

A Step towards Automated Support in Assessing Emotional Mirroring

Adriana-Mihaela Guran, Grigoreta-Sofia Cojocar, Dan Cojocar

Babeş-Bolyai University

1, M. Kogălniceanu Street, Cluj-Napoca, Romania

E-mail: {adriana.guran, grigoreta.cojocar, dan.cojocar}@ubbcluj.ro

Abstract. Empathy is a central construct in social cognition and is defined as the ability to recognize and adequately react emotionally to an affective message transferred by a human counterpart by sharing their emotion (mirroring). In the therapy process, empathy plays an important role in building a strong connection between the participants. In this paper, we describe a support system that provides visual representations of the emotional mirroring between two people based on the analysis and correlation of video and audio data and EMG sensor outputs. A set of metrics related to the emotion evolution of the two participants is proposed, and the identified correlations on a set of 10 dyads are presented.

Keywords: AI, emotions, mirroring, empathy, automation

DOI:10.37789/ijusi.2024.17.4.1

1. Introduction

Mirroring is the behavior in which one person subconsciously imitates the gesture, speech pattern, or attitude of another, as defined by Chartrand & Bargh (1999).

Imitating and comprehending others' activities is particularly helpful for young children's learning skills, such as speech, movement, and play. Mirror neurons also appear to provide sufficient information to predict why someone is performing the behavior they are engaged in, and this is a powerful mechanism for emotional understanding (Rasmussen & Bliss, 2014; Cook et al., 2014). While observing emotional information, the mirror neuron system makes it possible to generate a brain state that matches that of the person being observed, providing an automatic share of their experience and supporting empathy (Penagos-Corzo et al., 2022).

Empathy is a central construct in social cognition and is defined as the ability to recognize and adequately react emotionally to an affective message

transferred by a human counterpart by sharing, to a certain degree, their emotion (de Vignemont & Singer, 2006). Empathy plays a vital role in the counselling process, as it fosters a therapeutic environment where clients feel understood, supported, and validated. Empathy forms the foundation of a strong therapeutic alliance. When clients perceive that their counselor understands and empathizes with their experiences, they are more likely to establish trust and open their concerns (Bohart & Greenberg, 1997). A positive therapeutic relationship enables clients to feel supported and increases their willingness to collaborate on treatment goals. A positive therapeutic relationship characterized by empathy also contributes to improved treatment outcomes, client satisfaction, and adherence to therapy (Elliott et al., 2018).

Assessing emotional mirroring includes methods such as observation, use of EMG for facial muscles (Scarpazza et al., 2018), SMT (Self-Mirroring Technique) (Vinai et al., 2015), use of IRM images, and the corresponding brain regions that are activated. In this paper, we propose an interactive support system that automatically gathers information about two people's emotions from their facial expressions, spoken text, EMG, and eye-tracking. We also introduce some metrics that could support the assessment of emotional mirroring in two people engaged in a video-recorded conversation.

2. Related work

In the literature, there are a few approaches that treat people's interaction as a multi-stream dynamic process. Usually, the audio, visual, and text are used to assess the interaction quality. Most of the existing approaches integrate visual and audio information to achieve multimodal emotion recognition (Datu& Rothkrantz, 2015; Kraus & Chen, 2013), while Tan et al. (2021) add the use of EEG signals to complement the audio and visual information. The present work uses the auditory, visual, and EMG information to perform multimodal emotion recognition and focuses also on eye contact between the two partners. To our knowledge, there is no similar approach in the literature.

The evolution of Artificial Intelligence has advanced the progress in multimodal emotion recognition, with researchers developing models based on neural networks, graphs, and SVM to model the cross-modal emotions dynamics or to classify input signals from different sources (Joshi et al., 2022; Pandeya, 2021). Other approaches have integrated Transformers in

multimodal emotion recognition (Lian et al., 2021; Siriwardhana et al., 2021). The most recent approaches use Multimodal Large Language Models (MLLMs) to assess empathy (Zhang et al., 2025). These models need large datasets for training and validation, like IEMOCAP (Busso, 2008) or MOSEI (Zadeh et al., 2018). The intention of our preliminary study is to use free and offline existing models to assess their results for identifying emotion mirroring.

