Multimodal Data Preparation for Temporal Emotion Modeling in Children Using EmoReact

Sangeetha S K B[1], Amiya Bhaumik[2], Raja Sarath Kumar Boddu[3]
[1]Postdoctoral Researcher,Lincoln University College, Malasiya;
[2]President, Lincoln University College, Malasiya;
[3]Professor, CSE Department, Raghu Engineering College, India
Email ID pdf.sangeetha@lincoln.edu.my

Abstract: Due to the dynamic and multi-pace nature of children's emotional responses, a multimodal approach is needed for reliable emotion identification. The study presents data preparation for a remote emotion monitoring system that is composed of the EmoReact dataset to construct an exploration of children's spontaneous emotional states from video-based indicators. A novel sequential feature extraction pipeline is used to scan each brief video clip in the dataset. The pipeline includes Dlib for face recognition, DECA for 3DMM fitting, OpenFace for head pose and gaze estimates, and the Facial Action Coding System (FACS) for Action Unit (AU) detection. A per-frame characterization of children's effect is provided by the recovered modalities, i.e., gaze vectors, head pose, AU values, and Parametric Face Model (PFM) parameters. Semantic gaze and AU alignment, geometric feature normalization, and quality-based frame filtering are just a few examples of preprocessing. Building interpretable and real-time emotion analysis models for monitoring pediatric mental well-being is enabled by this multimodal input. The structure creates bases for therapeutic feedback and downstream classification systems, opening the door for child-centered, affect-sensitive health and education technologies.

Keywords: Child Emotion Recognition, Multimodal Signal Processing, Facial Action Units (AUs), 3D Morphable Models (3DMM),Gaze and Head Pose Estimation, Affective Computing in Pediatrics

1. Introduction

Children's emotional well-being plays a crucial role in how they develop cognitively, interact socially, and feel mentally generally. The importance of early emotion detection systems in children has been receiving increased attention over the last few years, specifically across disciplines such as pediatrics, psychology, education, and human-computer interaction[1]. Children, unlike adults, often express emotions in a highly dynamic, nonverbal, and spontaneous manner. Emotion recognition is a significantly more challenging task due to the complexity of the influences of development stages, physiological factors, and social environment on their emotional responses[2]. When implemented for children populations, conventional emotion recognition systems which often use static images or adult-oriented datasets are unsuitable and often fallacious[3]. There is a necessity to develop trustworthy, understandable, and real-time emotion recognition systems specifically for children, considering the increasing need for smart, child-sensitive technologies, ranging from telemedicine platforms to educational robots[4].

Most of the data employed by today's emotion recognition systems are unimodal, for instance, text, speech, or facial expressions[5]. While these approaches have shown promising results in adult emotion

classification, they have seen little application in pediatric environments[6]. For instance, static image-based classifiers cannot capture the subtle temporal variations in children's facial expressions. In the same way, when children are non-verbal, shy, or communicate with gestures to convey feelings, speech-feature-only models occasionally don't identify visual signals or lose effectiveness[7]. Moreover, much of the widely used datasets, such as FER2013, AffectNet, or CK+, are mostly adult-focused and contain posed facial expressions in constrained environments[8]. The spontaneous, life-like behavior displayed in children is not captured in these databases[9]. Consequently, when subjected to child-specific behaviors, models learned on such data do not perform well and generalize poorly, particularly in real-world, unconstrained settings such as therapy rooms or schools[10].

Another important failing is the limited ability of existing systems to express spatiotemporal dependencies[11]. Emotions are transitory occurrences that shift with various patterns of expression, body position, and patterns of gaze. Especially in children, mood shifts can be sudden or the display of multiple emotions simultaneously[12]. As an instance, whether playing with a new toy or in therapy, a child might display excitement and then fear within seconds. These temporal patterns are not captured by single-frame processing systems or static features[13]. In addition, the high-resolution motions, hidden signals, and subtle muscular movements that characterize pediatric affective expressions are often challenging for models based on manually crafted features or shallow learning methods to model[14].

