HỆ THỐNG NHẬN DIỆN CẢM XÚC SỬ DỤNG CÁC THIẾT BỊ IOT HIỆU NĂNG YẾU TRONG MÔI TRƯỜNG LỚP HỌC

Vương Thị Uyên, **Nguyễn Nam Hải**, **Đào Trọng Hiếu**, **Nguyễn Việt Hưng**, **Trần Tiến Công**
Học Viện Công Nghệ Bưu Chính Viễn Thông

Tóm tắt: **Bài báo này phân tích vai trò của công nghệ nhận diện cảm xúc trong việc nâng cao hiệu quả học tập của học sinh thông qua giám sát trạng thái cảm xúc theo thời gian thực. Chúng tôi đề xuất một hệ thống nhận diện cảm xúc đa khuôn mặt tự động, có khả năng phân tích đồng thời cảm xúc của nhiều học sinh, được đánh giá trên bộ dữ liệu Emotion PTIT – bao gồm 1.500 hình ảnh và 71 video thu thập từ môi trường lớp học và được gán nhãn thủ công nhằm hỗ trợ các nhiệm vụ nhận diện khuôn mặt và đánh giá tương tác trong lớp học. Để tối ưu hóa sự năng động trong lớp học, chúng tôi phát triển một công thức mới nhằm đánh giá mức độ tương tác lớp học dựa trên dữ liệu cảm xúc của học sinh. Hệ thống được thiết kế cho các thiết bị IoT tiết kiệm năng lượng, giải quyết các thách thức về cân bằng tải và giảm độ trễ khi xử lý luồng video trực tiếp, nhằm đánh giá hiệu quả tương tác trong lớp học. Hệ thống của chúng tôi có khả năng phát hiện đồng thời lên tới 50 khuôn mặt với tốc độ 25 khung hình/giây, đạt độ chính xác 88,68%.**

Từ khóa: **Hệ thống nhận diện cảm xúc, thiết bị IoT, chất lượng lớp học, hệ thống thời gian thực**

I. GIỚI THIỆU

Sự phát triển của công nghệ nhận diện cảm xúc trong lĩnh vực giáo dục đã mở ra những hiểu biết sâu sắc về trạng thái cảm xúc của con người, tạo nên tiềm năng lớn trong việc nâng cao trải nghiệm học tập. Cảm xúc đóng vai trò thiết yếu trong quá trình học tập, ảnh hưởng trực tiếp đến khả năng ghi nhớ kiến thức và trạng thái tâm lý của học sinh [1]. Tuy nhiên, trong bối cảnh giáo dục hiện đại, việc nhận diện và phản hồi chính xác các trạng thái cảm xúc của học sinh đang trở nên ngày càng phức tạp [2].

Nhận diện cảm xúc là quá trình giải mã cảm xúc thông qua các biểu hiện khuôn mặt, ngôn ngữ cơ thể, mẫu giọng nói, giao tiếp văn bản, và các chỉ số sinh lý [3]. Lĩnh vực này đang phát triển mạnh mẽ, với các ứng dụng quan trọng trong chăm sóc sức khỏe, giải trí, an ninh và giáo dục [4]. Đặc biệt trong giáo dục, công nghệ nhận diện cảm xúc tạo ra những ảnh hưởng đáng kể lên quá trình học tập của học sinh, giúp hình thành các quy trình giáo dục tùy chỉnh và xây dựng môi trường học tập lấy cảm xúc làm trung tâm [5].

Các nghiên cứu gần đây đã đưa ra các phương pháp tiếp cận đa chiều, kết hợp dữ liệu khuôn mặt, cơ thể, giọng nói và văn bản nhằm nâng cao độ chính xác của nhận diện cảm xúc trong bối cảnh giáo dục [6]. Mặc dù đã có những tiến bộ, việc phân tích toàn diện về sự tham gia của học sinh trong lớp học thông qua phản hồi cảm xúc vẫn chưa được nghiên cứu đầy đủ. Để lấp đầy khoảng trống này, chúng tôi đề xuất một công thức mới để phân loại dữ liệu cảm xúc, nhằm đánh giá chất lượng lớp học dựa trên các cảm xúc của học sinh như trạng thái trung lập và ngạc nhiên, biểu thị mức độ tập trung, nhiệt tình và tích cực. Chúng tôi cũng phát triển các kịch bản lớp học để đánh giá tác động của cảm xúc đối với mức độ tham gia và hứng thú của học sinh.

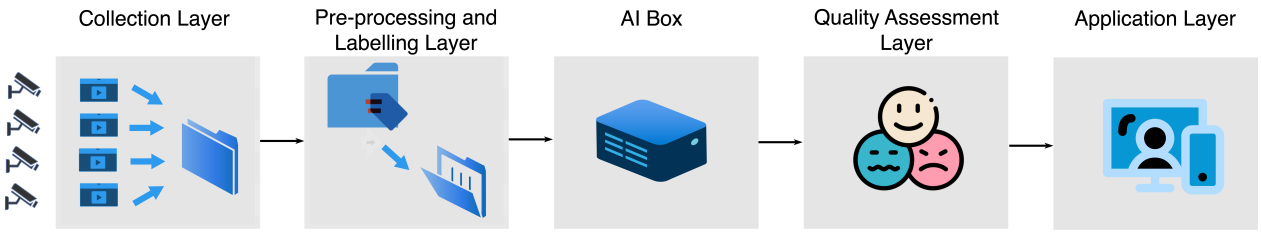
Hệ thống của chúng tôi khởi đầu bằng việc ghi lại các biểu hiện khuôn mặt thông qua video, sau đó thực hiện tiền xử lý dữ liệu, gắn nhãn, phân loại cảm xúc dựa trên công nghệ học sâu tích hợp trên các thiết bị IoT và tiến hành đánh giá chất lượng lớp học. Kết quả được hiển thị thông qua một ứng dụng web, cung cấp những thông tin hữu ích và có giá trị cho giáo viên (Hình 1).

Các đóng góp mới của chúng tôi được tóm tắt như sau:

- Thu thập và công bố một tập dữ liệu mới, mang tên "Emotion PTIT,"đây là tập dữ liệu đầu tiên cho phép đánh giá cảm xúc của nhiều cá nhân cùng lúc.
- Thiết kế và đề xuất một khung làm việc toàn diện dựa trên công nghệ IoT để đánh giá chất lượng lớp học.
- Đề xuất một công thức mới để đo lường chất lượng học tập trong lớp học thông minh.
- Tiến hành các thí nghiệm toàn diện trên các tập dữ liệu thực tế, chứng minh hiệu quả của khung làm việc được đề xuất.

Nghiên cứu này được cấu trúc như sau: Phần II trình bày các nghiên cứu liên quan, Phần III mô tả chi tiết quy trình thu thập dữ liệu và triển khai hệ thống, trong khi Phần IV tập trung vào việc đánh giá các phương pháp được đề xuất.