3. Our application

In the therapy process, it is important to build a relationship between the client and the therapist. Assessing the quality of communication between the two participants is important to decide the future of their collaboration. Information for the task of assessing the matching between the two actors can be extracted from various sources, like their mimic, their gestures, the information they exchange verbally (their dialogue), and the duration of eye contact. Affective states are associated with physiological responses across the body, so the use of facial electromyography (EMG) provides additional support in investigating the presence of these affective states. Information on the affective states (emotions) of the participants and the EMG information could guide the process of automatically assessing the intention of future collaboration.

Our supporting application uses two video recordings and EMG data from two facial muscles (the zygomatic and corrugator muscles) of the two participants. From the video, we extract information for eye contact detection, transcription of the dialogue, and emotions from facial expressions. This information, together with EMG data, supports the evaluation of emotional mirroring. The following steps are performed by the developed application: obtaining the text of the conversation between the therapist and the client and analyzing it to identify emotions or sentiments in the speech, the usage/repeating of certain words, or the frequency of some words; automatically recognizing some of the most encountered emotions from the participants' faces: happiness, sadness, anger, disgust, fear, surprise, and neutrality; tracking the eye movements of both client and therapist during the conversation and merging the obtained data and establishing the correlations between the two participants in the dialogue.

To analyze the conversation text, the audio of the recording is used; to

automatically detect emotions, the images from the recording are used; and to monitor the eyes' movements, the images are used.

To process the data from the recordings, a package of modules was developed, each module handling one source of information (video, sound, text). The information from all sources is combined to accurately establish correlations between data. That is why an additional module that integrates the results obtained from all the sources was also developed. The dedicated modules have integrated machine learning methods for data processing. The package contains the following modules (as shown in Figure 1):

- A module for voice recognition and transformation of audio signals into textual information, such as the detection of interlocutors who spoke, the transformation of the audio signal into text, the automatic identification of text's polarity (positive, negative, or neutral), and the computation of some statistics;
- A module for the recognition of emotions in human interactions in images/video-audio with two people and the creation of a chronological unfolding corresponding to the evolution of the identified emotions;
- A module for the detection and monitoring of eye movements (to detect eye contact), which allows the subsequent processing and analysis of the obtained information;
- A module that allows for the integration of the results obtained in previous steps (modules) and the EMG data and their in-depth analysis, which also provides a chronological graphic representation.

The results from each module are integrated into an Excel file used to further analyze the raw data obtained. Each module uses third party free offline libraries for its task. The selected libraries have been identified after a broad research process on existing solutions. Based on the results obtained during the testing phase, we have chosen the following offline and free models for each specific task: Whisper model for text transcription (Radford et al., 2023); Bert base uncased poems model for text polarity (Bert); Haar Cascade model for face recognition (Viola, 2001); Poster++ model for emotion recognition from face images (Mao et al., 2024); and OpenVino model for eye-tracking (OpenVino, 2024).

Using the selected models, each module (text, emotion recognition from images, and eye-tracking) outputs the data in a human-readable format that is

further processed by the integration module into an Excel file.

For voice recognition and sentiment analysis from written text the output contains the following data: the starting time of a spoken text (sentence, paragraph) in seconds, the ending time of a spoken text (sentence, paragraph) in seconds, the speaker of the sentence, the spoken text, the polarity of the sentence (three possible labels: positive, negative, or neutral) and a confidence score for polarity classification (a value between 0 and 1, where 1 means fully confident).

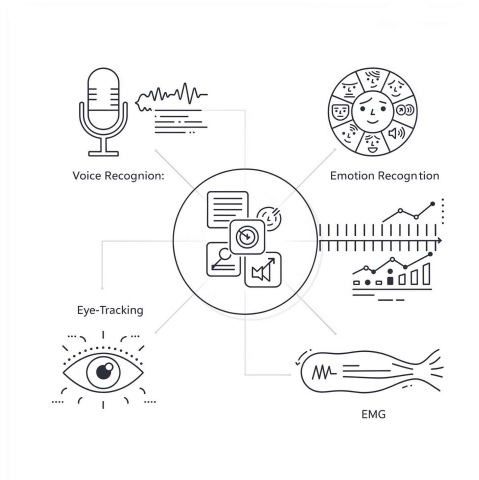


Figure 1. The architecture of the system

For emotion recognition from video, first, we extract the frames from the video, using OpenCV (OpenCV, n.d.). Afterwards, each frame is analyzed to detect the faces in it, using the Haar Cascade model. Then, on each detected face, the Poster++ emotion recognition model is applied to identify the emotions. For each analyzed frame, the output contains the seven emotions (surprise, fear, disgust, happiness, sadness, anger, and neutral), a probability for each emotion, the dominant emotion (the emotion with the highest probability), and a timestamp for the frame (in seconds).