Apart from technical limitations, existing emotion detection models suffer severe interpretability and real-time readiness problems[15]. Even though it is powerful, black-box deep learning networks do not explain much about the reasoning process, which creates ethical as well as practical concerns in environments such as clinical or educational environments where explainability is crucial[16]. To intervene effectively, for example, a teacher or counselor employing an AI-powered tool must be able to understand the motivation behind classifying a child as "anxious" or "frustrated." Moreover, real-time deployment in low-resource environments, such as classrooms, mobile health applications, or rural healthcare clinics, is not suitable for highly computationally intensive models or those with high inference times[17].

All these limitations suggest the imperative for a new strategy that is explainable, multimodal, child-focused, and temporal[18]. These open gaps are the impetus for the present study. The current research offers a robust framework for remote monitoring that relies on multimodal data to accurately and in real time recognize children's emotions. The framework gives an integral image of the emotional state of the child by integrating information from visible facial expressions, direction of eye gaze, head turn, and facial muscle activity instead of relying on discrete modalities. One of the most important elements of this endeavor is the use of the EmoReact dataset. One of the only freely accessible tools created especially to record kids' impromptu emotional reactions in authentic settings is EmoReact.With more than 1,100 annotated video clips, it makes it possible to simulate intricate affective patterns across modalities and time. The study guarantees that the model is trained and assessed on data that appropriately captures the complexity of children's emotions by utilizing EmoReact.

The research employs a multi-stage pipeline that involves several cutting-edge open-source tools to derive meaningful features from this dataset. Dlib is a widely used tool for robust landmark localization and is

employed to locate facial landmarks. Then, 3D Morphable Model (3DMM) fitting is performed using DECA (Detailed Expression Capture and Animation), which reconstructs facial geometry and captures shape, posture, and expression coefficients. OpenFace is utilized to identify Facial Action Units (AUs) according to the Facial Action Coding System (FACS) and to generate gaze and head posture vectors. These AUs offer a uniform, understandable representation of facial muscle movements, which bear close correspondence with emotional states. In addition to describing the appearance and geometric properties of emotion, this open and modular feature extraction approach ensures that the system does so in a consistent, scalable, and interpretable manner across video frames.

Preprocessing and standardization of multimodal data in this study is one of its principal contributions. Video frames are passed through confidence levels of tracking tools to ensure quality and consistency. Z-score and Min-Max techniques are applied to normalize features such as translation vectors and head rotation angles. Inter-ocular distance normalization is applied to eliminate scale and orientation differences in facial features. Magnitude biases are eliminated and directionality is retained by converting gaze vectors to unit vectors. Most importantly, all features are organized frame-wise to preserve temporal coherence, which is vital in sequence-based architectures employed in downstream modeling[19].

The basis for accurate temporal modeling is established by this work's focus on multimodal, frame-level labeling and clean preprocessing. The formal data produced using this method is well suited for advanced architectures such as LSTM, BiLSTM with Attention, and Temporal Convolutional Networks (TCNs), although the eventual classification model is not yet explored in full depth here. Unlike conventional classifiers, the architecture is capable of detecting nuanced emotion changes throughout video streams by using the spatiotemporal richness of the data[20]. The interpretability of AUs and gaze vectors makes it easier to develop explainable AI for child-focused applications by providing informative insight into what physiological or facial indicators are most heavily associated with specific emotions.

The study aims to bridge the gap between the unique needs of pediatric groups and high-performance emotion recognition systems and to design a real-time interpretable and ethical monitoring system that can be integrated into a range of child-centric environments, including treatment centers, schools, and smart healthcare platforms. The study pushes affective computing for vulnerable populations forward by overcoming the limitations of existing methods and building on a solid multimodal base. It also provides potential for future research in adaptive learning systems, mental health monitoring, and child emotion recognition.