Tác giả liên hệ: Trần Tiến Công,
Email:congtt@ptit.edu.vn.
Đến tòa soạn: 10/2024, chỉnh sửa: 11/2024, chấp nhận đăng: 12/2024.



Hình 1: Sơ đồ hệ thống phân tích chất lượng bài giảng dựa trên nhận diện cảm xúc.

II. NGHIÊN CỨU LIÊN QUAN

Nghiên cứu của chúng tôi tập trung vào ba lĩnh vực chính: nhận diện khuôn mặt, theo dõi cảm xúc và vai trò của cảm xúc trong quá trình giáo dục.

Nhận diện khuôn mặt đã có những tiến bộ đáng kể, chuyển từ các kỹ thuật học máy truyền thống sang các phương pháp phức tạp hơn dựa trên học sâu. Các phương pháp học máy ban đầu, như Haar Cascades - một phương pháp sử dụng các đặc trưng trực quan đơn giản để phát hiện khuôn mặt [7] được giới thiệu bởi Viola và Jones, sử dụng một chuỗi các bộ phân loại để nhận diện khuôn mặt một cách nhanh chóng và hiệu quả. Tương tự, phương pháp HOG+SVM (Histogram of Oriented Gradients kết hợp với Support Vector Machines) là kỹ thuật kết hợp giữa việc trích xuất đặc trưng dựa trên hướng gradient của ảnh và thuật toán phân loại máy vector hỗ trợ [8], đạt được độ chính xác trung bình trong khi vẫn đảm bảo hiệu quả tính toán. Mặc dù có những ưu điểm, các phương pháp truyền thống này vẫn được coi là hạn chế trong việc xử lý các tác vụ nhận diện phức tạp và đầy thách thức hơn.

Với sự ra đời của học sâu, lĩnh vực nhận diện khuôn mặt đã trải qua một bước chuyển đổi quan trọng, đạt được độ chính xác và khả năng phục hồi cao hơn trong nhiều điều kiện khác nhau. Các mạng nơ-ron tích chập (CNN) [9] đã trở thành nền tảng của các mô hình tiên tiến, như MTCNN (Multi-task Cascaded Convolutional Networks) - một mạng CNN đa nhiệm vụ tổ chức theo dạng phân tầng [10], nổi bật trong các tình huống phức tạp nhờ vào quy trình xử lý đa giai đoạn. RetinaFace - một framework nhận diện khuôn mặt đa nhiệm vụ dựa trên mạng một lần duyệt (single-stage) [11], áp dụng chiến lược học đa nhiệm, đã cải thiện hiệu suất nhận diện đồng thời duy trì khả năng xử lý thời gian thực thông qua các kiến trúc mạng back bone như MobileNet [12] và ResNet [13]. Trong các môi trường hạn chế tài nguyên, mô hình ULFG (Ultra-Light-Fast-Generic Face Detector) - một bộ phát hiện khuôn mặt siêu nhẹ [14], được tối ưu hóa cho các thiết bị IoT thông qua lượng hóa mô hình, đã đạt được sự cân bằng hiệu quả giữa tốc độ và độ chính xác, tùy thuộc vào độ phân giải đầu vào. Những tiến bộ này làm nổi bật xu hướng chuyển dịch sang các phương pháp nhận diện khuôn mặt có hiệu suất cao, đặc biệt trong các bối cảnh yêu cầu cả hiệu quả và độ chính xác.