For eye-tracking, the OpenVINO model is used to determine the gaze vector of each person in the conversation recording, and then, for each frame from the video, using some heuristics, the modules determine if the two people are looking at each other or not. Also, there is always the possibility that the algorithm cannot determine where a person is looking. In such cases,

a particular value is used to signal this case. For each frame, the output contains a value indicating whether a person is looking at the other person, or not, or the special case, and a timestamp for the frame (in milliseconds).

To correlate the output obtained from all modules, an Excel file is created, which contains the data from all three modules: a time axis (in milliseconds), the value of the EMG sensors (corrugator and zygomatic) for each person, the probabilities for considered emotion (happiness, sadness, anger, fear, disgust, surprise and neutral), the dominant emotion (decided by the emotion recognition model used), the text transcript and its polarity for each person, and the value for the eye-direction for each person. This file is then used for in depth analysis (graphics, Pearson correlation computation, and other heuristic predictors).

The information gathered in the Excel file is used to predict the intention for further interaction. We propose a set of predictors that synthesizes the key elements in the interaction. The proposed predictors are grouped into four classes based on the source of the information: EMG based predictors (see Table 1) like the duration of smiles or frowns for one person in the interaction; emotion recognition-based predictors (Table 2) like the number of frames a participant had a dominant positive emotion; eye-tracking predictors (Table 3) related to how long do the participants make or do not make eye contact; and text based predictors (Table 4) like the time when one person speaks related to the duration of the conversation.

Table 1 presents the list of EMG-based predictors for one person (P1) and their meaning. Each predictor has been computed for both participants in the interaction, meaning that the corresponding predictors have been computed for the other participant in the interaction (P2), too.

Table 1 – Proposed EMG-based predictors for empathy assessment

Predictor's name	Predictor's description
P1-EMG-(Smile Frown)-length	Computes how long the smiles/frowns were in comparison with the total length (time) of the smiles/frowns for person P1
P1-EMG-(Smile Frown)-frequency	Computes how long the (smiles/frowns) were in comparison with the total length (time) of the smiles / frowns for person P1
P1-EMG-(Smile Frown)-number	Computes the number of smiles/frowns for person P1

Due to space constraints, we have synthesized similar predictors by surrounding all possibilities of predictors. For example, P1-EMG-(Smile|Frown)-length means that we have two predictors, P1-EMG-Smile-

length and P1-EMG-Frown-length.

Table 2 presents the list of facial emotion recognition-based predictors for one person (P1) and their meaning. Each predictor has been computed for both participants in the interaction, meaning that the corresponding predictors have been computed for the other participant in the interaction (P2), too.

Table 2 – Proposed facial emotion recognition-based predictors for empathy assessment

Predictor's name	Predictor's description
P1-Emo-(Poz Neg Neutral)	Computes the ratio of frames where P1-Emo-Poz P1-Emo-Neg P1-Emo-Neutral is the biggest, from all analyzed frames for person P1
P1-Emo-Do-(Surprise Anger Fear Happiness Disgust Sadness Neutral)	Computes the ratio of frames when person P1's Emo-Surprise/ P1's Emo-Anger/ P1's Emo-Fear/ P1's Emo-Happiness/ P1's Emo-Disgust/ P1's Emo-Sadness/ P1's Emo-Neutral was the dominant emotion from all analyzed frames for person P1
P1-Emo-Do-Negative	Computes the ratio of frames when person P1's Emo-Negative was the dominant emotion from all analyzed frames for person P1. Emo-Neg is considered to be the sum of all four negative emotions when dominant: fear, disgust, sadness, and anger.
P1-Emo-Do-Positive	Computes the ratio of frames when person P1's Emo-Positive was the dominant emotion from all analyzed frames for person P1. Emo-Positive is considered the happiness emotion when dominant.