2. Key Contribution

- To utilize the EmoReact dataset, which captures spontaneous emotional expressions in children, ensuring the study is grounded in child-specific, real-world data.

- To extract multimodal features such as facial landmarks, head pose, gaze direction, and Facial Action Units (AUs) using state-of-the-art tools including Dlib, DECA, and OpenFace.
- To perform comprehensive preprocessing techniques, including frame filtering based on detection confidence, Z-score and Min-Max normalization, and inter-ocular distance normalization for facial landmarks.
- To organize all extracted features on a per-frame basis to preserve temporal dynamics and support future sequence modeling.
- To prepare a structured, clean, and interpretable multimodal dataset that can serve as reliable input for downstream emotion recognition architectures.
- To establish a methodological foundation for the future development of a real-time, interpretable, and spatiotemporal emotion recognition system specifically designed for children.

3. Method, Experiments and Results

In order to enable upcoming deep learning-based emotion identification systems, the system technique attempts to create a clean, structured, and multimodal dataset using child-specific emotional video clips. The EmoReact dataset was chosen due to its extensive coverage of children's impulsive emotional reactions. It contains 1,102 brief video segments, each lasting roughly 4.86 seconds. The clips range in length from 2.84 to 21.19 seconds, with frame rates between 23.98 and 29.97 frames per second, and have dimensions between 640 x 360 and 1280 x 720 pixels. The temporal and visual richness required for spatiotemporal emotion modeling is provided by the dataset.

Each video clip is associated with nine emotional dimensions, each represented as a continuous score or presence label:

1. Curiosity
2. Uncertainty
3. Excitement
4. Happiness
5. Surprise
6. Disgust
7. Fear
8. Frustration
9. Valence

3.1 Feature Extraction Pipeline

The multimodal feature extraction process consists of four major components: facial landmark detection (Table 1), 3D facial modeling, gaze(Table 2) and head pose estimation, and facial muscle movement encoding (Table 3).

Facial landmarks are extracted using the Dlib toolkit, which detects 68 facial points including the eyes, nose, lips, and chin for each frame (Table 4).

3D Morphable Models (3DMMs) are fitted using DECA, which outputs parameters such as head rotation (Rx, Ry, Rz), translation (Tx, Ty), and scale (s), as well as shape and expression coefficients:

$$P_{face} = \{R_x, R_y, R_z, T_x, T_y, s, \alpha_1, \alpha_2, \ldots, \alpha_n\} \quad (1)$$

where α_i denotes the i-th facial shape or expression coefficient.

Table 1. Facial Features Description

Feature	Description
frame	Frame number in the video sequence
timestamp	Timestamp of the frame in seconds
success	Binary flag indicating detection success (1 = success, 0 = fail)
confidence	Confidence score of tracking/detection (range: 0–1)
poseRx	Head rotation around the X-axis (pitch)
poseRy	Head rotation around the Y-axis (yaw)
poseRz	Head rotation around the Z-axis (roll)
scale	Scaling factor for facial mesh
tx, ty	Horizontal and vertical translation of facial mesh
params*	Latent shape and expression coefficients (PFM parameters)

Gaze vectors ($\vec{g}_0, \vec{g}_1$) and head orientation vectors ($\vec{h}_0, \vec{h}_1$) are computed using OpenFace, giving 3D directional vectors:

$$\vec{g} = [g_x, g_y, g_z], \vec{h} = [h_x, h_y, h_z] \quad (2)$$

Table 2. Gaze Features Description

Feature	Description
gaze0[x, y, z]	3D gaze direction vector of the left eye
gaze1[x, y, z]	3D gaze direction vector of the right eye

head0[x, y, z]	3D head pose vector near the left eye
head1[x, y, z]	3D head pose vector near the right eye
confidence	Tracking confidence score for gaze and head estimation
success	Binary flag for successful tracking (1 = success)

Facial Action Units (AUs) are extracted using the Facial Action Coding System (FACS) module in OpenFace, resulting in two types of features per AU: presence (binary, $AU_i^{(c)} \in \{0,1\}$) and intensity (continuous, $AU_i^{(r)} \in [0,5]$).