III. METHODOLOGY

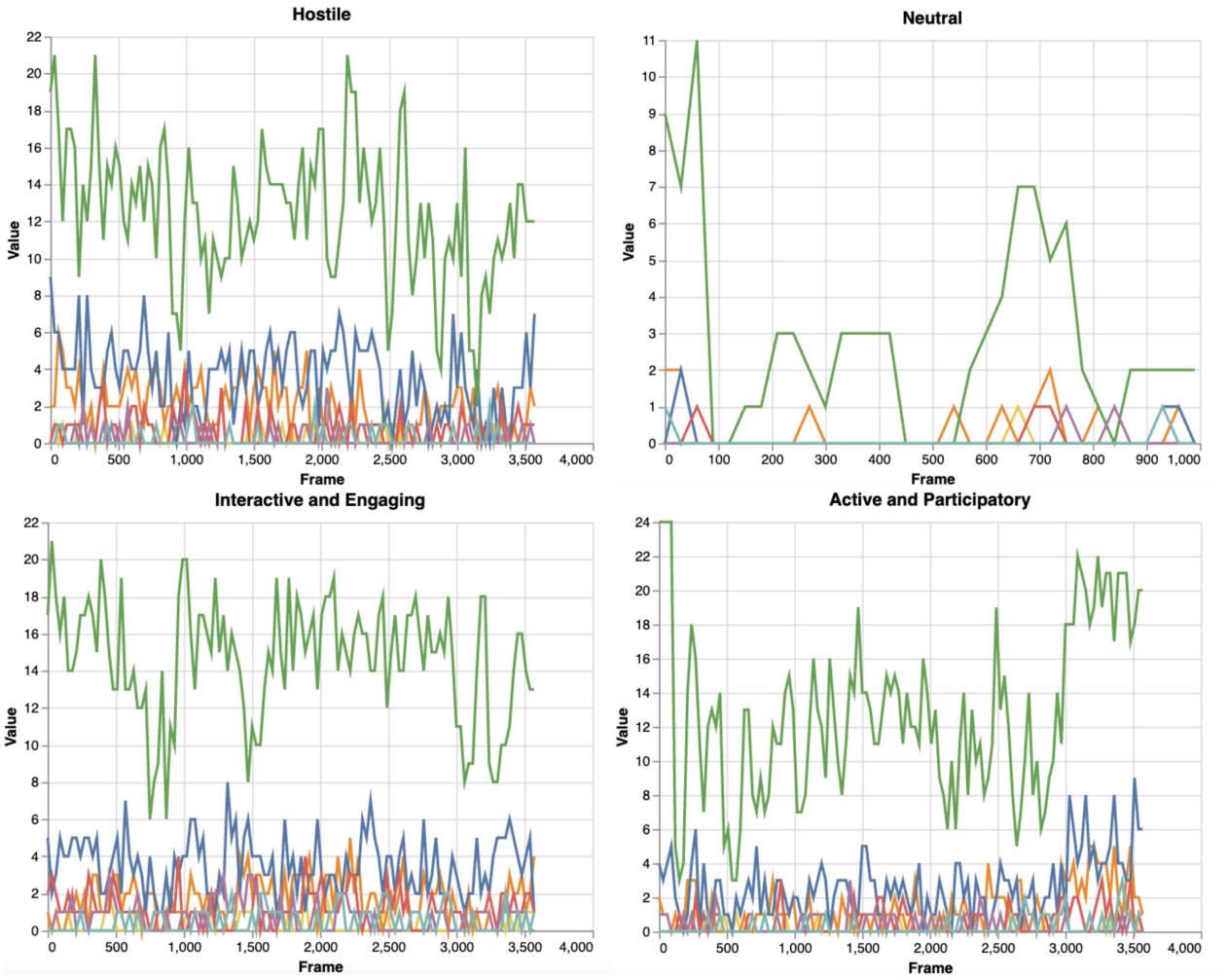
A. Lớp Tiền Xử Lý và Lớp Gán Nhãn

1) Thu thập dữ liệu: Chúng tôi sử dụng bốn IP camera để ghi lại video của một lớp học tại trường tiểu học Nguyễn Trãi. Dữ liệu thu thập bao gồm các cảm xúc cơ bản như trung lập, vui vẻ, ngạc nhiên, buồn bã, giận dữ, ghê tởm và sợ hãi. Mục tiêu của việc thu thập dữ liệu này là để đánh giá chất lượng các buổi học thông qua sự thay đổi cảm xúc của học sinh. Mỗi sự thay đổi trong cảm xúc cho thấy liệu môi trường lớp học đang có động lực tích cực hay tiêu cực. Tập dữ liệu này, mang tên "Emotion PTIT"là một đóng góp mới cho lĩnh vực này, vì chưa có tập dữ liệu nào trước đó được công bố đánh giá cảm xúc của nhiều cá nhân đồng thời.

2) Gán nhãn dữ liệu: Các nhãn được sử dụng để đánh giá chất lượng lớp học bao gồm bốn loại: Căng Thẳng, Trung Lập, Tương Tác và Hứng Thú, Chủ Động và Tham Gia, được định nghĩa như sau:

- **Căng thẳng**: Cảm xúc tiêu cực (giận dữ, ghê tởm) xuất hiện với tần suất cao và chiếm ưu thế so với trạng thái trung lập, tạo ra một môi trường lớp học đầy căng thẳng.
- **Trung lập**: Trạng thái trung lập chiếm ưu thế, với rất ít sự xuất hiện của cảm xúc tích cực hoặc tiêu cực, thể hiện một trạng thái ổn định nhưng thiếu sự nhiệt tình.
- **Tương tác và hứng thú**: Trạng thái trung lập vẫn chiếm ưu thế, nhưng các cảm xúc tích cực như vui vẻ và ngạc nhiên xuất hiện thường xuyên hơn, thể hiện sự tham gia tích cực và hứng thú từ phía học sinh.
- **Chủ động và tham gia**: Trạng thái trung lập chiếm ưu thế, với mức độ xuất hiện của cảm xúc tích cực ở mức thấp hơn, phản ánh sự tham gia chủ động nhưng không quá sôi nổi.

Chúng tôi tiến hành gán nhãn cho các video với độ dài mỗi video khoảng 2 phút. Quy trình gán nhãn được thực hiện bằng cách chia video thành các khung hình với tốc độ xấp xỉ 30 khung hình mỗi giây (fps). Đối với mỗi khung hình, chúng tôi ghi lại số lượng cảm xúc, số thứ tự khung hình hiện tại, số lượng khuôn mặt trong khung hình, cùng với số lượng khuôn mặt được nhận diện thành công và lưu trữ dữ liệu này vào tệp CSV. Tuy nhiên, không phải tất cả các khung hình đều được giữ lại; chúng tôi so sánh mỗi khung hình hiện tại với khung



Hình 2: Biểu đồ thể hiện cảm xúc của học sinh theo thời gian với các màu tương ứng: xanh dương (giận dữ), cam (ghê tởm), đỏ (sợ hãi), xanh dương nhạt (vui vẻ), xanh lá cây (trung lập), vàng (buồn bã) và tím (ngạc nhiên), giúp suy luận tình trạng tổng thể của lớp học.

hình trước đó và loại bỏ các khung hình không có sự thay đổi trong số lượng cảm xúc.

Chúng tôi thiết lập một nhóm gán nhãn gồm 22 người được đào tạo chuyên sâu về đánh giá chất lượng lớp học dựa trên số lượng cảm xúc và tình trạng chung của học sinh trong buổi học thu thập. Bên cạnh đó, năm người được phân công giám sát quá trình gán nhãn dữ liệu. Thống kê của một video mẫu được minh họa trong Hình 2 với với trục hoành thể hiện số khung hình và trục tung biểu thị số lượng khuôn mặt thể hiện cảm xúc. Mỗi đường màu trong biểu đồ đại diện cho một loại cảm xúc khác nhau. Trong trạng thái **Tương tác và hứng thú**, trạng thái trung lập chiếm ưu thế với dao động từ 8-20, kèm theo sự xuất hiện đều đặn của cảm xúc vui vẻ và thỉnh thoảng có cảm xúc ngạc nhiên. Đối với trạng thái **Căng thẳng**, cảm xúc giận dữ chiếm khoảng 30%, trạng thái trung lập dao động mạnh chiếm 50% và các cảm xúc tiêu cực như ghê tởm và sợ hãi chiếm khoảng 20%. Ở trạng thái **Trung lập**, biểu đồ cho thấy sự áp đảo của trạng thái trung lập với hơn 80%, trong khi các cảm xúc khác chỉ xuất hiện với tỷ lệ dưới 20%. Cuối cùng, trong trạng thái **Chủ động và tham gia**, trạng thái trung lập vẫn chiếm ưu thế với khoảng 70%, đi kèm với sự xuất hiện đều đặn của các cảm xúc tích cực như vui vẻ và ngạc nhiên chiếm khoảng 30%.

B. Lớp AI

Trong phần này, chúng tôi chi tiết hóa phương pháp được áp dụng trong hệ thống nhận diện biểu cảm khuôn mặt, như được minh họa trong sơ đồ quy trình (Hình 3).

Dưới đây là các bước của hệ thống:

1) **Dữ liệu đầu vào trực tiếp:** Các luồng video liên tục ghi lại biểu cảm khuôn mặt động của các chủ thể, cung cấp nguồn dữ liệu hình ảnh trực tiếp.
2) **Tiền xử lý:** Dữ liệu đầu vào được chuẩn hóa thông qua các kỹ thuật như điều chỉnh ánh sáng và căn chỉnh khuôn mặt, nhằm nâng cao chất lượng dữ liệu cho các bước xử lý tiếp theo.
3) **Nhận diện khuôn mặt:** Khuôn mặt được xác định và định vị trong dữ liệu hình ảnh thông qua các thuật toán có khả năng nhận diện cấu trúc khuôn mặt và các điểm mốc quan trọng.

