterns, we extracted frame-level acoustic features from each participant's audio track using `librosa`. Audio files were loaded at a sampling rate of 22,050 Hz, and features were computed with a hop length of 512 samples (approximately 23 ms per frame).

The following features were extracted to capture prosodic, spectral, and temporal characteristics of speech:

- **Energy**: Root mean square (RMS) energy converted to decibels, reflecting vocal intensity and amplitude modulation.
- **Pitch (F0)**: Fundamental frequency estimated using the probabilistic YIN (PYIN) algorithm [], with a valid range of C2 to C7 (approximately 65–2093 Hz). Pitch confidence scores were retained to account for voicing probability.
- **Speaking rate proxy**: Onset strength (spectral flux) was extracted as a frame-level feature reflecting the rate of acoustic change. While true speaking rate requires phonetic transcription, onset strength provides a continuous signal-processing approximation that captures temporal articulation dynamics without requiring speech-to-text alignment.

All features were temporally aligned to ensure consistent frame counts, and frame timestamps were computed using `librosa. frames_to_time`. The resulting feature set was exported as a time-indexed CSV file for each pair.

C. 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 [7, 6].

D. Linguistic Sentiment Analysis

To capture the affective content of spoken language, we extracted second-level sentiment scores from timestamped transcriptions of each conversation. Transcripts were structured in blocks containing speaker identity, temporal boundaries (start and end timestamps in HH:MM:SS.mmm format), and utterance text.

Sentiment analysis was performed using VADER (Valence Aware Dictionary and sEntiment Reasoner), a lexicon-based sentiment analysis tool optimized for social media and conversational text [8]. For each utterance, VADER computed four sentiment scores: positive (pos), neutral (neu), negative (neg), and compound (a normalized aggregate score ranging from -1 to +1).

To enable temporal alignment with other modalities, sentiment scores were expanded from utterance-level to second-level resolution. Each whole second covered by an utterance was assigned the sentiment scores of that utterance. When multiple utterances overlapped within a single second, the most recent utterance's sentiment was retained to avoid redundancy. The resulting time-indexed CSV files contained per-second sentiment scores, speaker labels, and the corresponding utterance text, facilitating multimodal fusion with acoustic and visual features at 1 Hz temporal resolution.

IV. 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.

A. 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**
- 4 **Hand and upper-body movement features for each person**
- 3 **Acoustic features for the date**
- 4 **Sentiment features for each person**

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

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

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.

B. 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.

C. 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.

D. Baseline Model for Comparison

To evaluate the contribution of temporal sequence modeling (RQ2), we compare the LSTM-based approach against a baseline that uses time-averaged features. For each 15-second sequence $\mathbf{X} \in \mathbb{R}^{15 \times 28}$, we compute the temporal mean:

$$\bar{\mathbf{x}} = \frac{1}{T} \sum_{t=1}^{T} \mathbf{x}_t \in \mathbb{R}^{28}.$$

This averaged feature vector is used to train a logistic regression classifier with balanced class weights and maximum 1000 iterations. The baseline model has access to the same multimodal information but lacks temporal structure, allowing us to isolate the contribution of sequence modeling.

E. 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).

F. 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.

G. Summary

This methodology integrates multimodal behavioral signals into a sequential modeling framework. The approach adheres to best practices in affective computing and sequence-based behavioral modeling [6, 5].

V. Related Work

Our work lies at the intersection of multimodal affective computing [7, 6, 5] and the psychology of romantic attraction [2, 3, 1]. In contrast to classical studies that examine isolated channels (e.g., only facial attractiveness [9] or only body movement [4]), we adopt an integrated perspective combining facial expressions, body movement, voice, and spoken language over time [7, 2]. 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.

A. 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 [5, 6]. Architectures such as the Tensor Fusion Network explicitly model unimodal, bimodal, and trimodal interactions between vision, audio, and language to improve sentiment analysis [10], while more recent approaches introduce attention-based mechanisms and feature reuse strategies to enhance multimodal emotion recognition from audiovisual data [11].

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 [12].

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 [13, 6, 7], our prediction task is shifted from intra-individual emotion classification to dyadic attraction estimation.