Table 3 presents the list of eye-contact-based predictors for one person (P1) and their meaning. Each predictor has been computed for both participants.

Table 3 – Proposed eye-contact-based predictors for empathy assessment

Predictor's name	Predictor's description
P1-Eye-(Yes No)	Computes the ratio between the number of frames when person P1 does/does not look towards P2 over the total number of frames when P1 looks/does not look towards P2 (the frames when P1's eye could not be detected are not considered)
Eye-Yes-Both	Computes the ratio between the cardinality of the set of frames when P1 looks towards P2 and P2 looks towards P1, and the total number of analyzed frames. It measures how long the eye contact was between P1 and P2
P1-Eye-Null	Computes the ratio between the number of frames when P1's eyes were not detected and the total number of analyzed frames.

Finally, Table 4 presents the list of text-based predictors for one person (P1) and their meaning. Each predictor has been computed for both participants.

Table 4 – Proposed eye contact-based predictors for empathy assessment

Predictor's name	Predictor's description
Tra-P1-Len	Computes the ratio between the number of milliseconds that P1 speaks and the number of milliseconds of the entire discussion
Tra-P1-Wrd	Computes the number of times when P1 repeats P2's relevant words
Pol-P1- (Positive Negative Neutral)	Computes the ratio between the number of milliseconds P1 spoke, and the text polarity is positive/negative/neutral, and the total number of milliseconds when P1 spoke

4. Application assessment

To assess the suitability of our application, ten pairs of young people (girl and boys students at our university) have been invited in laboratory settings to have a video-recorded conversation on given general topics. Each conversation lasted about 10 minutes, and each pair of speakers was formed by a girl and a boy. The conversation was video recorded, and both participants had the EMG sensors mounted on their faces. Two identical cameras have been used to obtain video recordings of each participant's face. At the end of the conversation session, two video files that we refer to as a dyad (one for each participant) and four EMG data files (one for the zygomatic muscle and one for the corrugator muscle of each participant) have been obtained and used for further analysis.

The collected data were analyzed by a team of psychologists using Spearman's ρ correlation (Argyrous, 2011) to assess the strength and direction of the monotonic relationships between the ranked variables (in our case, the predictors' values). This non-parametric method was chosen due to its robustness against outliers and its applicability to ordinal data, making it suitable for the variables under study. The analysis involved ranking the data points and calculating the Spearman correlation coefficient for each pair of variables. The resulting ρ values ranged from -1 to +1, with values closer to +1 indicating a strong positive association, values closer to -1 indicating a strong negative association, and values around 0 suggesting no significant association. This approach allowed for a comprehensive evaluation of the relationships between variables without the need for assumptions of normality and linearity, providing a more accurate representation of the

underlying data patterns.

EMG-measured smiles and frowns

There was a large negative correlation between EMG-measured smile and frown frequency, confirming the reliability of this measure in discriminating between the expression of positive and negative emotions. In which between-participant correlations are concerned, P1 positive-speech was positively related to P2 smile frequency ($r = .76$, $r = -.68$, $p < .05$ ¹) and negatively related to P2 frown frequency ($r = -.76$). As for within-participant results, P1 frown number was negatively correlated with P1 positive speech ($r = -.57$, $p < .05$). Thus, as also found in previous studies, facial EMG can provide useful indices of emotional experiences (Hess et al., 2007; Tassinari and Cacioppo, 2019; van Boxtel, 2010; Kraus and Chen, 2013).

Facial expressions of emotions (video-based)

Like EMG-measured positive emotions, facial expressions of positive emotions in P2 were positively related to positive speech in P1 ($r = .71$; $p < .05$) and were negatively related to negative speech in P2 ($r = -.58$; $p < .05$). Some authors suggest that such relations may partly be due to the contagious nature of positive affect, which can induce a similar emotional state in others, thereby broadening their thought-action repertoires and enhancing their use of inclusive, creative, and empathetic communication (Fredrickson, 2001).

Additionally, conversations initiated with high emotional positivity often led to increased use of words related to social processes, such as "friends" and "family," and positive emotion words like "happy" and "love" by the other participant (Pennebaker, Mehl, & Niederhoffer, 2003). Additionally, P1's facial expressions of positive emotions were significantly linked to P1 repeating words used by P2 ($r = .65$; $p < .05$), also shown in (Giles et al., 1991).

