TẠP CHÍ KHOA HỌC
TRƯỜNG ĐẠI HỌC SƯ PHẠM TP HỒ CHÍ MINH
Tập 23, Số 2 (2026): 237-248
Website: https://journal.hcmue.edu.vn
ISSN: 2734-9918

HO CHI MINH CITY UNIVERSITY OF EDUCATION
JOURNAL OF SCIENCE
Vol. 23, No. 2 (2026): 237-248
https://doi.org/10.54607/hcmue.js.23.2.5044(2026)

Bài báo nghiên cứu

TĂNG CƯỜNG NHẬN DIỆN CẢM XÚC THÔNG QUA TÍCH HỢP ĐẶC TRƯNG NGỮ CẢNH ĐA PHƯƠNG THỨC

Nguyễn Viết Hưng, Trần Thanh Nhã*, Nguyễn Quốc Hưng, Lý Nguyễn Tiến Đạt, Nguyễn Quốc Trọng, Tạ Công Phi

Trường Đại học Sư phạm Thành phố Hồ Chí Minh, Việt Nam

**Tác giả liên hệ: Trần Thanh Nhã – Email: nhatt@hcmue.edu.vn*

Ngày nhận bài: 06-7-2025; Ngày nhận bài sửa: 08-08-2025; Ngày duyệt đăng: 03-9-2025

TÓM TẮT

Trong kỉ nguyên số, nhu cầu về các hệ thống thông minh có khả năng thấu cảm với cảm xúc người dùng ngày càng tăng cao. Tuy nhiên, các phương pháp nhận diện cảm xúc hiện có, dù là đơn phương thức hay đa phương thức, thường chưa thể tích hợp thông tin từ nhiều nguồn một cách chặt chẽ và tận dụng ngữ cảnh một cách hiệu quả. Điều này khiến các mô hình dễ bị ảnh hưởng bởi nhiễu hoặc thông tin thiếu sót từ dữ liệu đầu vào. Để khắc phục hạn chế này, nghiên cứu này giới thiệu MCFF (Multi-Modal Contextual Feature Fusion), một kiến trúc học sâu đa phương thức được thiết kế để khai thác đồng thời thông tin hình ảnh, âm thanh và văn bản. Kết quả thực nghiệm trên bộ dữ liệu IEMOCAP đạt 82,89% Accuracy và 82,86% F1-score, cho thấy MCFF có hiệu suất cạnh tranh mạnh mẽ so với các phương pháp tiên tiến khác. MCFF cho thấy tiềm năng ứng dụng rộng rãi trong các hệ thống tương tác thông minh, từ cải thiện trải nghiệm trong giáo dục trực tuyến và trợ lí ảo cho đến hỗ trợ quan trọng trong lĩnh vực chăm sóc sức khỏe tâm thần.

Từ khóa: thị giác máy tính; học sâu; nhận diện cảm xúc; đa phương thức

1. Giới thiệu

Trong những năm gần đây, công nghệ số đã định hình lại căn bản cách chúng ta tương tác với thế giới, từ học tập trực tuyến, làm việc từ xa đến giao tiếp với các trợ lí ảo. Tuy nhiên, việc thiếu hụt kết nối cảm xúc giữa người dùng và nền tảng số vẫn là một thách thức lớn. Cảm giác cô lập và thiếu sự thấu hiểu này không chỉ ảnh hưởng đến trải nghiệm mà còn trực tiếp làm giảm hiệu quả học tập, năng suất lao động và chất lượng cuộc sống (Goswami et al., 2024; Luria et al., 2019). Do đó, việc nắm bắt chính xác cảm xúc của người dùng trở nên vô cùng quan trọng. Việc nhận diện cảm xúc mang lại tiềm năng ứng dụng rộng lớn, từ việc phát triển các hệ thống tương tác trực quan, giáo dục thông minh, an toàn giao thông, chẩn đoán tâm lí (Tran et al., 2025) cho đến việc cá nhân hóa trải nghiệm người dùng như gợi ý âm nhạc Nguyen et al., 2024).

Cite this article as: Nguyen, V. H., Tran, T. N., Nguyen, Q. H., Ly, N. T. D., Nguyen, Q. T., & Ta, C. P. (2026). Enhancing emotion recognition through multimodal contextual feature integration. *Ho Chi Minh City University of Education Journal of Science, 23*(2), 237-248. https://doi.org/10.54607/hcmue.js.23.2.5044(2026)

Các phương pháp nhận diện cảm xúc truyền thống chủ yếu tiếp cận theo hướng đơn phương thức, tập trung khai thác độc lập từng loại tín hiệu như đặc trưng khuôn mặt (Ly et al., 2025), chuỗi tín hiệu âm thanh (Fu et al., 2023), hoặc thông tin ngữ nghĩa từ văn bản (Adoma et al., 2020). Tuy nhiên, các phương pháp tiếp cận đơn phương thức truyền thống thường không nắm bắt được sự phức tạp của các biểu hiện cảm xúc trong thế giới thực, vốn có bản chất đa phương thức (Cheng et al., 2024).

Để giải quyết thách thức này, nghiên cứu này đề xuất MCFF (Multi-Modal Contextual Feature Fusion) – một phương pháp nhận dạng cảm xúc đa phương thức tiên tiến. MCFF được thiết kế để không chỉ tận dụng tối đa thông tin đa dạng từ ba nguồn đầu vào chính là âm thanh, văn bản và hình ảnh, mà còn đặc biệt chú trọng vào việc tích hợp chặt chẽ các thông tin ngữ cảnh giữa chúng. Bằng cách tận dụng sức mạnh của các mô hình tiền huấn luyện (pre-trained models) mạnh mẽ cho từng phương thức riêng biệt, bao gồm Wav2vec 2.0 cho âm thanh, DistilBERT cho văn bản và Swin Transformer V2 cho hình ảnh. Các đặc trưng được trích xuất sau đó được kết hợp thông qua một kiến trúc dựa trên Transformer, cho phép mô hình học được mối quan hệ phức tạp và tương tác lẫn nhau giữa các phương thức. Nhờ đó, MCFF đã chứng minh hiệu suất phân loại mạnh mẽ, cạnh tranh với các nghiên cứu tiên tiến nhất trong lĩnh vực này.