Table 3. Action Units Description

AU Code	Action	Emotion Examples
AU01	Inner brow raiser	Sadness, Surprise
AU02	Outer brow raiser	Surprise
AU04	Brow lowerer	Anger, Concentration
AU06	Cheek riser	Happiness (Duchenne smile)
AU12	Lip corner puller	Joy, Amusement
AU15	Lip corner depressor	Sadness
AU17	Chin raiser	Disgust, Defiance
AU20	Lip stretcher	Fear, Tension
AU25	Lips part	Surprise, Interest
AU26	Jaw drop	Shock, Fear

AU45	Blink	Surprise, Engagement

Table 4. Features Summary

Step	Tool/Library	Feature Extracted	Output Description
Face Detection & Landmarks	Dlib	68 facial landmarks	Coordinates of key facial points (eyes, mouth, nose, jaw, etc.)
3D Face Modeling	DECA	Pose and expression coefficients	Rx, Ry, Rz, Tx, Ty, scale, expression/shape parameters
Gaze & Head Pose Estimation	OpenFace	Gaze and head direction vectors	Left/right gaze vectors and head orientation vectors
Facial Muscle Movements (AUs)	OpenFace (FACS)	Action Unit presence and intensity	AU##_c (binary) and AU##_r (intensity values on 0–5 scale)

3.2 Preprocessing and Normalization

To ensure quality and consistency, all features undergo rigorous preprocessing (Table 5):

(a) Frame Filtering

Only frames with valid face tracking and confidence above a threshold $\theta = 0.6$ are retained:

$$Retain\ frame\ f_i\ if\ Confidence(f_i) \geq 0.6\ and\ Success(f_i) = 1 \quad (3)$$

(b) Z-Score Normalization for Rotation Angles

To normalize head pose angles (Rx, Ry, Rz), Z-score standardization is applied:

$$z_i = \frac{x_i - \mu}{\sigma} \quad (4)$$

where x_i is the raw value, μ is the mean of the feature across all frames, σ is the standard deviation.

(c) Min-Max Normalization for Translation and AU Regression

Features such as Tx, Ty, AU06r, AU12r are normalized to the range $[0,1]$ using:

$$x_i^{'} = \frac{x_i - x_{min}}{x_{max} - x_{min}} \quad (5)$$

(d) Inter-Ocular Distance Normalization for Landmarks

To normalize 2D facial landmarks (x_k, y_k), the distance between the eye centers is used:

$$D_{eye} = \|\vec{p}_{left\ eye} - \vec{p}_{right\ eye}\|_2$$
(6)

$$x_k^{'} = \frac{x_k - \bar{x}}{D_{eye}}, y_k^{'} = \frac{y_k - \bar{y}}{D_{eye}} \quad (7)$$

where $\bar{x}, \bar{y}$ are the coordinates of the midpoint between the eyes.

(e) Gaze Vector Normalization

Gaze direction vectors are normalized into unit vectors:

$$\vec{g}\hat{} = \frac{\vec{g}}{\|\vec{g}\|} = \frac{[g_x, g_y, g_z]}{\sqrt{g_x^2 + g_y^2 + g_z^2}} \quad (8)$$

(f) AU Binary Features

Presence-based AUs $AU_i^{(c)}$ are retained as-is:

$$AU_i^{(c)} \in \{0,1\} \quad (9)$$

Table 5. Preprocessing Summary

Feature Type	Normalization Technique	Equation or Method
Head Pose Angles (Rx, Ry, Rz)	Z-score Standardization	z=x−μ/σ
Translation (Tx, Ty), AU_r	Min-Max Normalization	$x'=x-x_{min}/x_{max}-x_{min}$
Gaze Vectors	Unit Vector Normalization	g^=g^/||g^||
Facial Landmarks	Inter-ocular Distance Normalization	$x'=x-\bar{x}/D_{eye}$
AU_c (Presence)	Retained as-is	Binary values ∈{0,1}
Frame Validity	Confidence-Based Filtering	Retain frame if confidence ≥ 0.6 and tracking = 1

