Enhancing Contextual Emotion Recognition Using Large Vision-Language Models: A Review

Vaishnavi Chevale*, Santosh Gaikwad**, Arshiya Khan***, R.S. Deshpande****
*Department Of Computer Science and Application, JSPM UNIVERSITY
Vaishnavichevale42@gmail.com
**Associate Professor, Faculty of Science and Technology, JSPM University Pune
santosh.gaikwadcsit@gmail.com
***Assistant Professor, Faculty of Science and Technology, JSPM University Pune
ak.scos@jspmuni.ac.in
****Professor and Dean, Faculty of Science and Technology, JSPM University Pune
dean-fst@jspmuni.ac.in

***Abstract*—Emotion recognition has become increasingly relevant in human-computer interaction, mental health assessment, and social robotics. Traditional models often rely on facial expressions or speech to identify emotions, ignoring the contextual subtleties that drive real emotional states. The emergence of Large Vision-Language Models (VLMs), which integrate multi-modal input to derive semantic understanding, presents an opportunity to enhance the performance and contextual sensitivity of emotion recognition systems. This review paper explores the role of VLMs in contextual emotion recognition, discussing foundational architectures like CLIP, Flamingo, and GPT-Vision hybrids, and their application to understanding emotion in visual and linguistic settings. We also investigate key challenges, datasets, metrics, and future prospects in building robust, real-time, and ethically grounded emotion recognition models.**

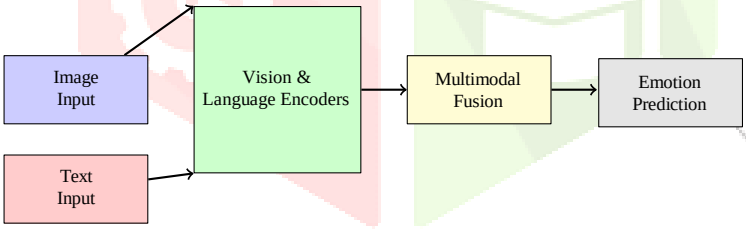

Fig. 1: VLM-Based Emotion Recognition Pipeline

I. Introduction

Emotion recognition is central to developing machines that can understand and interact with humans naturally. Traditional emotion recognition systems use facial expression analysis, audio signal processing, or physiological signals to detect affective states [4], [6]. However, human emotion is deeply contextual, depending on background information, situational cues, and semantic interpretation of surroundings.

Recent advancements in artificial intelligence, particularly Large Vision-Language Models (VLMs), offer a promising solution [1]–[3]. These models integrate visual and textual data to produce richer and more accurate semantic representations. By considering both what is seen and what is said, VLMs can interpret emotional context in a way that mimics human understanding more closely.

Large-scale pretraining, zero-shot transfer capabilities, and fine-tuning on emotion-specific datasets allow these models to outperform traditional unimodal models in contextual emotion recognition [5], [6]. Vision-language models such as CLIP [1], Flamingo [2], and LLaVA [3] demonstrate the potential of joint image-text understanding for nuanced tasks like emotion classification.

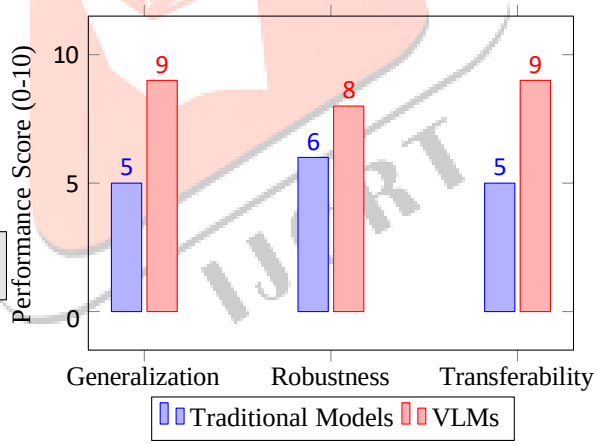

Fig. 2: Performance Comparison: Traditional vs. Vision-Language Models

II. Background and Motivation

Emotion recognition has long relied on Ekman's six basic emotions (happiness, sadness, anger, surprise, fear, disgust) and extended frameworks like Plutchik's wheel. However, conventional machine learning approaches often fail to capture the deeper semantic meaning or the environment in which emotions are expressed.