2. Nội dung nghiên cứu

2.1. Một số công trình liên quan

Cảm xúc con người là một trạng thái tâm lí phức tạp, được biểu hiện thông qua nhiều kênh thông tin đa dạng như giọng nói, văn bản hoặc các thông tin phi ngôn ngữ (biểu cảm khuôn mặt, tư thế). Trong lĩnh vực nhận dạng cảm xúc bằng giọng nói, nhiều nghiên cứu đã khai thác các tín hiệu của âm thanh như ngữ điệu, đặc trưng phổ tần và cường độ âm thanh (Bhosale et al., 2020; Hsu et al., 2021; Naderi & Nasersharif, 2023). Nghiên cứu (Bhosale et al., 2020) đã đề xuất một mô hình kết hợp các lớp tích chập (CNN) và cơ chế tự chú ý đa đầu (multi-head self-attention) để tích hợp các đặc trưng ngôn ngữ mã hóa sâu và đặc trưng biểu diễn phổ tần âm thanh. Trong một nghiên cứu khác, (Hsu et al., 2021) đã đề xuất mô hình HuBERT, một phương pháp học biểu diễn giọng nói tự giám sát, thông qua việc sử dụng một bước phân cụm ngoại tuyến để tạo ra các nhãn mục tiêu căn chỉnh cho một tác vụ dự đoán giống BERT.

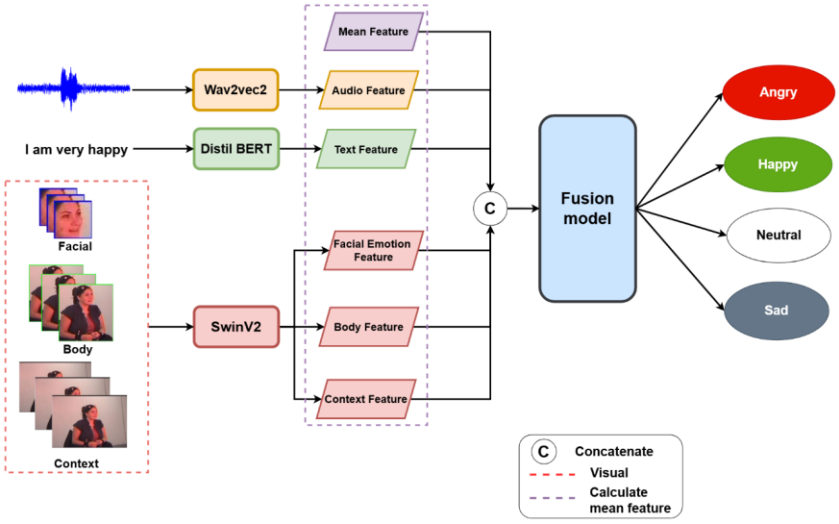
Trong bối cảnh đó, lĩnh vực nhận diện cảm xúc dựa trên văn bản đã có những bước tiến vượt bậc nhờ vào sự phát triển của mô hình Transformer (Acheampong et al., 2021; Ding et al., 2023). (Acheampong et al., 2021) đã tổng hợp và đánh giá các phương pháp nhận diện cảm xúc dựa trên BERT và chỉ ra rằng việc khai thác deep embedding từ các biến thể BERT hỗ trợ hiệu quả cho việc nắm bắt ngữ nghĩa tinh vi và cảm xúc tiềm ẩn trong văn bản. Ngoài ra, (Ding et al., 2023) đề xuất DialogueINAB để học sự tương tác giữa thái độ và hành vi lời nói của người tham gia thông qua văn bản.

Biểu cảm khuôn mặt được xem là một trong những kênh truyền tải thông tin cảm xúc trực quan và dễ nhận diện, các kĩ thuật về FER (nhận diện biểu cảm khuôn mặt) đã thu hút sự quan tâm lớn của cộng đồng nghiên cứu. (Zhang et al., 2024) đề xuất Multi-Scale Feature Fusion CNN, trong đó các bộ lọc kích cỡ khác nhau được kết hợp qua cơ chế attention để tập trung vào cả chi tiết vi mô và tổng thể. (Roy et al., 2024) phát triển ResemoteNet, một kiến trúc CNN kết hợp hàm mất mát tùy biến nhằm đồng thời tối ưu hóa độ chính xác và giảm thiểu lệch dự đoán, cho thấy hiệu quả vượt trội trên các bộ dữ liệu FER tiêu chuẩn với việc giảm đáng kể top-k loss.

Nhìn chung, các phương pháp nhận diện cảm xúc đơn phương đều cho thấy tiềm năng. Tuy nhiên, do bị hạn chế bởi sự thiếu hụt các tín hiệu từ các phương thức khác, dễ bị ảnh hưởng bởi nhiễu, yếu tố bên ngoài hoặc thông tin bị che khuất, dẫn đến hiệu suất chưa tối ưu. Để khắc phục những hạn chế trên, các phương pháp đa phương thức đã chứng minh hiệu quả vượt trội khi kết hợp nhiều nguồn dữ liệu. Điển hình, (Patamia et al., 2023) tận dụng các mô hình tiền huấn luyện dựa trên Tranformer như Wav2vec2, Bert để trích xuất các đặc trưng đa phương thức và đạt hiệu suất ấn tượng trên bộ dữ diệu IEMOCAP với độ chính xác 77,58%. Bên cạnh đó, (Nguyen et al., 2023) ứng dụng đồ thị thời gian để mô hình hóa ngữ cảnh hội thoại. Dù vậy, các mô hình này vẫn chưa khai thác trọn vẹn bối cảnh tương tác và thường dùng chiến lược trích xuất khung hình đơn giản, làm bỏ sót nhiều tín hiệu cảm xúc tinh tế.