4) **Trích xuất đặc trưng:** Các đặc trưng nổi bật được trích xuất từ các khuôn mặt đã nhận diện, tập trung vào các yếu tố hình học và diện mạo cần thiết để nhận dạng biểu cảm khuôn mặt.
5) **Bộ phân loại CNN:** Các đặc trưng đã trích xuất được phân loại vào các trạng thái cảm xúc riêng biệt bằng mạng nơ-ron tích chập (CNN), được huấn luyện để phân biệt giữa các biểu cảm khác nhau.
6) **Đầu ra cảm xúc:** Giai đoạn cuối cùng xuất ra phân loại các biểu cảm khuôn mặt của các chủ thể thành các danh mục như vui vẻ, buồn bã, ngạc nhiên và các trạng thái khác, được xác định bởi CNN.

1) Nhận diện khuôn mặt.: Hệ thống của chúng tôi được thiết kế nhằm giám sát lớp học thông qua việc xử lý các khung hình ghi lại từ nhiều camera được bố trí chiến lược để bao quát toàn bộ học sinh. Khuôn mặt của từng học sinh sau đó được nhận diện và trích xuất từ cảnh quay thu được. Để chính thức hóa quy trình này, chúng tôi giới thiệu công thức toán học sau đây.

Giả sử k là tổng số thiết lập camera trong lớp học. Mỗi thiết lập camera ghi lại một chuỗi khung hình. Đối với mỗi thiết lập camera j, trong đó $j \in [1, k]$, để S_j đại diện cho chuỗi khung hình được ghi lại bởi camera thứ j. Chuỗi này được biểu diễn như sau:

$$\mathcal{S}_j = \{a_{ij} \mid a_{ij} \in \mathbb{R}^{H \times W \times C},\ i \in [1, e_j]\} \tag{1}$$

trong đó a_{ij} đại diện cho khung hình thứ i được ghi lại bởi camera thứ j, H và W là chiều cao và chiều rộng của mỗi khung hình, C là số lượng kênh màu (ví dụ: kênh RGB), và e_j là tổng số khung hình được ghi lại bởi camera thứ j.

Do đó, tập hợp toàn bộ các khung hình được ghi lại bởi tất cả k camera là:

$$\mathcal{S} = \bigcup_{j=1}^{k} \mathcal{S}_j \tag{2}$$

Nhiệm vụ nhận diện khuôn mặt bao gồm việc xác định khuôn mặt trong mỗi khung hình. Điều này được thực hiện thông qua một hàm nhận diện f_d xử lý một khung hình a_{ij} từ tập S_j để dự đoán $\mathcal{Y}_{\text{bb}}$, chứa các hộp bao và điểm số độ tin cậy đã được nhận diện:

$$\mathcal{Y}_{\text{bb}} = f_d(a_{ij}) \tag{3}$$

Cụ thể,

$$\mathcal{Y}_{\text{bb}} = \{(t_{x_l}, t_{y_l}, w_l, h_l, p_l) \mid l \in [1, m], p_l > \theta\}$$

trong đó, (t_{x_l}, t_{y_l}) là tọa độ của tâm hộp bao thứ l, (w_l, h_l) là chiều rộng và chiều cao của hộp bao thứ l, p_l là điểm số độ tin cậy liên quan đến hộp bao thứ l, m là số lượng khuôn mặt được nhận diện và θ là ngưỡng độ tin cậy.

2) Phân loại cảm xúc.: Sau khi nhận diện khuôn mặt, hệ thống trích xuất các hộp bao cắt từ từng hình ảnh, tạo thành m hình ảnh khuôn mặt. Mục tiêu là phân loại từng khuôn mặt vào một trong bảy loại cảm xúc: trung lập, vui vẻ, ngạc nhiên, buồn bã, giận dữ, ghê tởm và sợ hãi. Giả sử $\mathcal{I} = \{I_1, I_2, \ldots, I_m\}$ đại diện cho tập hợp các hình ảnh khuôn mặt đã cắt, trong đó mỗi hình ảnh $I_o \in \mathbb{R}^{H' \times W' \times C}$ có kích thước $H' \times W'$ với C kênh màu.

Việc phân loại được thực hiện bằng hàm Mạng Nơ-ron Tích Chập (CNN) f_{CNN}, hàm này ánh xạ trực tiếp mỗi hình ảnh đầu vào I_o thành một nhãn cảm xúc dự đoán $\hat{y}$ từ tập hợp các lớp cảm xúc được định sẵn $C_1, C_2, \ldots, C_7$:

$$\hat{y} = f_{\text{CNN}}(I_o), \tag{4}$$

trong đó $\hat{y} \in C_1, C_2, \ldots, C_7$ đại diện cho lớp cảm xúc được dự đoán.

C. LỚP ĐÁNH GIÁ CHẤT LƯỢNG

1) Bộ phân loại: **Tính toán tần suất cảm xúc**

Dựa trên kết quả phân loại cảm xúc từ Lớp Dịch Vụ AI, trong đó mỗi khuôn mặt b trong khung hình τ được gán một nhãn cảm xúc dự đoán $\hat{y}_b^{(\tau)}$, bước đầu tiên là tính tần suất tương đối $\gamma_q^{(\tau)}$ của mỗi lớp cảm xúc q trong khung hình . Tần suất được tính như sau:

$$\gamma_q^{(\tau)} = \frac{1}{F^{(\tau)}} \sum_{b=1}^{F^{(\tau)}} \mathbb{I}\left(\hat{y}_b^{(\tau)} = C_q\right), \tag{5}$$

trong đó $C_q \in \{C_1, C_2, \ldots, C_7\}$ đại diện cho các lớp cảm xúc được định sẵn, $\mathbb{I}(\cdot)$ là hàm tính tần suất cảm xúc.

Tính điểm khung hình

Sau khi xác định tần suất cảm xúc $\gamma_q^{(\tau)}$ cho mỗi khung hình τ, điểm cảm xúc tổng thể $\Psi^{(\tau)}$ cho khung hình được tính. Điểm này được tính bằng cách sử dụng tổng có trọng số của các tần suất cảm xúc:

$$\Psi^{(\tau)} = \sum_{q=1}^{7} \left(\gamma_q^{(\tau)} \beta_q\right) + \eta, \tag{6}$$

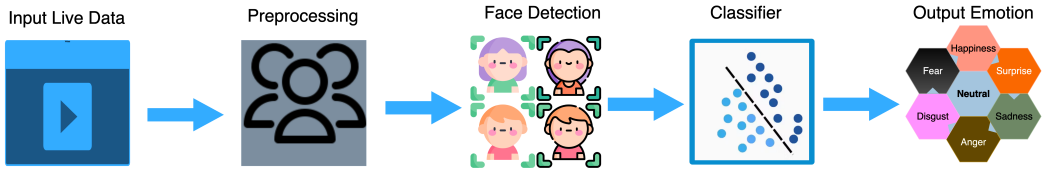
trong đó, β_q là trọng số được gán cho mỗi lớp cảm xúc, và η là một hằng số điều chỉnh.