Speech content and conversation-related variables

In addition to the findings described above, the number of P2's words repeated by P1 was positively related to P2's positive speech ($r = .63$; $p < .05$).

¹ r - correlation coefficient, the p -value – probability value is a measure that helps determine the significance of the results obtained from a hypothesis test

Eye contact

Eye contact is a fundamental component of nonverbal communication and social interaction, and gaze plays an important role in initiating and regulating social interactions (Hessels, 2020). In this study, two significant relations were found in terms of eye contact: P1's eye-contact was positively associated with mutual visual contact ($r = .70$; $p < .05$), and negatively related to P2 negative speech ($r = -.61$; $p < .05$). Previous research shows that when individuals engage in eye contact, it often leads to synchronized gaze behavior and enhanced communication (Koyke et al, 2016; Wohltjen & Wheatley, 2021). In which the relation to speech, a widespread belief in many cultures is that direct eye contact reflects honest speech, and that people avert their gaze when they are not being honest or in cases of negative and uncomfortable verbal interactions (Williams et al., 2009).

Interpersonal synchrony

Based on recent data regarding the role of interpersonal synchrony in influencing the quality of interpersonal interactions (Davis et al., 2018), we were also interested if correlations between P1's and P2's facial expressions of positive and negative emotions, and EMG-measured frowns and smiles were related to self-rated pleasantness of the interaction and with the wish for future interaction.

The correlation in smiles was positively associated with P2 positive speech ($r = .66$; $p < .05$). These results align with data showing that positive emotional expressions are often interpreted as indicators of friendliness, warmth, and approachability (Finkel & Eastwick, 2015). When individuals perceive these cues from their interaction partner, they are more likely to develop positive attitudes towards them (Finkel & Eastwick, 2015). This increased liking further motivates individuals to seek out additional interactions, as they anticipate enjoyable and rewarding experiences (Hatfield et al., 1993). Moreover, when individuals share positive emotional expressions, they are more likely to experience increased rapport and connection (Gonzaga et al., 2001). This phenomenon can be attributed to the intrinsic human tendency to respond to positive emotional cues with corresponding positive emotions (Hess & Fischer, 2013). The support system also provides the functionality of graphically representing the correlation of the proposed predictors for two people, to visually assess their emotional mirroring. Figure 2 and Figure 3 describe a graphical representation of two different situations: in Figure 2 we

have a high emotional mirroring between two persons regarding the happiness emotion, and in Figure 3 is depicted where the emotional mirroring is low.

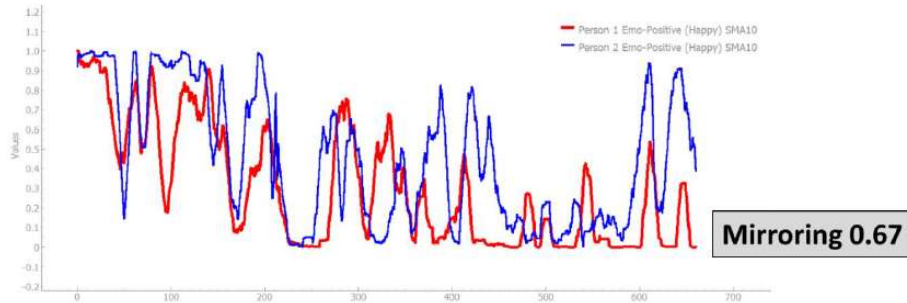


Figure 2. Emotional mirroring view for happy emotion

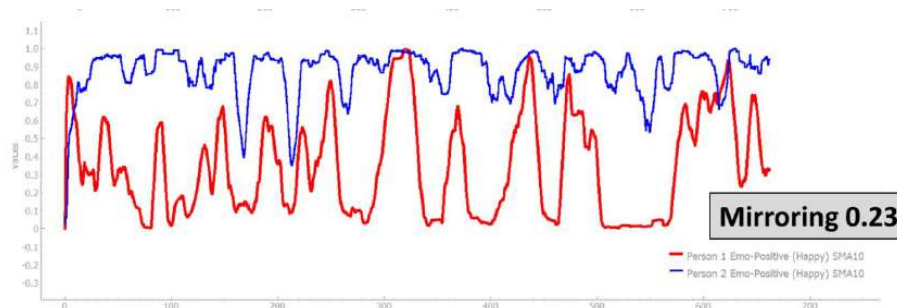


Figure 3. Emotional mirroring view for happy emotion

Limitations

Limitations have been identified at the data gathering step, including technical constraints determined by the AI models that have been used, and at the analysis step, as will be further detailed.