2.2. Phương pháp đề xuất

Giả sử $s_i = (a_i, t_i, v_i)$ biểu thị cho mẫu dữ liệu thứ i trong bộ dữ liệu, trong đó a_i biểu thị cho đặc trưng âm thanh, t_i biểu thị cho đặc trưng văn bản và v_i biểu thị cho đặc trưng hình ảnh. Các phần tiếp theo sẽ đi sâu vào chi tiết từng nhánh xử lí cho mỗi phương thức (a_i, t_i, v_i) và cách thức MCFF kết hợp chúng để đạt được hiệu suất tối ưu. Tổng quan kiến trúc của phương pháp này được minh họa chi tiết ở Hình 1.



Hình 1. *Kiến trúc của mô hình được đề xuất*

- *Nhánh âm thanh*

Mô hình wav2vec 2.0 được sử dụng để trích xuất đặc trưng trực tiếp từ dữ liệu sóng âm thô, giúp giảm phụ thuộc vào các bước tiền xử lí thủ công và vẫn nắm bắt được cả chi tiết âm vị lẫn ngữ cảnh toàn cục. Nhờ cơ chế học tự giám sát trên dữ liệu lớn, mô hình này đặc biệt hiệu quả trong nhận diện cảm xúc từ giọng nói, nơi dữ liệu gán nhãn hạn chế và yêu cầu phát hiện các biến đổi tinh tế về ngữ điệu và phổ tần (Islam et al., 2024). Đầu vào âm thanh a_i được xử lí qua các tầng convolutional và transformer của wav2vec 2.0. Từ các vector đặc trưng đầu ra của wav2vec 2.0, một lớp mean pooling được áp dụng để thu được vector $f_{a_{i1}}$ đại diện duy nhất cho toàn bộ đặc trưng âm thanh của mẫu. Tiếp theo, vector $f_{a_{i1}}$ được đưa qua 1 lớp Fully Connected (FC) và chuẩn hóa bởi một lớp Batch Normalization (BN). Quá trình này giúp ổn định phân phối dữ liệu và tạo ra vector đặc trưng $f_{a_{i2}}$.

$$f_{a_{i2}} = BN\left(FC\left(f_{a_{i1}}\right)\right) \quad (1)$$

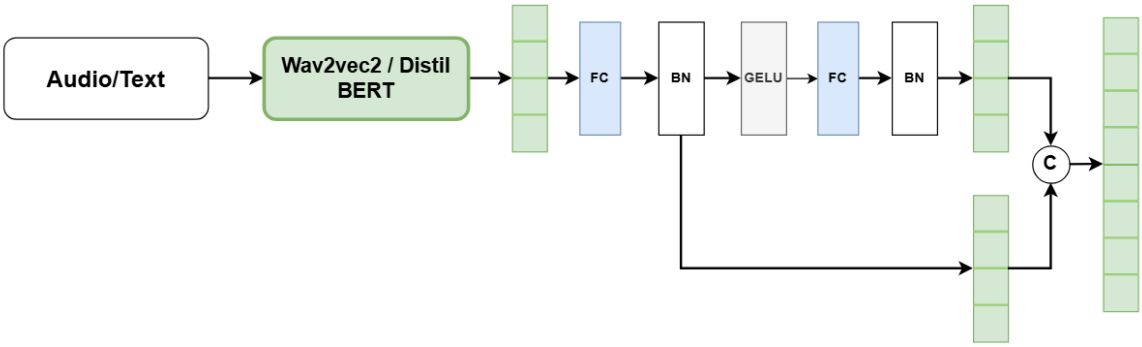
Đầu ra $f_{a_{i2}}$ tiếp tục được truyền qua hàm kích hoạt GELU để tạo ra các biểu diễn phi tuyến. Sau đó, kết quả này một lần nữa được đưa qua một lớp FC và BN, tạo thành đặc trưng $f_{a_{i3}}$:

$$f_{a_{i3}} = BN\left(FC\left(GELU\left(f_{a_{i2}}\right)\right)\right) \quad (2)$$

Cuối cùng, hai đặc trưng $f_{a_{i2}}$ và $f_{a_{i3}}$ được ghép nối với nhau để tạo thành biểu diễn bậc cao của đặc trưng âm thanh cho mẫu dữ liệu.

$$f_{a_i} = f_{a_{i2}} || f_{a_{i3}} \quad (3)$$

Đặc trưng f_{a_i} này sau đó được sử dụng làm đầu vào khi kết hợp với các phương thức khác.



Hình 2. *Kiến trúc trích xuất thông tin âm thanh/văn bản*

- *Nhánh văn bản*

Các dữ liệu văn bản f_{t_i} sẽ được đưa qua mô hình DistilBERT (Sanh et al., 2020) được huấn luyện trước để trích xuất đặc trưng. Đây là phiên bản rút gọn của BERT thông qua kĩ thuật *knowledge distillation*, giúp giảm số lượng tham số và chi phí tính toán nhưng vẫn duy trì hiệu năng tương đương, đặc biệt phù hợp cho các ứng dụng có hạn chế về tài nguyên (Ta et al., 2025). Chúng tôi trích xuất token [CLS] của DistilBERT để tạo thành đặc trưng $f_{t_{i1}}$ đại diện cho toàn bộ câu. Tương tự như quá trình xử lí đặc trưng âm thanh, vector $f_{t_{i1}}$ này

được đưa qua một lớp FC và chuẩn hóa bởi một lớp BN. Bước này nhằm ổn định phân phối dữ liệu và thu được vector đặc trưng $f_{t_{i2}}$.

$$f_{t_{i2}} = BN\left(FC\left(f_{t_{i1}}\right)\right) \quad (4)$$

Vector $f_{t_{i2}}$ được truyền qua hàm kích hoạt GELU để tạo ra các biểu diễn phi tuyến, trước khi qua một lớp FC khác và được chuẩn hóa lần nữa bởi BN, tạo thành vector đặc trưng $f_{t_{i3}}$.

$$f_{t_{i3}} = BN\left(FC\left(GELU\left(f_{t_{i2}}\right)\right)\right) \quad (5)$$

Cuối cùng, hai đặc trưng $f_{t_{i2}}$ và $f_{t_{i3}}$ được ghép nối với nhau để tạo thành biểu diễn bậc cao f_{t_i} của đặc trưng văn bản, sẵn sàng cho việc kết hợp với các phương thức khác:

$$f_{t_i} = f_{t_{i2}} || f_{t_{i3}} \quad (6)$$

Hình 2 minh họa tổng quát kiến trúc của mô hình trích xuất và xử lí đặc trưng âm thanh hoặc văn bản.