The motivation for using VLMs stems from their ability to link textual descriptions with visual content, allowing for a more comprehensive understanding of emotional states. For example, interpreting a smile in a funeral scene requires contextual awareness — something traditional facial recognition

models might misinterpret. VLMs help to bridge that semantic gap by leveraging paired image-text datasets and powerful attention mechanisms.

III. Theoretical Foundations

Contextual emotion recognition is grounded in theories of affective computing and multimodal semantics. The key theoretical frameworks include:

- **Appraisal Theory:** Emotions arise from evaluations of events based on relevance, implications, and coping potential.
- **Multimodal Communication Theory:** Emotions are expressed through gestures, expressions, speech, and language, requiring unified processing.
- **Cognitive-Affective Integration:** Emotion understanding requires cognitive models capable of interpreting environmental cues and past experience.

IV. Vision-Language Model Architectures

A. CLIP (Contrastive Language-Image Pretraining)

Developed by OpenAI, CLIP learns visual concepts from natural language supervision. It aligns image and text embeddings in a shared latent space, enabling zero-shot classification. Though not emotion-specific, CLIP's general understanding enables it to infer emotions when fine-tuned.

B. Flamingo

Flamingo is a few-shot learning model trained on vast web-scale image-text data. It incorporates vision encoders and autoregressive language decoders for tasks requiring fine-grained understanding, including emotion detection in social scenes.

C. LLaVA and GPT-Vision Models

LLaVA and other GPT-4V models extend LLM capabilities to visual inputs, supporting question answering, captioning, and contextual understanding. These models have the potential to extract emotional states based on dialogues, scene settings, and character expressions.

V. Fusion Techniques for Contextual Emotion Understanding

- **Early Fusion:** Combines image and text inputs at the feature level. Suitable for emotion classification in multimodal datasets.
- **Late Fusion:** Processes visual and textual data separately and merges predictions. Useful in ensemble learning.
- **Cross-Attention:** Allows one modality to attend to the other, as in Transformer-based architectures.

VI. Datasets

TABLE I: Common Datasets for Vision-Language Emotion Recognition

Dataset	Modalities	Application
EMOTIC	Image + Context Labels	Visual emotion detection
MELD	Video + Dialogue	Multimodal emotion in conversations
AVE	Audio + Video + Text	Affective event detection
GHOUL	Image + Captions	Social scene understanding

VII. Evaluation Metrics

To evaluate contextual emotion recognition, several metrics are used:

- **Accuracy and F1-Score:** Measure overall correctness and balance in class predictions.
- **Mean Average Precision (mAP):** For multi-label classification problems.
- **Emotion-wise Confusion Matrix:** Helps identify confusion among similar emotions.
- **BLEU, ROUGE, CIDEr:** For evaluating descriptive emotional captioning.

VIII. Applications

- **Healthcare:** Monitoring patient mood or detecting depression from images and messages.
- **Education:** Adapting content delivery based on student emotion.
- **Social Media:** Analyzing user sentiment based on image-caption pairs.
- **Robotics:** Emotional navigation in human-robot interactions.
- **Security and Surveillance:** Detecting agitation, stress, or threat levels in public spaces.

IX. Case Studies

A. Case Study 1: Emotion-aware Tutoring System

A multimodal VLM-based tutoring system dynamically adapts explanations based on visual feedback of students' confusion or engagement. Early trials show increased retention and satisfaction rates.

B. Case Study 2: Mental Health Companion App

An app integrated with GPT-Vision assesses facial and textual cues from journal entries and selfies to provide mood summaries and well-being tips.

X. Comparative Analysis

Compared to traditional CNN+LSTM pipelines, VLMs offer:

- **Better generalization:** Especially in unseen domains.
- **Improved robustness:** Against background noise and ambiguity.
- **Transfer learning potential:** Easily adapted to new tasks with minimal data.