Phân loại trạng thái lớp học

Để xác định trạng thái tổng thể của lớp học, chúng tôi xem xét điểm cảm xúc trung bình trên tất cả các khung hình trong buổi học. Giả sử $\mathcal{F} = \{\tau_1, \tau_2, \ldots, \tau_T\}$ đại diện cho tập hợp các khung hình trong buổi học. Trạng thái tổng thể của lớp học Λ^* được xác định bởi phương trình sau:

$$\Lambda^* = \mathcal{G}\left(\frac{1}{T} \sum_{\tau \in \mathcal{F}} \Psi^{(\tau)}\right), \tag{7}$$

trong đó $\mathcal{G}(\cdot)$ là hàm phân loại, gán trạng thái tổng thể của lớp học dựa trên các ngưỡng được xác định trước.



Hình 3: Flow diagram of the AI service for facial expression recognition.

2) Các nhãn: Hiểu rõ trạng thái cảm xúc của học sinh là yếu tố quan trọng trong việc thúc đẩy một môi trường học tập hiệu quả. Các nhãn được sử dụng để mô tả động lực cảm xúc trong lớp học bao gồm: Tương Tác và Hứng Thú, Chủ Động và Tham Gia, Trung Lập và Căng Thẳng.

IV. KẾT QUẢ THỰC NGHIỆM

A. Các phương pháp so sánh và Chỉ số hiệu suất

Hệ thống đề xuất bao gồm ba thành phần cốt lõi: nhận diện khuôn mặt, phát hiện cảm xúc và đánh giá chất lượng lớp học. Đối với AI Box trong Hình 1, chúng tôi phát triển một thiết bị IoT tiết kiệm năng lượng dựa trên *NVIDIA Jetson TX2*. Thiết bị Jetson TX2 được trang bị GPU NVIDIA Pascal™ với 256 lõi CUDA, mang lại khả năng xử lý hiệu quả các tác vụ tính toán song song. Ngoài ra, thiết bị này còn sử dụng CPU lõi kép NVIDIA Denver 2 64-bit kết hợp với lõi tứ ARM Cortex-A57 MPCore, đảm bảo sự cân bằng tối ưu giữa hiệu suất năng lượng và khả năng xử lý.

Do khả năng tính toán hạn chế của AI box, chúng tôi đã đánh giá thành phần nhận diện khuôn mặt bằng cách sử dụng hai mô hình nhẹ, gồm *Ultra-Light-Fast-Generic-Face-Detector* [14] và *RetinaFace* [11]. Đối với nhiệm vụ phân loại cảm xúc, chúng tôi tiến hành thí nghiệm trên bốn mô hình bao gồm: *ResNet50, VGG13, VGG16*, và *MobileNetV2* [13], [15], [16].

Chúng tôi đánh giá hiệu suất nhận diện đối tượng bằng chỉ số Trung bình Độ chính xác Trung bình (mAP), theo phương pháp của Henderson et al. [17], nhằm đo lường độ chính xác và độ phủ trên các ngưỡng Giao điểm trên Hợp (IoU) khác nhau. Đối với nhiệm vụ phân loại cảm xúc và đánh giá chất lượng lớp học, hiệu suất mô hình được đo bằng độ chính xác, đảm bảo tính nhất quán trong việc đánh giá cả hai nhiệm vụ.

B. Thiết lập Thực nghiệm

Quá trình huấn luyện được thực hiện trên GPU Nvidia 4090 với các siêu tham số sau: kích thước batch là 32, 100 epochs, tốc độ học là 10^{-4}, tỷ lệ dropout là 0.4 và tỷ lệ điều chuẩn là 0.1. Cơ chế dừng sớm được áp dụng với ngưỡng kiên nhẫn là 15 epochs, và giảm tốc độ học khi đạt ngưỡng plateau được áp dụng với ngưỡng kiên nhẫn là 10 epochs và hệ số giảm là 0.1.

Phương pháp gán trọng số cảm xúc:

1) Mô hình toán học: Giá trị trung bình của cảm xúc được tính bằng công thức:

$$\Psi_{\text{avg}} = \frac{1}{T} \sum_{\tau \in \mathcal{F}} \Psi^{(\tau)} \tag{8}$$

2) Bảng trọng số: Trọng số cho từng cảm xúc được xác định như sau:

$$w_i = \begin{cases} 0.9 & \text{sợ hãi} \\ 0.85 & \text{giận dữ} \\ 0.8 & \text{ghê tởm} \\ 0.75 & \text{buồn bã} \\ 0.7 & \text{vui vẻ} \\ 0.6 & \text{ngạc nhiên} \\ 0.5 & \text{trung lập} \end{cases} \tag{9}$$

3) Nguyên lý tối ưu: Mối quan hệ giữa các thành phần:

$$w_i \propto \frac{I_i}{f_i} \tag{10}$$

trong đó:

- w_i: trọng số cảm xúc
- f_i: tần suất xuất hiện
- I_i: hệ số tác động

4) Phân loại trạng thái: Trạng thái lớp học được xác định theo ngưỡng:

$$\text{Trạng thái} = \begin{cases} \text{Căng thẳng} & \Psi_{\text{avg}} \geq 0.64 \\ \text{Chủ động và tham gia} & 0.61 \leq \Psi_{\text{avg}} < 0.64 \\ \text{Tương tác và hứng thú} & 0.58 \leq \Psi_{\text{avg}} < 0.61 \\ \text{Trung lập} & \Psi_{\text{avg}} < 0.58 \end{cases} \tag{11}$$

C. Ví dụ minh họa tính toán điểm cảm xúc từ video

1) Dữ liệu đầu vào: Giả sử chúng ta có một video gồm 3 khung hình ($T = 3$) và mỗi khung hình chứa nhiều khuôn mặt với các cảm xúc được phát hiện như sau:

- **Khung hình 1**:
  - Tổng số khuôn mặt: 5
  - Cảm xúc phát hiện: Trung tính (2), Hạnh phúc (1), Giận (1), Buồn (1)
  - Trọng số: 0.5, 0.7, 0.85, 0.75
- **Khung hình 2**:
  - Tổng số khuôn mặt: 3

  - Cảm xúc phát hiện: Trung tính (1), Ngạc nhiên (1), Sợ hãi (1)
  - Trọng số: 0.5, 0.6, 0.9
- **Khung hình 3**:
  - Tổng số khuôn mặt: 4
  - Cảm xúc phát hiện: Trung tính (2), Hạnh phúc (1), Ghê tởm (1)
  - Trọng số: 0.5, 0.7, 0.8

2) Tính toán tần suất cảm xúc: Tần suất của từng cảm xúc trong mỗi khung hình ($\gamma_q^{(\tau)}$) được tính bằng công thức:

$$\gamma_q^{(\tau)} = \frac{1}{F^{(\tau)}} \sum_{b=1}^{F^{(\tau)}} I\left(y_b^{(\tau)} = C_q\right)$$