- *Nhánh video*

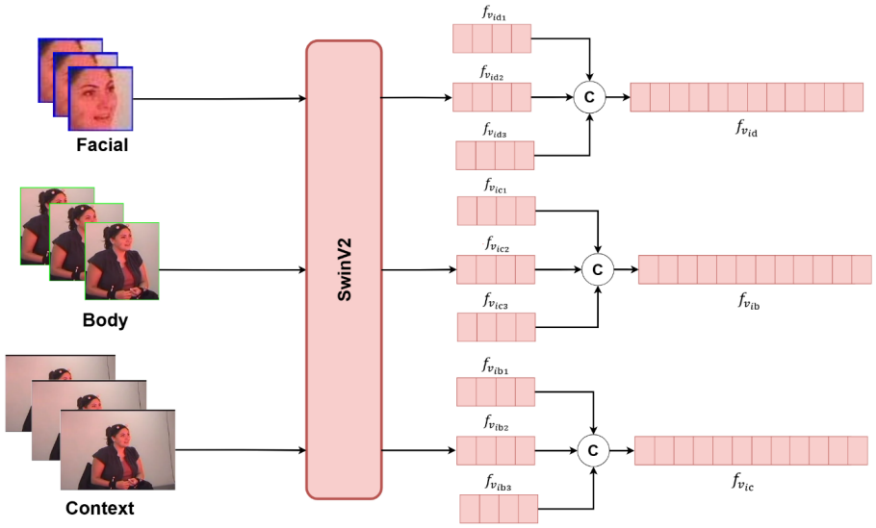
Thông tin về hình ảnh v_i bao gồm các thông tin chi tiết v_{ij}, trong đó $j \in \{b, c, d\}$. Cụ thể, b biểu thị thông tin về cơ thể (Body), c biểu thị thông tin về ngữ cảnh (Context) và d biểu thị thông tin về khuôn mặt (Facial).

Đầu tiên, mỗi loại thông tin v_{ij} được đưa vào Swin Transformer V2 (Liu et al., 2022) được huấn luyện trước để trích xuất các đặc trưng như cảm xúc khuôn mặt, tư thế cơ thể và ngữ cảnh của cảnh quay. Nhờ cơ chế chia ảnh thành các cửa sổ (window) và dịch chuyển cửa sổ (shifted windows), mô hình có thể xử lí ảnh độ phân giải cao với số lượng tham số và chi phí tính toán thấp hơn, đồng thời khai thác hiệu quả cả thông tin cục bộ và toàn cục, giúp nâng cao độ chính xác trong nhận diện cảm xúc từ hình ảnh (Islam et al., 2024). Để có một biểu diễn tổng quát cho, vector đặc trưng trung bình (Mean seq frame) $f_{v_{ij1}}$ được tính toán bằng cách lấy trung bình tất cả các vector đặc trưng của các frame trong chuỗi. Tuy nhiên, do các frame ở đầu hoặc cuối video có thể không phản ánh chính xác nhãn dữ liệu hoặc chỉ một phần của video chứa thông tin quan trọng. Điều này có thể làm suy giảm tính đại diện của vector đặc trưng trung bình. Để khắc phục vấn đề này và đảm bảo mô hình nắm bắt được các thông tin quan trọng nhất, chúng tôi không chỉ sử dụng đặc trưng từ Mean seq frame $f_{v_{ij1}}$, mà còn bổ sung thêm đặc trưng từ frame đầu tiên $f_{v_{ij2}}$ và frame cuối cùng $f_{v_{ij3}}$ của mỗi thông tin đầu vào.

Các đặc trưng $f_{v_{ij1}}$, $f_{v_{ij2}}$và $f_{v_{ij3}}$ sau đó được ghép nối lại để tạo thành đặc trưng $f_{v_{ij}}$ duy nhất cho từng loại thông tin (cảm xúc khuôn mặt, tư thế cơ thể và ngữ cảnh riêng biệt):

$$f_{v_{ij}} = f_{v_{ij1}} \parallel f_{v_{ij2}} \parallel f_{v_{ij3}} \quad (7)$$

Phương pháp này giúp giảm thiểu tác động của các frame không phù hợp và tăng cường khả năng nhận diện các yếu tố then chốt, từ đó cải thiện chất lượng của biểu diễn các đặc trưng trong video. Hình 3 minh họa kiến trúc mô hình trích xuất các đặc trưng này từ video.



Hình 3. *Kiến trúc của nhánh video*

- *Phương pháp kết hợp*

Sau khi đã trích xuất và xử lí riêng biệt các đặc trưng từ từng phương thức, chúng tôi tiến hành tổng hợp các biểu diễn này. Đầu tiên, để có được một biểu diễn ban đầu đại diện cho toàn bộ các phương thức, chúng tôi thực hiện phép cộng trung bình (average pooling) trên các vector đặc trưng đã được tổng hợp, tạo nên vector đặc trưng f_{mean_i}.

$$f_{mean_i} = \frac{1}{13}\left(\sum_{m\in\{2,3\}} f_{a_{im}} + \sum_{n\in\{2,3\}} f_{t_{in}} + \sum_{p\in\{1,2,3\}} f_{v_{ijp}}\right) \quad (8)$$

Trong đó $f_{a_{im}}$ là tập hợp các đặc trưng âm thanh, $f_{t_{in}}$ là tập hợp các đặc trưng văn bản và $f_{v_{ijp}}$ là tập hợp các đặc trưng hình ảnh của mẫu dữ liệu i.