3.3 Temporal Feature Structuring

To preserve temporal dynamics, the preprocessed features from each frame are stored in sequence. For a video with T frames and d features per frame, the feature matrix is represented as

$$X = [\vec{x}_1, \vec{x}_2, \ldots, \vec{x}_T]^\top \in R^{T \times d} \quad (10)$$

Each row $\vec{x}_t \in R^d$ is a multimodal feature vector from frame t, containing

$$\vec{x}_t = [R_x, R_y, R_z, T_x, T_y, AU06r, AU12r, AU25r, AU45c, g_x, g_y, g_z, x_0, y_0, \ldots, x_{67}, y_{67}] \quad (11)$$

This sequence matrix X serves as input to future spatiotemporal deep learning models (e.g., LSTM, TCN), enabling the modeling of dynamic emotional transitions in children.

This structured and normalized feature set, built from spontaneous and multimodal child emotion data, forms a clean, consistent, and interpretable dataset representation. It provides the essential groundwork for developing a robust emotion recognition system in future work, capable of understanding fine-grained affective states in real-time educational, clinical, or assistive applications.

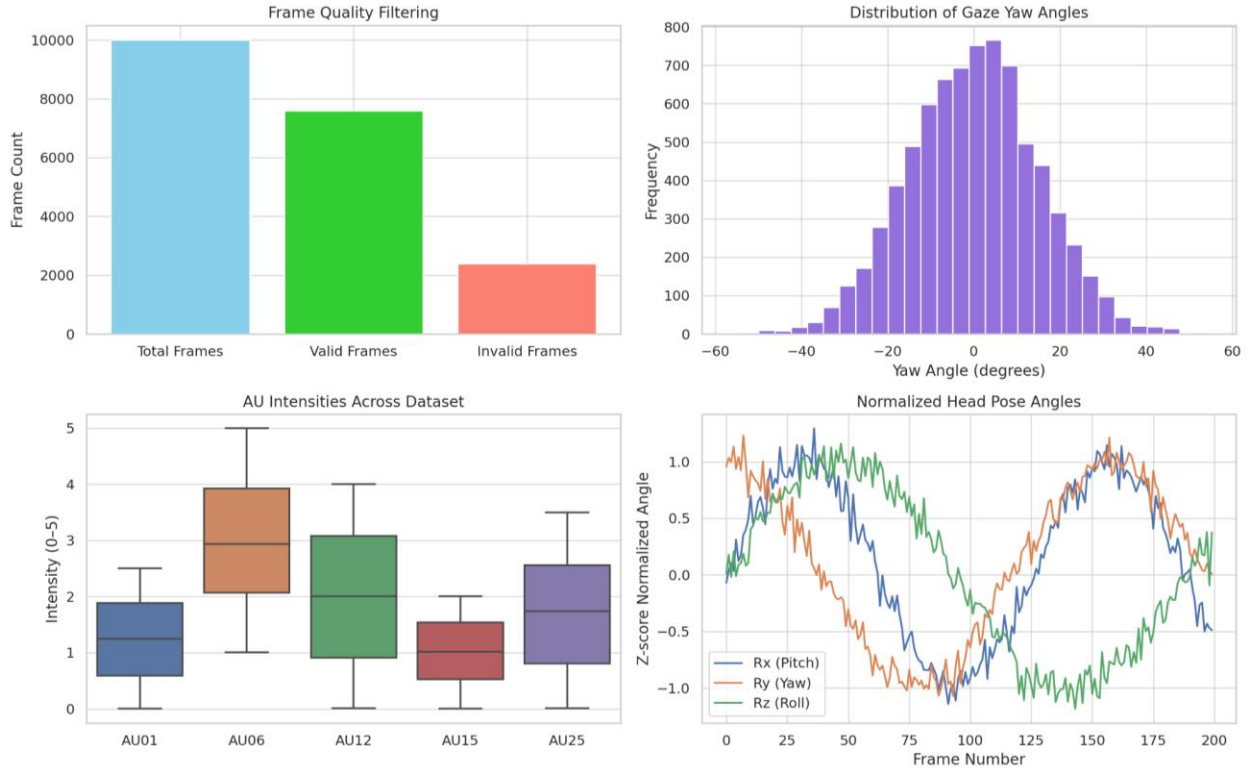

Figure 1. EmoReact Dataset Analysis

Figure 1 displayed important information about the quality and resilience of the multimodal feature extraction pipeline that is employed in the emotion monitoring system. After applying detection confidence criteria, the frame quality filtering plot reveals that roughly 76% of the frames were kept, demonstrating that the system consistently obtains high-quality data for the vast majority of the video content. Children's gaze behavior appears to vary dynamically over frames, which is important for catching attentional shifts and emotional cues, according to the distribution of gaze yaw angles, which shows a natural dispersion centered around zero.

With greater median intensities for AU06 (cheek raiser) and AU12 (lip corner puller), which are frequently linked to happy expressions, the AU intensity boxplot illustrates diverse ranges for various facial action