Áp dụng cho từng khung hình:

- **Khung hình 1**:

$$\gamma^{(1)}_{\text{Trung tính}} = \frac{2}{5}, \quad \gamma^{(1)}_{\text{Hạnh phúc}} = \frac{1}{5}, \quad \gamma^{(1)}_{\text{Giận}} = \frac{1}{5},$$

$$\gamma^{(1)}_{\text{Buồn}} = \frac{1}{5}$$

- **Khung hình 2**:

$$\gamma^{(2)}_{\text{Trung tính}} = \frac{1}{3}, \quad \gamma^{(2)}_{\text{Ngạc nhiên}} = \frac{1}{3}, \quad \gamma^{(2)}_{\text{Sợ hãi}} = \frac{1}{3}$$

- **Khung hình 3**:

$$\gamma^{(3)}_{\text{Trung tính}} = \frac{2}{4}, \quad \gamma^{(3)}_{\text{Hạnh phúc}} = \frac{1}{4}, \quad \gamma^{(3)}_{\text{Ghê tởm}} = \frac{1}{4}$$

3) Tính điểm cảm xúc tổng quan cho từng khung hình: Điểm cảm xúc tổng quan ($\Psi^{(\tau)}$) được tính bằng:

$$\Psi^{(\tau)} = \sum_{q=1}^{7} \gamma_q^{(\tau)} \cdot \beta_q + \eta$$

Giả sử tham số bù $\eta = 0$, ta có:

- **Khung hình 1**:

$$\Psi^{(1)} = \left(\frac{2}{5} \cdot 0.5\right) + \left(\frac{1}{5} \cdot 0.7\right) + \left(\frac{1}{5} \cdot 0.85\right) + \left(\frac{1}{5} \cdot 0.75\right) = 0.63$$

- **Khung hình 2**:

$$\Psi^{(2)} = \left(\frac{1}{3} \cdot 0.5\right) + \left(\frac{1}{3} \cdot 0.6\right) + \left(\frac{1}{3} \cdot 0.9\right) = 0.67$$

- **Khung hình 3**:

$$\Psi^{(3)} = \left(\frac{2}{4} \cdot 0.5\right) + \left(\frac{1}{4} \cdot 0.7\right) + \left(\frac{1}{4} \cdot 0.8\right) = 0.63$$

4) Tính điểm tổng quan của video: Điểm tổng quan của video (Ψ_{avg}) được tính bằng trung bình cộng các điểm cảm xúc của tất cả khung hình:

$$\Psi_{\text{avg}} = \frac{1}{T} \sum_{\tau=1}^{T} \Psi^{(\tau)}$$

Áp dụng cho video:

$$\Psi_{\text{avg}} = \frac{\Psi^{(1)} + \Psi^{(2)} + \Psi^{(3)}}{3} = \frac{0.63 + 0.67 + 0.63}{3} = 0.643$$

5) Phân loại trạng thái lớp học: Dựa vào giá trị $\Psi_{\text{avg}} = 0.643$, trạng thái lớp học được phân loại là:

Căng thẳng $(\Psi_{\text{avg}} \geq 0.64)$

D. Kết quả Thực nghiệm

Trong phần này, chúng tôi tập trung vào bốn câu hỏi nghiên cứu chính và thiết kế các thí nghiệm nhằm tìm hiểu câu trả lời cho các câu hỏi này.

Câu hỏi 1. Tại sao chọn mô hình nhận diện khuôn mặt siêu nhẹ cho việc nhận diện khuôn mặt?

Câu hỏi 2. Làm thế nào để chọn mô hình phân loại cảm xúc khuôn mặt tốt nhất?

Câu hỏi 3. Làm thế nào để chọn mô hình tốt nhất cho việc đánh giá chất lượng lớp học?

Câu hỏi 4. Hệ thống đề xuất hoạt động như thế nào trong các môi trường lớp học thực tế?

Bảng I: So sánh hiệu suất của ULFG và RetinaFace

Mô hình	mAP	Thời gian suy luận	Kích thước mô hình (MB)
ULFG phiên bản-RFB	0.928	0.02	1.11
ULFG phiên bản-slim	0.91	**0.018**	**1.04**
RetinaFace	**0.95**	0.04	1.7

1) Tại sao chọn mô hình nhận diện khuôn mặt siêu nhẹ cho việc nhận diện khuôn mặt? (Câu hỏi 1): Trong nghiên cứu của chúng tôi, chúng tôi sử dụng tập dữ liệu Emotion PTIT cho cả mục đích huấn luyện và đánh giá nhờ vào sự đa dạng và phong phú về biểu cảm khuôn mặt của nó. Tập dữ liệu toàn diện này cung cấp cơ sở đáng tin cậy hơn để đánh giá hiệu suất của mô hình. Chúng tôi chia tập dữ liệu thành ba phần: 70% để huấn luyện, 20% để xác thực và 10% để kiểm tra. Đối với mỗi mô hình, chúng tôi đánh giá hiệu suất bằng hai chỉ số chính: mAP và thời gian suy luận.

Trong Bảng I, chúng tôi trình bày hiệu suất của ba mô hình: ULFG phiên bản-RFB, ULFG phiên bản-slim và RetinaFace, theo độ chính xác (được đo bằng mAP) và thời gian suy luận. Mô hình đề xuất của chúng tôi, ULFG phiên bản-RFB, cho thấy hiệu suất cạnh tranh so với cả hai mô hình cơ sở, đồng thời thể hiện hiệu quả về kích thước mô hình và thời gian suy luận. Chúng tôi thảo luận một số điểm chính rút ra từ các kết quả này:

Độ chính xác cân bằng: Mặc dù RetinaFace đạt mAP cao nhất ở mức 0.95, ULFG phiên bản-RFB bám sát với

Bảng II: Các chỉ số hiệu suất của các mô hình học sâu khác nhau cho phân loại cảm xúc khuôn mặt.

Mô hình	Độ chính xác	Điểm F1 (F1 Score)
ResNet50	0.663	0.6607
VGG13	0.687	0.6865
VGG16	0.676	0.6740
MobileNetV2	**0.843**	**0.8400**

mAP là 0.928, cho thấy độ chính xác cạnh tranh. Sự đánh đổi nhỏ này về độ chính xác là chấp nhận được, khi xem xét các lợi ích khác mà ULFG phiên bản-RFB mang lại.