Tiếp theo, để tạo ra một biểu diễn thống nhất và bậc cao từ tất cả các phương thức, chúng tôi sử dụng phép ghép nối như được minh họa ở Hình 4. Các đặc trưng đã được xử lí từ âm thanh (f_{a_i}), văn bản (f_{t_i}), và các đặc trưng hình ảnh ($f_{v_{ij}}$) tương ứng với tư thế, ngữ cảnh, khuôn mặt từ hình ảnh) được ghép nối lại với nhau để tạo thành một đặc trưng $f_{combined}$ duy nhất, đại diện cho thông tin đa phương thức của mẫu dữ liệu:

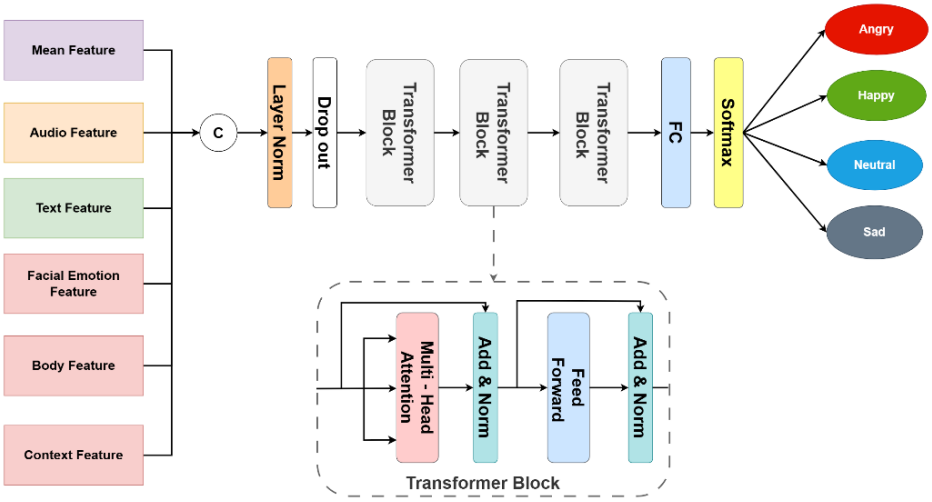
$$f_{combined} = f_{mean_i} \parallel f_{a_i} \parallel f_{t_i} \parallel f_{v_{ij}}$$

Đặc trưng tổng hợp được đưa vào khối Transformer gồm một lớp Layer Normalization (Dropout = 0.1) và ba lớp Transformer liên tiếp, giúp học các mối quan hệ phức tạp và tương tác ngữ cảnh giữa các phương thức. Với khả năng xử lí và kết hợp đồng thời nhiều loại dữ liệu như thị giác, thính giác và ngôn ngữ, Transformer đặc biệt hiệu quả trong các tác vụ đa phương thức như phân loại, phân đoạn, mô tả ảnh hay suy luận ngữ nghĩa (Islam et al., 2024).

$$f' = Trans(f_{combined}) \quad (10)$$

Đầu ra f' thu được từ Transformer block tiếp tục được đưa qua một lớp FC kết hợp với hàm kích hoạt Softmax. Từ đó thu được $\hat{y}$ chứa phân bố xác suất dùng để phân loại cảm xúc.

$$\hat{y} = Softmax\big(FC(f')\big) \quad (11)$$



Hình 4. *Kiến trúc phương pháp kết hợp*

3. Thực nghiệm và so sánh

3.1. Bộ dữ liệu thực nghiệm

Nghiên cứu này sử dụng bộ dữ liệu IEMOCAP (Interactive Emotional Dyadic Motion Capture) (Busso et al., 2008) để huấn luyện và đánh giá hiệu suất mô hình. IEMOCAP là một bộ dữ liệu đa phương thức bao gồm âm thanh, văn bản và video. Để đảm bảo tính so sánh với các công trình trước, nghiên cứu tập trung vào bốn cảm xúc chính: tức giận (1103 mẫu), buồn bã (1084 mẫu), trung tính (1708 mẫu) và vui vẻ (1363 mẫu). Trong quá trình xử lí dữ liệu, các phương thức được xử lí riêng biệt. Đối với dữ liệu văn bản, các bản ghi lời cuộc đối thoại được đệm (padding) để đảm bảo tất cả các câu trong bộ dữ liệu có cùng một chiều dài. Với video, mô hình YOLOv5s được sử dụng để phát hiện các đối tượng trong mỗi frame. Mỗi frame sẽ bao gồm ba loại ảnh: ảnh khuôn mặt, ảnh người nói chính và ảnh tổng thể ngữ cảnh.

3.2. Môi trường thực nghiệm và phương pháp đánh giá

Nghiên cứu này được thực nghiệm bằng ngôn ngữ lập trình Python phiên bản 3.7, sử dụng các thư viện chính như PyTorch và Hugging Face Transformers. Các thí nghiệm được tiến hành trên môi trường Kaggle với cấu hình phần cứng gồm CPU Intel(R) Xeon(R) CPU @ 2.20GHz, bộ nhớ 29GB RAM và GPU Tesla P100 16GB. Hiệu suất của mô hình được đánh giá thông qua độ đo Accuracy và F1-score.