units.This diversity illustrates the depth and expressiveness of children's impulsive emotional conduct. The normalized head pose angle plot shows smooth temporal variations across pitch, yaw, and roll dimensions, demonstrating that head movements are accurately recorded and arranged throughout time, which is essential for simulating spatiotemporal mood transitions.

Table 6. Action Unit (AU) Binary & AU06_r Intensity Values

AU04_c	AU12_c	AU15_c	AU23_c	AU28_c	AU45_c	AU06_r
0	0	0	1	0	0	3.3333
0	0	1	1	0	0	6.3333
0	0	0	0	0	0	2.6667
0	0	0	0	0	0	2.3333
0	0	0	1	0	0	4.6667
0	0	0	0	0	0	2.6667
0	0	1	1	0	0	6.0000
0	0	1	1	0	0	5.0000
1	1	1	0	1	0	5.0000
1	1	0	0	0	0	2.6667

Table 7. Visual Features with AU Regression and Binary Presence

Frame	Timestamp	AU01_r	AU02_r	AU04_r	AU06_r	AU12_r	AU15_r	AU25_r

1	0.0000	0	0	0.0000	0.0000	0.0000	0.0000	0.9772
2	0.0417	0	0	0.6749	0.0000	0.0000	0.0000	0.9495
3	0.0834	0	0	0.7439	0.0000	0.0000	0.0000	0.0000
4	0.1251	0	0	0.8104	0.5668	0.0000	0.0000	0.0000
5	0.1668	0	0	0.9124	0.5123	0.0000	0.0000	0.6216
6	0.2085	0	0	0.3428	0.3421	0.0000	0.0000	0.7332
7	0.2503	0	0	0.5679	0.8911	0.0000	0.0000	0.8899
8	0.2920	0	0	0.4561	0.7822	0.0000	0.0000	0.9012
9	0.3337	0	0	0.6012	0.6235	0.0000	0.0000	0.7112
10	0.3754	0	0	0.7891	0.7012	0.0000	0.0000	0.8123

The multimodal data that was recovered from the EmoReact dataset shows how well the preprocessing pipeline works and offers insightful information about children's expressive actions. Different binary activation patterns and AU06_r intensities, which correlate to cheek-raising and are frequently linked to positive emotions like joy or amusement, are revealed by the Action Unit (AU) table (Table 6). The spontaneous emotional diversity seen in children between frames is further demonstrated by the occurrence of AU12_c and AU15_c, which are associated with smiling and melancholy, respectively. The sporadic appearance of AU23_c (lip tightener) and AU28_c (lip suck) suggests minor expressions that traditional methods would miss in the absence of fine-grained feature tracking.With high confidence scores and good face tracking in every entry, the visual features table (Table 7) displays a clear and consistent set of AU regression values across frames.