XI. Limitations of Existing Work

- **Overfitting to Visual Cues:** Some models neglect the linguistic context.
- **Cultural and Linguistic Biases:** Pretrained datasets often lack diversity.
- **Temporal Resolution:** Emotion dynamics over time are poorly captured.

XII. Real-Time Emotion Recognition

To meet real-time demands in AR/VR or live interactions, models must be optimized for:

- Latency reduction via quantization or pruning.
- Edge deployment using TensorRT or ONNX runtimes.
- Streaming data pipelines for continuous input.

XIII. Role of Multimodal Transformers

Multimodal Transformers like Perceiver and BLIP-2 use unified attention blocks to handle multi-modal fusion. Their architectural flexibility allows real-time contextual emotion reasoning and personalized emotion summarization across sessions.

XIV. Multilingual and Multicultural Emotion Analysis

Cross-cultural emotion recognition remains underexplored. Language and visual cues differ significantly across cultures. Multilingual models and culturally inclusive datasets are vital to avoiding bias and achieving broader generalization.

XV. Ethical and Societal Considerations

Building ethical emotion recognition systems requires attention to fairness, transparency, and informed consent. Misclassification may lead to emotional harm or discriminatory outcomes. Inclusive dataset curation and explainable AI approaches are essential for building trust.

XVI. Future Directions

- Integration of physiological signals (EEG, GSR) with vision-language models.
- Federated emotion learning to preserve privacy across devices.
- Emotion-aware conversational agents with long-term memory.
- Building multilingual, multicultural emotion datasets.
- Use of synthetic data for emotion diversity augmentation.

XVII. Conclusion

Large Vision-Language Models mark a new frontier in contextual emotion recognition. By combining multimodal inputs and powerful pretraining strategies, they enhance emotion detection accuracy and relevance. Continued research is needed to make these models more interpretable, fair, and scalable for real-world applications.

References

[1] Radford, A., et al. (2021). Learning Transferable Visual Models From Natural Language Supervision. ICML.
[2] Alayrac, J.-B., et al. (2022). Flamingo: A Visual Language Model for Few-Shot Learning. DeepMind.
[3] Liu, H., et al. (2023). LLaVA: Large Language and Vision Assistant. arXiv preprint.
[4] Zadeh, A., et al. (2018). CMU-MOSEI: Multimodal Language Analysis in the Wild. ACL.
[5] Kosti, R., et al. (2019). Context-Aware Emotion Recognition Using EMOTIC Dataset. TPAMI.
[6] Poria, S., et al. (2019). MELD: Multimodal EmotionLines Dataset. ACL.
[7] Wang, H., et al. (2020). AVE: An Audio-Visual Emotion Dataset. IEEE Access.
[8] Yin, Z., et al. (2023). GHOUL: Understanding Group Emotion in Social Images. CVPR.
[9] Li, Y., et al. (2021). Integrating Multimodal Fusion and Context-Awareness in Emotion Recognition. IJCAI.
[10] Tsai, Y. H. H., et al. (2019). Multimodal Transformer for Emotion Recognition. EMNLP.
[11] Baltrus˘aitis, T., Ahuja, C., Morency, L. P. (2019). Multimodal Machine Learning: A Survey and Taxonomy. IEEE TPAMI.
[12] Devlin, J., et al. (2018). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. NAACL.
[13] Dosovitskiy, A., et al. (2020). An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. ICLR.
[14] Yang, Z., et al. (2021). Multimodal Sentiment Analysis using Visual and Textual Fusion. IEEE Transactions.
[15] Hu, R., Singh, A. (2021). ViLT: Vision-and-Language Transformer Without Convolution or Region Supervision. ICML.
[16] Akbari, H., et al. (2021). VATT: Transformers for Multimodal Self-Supervised Learning. NeurIPS.
[17] Li, X., et al. (2022). BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation. arXiv.
[18] Chen, M., et al. (2020). UNITER: Learning UNiversal Image-TExt Representations. ECCV.
[19] Zhang, Z., et al. (2020). ResMLP: Feedforward Networks for Image Classification with Data-efficient Training. arXiv.
[20] Rashkin, H., et al. (2019). Towards Empathetic Open-domain Conversation Models: A New Benchmark and Dataset. ACL.