Perceived emotion recognition from nonverbal communication cues in images and videos

Willams de Lima Costa[1], **Estefania T. Martinez**[2],
Lucas S. Figueiredo[1], **Veronica Teichrieb**[1]

[1]Voxar Labs, Centro de Informática,
Universidade Federal de Pernambuco, BR

[2]Data Management and Biometrics Group,
University of Twente, NL

{wlc2,lsf,vt}@cin.ufpe.br, e.talaveramartinez@utwente.nl

Abstract. *Identifying emotions enables intelligent systems to monitor user behavior, leading to a deeper understanding of the person. Perceiving emotion occurs naturally in humans through the communication of nonverbal cues, in which emotional features are communicated implicitly through multiple channels. In this thesis, we propose three automatic frameworks supported by evidence from the behavioral psychology literature for emotion recognition: (1) an emotion recognition approach based solely on situational context, (2) a body-language model that uses gait features to predict emotion from walking styles from videos, and (3) a multi-cue model that combines facial expression, situational context, and body language to perceive emotions in images. The obtained results by our proposed models equal the state of the art but with severe improvements related to computational cost.*

Resumo. *Identificar emoções permite que sistemas inteligentes monitorem o comportamento de usuários, levando a uma compreensão mais profunda da pessoa. A percepção emocional é um processo que ocorre naturalmente em humanos por meio da comunicação de sinais não verbais, nos quais características emocionais são comunicadas implicitamente por meio de múltiplos canais. Nesta tese, propomos três técnicas que são respaldadas por evidências da literatura de psicologia de comportamento para o reconhecimento de emoções: (1) uma abordagem de reconhecimento de emoções baseada exclusivamente no contexto situacional, (2) um modelo de linguagem corporal que utiliza características de marcha para prever emoções a partir de estilos de caminhada em vídeos, e (3) um modelo que recebe múltiplos sinais extraídos de expressões faciais, contexto situacional e linguagem corporal para perceber emoções em imagens. Os resultados obtidos por nossos modelos se igualam ao estado da arte, mas com melhorias significativas relacionadas ao custo computacional.*

1. Introduction

Twenty-five years after Picard's seminal book *Affective Computing* [Picard 2000], many of the technical limitations highlighted, such as memory constraints and inefficient computing, have been largely solved. However, the central argument remains unsolved:

computers still struggle to recognize and respond to human affective behavior. As Human-Computer Interaction (HCI) moves toward immersive and natural interaction paradigms [Valli 2007, Valli 2008], the need to decode affective states from human behavior becomes critical.

Applications in virtual, augmented, and mixed reality rely increasingly on the ability of machines to understand users in a human-like way, beyond traditional input modalities. Recognizing emotion from visual cues, such as facial expressions, posture, and contextual information, enables more responsive, adaptive, and immersive systems.

This task, however, extends beyond classical perception tasks like object detection. Emotion recognition involves cognitive inference [Halford and Hine 2016, Nes et al. 2023], where models must decode internal, subjective states based on external nonverbal cues. In this thesis, we propose a strong definition of emotion recognition as the task of inferring perceived affective states from nonverbal cues, including emotion, mood, and thought. We argue that vision-based systems capable of extracting such cues are essential to enabling truly immersive human-machine experiences.

This thesis introduces a multi-faceted investigation of emotion recognition through computer vision, presenting contributions such as a psychology-informed theoretical framework for understanding nonverbal cues in affective behavior, a novel benchmark focused on Latin America, and three approaches suitable for XR integration. These contributions advance affective computing toward more realistic and emotionally aware XR environments, enabling next-generation applications such as emotionally responsive avatars, adaptive training simulators, and empathic virtual agents.

2. The Human Perspective on Emotion

Human communication relies not only on what is said, but also on how it is expressed through subtle, unintentional signals. Nonverbal communication, such as facial expressions, body posture, and motion, plays a central role in shaping how we perceive others' emotions. These cues often emerge spontaneously and are decoded naturally and rapidly by observers [Jacob et al. 2016, Buck 1991]. In this work, we focus on three primary nonverbal cues: facial expressions, situational context, and body language.

Finally, emotion perception is not purely biological nor purely cultural; it emerges from the interaction of both. While studies with isolated or blind participants support the universality of certain expressions [Ekman 1993, Tracy and Matsumoto 2008], other works reveal cultural shaping of emotional behavior [Mesquita et al. 2017]. Cultural norms influence how emotions are expressed, perceived, and valued [Mesquita and Frijda 1992], raising questions about the generalizability of datasets collected in North America and Europe.