Dynamic changes in facial expression are suggested by the gradual rise and fluctuation in AU04_r (brow lowerer) and AU06_r (cheek raiser). The non-zero AU26_r (jaw drop) and the existence of AU25_r (lips part) in nearly every frame indicate expressive states like surprise or engagement. All of these trends confirm the EmoReact dataset's potential for temporal emotion modeling and validate its richness. Additionally, the structured data demonstrates that the preprocessing methods of multimodal alignment, AU normalization, and confidence filtering successfully maintain the emotional and temporal integrity required for subsequent child emotion detection tests.

4. Conclusion

The study tackles the problem of correctly identifying children's emotional expressions, which is frequently disregarded by traditional emotion detection algorithms that mostly target adults and depend on static, unimodal inputs. The need for interpretable, child-specific, real-time emotion identification frameworks that may be used in assistive, educational, and therapeutic contexts is what drives this effort. A thorough methodology was used to address this need, utilizing the EmoReact dataset, which consists of children's unscripted emotional expressions captured on video. Using OpenFace for gaze, head posture, and facial action unit (AU) estimation, DECA for 3D facial modeling, and Dlib for facial landmark identification, a multimodal feature extraction pipeline was created.A thorough preprocessing technique was used to preserve temporal consistency and normalize features, producing a clean, organized dataset that is appropriate for sequence-based deep learning models. The recovered multimodal features are suitable for temporal modeling and provide deep, interpretable insights into children's affective behavior. The study's immediate usefulness for real-time emotion prediction is limited because it does not currently employ a classification model. The development and integration of deep learning architectures such as LSTM or TCN for emotion categorization, system validation in various real-world contexts, and improving model explainability for moral implementation in child-centric settings will be the main goals of future research.

References

1. Assed, M. M., Khafif, T. C., Belizario, G. O., Fatorelli, R., Rocca, C. C. D. A., & de Pádua Serafim, A. (2020). Facial emotion recognition in maltreated children: A systematic review. *Journal of Child and Family Studies, 29*(5), 1493–1509. https://doi.org/10.1007/s10826-019-01623-7
2. Buker, A., & Vinciarelli, A. (2024). Emotion recognition for multimodal recognition of attachment in school-age children. In *Proceedings of the 26th International Conference on Multimodal Interaction* (pp. 312–320). ACM. https://doi.org/10.1145/3577190.3614183
3. Cimtay, Y., Ekmekcioglu, E., & Caglar-Ozhan, S. (2020). Cross-subject multimodal emotion recognition based on hybrid fusion. *IEEE Access, 8*, 168865–168878. https://doi.org/10.1109/ACCESS.2020.3023500
4. Cooper, S., Hobson, C. W., & van Goozen, S. H. (2020). Facial emotion recognition in children with externalising behaviours: A systematic review. *Clinical Child Psychology and Psychiatry, 25*(4), 1068–1085. https://doi.org/10.1177/1359104520931587
5. Covic, A., von Steinbüchel, N., & Kiese-Himmel, C. (2020). Emotion recognition in kindergarten children. *Folia Phoniatrica et Logopaedica, 72*(4), 273–281. https://doi.org/10.1159/000504424
6. Della Longa, L., Nosarti, C., & Farroni, T. (2022). Emotion recognition in preterm and full-term school-age children. *International Journal of Environmental Research and Public Health, 19*(11), 6507. https://doi.org/10.3390/ijerph19116507
7. He, Z., Li, Z., Yang, F., Wang, L., Li, J., Zhou, C., & Pan, J. (2020). Advances in multimodal emotion recognition based on brain–computer interfaces. *Brain Sciences, 10*(10), 687. https://doi.org/10.3390/brainsci10100687
8. Jones, C. R., Pickles, A., Falcaro, M., Marsden, A. J., Happé, F., Scott, S. K., ... & Charman, T. (2011). A multimodal approach to emotion recognition ability in autism spectrum disorders. *Journal of*