We have also identified problems with the used models that need to be further addressed by the AI algorithms, and we present them in the following.

- The voice recognition and speaker diarization models do not perform well if two people are speaking simultaneously. Only one speaker is recognized, and the transcript will contain the text from both speakers.
- The voice recognition may generate errors if both social actors are

male/female. It is recommended to use a male interacting with and a female for better accuracy.

- The transcript model does not perform well when two words spell differently, but they sound the same (i.e., ate-eight, buy-by, for-four, here-hear, right-write). This may influence the results of sentiment analysis from the text.
- The chosen transcript and speaker diarization models sometimes give as result a text repeated many times (ex. I may not. I may not. I may not).
- The face detection model can detect two faces in the same frame, and the emotion recognition model will analyze them, but this may affect the results of the data correlation.
- The selected emotion recognition models cannot be calibrated, based on a person's face. For example, a detected face had the model predict a dominant sadness emotion, but the human experts considered it to be a neutral one for that specific example.
- For eye-tracking and eye-contact calibration methods must be carefully performed to obtain relevant results.

Another limitation of the current study is the reduced sample size, which limits the possibility of generalizing results. Despite this limitation, we believe the current research has focused on an important question, which was approached using an original combination of methods (i.e., behavioral, physiological, AI-based) that, to our knowledge, has not been attempted before in the literature.

5. Conclusions and further work

In this paper, we have presented our approach in developing an application to support automatic emotional mirroring assessment using different input sources. We integrate data from sources such as EMG, video, and audio recordings to automatically detect emotions from sensors, facial expressions, and dialogue (the transcript of the conversation), to identify facial and verbal mirroring and eye contact. We have also proposed a set of predictors to assess emotional mirroring that can be automatically computed based on the collected data. In the future we intend to refine our approach by testing new AI models with improved performance and to include new emotions

recognition sources, such as voice or body posture.

Acknowledgements

The application was developed in the framework of the Babeş-Bolyai University project no. 856/08.11.2023. We thank the six students from the Faculty of Mathematics and Computer Science from Babeş-Bolyai University for their involvement in the design and implementation of the application.

This work would not have been possible without the support of our colleagues from the Faculty of Psychology and Educational Sciences from the Babeş-Bolyai University. We are truly grateful for their hands-on involvement in the interview process and for the patience and wisdom they shared during the data analysis phase. Their guidance in interpreting the findings from multiple sources provided us with a broader perspective and a more nuanced understanding of the subject matter. Their involvement has been vital to the successful completion of this project.

References

- Argyrous, G. (2011). *Statistics for research: With a guide to SPSS* (3rd ed.). SAGE Publications.
- Bohart, A. C., & Greenberg, L. S. (1997). Empathy reconsidered: New directions in psychotherapy. *American Psychologist*, *52*(7), 749-755
- van Boxtel, A. (2010). Facial EMG as a tool for inferring affective states. In A. J. Spink et al. (Eds.), *Proceedings of Measuring Behavior 2010* (pp. 104–108). Noldus Information Technology.
- Busso, C., Bulut, M., Lee, C. C., Kazemzadeh, A., Mower, E., Kim, S., Chang, J. N., Lee, S., & Narayanan, S. (2008). IEMOCAP: Interactive emotional dyadic motion capture database. *Language Resources and Evaluation*, *42*(4), 335–359. <https://doi.org/10.1007/s10579-008-9076-6>
- Chartrand, T. L.; Bargh, J. A. (1999). The chameleon effect: The perception-behavior link and social interaction. *Journal of Personality and Social Psychology*. *76* (6): 893–910. doi:10.1037/0022-3514.76.6.893. PMID 10402679. S2CID 11818459. Archived from the original (PDF) on 2011-07-09
- Cook, R., Brewer, R., Shah, P., & Bird, G. (2014). Alexithymia, not autism, predicts poor recognition of emotional facial expressions. *Psychological Science*, *24*(5), 723–732. <https://doi.org/10.1177/0956797612463582>
- Datcu, D.; Rothkrantz, L.J. Semantic audiovisual data fusion for automatic emotion