Hiệu quả trong các ứng dụng thời gian thực: Mặc dù hơi chậm hơn so với ULFG phiên bản-slim (0.018 giây so với 0.02 giây trong thời gian suy luận), ULFG phiên bản-RFB vẫn đạt thời gian suy luận nhanh hơn đáng kể so với RetinaFace (0.04 giây). Đối với các ứng dụng thời gian thực, sự chênh lệch nhỏ giữa ULFG phiên bản-slim và phiên bản-RFB được bù đắp bởi độ chính xác cải thiện của phiên bản-RFB, khiến ULFG phiên bản-RFB trở thành lựa chọn cân bằng hơn cho các tác vụ đòi hỏi về thời gian.

Kích thước mô hình gọn nhẹ: Kích thước mô hình của ULFG phiên bản-RFB là 1.11 MB, gần với ULFG phiên bản-slim là 1.04 MB và nhỏ hơn đáng kể so với RetinaFace là 1.7 MB, khiến nó trở thành lựa chọn lý tưởng để triển khai trên các thiết bị IoT tiết kiệm năng lượng. Kích thước gọn nhẹ này giúp giảm mức sử dụng bộ nhớ và tiêu thụ năng lượng, cho phép các ứng dụng thời gian thực hiệu quả trên các thiết bị hạn chế về bộ nhớ và pin mà không làm giảm hiệu suất.

Kết luận, ULFG phiên bản-RFB mang lại sự kết hợp lý tưởng giữa độ chính xác, hiệu suất và tính nhỏ gọn, làm cho nó trở thành lựa chọn phù hợp trong các ứng dụng thực tiễn tại những môi trường hạn chế tài nguyên tính toán, như IoT hoặc các hệ thống nhúng.

Mô hình ULFG (phiên bản RFB) đạt mAP 92.8%, với thời gian suy luận 0.02 giây. Đây là lựa chọn tối ưu nhờ cân bằng giữa độ chính xác và tốc độ. So sánh với RetinaFace: RetinaFace đạt mAP cao hơn (95%), nhưng thời gian xử lý dài hơn (2.23 giây), không phù hợp với các ứng dụng thời gian thực.

2) Làm thế nào để chọn mô hình phân loại cảm xúc khuôn mặt tốt nhất? (Câu hỏi 2): Để xác định mô hình tối ưu cho phân loại cảm xúc khuôn mặt, chúng tôi đang đánh giá bốn kiến trúc học sâu được sử dụng rộng rãi: MobileNetV2, ResNet50, VGG13 và VGG16. Như được chỉ ra trong Bảng II, MobileNetV2 đạt độ chính xác cao nhất với 84.3%, tiếp theo là VGG13 với 68.7%, VGG16 với 67.6%, và ResNet50 với 66.3%. Các kết quả này cho thấy rằng MobileNetV2 vượt trội đáng kể so với các mô hình khác về độ chính xác phân loại.

Với hiệu suất vượt trội, MobileNetV2 trở thành mô hình hiệu quả nhất cho nhiệm vụ này. Hơn nữa, kiến trúc nhẹ và hiệu quả tính toán của nó khiến nó đặc biệt phù hợp cho việc nhận diện cảm xúc khuôn mặt theo thời gian thực trong các môi trường hạn chế về tài nguyên, như lớp học, nơi mà việc xử lý hiệu quả là yếu tố quan trọng cho triển khai.

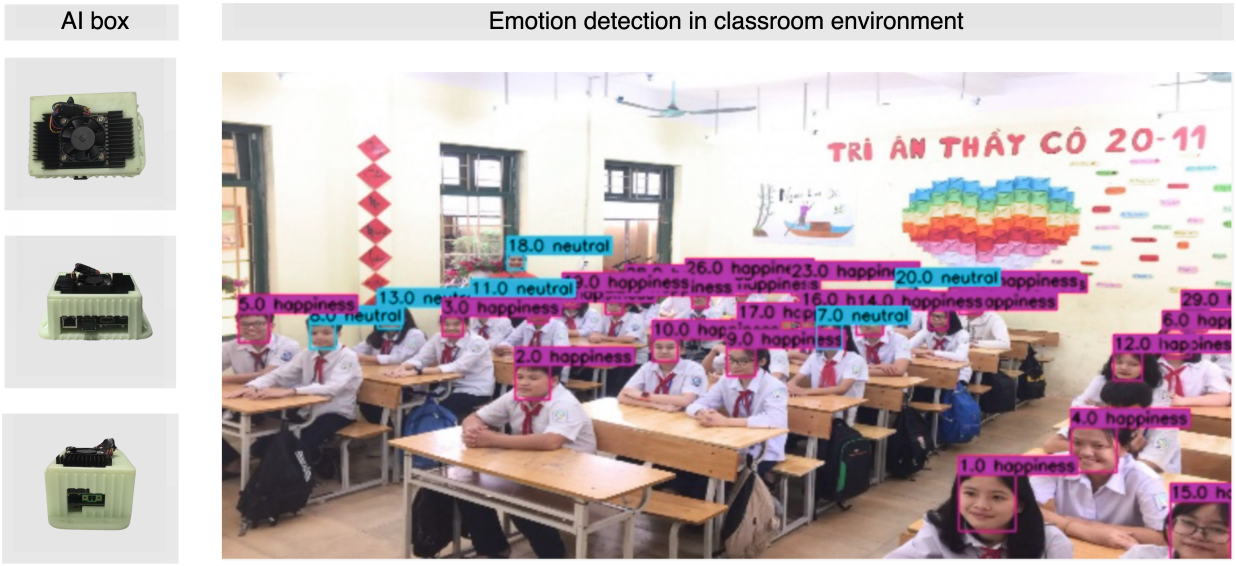
Ngược lại, mặc dù VGG13, VGG16, và ResNet50 cho thấy hiệu suất cạnh tranh, chúng yêu cầu nhiều tài nguyên tính toán hơn, điều này làm hạn chế tính thực tiễn của chúng trong các ứng dụng thời gian thực. Công việc trong tương lai nên khám phá việc tinh chỉnh các siêu tham số và áp dụng các kỹ thuật tối ưu hóa bổ sung để cải thiện hiệu suất của các mô hình khác mà không làm giảm hiệu quả.

Kết luận, ULFG phiên bản-RFB đạt được sự cân bằng tối ưu giữa độ chính xác, hiệu quả và tính gọn nhẹ, khiến nó trở thành lựa chọn lý tưởng cho các triển khai thực tế yêu cầu hiệu suất cao với tài nguyên tính toán hạn chế.

3) Làm thế nào để chọn mô hình tốt nhất cho việc đánh giá chất lượng lớp học? (Câu hỏi 3): Trong việc chọn mô hình tối ưu để đánh giá chất lượng lớp học, chúng tôi áp dụng một phương pháp có hệ thống để đánh giá cẩn thận các cơ chế trọng số và tham số ngưỡng. Chúng tôi áp dụng chiến lược giảm trọng số cho các nhãn xuất hiện thường xuyên, nhằm giảm bớt sự chi phối của chúng trong quá trình đánh giá tổng thể. Ngược lại, chúng tôi tăng trọng số cho các nhãn ít xuất hiện hơn để tăng cường ảnh hưởng của chúng, vì các nhãn tần suất thấp trong nghiên cứu của chúng tôi chủ yếu tương ứng với các cảm xúc tiêu cực, có tác động đáng kể đến môi trường lớp học.