B. Psychological Perspectives on Romantic Attraction

Psychological research on romantic attraction has identified multiple contributing factors, including physical appearance, voice, movement, similarity, and context [1]. 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, recent work [2] synthesizes 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. [4] show that synchronized body sway and movement patterns during speed dates predict romantic interest, highlighting the importance of nonverbal motor coordination. [14] 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 [13, 1].

C. Multimodal Attractiveness and Communication of Appeal

Beyond emotion recognition and laboratory paradigms, several studies explicitly examine how different modalities communicate attractiveness and desirability. [3] 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, [2] 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 [3, 9], we aim to predict attraction at the dyadic level: whether a specific participant expresses romantic interest in a specific interaction partner [4, 14]. This shift from population-level attractiveness to pair attraction introduces additional variability but brings the modeling task closer to the phenomenon of romantic choice.

VI. Results and Conclusions

The obtained results directly address the three research questions posed in this article.

A. Overall Performance and Temporal Modeling

TABLE I
PERFORMANCE COMPARISON: LSTM SEQUENCE MODEL VS. TIME-AVERAGED BASELINE

Model	Accuracy	Precision	Recall
Logistic Regression (time-avg.)	0.747	0.916	0.710
Stacked LSTM (sequences)	0.897	0.995	0.860
Absolute Improvement	+0.150	+0.079	+0.150
Relative Improvement	+20.1%	+8.6%	+21.1%

Regarding RQ1 (predictive power): The full multimodal model achieved a validation accuracy of 89.7% with precision of 99.5% and recall of 86.0% on the positive class (attraction), demonstrating that short multimodal behavioral sequences do contain predictive information about binary romantic attraction [5, 7]. The exceptionally high precision indicates that when the model predicts attraction, it is correct in nearly all cases, making it particularly suitable for applications where false positives must be minimized.

Regarding RQ2 (temporal modeling): To assess the value of temporal sequence modeling, we compared the LSTM against a logistic regression baseline using time-averaged features. The baseline achieved an accuracy of 74.7%, precision of 91.6%, and recall of 71.0%. The LSTM model outperformed the baseline by 15.0 percentage points in accuracy (20.1% relative improvement) and 15.0 percentage points in recall (21.1% relative improvement), confirming that temporal dependencies across the 15-second windows provide substantial predictive information beyond static feature summaries [12, 13]. The improvement in recall indicates that the LSTM is better at detecting subtle attraction cues that unfold over time, supporting the hypothesis that attraction signals are inherently dynamic and sequential in nature [4].

B. Multimodal Ablation Study

Regarding RQ3 (multimodal integration): To quantify the relative contribution of each modality, we conducted a systematic ablation study by training separate LSTM models on individual feature groups and comparing them to the full multimodal system. Table II presents the validation performance of each modality in isolation and in combination.

TABLE II
MODALITY ABLATION STUDY: VALIDATION PERFORMANCE

Modality	Features	Accuracy	F1	Precision	Recall
Hand Gestures	4	0.543	0.542	0.953	0.379
Action Units	17	0.667	0.709	0.939	0.570
Audio	3	0.687	0.746	0.885	0.645
Sentiment	4	0.783	0.865	0.779	0.972
All Combined	**28**	**0.897**	**0.922**	**0.995**	**0.860**

The ablation study reveals several important findings about the structure of attraction signaling in dyadic interactions:

Sentiment analysis emerges as the strongest individual predictor, achieving 78.3% accuracy and the highest single-modality F1 score of 0.865. Most notably, sentiment features attain 97.2% recall—substantially higher than any other modality—indicating that linguistic-emotional content is particularly effective at identifying when attraction is present. This suggests that verbal expressions of interest, positive affect, and conversational alignment play a central role in romantic attraction [1].

Audio features demonstrate robust mid-level performance (68.7% accuracy, F1 = 0.746) despite using only three acoustic parameters (energy, pitch, speaking rate). This finding aligns with prior work showing that paralinguistic vocal cues—such as prosody modulation and speech timing—convey interpersonal interest independently of linguistic content [2].

Facial action units achieve moderate discriminative power (66.7% accuracy, F1 = 0.709) with 17 features capturing smile intensity, eyebrow movements, and other facial expressions. While facial displays are theoretically central to attraction signaling [12], their standalone performance suggests they may be more informative when integrated with other modalities or that certain AUs contribute more than others.