- recognition. In *Emotion Recognition: A Pattern Analysis Approach*; John Wiley and Sons: Hoboken, NJ, USA, 2015; pp. 411–435., <https://doi.org/10.1002/9781118910566.ch16>
- Davis, M. H., Le, B., & Coy, A. E. (2018). Empathy and interpersonal relationships. In A. L. Vangelisti & D. Perlman (Eds.), *The Cambridge Handbook of Personal Relationships* (2nd ed., pp. 294–305). Cambridge University Press., <https://doi.org/10.1017/9781316417867>
- Elliott, R., Bohart, A. C., Watson, J. C., & Greenberg, L. S. (2018). Empathy. *Psychotherapy*, 55(4), 424–429, <https://doi.org/10.1037/a0022187>
- Fredrickson, B. L. (2001). The role of positive emotions in positive psychology: The broaden-and-build theory of positive emotions. *American Psychologist*, 56(3), 218–226. <https://doi.org/10.1037/0003-066X.56.3.218>
- Finkel, E. J., Simpson, J. A., & Eastwick, P. W. (2015). The psychology of close relationships: Fourteen core principles. *Annual Review of Psychology*, 66, 499–527.
- Giles, H., Coupland, J., & Coupland, N. (Eds.). (1991). *Contexts of accommodation: Developments in applied sociolinguistics*. Cambridge University Press. <https://doi.org/10.1017/CBO9780511663673>
- Gonzaga, G. C., Keltner, D., Londahl, E. A., & Smith, M. D. (2001). Love and the distinctiveness of affirmative cue display. *Journal of Personality and Social Psychology*, 81(1), 93–105. <https://doi.org/10.1037/0022-3514.81.1.93>
- Hatfield, E., Cacioppo, J. T., & Rapson, R. L. (1993). Emotional contagion. *Current Directions in Psychological Science*, 2(3), 96–100. <https://doi.org/10.1111/1467-8721.ep10770953>
- Hess, U., & Fischer, A. (2013). Emotional mimicry as social regulation. *Personality and Social Psychology Review*, 17(2), 142–157. <https://doi.org/10.1177/1088868312472607>
- Hessels, R. S. (2020). How does gaze to faces support face-to-face interaction? A review and perspective. *Psychonomic Bulletin & Review*, 27(5), 856–881. <https://doi.org/10.3758/s13423-020-01715-w>
- Joshi, A., Bhat, A., Jain, A., Singh, A., & Modi, A. (2022). COGMEN: Contextualized GNN-based multimodal emotion recognition. In *Proceedings of the 2022 conference of the North American chapter of the Association for Computational Linguistics: human language technologies*, 4148–4164. <https://doi.org/10.18653/v1/2022.naacl-main.306>
- Kraus, M. W., & Chen, S. (2013). A winning smile? Smile intensity, physical dominance, and fighter performance. *Psychological Science*, 24(3), 271–279. <https://doi.org/10.1177/09567976124>
- Huang, J.; Tao, J.; Liu, B.; Lian, Z.; Niu, M. Multimodal transformer fusion for continuous emotion recognition. In *Proceedings of the ICASSP 2020—2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Barcelona, Spain, 4–8 May 2020, 3507–3511 57783, <https://doi.org/10.1037/a0030745>
- Mao, J., Xu, R., Yin, X., Chang, Y., Nie, B., Huang, A., Wang, Y. (2024) POSTER++: A simpler and stronger facial expression recognition network, *Pattern Recognition*, 2024,