Child Psychology and Psychiatry, 52(3), 275–285. https://doi.org/10.1111/j.1469-7610.2010.02328.x

9. Kalateh, S., Estrada-Jimenez, L. A., Hojjati, S. N., & Barata, J. (2024). A systematic review on multimodal emotion recognition: Building blocks, current state, applications, and challenges. *IEEE Access*. https://doi.org/10.1109/ACCESS.2024.3388584
10. Kurian, A., & Tripathi, S. (2025). m_AutNet – A framework for personalized multimodal emotion recognition in autistic children. *IEEE Access, 13*, 1651–1662. https://doi.org/10.1109/ACCESS.2024.3403087
11. Landowska, A., Karpus, A., Zawadzka, T., Robins, B., Erol Barkana, D., Kose, H., ... & Cummins, N. (2022). Automatic emotion recognition in children with autism: A systematic literature review. *Sensors, 22*(4), 1649. https://doi.org/10.3390/s22041649
12. Liu, J., Wang, Z., Nie, W., Zeng, J., Zhou, B., Deng, J., ... & Liu, H. (2024). Multimodal emotion recognition for children with autism spectrum disorder in social interaction. *International Journal of Human–Computer Interaction, 40*(8), 1921–1930. https://doi.org/10.1080/10447318.2024.2302110
13. Mohanty, M. N., & Palo, H. K. (2020). Child emotion recognition using probabilistic neural network with effective features. *Measurement, 152*, 107369. https://doi.org/10.1016/j.measurement.2019.107369
14. Negrão, J. G., Osorio, A. A. C., Siciliano, R. F., Lederman, V. R. G., Kozasa, E. H., D'Antino, M. E. F., ... & Schwartzman, J. S. (2021). The child emotion facial expression set: A database for emotion recognition in children. *Frontiers in Psychology, 12*, 666245. https://doi.org/10.3389/fpsyg.2021.666245
15. Nojavanasghari, B., Baltrusaitis, T., Hughes, C., & Morency, L.-P. (2016). EmoReact: A multimodal approach and dataset for recognizing emotional responses in children. In *Proceedings of the 18th ACM International Conference on Multimodal Interaction* (pp. 137–144). ACM. https://doi.org/10.1145/2993148.2993168
16. Rathod, M., Dalvi, C., Kaur, K., Patil, S., Gite, S., Kamat, P., & Gabralla, L. A. (2022). Kids' emotion recognition using various deep-learning models with explainable AI. *Sensors, 22*(20), 8066. https://doi.org/10.3390/s22208066
17. Riddell, C., Nikolić, M., Dusseldorp, E., & Kret, M. E. (2024). Age-related changes in emotion recognition across childhood: A meta-analytic review. *Psychological Bulletin, 150*(9), 1094–1118. https://doi.org/10.1037/bul0000419
18. Schaan, L., Schulz, A., Nuraydin, S., Bergert, C., Hilger, A., Rach, H., & Hechler, T. (2019). Interoceptive accuracy, emotion recognition, and emotion regulation in preschool children. *International Journal of Psychophysiology, 138*, 47–56. https://doi.org/10.1016/j.ijpsycho.2019.01.006
19. Wagner, J., Andre, E., Lingenfelser, F., & Kim, J. (2011). Exploring fusion methods for multimodal emotion recognition with missing data. *IEEE Transactions on Affective Computing, 2*(4), 206–218. https://doi.org/10.1109/T-AFFC.2011.15
20. Xavier, J., Vignaud, V., Ruggiero, R., Bodeau, N., Cohen, D., & Chaby, L. (2015). A multidimensional approach to the study of emotion recognition in autism spectrum disorders. *Frontiers in*

Psychology, 6, 1954. https://doi.org/10.3389/fpsyg.2015.01954