3.3. Kết quả thực nghiệm và thảo luận

Bảng 1 tổng hợp kết quả kiểm thử của mô hình MCFF trên tập dữ liệu IEMOCAP, đồng thời trình bày kết quả của nhiều nghiên cứu trước đây với các đặc trưng và tổ hợp phương thức khác nhau. Các số liệu của các phương pháp trước được trích dẫn từ công bố gốc, đảm bảo tính nhất quán khi so sánh, trong khi các kết quả của MCFF được thu được từ quá trình thực nghiệm của chúng tôi. MCFF được đánh giá với các tổ hợp phương thức khác

nhau, gồm Audio + Text, Visual + Text, Audio + Visual và Audio + Text + Visual. Kết quả thực nghiệm cho thấy MCFF khi tích hợp cả ba phương thức (Text + Audio + Visual) đạt Accuracy 82,89% và F1-Score 82,86%, vượt xa các phương pháp chỉ sử dụng một hoặc hai phương thức. Một điểm đáng chú ý là sự đóng góp của dữ liệu hình ảnh (Visual) trong việc nâng cao hiệu suất của MCFF. Các tổ hợp có Visual, như Visual + Text (Accuracy 80,42%) và Audio + Visual (Accuracy 80,95%), đạt hiệu suất cao hơn đáng kể so với Audio + Text (Accuracy 70,19%). Điều này nhấn mạnh rằng dữ liệu hình ảnh, với các đặc trưng như biểu cảm khuôn mặt và ngôn ngữ cơ thể, cung cấp thông tin bổ sung quan trọng mà các phương pháp đơn phương thức hoặc tổ hợp Audio + Text không thể nắm bắt. Khi tích hợp cả ba phương thức, MCFF đạt hiệu suất tối ưu (Accuracy 82,89%), cho thấy rằng sự kết hợp đầy đủ các nguồn dữ liệu, cùng với cơ chế tích hợp ngữ cảnh, là yếu tố then chốt để đạt được hiệu quả nhận diện cảm xúc cao nhất.

Bảng 1. *So sánh kết quả của MCFF với các phương pháp khác trên bộ IEMOCAP*

Nghiên cứu	Đặc trưng	Accuracy	F1-Score
(Bhosale et al., 2020)	Audio	68,11%	-
(Khan et al., 2024)	Audio	72,75%	-
(Naderi & Nasersharif, 2023)	Audio	74,16%	-
(Ding et al., 2023)	Text	67,32%	67,22%
(Jia et al., 2022)	Text	69,60%	-
(Li et al., 2022)	Text + Visual	71,30%	-
(Sun et al., 2024)	Text + Audio	78,34%	-
(Khan et al., 2025)	Text + Audio	81,85%	-
(Patamia et al., 2023)	Text + Audio + Mocap	77,58%	-
(Joshi et al., 2022)	Text + Audio + Visual	-	84,50%
(Shayaninasab & Babaali, 2024)	Text + Audio + Visual	-	75,35%
(C.-V. T. Nguyen et al., 2023)	Text + Audio + Visual	84,73%	84,64%
Ours	Text + Audio	70,19%	70,21%
Ours	Text + Visual	80,42%	80,38%
Ours	Audio + Visual	80,95%	80,91%
Our (MCFF)	Text + Audio + Visual	82,89%	82,86%

So với các nghiên cứu khác, MCFF cho thấy sự vượt trội so với các phương pháp đơn phương thức khác, chẳng hạn như phương pháp dựa trên âm thanh như (Bhosale et al., 2020) chỉ đạt Accuracy 68,11%, (Khan et al., 2024) đạt 72,75% và (Naderi & Nasersharif, 2023) đạt 74,16%. Tương tự, các phương pháp dựa trên văn bản như (Ding et al., 2023) và (Jia et al., 2022) ghi nhận Accuracy lần lượt là 67,32% và 69,60%. Các phương pháp đơn phương thức phản ánh hạn chế của việc chỉ dựa vào một nguồn dữ liệu để nhận diện cảm xúc, thường không thể nắm bắt đầy đủ các sắc thái cảm xúc phức tạp, vốn được biểu đạt qua nhiều khía cạnh như ngữ điệu, từ ngữ hay biểu cảm khuôn mặt. Cách tiếp cận dễ bị ảnh hưởng bởi nhiễu và thiếu sót trong dữ liệu đầu vào. Các mô hình dựa trên âm thanh như (Naderi & Nasersharif, 2023) có thể gặp khó khăn trong việc phân biệt cảm xúc khi chất lượng âm thanh kém hoặc ngữ điệu không rõ ràng. Tương tự, các mô hình dựa trên văn bản như (Ding et al., 2023) bỏ qua các yếu tố phi ngôn ngữ, chẳng hạn như giọng điệu hoặc biểu cảm, dẫn đến việc hiểu

sai ý định cảm xúc. Hạn chế này đặc biệt rõ rệt trong các kịch bản thực tế, nơi cảm xúc thường được biểu đạt đồng thời qua nhiều kênh. Ngay cả tổ hợp Audio + Text của MCFF, dù đạt Accuracy thấp nhất trong các cấu hình của mô hình (70,19%), vẫn vượt qua các phương pháp đơn phương thức tốt nhất (Naderi & Nasersharif, 2023: 74,16%), chứng minh sư hiệu quả của việc kết hợp nhiều nguồn dữ liệu, dù chỉ là hai phương thức.

Hiệu suất vượt trội của MCFF không chỉ đến từ việc khai thác đồng thời nhiều nguồn dữ liệu (Audio, Text, Visual, Facial Emotion, Body, Context) mà còn từ cơ chế tích hợp đặc trưng tiên tiến, trong đó các Transformer Blocks với Multi-Head Attention cho phép mô hình khai thác sâu các mối quan hệ ngữ cảnh giữa các phương thức. Trái lại, nghiên cứu của (Patamia et al., 2023) chỉ kết hợp dữ liệu Text + Audio + MoCap bằng cách ghép nối đặc trưng (concatenation), làm hạn chế khả năng học biểu diễn đa phương thức. Tương tự, (Shayaninasab & Babaali, 2024) sử dụng SVM để học đặc trưng đa phương thức, vốn thiếu khả năng mô hình hóa các tương tác phức tạp và ngữ cảnh liên phương thức như MCFF.