- 110951, ISSN 0031-3203, <https://doi.org/10.1016/j.patcog.2024.110951>.
- Pandeya, Y.R.; Lee, J. (2021). Deep learning-based late fusion of multimodal information for emotion classification of music video. *Multimed. Tools Appl.* 2021, 80, 2887–2905., <https://doi.org/10.1007/s11042-020-08836>
- Penagos-Corzo, J. C., Cosio van Hasselt, M., Escobar, D., Vázquez-Roque, R. A., & Flores, G. (2022). Mirror neurons and empathy-related regions in psychopathy: Systematic review, meta-analysis, and a working model. *Social Neuroscience*, 17(5), 462–479. <https://doi.org/10.1080/17470919.2022.2128868>
- Pennebaker, J. W., Mehl, M. R., & Niederhoffer, K. G. (2003). Psychological aspects of natural language use: Our words, our selves. *Annual Review of Psychology*, 54(1), 547–577. <https://doi.org/10.1146/annurev.psych.54.101601.145041>
- Rasmussen, B., & Bliss, S. (2014). Beneath the surface: An exploration of neurobiological alterations in therapists working with trauma. *Smith College Studies in Social Work*, 84(2°3), 332–349. <https://doi.org/10.1080/00377317.2014.923714>
- Radford, A., Kim, J. W., Xu, T., Brockman, G., Mcleavey, C., Sutskever, I. (2023), Robust Speech Recognition via Large-Scale Weak Supervision, Proceedings of the 40th International Conference on Machine Learning, PMLR 202:28492-28518, 2023.
- Scarpazza, C., Làdavas, E. & Cattaneo, L. Invisible side of emotions: somato-motor responses to affective facial displays in alexithymia. *Exp Brain Res* 236, 195–206 (2018). <https://doi.org/10.1007/s00221-017-5118-x>
- Siriwardhana, S.; Kaluarachchi, T.; Billinghamurst, M.; Nanayakkara, S. (2020). Multimodal emotion recognition with transformer-based self-supervised feature fusion. *IEEE Access* 2020, 8, 176274–176285, <https://doi.org/10.1109/ACCESS.2020.3026823>
- Tan, Y.; Sun, Z.; Duan, F.; Solé-Casals, J.; Caiafa, C.F. (2021) A multimodal emotion recognition method based on facial expressions and electroencephalography. *Biomed. Signal Process. Control*, 70, 103029. <https://doi.org/10.1016/j.bspc.2021.103029>
- Tassinary, L. G., Cacioppo, J. T., & Vanman, E. J. (2019). The somatic system. In J. T. Cacioppo, L. G. Tassinary, & G. G. Berntson (Eds.), *Handbook of Psychophysiology* (4th ed., 151–182). Cambridge University Press. <https://doi.org/10.1017/9781107415782.008>
- Vinai P., Speciale M., Vinai L., Vinai P., Bruno C., Ambrosecchia M., et al. (2015). The clinical implications and neurophysiological background of using the self-mirroring technique to enhance the identification of emotional experiences: an example with rational emotive behavior therapy. *J. Rational Emot. Cogn. Behav. Ther.* 33 115–133. DOI: 10.1007/s10942-015-0205-z
- Viola, P., Jones, M. (2001) "Rapid object detection using a boosted cascade of simple features," Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2001, Kauai, HI, USA, 2001, pp. I-I, doi: 10.1109/CVPR.2001.990517. (HAAR) (n.d.). *HuggingFace*. Bert-Base-Uncased-Poems-Sentiment. <https://huggingface.co/nickwong64/bert-base-uncased-poems-sentiment>
- Williams, K. N., Herman, R. E., Gajewski, B., & Wilson, K. (2009). Elderspeak: Communication chain resistance in nursing home residents. *Journal of the American*

- Geriatrics Society*, 57(3), 488–491. <https://doi.org/10.1111/j.1532-5415.2008.02149.x>
- Wohltjen, S., & Wheatley, T. (2021). Eye contact marks the rise and fall of shared attention in conversation. *Proceedings of the National Academy of Sciences*, 118(37), e2106645118. <https://doi.org/10.1073/pnas.2106645118>
- Zadeh, A. B., Liang, P. P., Poria, S., Cambria, E., & Morency, L.-P. (2018). Multimodal language analysis in the wild: CMU-MOSEI dataset and interpretable dynamic fusion graph. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 2236–2246). Association for Computational Linguistics. <https://doi.org/10.18653/v1/P18-1208>
- Zhang et al. (2025). DialogueLLM: Context and Speaker Modeling for Conversational Emotion Recognition. *ACL Anthology*., <https://doi.org/10.1016/j.neunet.2025.107901>
- (n.d.). OpenVINO <https://docs.openvino.ai/2024/index.html>
- (n.d.) OpenCV. <https://opencv.org/>.