3. A Dataset for Emotion Recognition in Latin America

To address the cultural limitations of existing emotion recognition datasets, which are predominantly captured in North America and Europe, we introduce EiLA, a culturally-aware dataset focused on Brazilian and Latin American individuals in realistic and diverse scenarios. By incorporating situational context and emphasizing naturalistic behavior, EiLA enables emotion recognition approaches that extend beyond facial

Figure 1. Sample images from EiLA.

expressions, promoting models better suited for deployment in Latin American immersive environments.

The dataset comprises over 4,500 annotated frames across 15 minutes of video, featuring 78 participants with a balanced distribution of gender and skin tone diversity. Annotations were performed using Ekman's basic emotions and validated through multi-rater agreement, achieving a Fleiss' κ score significantly higher than standard datasets like EMOTIC, indicating strong inter-annotator consistency and reliability for future affective computing benchmarks.

4. Collecting and processing affective features

4.1. High-level context representation for emotion recognition

Building on the idea that intelligent systems must process nonverbal cues shaped by situational context, we argue that semantic, high-level representations generalize better to real-world data than low-level features. While prior works like EMOTIC [Kosti et al. 2017], CAER-Net [Lee et al. 2019], EmotiCon [Mittal et al. 2020], GLAMOR-Net [Le et al. 2021], and EmotionRAM [Costa et al. 2022] have leveraged contextual cues, our method focuses on summarizing emotional context without relying on fine-grained facial or behavioral signals. For example, in XR-based training simulations, detecting whether the virtual environment elicits confusion or engagement can guide adaptive scene transitions without needing to track each user's facial expression.

Methodology. We propose a method for high-level context representation in emotion recognition by leveraging semantic information extracted from image captions. We generate refined textual descriptions of images, remove noise and bias-prone elements, and map words to affective and semantic attributes using SenticNet and WordNet. These enriched descriptors are then structured into graphs, where nodes represent words, emotions, mood tags, and related concepts, and edges are weighted based on co-occurrence and sentiment polarity. Graphs are encoded using a GIN-based Graph Convolutional Network that jointly predicts categorical and continuous emotional dimensions. Our architecture is lightweight and suitable for deployment on low-power devices, achieving robust performance on the EMOTIC dataset while promoting semantic interpretability and cultural adaptability in immersive systems.

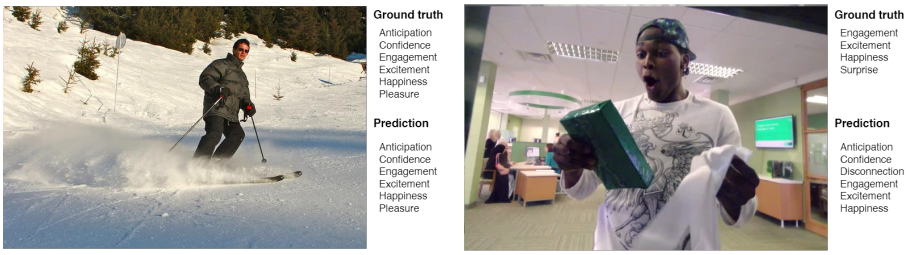

Figure 2. Sample results.

Results. Our proposed method achieves competitive performance in emotion recognition, outperforming several graph-based approaches such as Zhang et al. [Zhang et al. 2019] and variations of DRM/LEKG [Chen et al. 2023], while using only one nonverbal cue. Although other models have better quantitative performance, we only use one cue — context. We demonstrate that our approach remains effective and efficient, making it suitable for edge deployment, reaching up to 248 fps on a standard deep learning machine and 56 fps on a consumer-grade CPU (Intel i7-4790K). We show results of our model in Figure 2.

4.2. An analysis of gait for emotion recognition

Beyond its physical attributes, behavioral psychology research highlights gait as a social and emotional signal, which can also serve as a form of dynamic body language in naturalistic settings. Building on this, we hypothesize that gait patterns encode affective states and can be used as a reliable nonverbal cue for emotion recognition in unconstrained, real-world environments.

Methodology. Given a video sequence, we extract 3D body keypoints and represent the skeleton as a graph, where joints are vertices and bones are edges, enabling the use of Graph Convolutional Networks (GCNs) for gait analysis. Our pipeline encodes an adaptive joint topology, which allows for fine-grained modeling of gait patterns through spatial average pooling. To enrich the representation, our model also encodes affective features, such as joint angles, distances, and motion dynamics. This architecture enables robust emotion recognition from gait in unconstrained scenarios.

Results. Our model outperforms all other methods from the state-of-the-art in accuracy, but is also less ambiguous on classes such as Neutral and Happy. Besides the increase in accuracy, our model requires fewer computational resources.

4.3. Multiple cue processing in static domains

We propose EmotionRAM, a fast and efficient multi-cue framework for static emotion recognition. Our approach addresses two key limitations: the overreliance on facial expressions and the high computational cost of current approaches, which hinders large-scale deployment in smart environments. We process three nonverbal cues, namely facial expressions, scene context, and static body posture, to improve recognition. The simplicity and speed of our approach make it feasible for XR applications, as previously explored on SVR [Marinho et al. 2024].