Sự vượt trội của MCFF so với các phương pháp đơn phương thức nhấn mạnh tầm quan trọng của tích hợp đa phương thức trong việc xây dựng các hệ thống thông minh. Trong các ứng dụng thực tiễn như giáo dục trực tuyến, trợ lí ảo hoặc chăm sóc sức khỏe tâm thần, cảm xúc con người thường được biểu đạt qua nhiều kênh, đòi hỏi các mô hình phải có khả năng xử lí đồng thời nhiều nguồn dữ liệu. Dù vậy, MCFF vẫn còn tồn tại một số hạn chế, chẳng hạn như việc sử dụng các mô hình tiền huấn luyện với kích thước lớn để trích xuất các đặc trưng của từng phương thức làm gia tăng độ phức tạp và chi phí tính toán. Bên cạnh đó, kiến trúc nhiều tầng với ba mạng tiền huấn luyện lớn và các lớp Transformer liên tiếp khiến mô hình gặp hạn chế về khả năng diễn giải và giải thích, tạo ra một hộp đen khó theo dõi.

4. Kết luận

Nghiên cứu này đề xuất MCFF, một kiến trúc học sâu đa phương thức hiệu quả cho nhận dạng cảm xúc bằng cách tích hợp đồng thời ba luồng dữ liệu hình ảnh, âm thanh và văn bản. Thông qua việc khai thác triệt để các đặc trưng video đa chiều, tối ưu hóa trích xuất đặc trưng âm thanh bằng Wav2Vec2.0, cùng với việc áp dụng chiến lược xử lí và trích xuất đặc trưng toàn diện cho từng phương thức, MCFF đã đạt 82,89% Accuracy và 82,86% F1-score trên bộ dữ liệu IEMOCAP. Kết quả này không chỉ vượt trội so với tất cả các mô hình đơn và hai phương thức mà còn đạt hiệu suất tiệm cận với các phương pháp ba phương thức tiên tiến nhất, qua đó khẳng định hiệu quả của việc khai thác sâu ngữ cảnh đa phương thức và chiến lược cân bằng thông tin từ các chiều dữ liệu khác nhau. Tuy nhiên, mô hình hiện tại vẫn còn hạn chế do chi phí tính toán cao khi sử dụng các mô hình tiền huấn luyện cỡ lớn, cũng như khả năng giải thích kém do kiến trúc đa tầng phức tạp. Trong tương lai, chúng tôi định hướng tối ưu hóa mô hình cho các thiết bị biên, mở rộng huấn luyện trong đa dạng ngữ cảnh và nghiên cứu cơ chế tự điều chỉnh trọng số các phương thức theo thời gian nhằm tiếp tục nâng cao độ chính xác và tính thích nghi của hệ thống trong môi trường thực tế.

❖ ***Tuyên bố về quyền lợi:*** *Các tác giả xác nhận hoàn toàn không có xung đột về quyền lợi.*

❖ ***Lời cảm ơn:*** *Nghiên cứu này được tài trợ bởi Nguồn ngân sách khoa học và công nghệ Trường Đại học Sư phạm Thành phố Hồ Chí Minh trong đề tài mã số CS.2024.19.38*

TÀI LIỆU THAM KHẢO

Acheampong, F. A., Nunoo-Mensah, H., & Chen, W. (2021). Transformer models for text-based emotion detection: A review of BERT-based approaches. *Artificial Intelligence Review*, *54*(8), 5789–5829. https://doi.org/10.1007/s10462-021-09958-2

Adoma, A. F., Henry, N.-M., & Chen, W. (2020). Comparative analyses of bert, roberta, distilbert, and xlnet for text-based emotion recognition. *2020 17th International Computer Conference on Wavelet Active Media Technology and Information Processing (ICCWAMTIP)*, 117–121. https://ieeexplore.ieee.org/abstract/document/9317379/

Bhosale, S., Chakraborty, R., & Kopparapu, S. K. (2020). Deep encoded linguistic and acoustic cues for attention based end to end speech emotion recognition. *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 7189–7193. https://ieeexplore.ieee.org/abstract/document/9054621/

Busso, C., Bulut, M., Lee, C.-C., Kazemzadeh, A., Mower, E., Kim, S., Chang, J. N., Lee, S., & Narayanan, S. 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

Cheng, Z., Cheng, Z.-Q., He, J.-Y., Wang, K., Lin, Y., Lian, Z., Peng, X., & Hauptmann, A. (2024). Emotion-llama: Multimodal emotion recognition and reasoning with instruction tuning. *Advances in Neural Information Processing Systems*, *37*, 110805–110853.

Ding, J., Chen, X., Lu, P., Yang, Z., Li, X., & Du, Y. (2023). DialogueINAB: An interaction neural network based on attitudes and behaviors of interlocutors for dialogue emotion recognition. *The Journal of Supercomputing*, *79*(18), 20481–20514. https://doi.org/10.1007/s11227-023-05439-1

Fu, H., Zhuang, Z., Wang, Y., Huang, C., & Duan, W. (2023). Cross-corpus speech emotion recognition based on multi-task learning and subdomain adaptation. *Entropy*, *25*(1), 124.

Goswami, S. A., Dave, S., & Patel, K. C. (2024). The Need for Emotional Intelligence in Human-Computer Interactions. In *Harnessing Artificial Emotional Intelligence for Improved Human-Computer Interactions* (pp. 82–106). IGI Global. https://www.igi-global.com/chapter/the-need-for-emotional-intelligence-in-human-computer-interactions/349198

Hsu, W.-N., Bolte, B., Tsai, Y.-H. H., Lakhotia, K., Salakhutdinov, R., & Mohamed, A. (2021). Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *29*, 3451–3460.

Islam, S., Elmekki, H., Elsebai, A., Bentahar, J., Drawel, N., Rjoub, G., & Pedrycz, W. (2024). A comprehensive survey on applications of transformers for deep learning tasks. *Expert Systems with Applications*, *241*, 122666.

Jia, N., Zheng, C., & Sun, W. (2022). A multimodal emotion recognition model integrating speech, video and MoCAP. *Multimedia Tools and Applications*, *81*(22), 32265–32286. https://doi.org/10.1007/s11042-022-13091-9

Joshi, A., Bhat, A., Jain, A., Singh, A. V., & Modi, A. (2022). *COGMEN: COntextualized GNN based Multimodal Emotion recognitioN* (No. arXiv:2205.02455). arXiv. https://doi.org/10.48550/arXiv.2205.02455

Khan, M., Gueaieb, W., El Saddik, A., & Kwon, S. (2024). MSER: Multimodal speech emotion recognition using cross-attention with deep fusion. *Expert Systems with Applications*, *245*, 122946.