Hand gestures show the weakest individual performance (54.3% accuracy, F1 = 0.542), barely exceeding chance. The low recall (37.9%) indicates that hand movement patterns alone rarely signal attraction reliably. However, gestures may still contribute contextual information when combined with other cues [15].

Multimodal integration yields substantial performance gains. The full 28-feature model achieves 89.7% accuracy and an F1 score of 0.922, representing an 11.4% improvement over the best single modality (sentiment). Crucially, multimodal fusion increases precision from 77.9% to 99.5%—a 27.7% relative improvement—enabling near-perfect positive prediction accuracy while maintaining competitive recall (86.0%). This demonstrates that although sentiment analysis alone captures many attraction signals, integrating facial, bodily, and acoustic modalities substantially reduces false positives and provides a more robust predictive model [5, 10].

The complementary nature of the modalities is evident in their precision-recall trade-offs: sentiment maximizes recall (detecting most attraction cases), while multimodal integration maximizes precision (ensuring high confidence in positive predictions). This pattern suggests that verbal-emotional signals provide broad coverage of attraction indicators, while nonverbal cues (facial expressions, voice, gestures) add specificity and contextual validation [7, 11].

C. Limitations and Future Work

Several limitations should be noted. First, the dataset exhibits class imbalance (71% attracted vs. 29% not attracted), which may influence model predictions despite class-weight balancing. Second, the relatively small validation set ($\sim$300 sequences across 18 individuals) limits the strength of generalization claims. Third, while the ablation study quantifies modality contributions, we have not yet examined attention mechanisms or feature-level importance within modalities, which could reveal which specific behavioral cues drive predictions.

Several aspects of the experimental design support the validity of these conclusions. Temporal segmentation and per-second aggregation ensure that predictions are based on dynamic interaction patterns rather than isolated frames [15]. Participant-level data splitting prevents information leakage between training and validation sets, promoting generalization to unseen individuals. Furthermore, deterministic handling of missing modalities through zero filling avoids introducing artificial correlations via imputation. Together, these design choices strengthen confidence that the reported performance reflects learned multimodal behavioral patterns rather than artifacts of preprocessing or evaluation.

In conclusion, this work provides evidence that multimodal, sequence-based modeling offers a promising approach to predicting romantic attraction from short segments of dyadic interaction [7, 13]. The systematic ablation study confirms that while individual modalities—particularly sentiment and audio—carry substantial predictive power, their integration yields significant performance gains, validating the multimodal hypothesis central to contemporary theories of interpersonal attraction [1, 2]. Future work should include attention-based architectures for interpretability [10], cross-validation across multiple data splits, testing on independent datasets, and investigation of feature-level importance within each modality to identify the specific behavioral cues that most reliably signal romantic interest [11].

VII. References

References

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

[2] 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. *Brit. J. Psychol.*, 116(3):684–701, 2025.

[3] 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.

[4] A. Chang, H. E. Kragness, W. Tsou, D. J. Bosnyak, A. Thiede, and L. J. Trainor. Body sway predicts romantic interest in speed dating. *Soc. Cogn. Affect. Neurosci.*, 16(1–2):185–192, 2021.

[5] Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. Multimodal machine learning: A survey and taxonomy. 2017. Preprint of [5].

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

[7] K. Kraack. A multimodal emotion recognition system: Integrating facial expressions, body movement, speech, and spoken language. *arXiv preprint arXiv:2412.17907*, 2023.

[8] C. J. Hutto and Eric Gilbert. Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Proc. Int. AAAI Conf. Web Soc. Media (ICWSM)*, volume 8, pages 216–225, Ann Arbor, MI, USA, 2014.

[9] T. Sano and H. Kawabata. A computational approach to investigating facial attractiveness factors using geometric morphometric analysis and deep learning. *Sci. Rep.*, 13(1):19797, 2023.

[10] Amir Zadeh, Minghai Chen, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. Tensor fusion network for multimodal sentiment analysis. In *Proc. Empirical Methods Natural Lang. Process. (EMNLP)*, pages 1103–1114, Copenhagen, Denmark, 2017.

[11] 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. In *Proc. Interspeech*, pages 2433–2437, Rhodes Island, Greece, 2023.

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

[13] 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.

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

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