Methodology. Given an image, we extract three different cues. Firt, we employ a face encoding stream that extracts features using the 2D position of the face bounding

(a) Fear (b) Sad (c) Surprise

Figure 3. Visualization of which regions of the context are more important for the classification. Reduced for brevity to three classes.

box through a simple Convolutional Neural Network (CNN). Then, we occlude the face using a black rectangle and feed this new image to a context encoding stream, forcing this CNN to learn representations from the background. Finally, we extract the 2D body pose of the person and learn body language representations. With these features, we adaptively learn each one's importance (e.g., reducing importance in cases where the context is not representative, or the face is not visible) to classify the emotion of someone.

Results. Our method is 0.12% worse than the state-of-the-art, GLAMOR-Net, but also 9x faster. This highlights a strong positive trade-off regarding accuracy and inference time, which allows for stable classification over stable performance. We show a few results in Figure 3.

5. Conclusion

This work presents a comprehensive investigation into vision-based emotion recognition through nonverbal cues, contributing new models, datasets, and insights that enhance the development of emotionally-aware interactive systems. Our contributions not only improve the accuracy and generalizability of emotion recognition across cultural contexts, but also promote feasible integration into immersive environments such as XR systems. For future works, we will continue exploring the intersection of immersive technologies and behavior through AI agents for everyday support systems.

Acknowledgements

This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Finance Code 001.

References

Buck, R. (1991). Motivation, emotion and cognition: A developmental-interactionist view. *International review of studies on emotion*, 1:101–142.

Chen, J., Yang, T., Huang, Z., Wang, K., Liu, M., and Lyu, C. (2023). Incorporating structured emotion commonsense knowledge and interpersonal relation into context-aware emotion recognition. *Applied Intelligence*, 53(4):4201–4217.

Costa, W., Macêdo, D., Zanchettin, C., Talavera, E., Figueiredo, L. S., and Teichrieb, V. (2022). A fast multiple cue fusing approach for human emotion recognition. *SSRN preprint 4255748*.

Ekman, P. (1993). Facial expression and emotion. *American psychologist*, 48(4):384.

Halford, G. S. and Hine, T. J. (2016). Fundamental differences between perception and cognition aside from cognitive penetrability. *Behavioral and Brain Sciences*, 39.

Jacob, H., Kreifelts, B., Nizielski, S., Schütz, A., and Wildgruber, D. (2016). Effects of emotional intelligence on the impression of irony created by the mismatch between verbal and nonverbal cues. *PloS one*, 11(10):e0163211.

Kosti, R., Alvarez, J. M., Recasens, A., and Lapedriza, A. (2017). Emotion recognition in context. *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1667–1675.

Le, N., Nguyen, K., Nguyen, A., and Le, B. (2021). Global-local attention for emotion recognition. *Neural Computing and Applications*, pages 1–15.

Lee, J., Kim, S., Kim, S., Park, J., and Sohn, K. (2019). Context-aware emotion recognition networks. *IEEE International Conference on Computer Vision*, pages 10143–10152.

Marinho, I., Padilha, R., Vitorino, G., Batista, M., Oliveira, Y., Almeida, J. V., Costa, W., Araújo, C., and Teichrieb, V. (2024). Eyes of fear: Leveraging emotion recognition for virtual reality experience. In *Proceedings of the 26th Symposium on Virtual and Augmented Reality*, pages 90–96.

Mesquita, B., Boiger, M., and De Leersnyder, J. (2017). Doing emotions: The role of culture in everyday emotions. *European Review of Social Psychology*, 28(1):95–133.

Mesquita, B. and Frijda, N. H. (1992). Cultural variations in emotions: a review. *Psychological bulletin*, 112(2):179.

Mittal, T., Guhan, P., Bhattacharya, U., Chandra, R., Bera, A., and Manocha, D. (2020). Emoticon: Context-aware multimodal emotion recognition using frege's principle.

Nes, A., Sundberg, K., and Watzl, S. (2023). The perception/cognition distinction. *Inquiry*, 66(2):165–195.

Picard, R. W. (2000). *Affective computing*. MIT press.

Tracy, J. L. and Matsumoto, D. (2008). The spontaneous expression of pride and shame: Evidence for biologically innate nonverbal displays. *Proceedings of the National Academy of Sciences*, 105(33):11655–11660.

Valli, A. (2007). Natural interaction. *White Paper*.

Valli, A. (2008). *Alessandro Valli - Notes on Natural Interaction*.

Zhang, M., Liang, Y., and Ma, H. (2019). Context-aware affective graph reasoning for emotion recognition. In *2019 IEEE International Conference on Multimedia and Expo (ICME)*, pages 151–156. IEEE.