Khan, M., Tran, P.-N., Pham, N. T., El Saddik, A., & Othmani, A. (2025). MemoCMT: Multimodal emotion recognition using cross-modal transformer-based feature fusion. *Scientific Reports*, *15*(1), 5473.

Li, Z., Tang, F., Zhao, M., & Zhu, Y. (2022). *EmoCaps: Emotion Capsule based Model for Conversational Emotion Recognition* (No. arXiv:2203.13504). arXiv. https://doi.org/10.48550/arXiv.2203.13504

Liu, Z., Hu, H., Lin, Y., Yao, Z., Xie, Z., Wei, Y., Ning, J., Cao, Y., Zhang, Z., & Dong, L. (2022). Swin transformer v2: Scaling up capacity and resolution. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 12009–12019. http://openaccess.thecvf.com/content/CVPR2022/html/Liu_Swin_Transformer_V2_Scaling_Up_Capacity_and_Resolution_CVPR_2022_paper.html

Luria, M., Zoran, A., & Forlizzi, J. (2019). *Challenges of Designing HCI for Negative Emotions* (No. arXiv:1908.07577). arXiv. https://doi.org/10.48550/arXiv.1908.07577

Ly, D., Tran, N., Nguyen, H. Q., Nguyen, T., Nguyen, L., & Nguyen, H. (2025). A Graph Attention Network-Enhanced Approach to Facial Expression Recognition Using Hybrid Pixel-Geometry Features. *International Journal of Intelligent Engineering & Systems*, *18*(5).

Naderi, N., & Nasersharif, B. (2023). Cross corpus speech emotion recognition using transfer learning and attention-based fusion of wav2vec2 and prosody features. *Knowledge-Based Systems*, *277*, 110814.

Nguyen, C.-V. T., Mai, A.-T., Le, T.-S., Kieu, H.-D., & Le, D.-T. (2023). Conversation Understanding using Relational Temporal Graph Neural Networks with Auxiliary Cross-Modality Interaction. *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 15154–15167. https://doi.org/10.18653/v1/2023.emnlp-main.937

Nguyen, H., Tran, N., Ly, D., Tran, A., Nguyen, A., & Vo, H. (2024). A Model for Song Recommendation Based on Facial Emotion Analysis and Musical Emotion. *International Journal of Intelligent Engineering & Systems*, *17*(4). https://inass.org/wp-content/uploads/2024/03/2024083177-2.pdf

Patamia, R. A., Santos, P. E., Acheampong, K. N., Ekong, F., Sarpong, K., & Kun, S. (2023). Multimodal speech emotion recognition using modality-specific self-supervised frameworks. *2023 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, 4134–4141. https://ieeexplore.ieee.org/abstract/document/10394418/

Roy, A. K., Kathania, H. K., Sharma, A., Dey, A., & Ansari, M. S. A. (2024). ResEmoteNet: Bridging accuracy and loss reduction in facial emotion recognition. *IEEE Signal Processing Letters*. https://ieeexplore.ieee.org/abstract/document/10812829/

Sanh, V., Debut, L., Chaumond, J., & Wolf, T. (2020). *DistilBERT, a distilled version of BERT: Smaller, faster, cheaper and lighter* (No. arXiv:1910.01108). arXiv. https://doi.org/10.48550/arXiv.1910.01108

Shayaninasab, M., & Babaali, B. (2024). *Multi-Modal Emotion Recognition by Text, Speech and Video Using Pretrained Transformers* (No. arXiv:2402.07327). arXiv. https://doi.org/10.48550/arXiv.2402.07327

Ta, P., Tran, N., Nguyen, H., & Nguyen, H. D. (2025). Detecting signs of depression on social media: A machine learning analysis and evaluation. *Sustainable Futures*, 100827.

Tran, N., Ta, P., Nguyen, H., Nguyen, H. D., & Le, A.-C. (2025). Hybrid contextual and sentiment-based machine learning model for identifying depression risk in social media. *Expert Systems with Applications*, *291*, 128505.

Zhang, X., Fu, X., Qi, G., & Zhang, N. (2024). A multi-scale feature fusion convolutional neural network for facial expression recognition. *Expert Systems*, *41*(4), e13517. https://doi.org/10.1111/exsy.13517

ENHANCING EMOTION RECOGNITION THROUGH MULTIMODAL CONTEXTUAL FEATURE INTEGRATION

Nguyen Viet Hung, Tran Thanh Nha*, Nguyen Quoc Hung,
Ly Nguyen Tien Dat, Nguyen Quoc Trong, Ta Cong Phi
Ho Chi Minh City University of Education, Vietnam
**Corresponding Author: Tran Thanh Nha – Email: nhatt@hcmue.edu.vn*
Received: July 06, 2025; Revised: August 08, 2025; Accepted: September 03, 2025

ABSTRACT

In the digital age, the demand for intelligent systems capable of understanding users' emotions is continuously increasing. However, existing emotion recognition methods, whether unimodal or multimodal, often struggle to integrate information from multiple sources cohesively and leverage contextual cues effectively. This limitation makes models susceptible to noise or incomplete information from input data. To address this limitation, this research introduces MCFF (Multi-Modal Contextual Feature Fusion), a multimodal deep learning architecture designed to simultaneously leverage visual, audio, and textual information. Experimental results on the IEMOCAP dataset yielded an accuracy of 82.89% and an F1-score of 82.86%, demonstrating MCFF's competitive strong performance compared with other state-of-the-art methods. MCFF exhibits broad potential for application in intelligent interactive systems, ranging from enhancing experiences in online education and virtual assistants to providing crucial support in mental healthcare.

Keywords: computer vision; deep learning; emotion recognition; multimodal