Ngoài ra, chúng tôi thử nghiệm với các thiết lập ngưỡng khác nhau, điều chỉnh các ngưỡng này dựa trên các trọng số cảm xúc đã được hiệu chỉnh lại. Các ngưỡng này được kiểm tra nghiêm ngặt trong một khoảng giá trị cụ thể và áp dụng cho tập dữ liệu đã được gán nhãn của chúng tôi. Kết quả cho thấy hiệu suất mạnh mẽ, với độ chính xác của mô hình đạt 88%. Điều này cho thấy rằng việc tinh chỉnh phân bố trọng số và xác định ngưỡng đóng góp đáng kể vào việc phát hiện chính xác cảm xúc của học sinh và đánh giá toàn diện chất lượng lớp học.

4) Hệ thống đề xuất hoạt động như thế nào trong môi trường lớp học thực tế? (Câu hỏi 4): Chúng tôi tiến hành một nghiên cứu định tính, hiển thị kết quả từ hệ thống AI thông qua các hình ảnh thu được từ camera (Hình 4). Phân phối cảm xúc cung cấp cơ sở để đánh giá không khí lớp học và mức độ tham gia của học sinh, là các yếu tố quan trọng trong việc đánh giá chất lượng giảng dạy. Hơn nữa, chúng tôi tích hợp mô hình vào một hộp AI dựa trên IoT, cho phép xử lý dữ liệu trực tiếp trên các thiết bị biên, giúp giảm tải cho máy chủ và giảm độ trễ, đồng thời nâng cao hiệu quả chi phí cho các triển khai quy mô lớn. Hộp AI đảm bảo tính sẵn sàng cao, hoạt động độc lập với điều kiện mạng, duy trì hoạt động liên



Hình 4: Ứng dụng thực tế của công nghệ nhận diện cảm xúc

tục ngay cả khi các thiết bị khác gặp sự cố. Thiết lập này cải thiện khả năng mở rộng và hỗ trợ đánh giá thời gian thực trong môi trường lớp học.

5) NVIDIA Jetson TX2 có thể đáp ứng như thế nào về hiệu suất xử lý thời gian thực và mức tiêu thụ năng lượng khi triển khai trên lớp học thực tế?: Để đánh giá hiệu suất thời gian thực và mức tiêu thụ năng lượng của thiết bị IoT được sử dụng trong hệ thống nhận diện cảm xúc, chúng tôi tiến hành kiểm tra NVIDIA Jetson TX2. Các thông số kỹ thuật và kết quả đánh giá được trình bày trong Bảng III.

Bảng III: Hiệu suất thời gian thực và mức tiêu thụ năng lượng của Jetson TX2 trên thiết bị IoT

Thông số	Jetson TX2
CPU	Dual-core Denver 2 + Quad-core ARM Cortex-A57 (2 GHz)
GPU	NVIDIA Pascal với 256 CUDA cores và 64 Tensor cores
Bộ nhớ RAM	8 GB LPDDR4 (tùy chọn hỗ trợ ECC)
Bộ nhớ lưu trữ	32 GB eMMC 5.1
Tiêu thụ điện năng tối đa	15 W (với chế độ tiết kiệm năng lượng)
Hiệu suất xử lý CPU	1.8 GHz
Hiệu suất xử lý GPU	1.3 TFLOPS
Tốc độ xử lý hình ảnh	25 FPS
Độ trễ xử lý	40 ms/frame

a) Đánh Giá Hiệu Suất Thời Gian Thực:
Jetson TX2 đã được sử dụng để xử lý và phân tích các luồng video từ các camera trong hệ thống nhận diện cảm xúc. Dưới đây là một số kết quả chính từ các thử nghiệm thực nghiệm:

Tốc độ Xử Lý Hình Ảnh: Hệ thống của chúng tôi có khả năng xử lý đồng thời tới 50 khuôn mặt với tốc độ 25 khung hình/giây (FPS), đảm bảo khả năng phản hồi nhanh chóng và chính xác trong thời gian thực.

Độ Chính Xác: Hệ thống đạt độ chính xác 88,68% trong việc phân loại cảm xúc của các khuôn mặt được nhận diện.

Độ Trễ Xử Lý: Độ trễ trung bình khi xử lý mỗi khung hình là 40 ms/frame, đảm bảo phản hồi gần như ngay lập tức trong các ứng dụng thời gian thực.

b) Mức Tiêu Thụ Năng Lượng:
Jetson TX2 được tối ưu hóa để hoạt động hiệu quả hơn trong các chế độ tiết kiệm năng lượng, giúp kéo dài thời lượng pin và giảm nhiệt độ hoạt động. Dưới đây là các kết quả về mức tiêu thụ năng lượng:

Tiêu Thụ Điện Năng Tối Đa: 15 W trong chế độ hoạt động cao.

Chế Độ Tiết Kiệm Năng Lượng: Khi sử dụng các chế độ tiết kiệm năng lượng, Jetson TX2 có thể giảm mức tiêu thụ điện năng xuống còn 10 W, giúp tăng thời gian hoạt động cho các ứng dụng di động và liên tục.

Quản Lý Nhiệt: Hệ thống làm mát hiệu quả giúp duy trì nhiệt độ hoạt động ổn định, giảm nguy cơ quá nhiệt trong các thiết bị IoT di động.

c) Ảnh Hưởng của Hiệu Suất và Tiêu Thụ Năng Lượng đến Hệ Thống IoT:
Sự cải thiện về hiệu suất thời gian thực và mức tiêu thụ năng lượng của Jetson TX2 đã mang lại nhiều lợi ích cho hệ thống nhận diện cảm xúc trong môi trường lớp học, bao gồm:

Xử Lý Mượt Mà: Khả năng xử lý nhanh chóng và đồng thời nhiều luồng video đảm bảo hệ thống luôn cung cấp các đánh giá cảm xúc chính xác và kịp thời.

Tiết Kiệm Năng Lượng: Chế độ tiết kiệm năng lượng giúp hệ thống hoạt động liên tục trong thời gian dài mà không cần nguồn điện ổn định, thích hợp cho các thiết bị IoT di động.

Tính Linh Hoạt và Mở Rộng: Với khả năng kết nối và mở rộng linh hoạt, Jetson TX2 cho phép tích hợp thêm các thiết bị ngoại vi và mở rộng hệ thống khi cần thiết mà không làm giảm hiệu suất.

V. KẾT LUẬN

Trong nghiên cứu này, chúng tôi đã giới thiệu một hệ thống nhận diện cảm xúc được thiết kế để xác định và phân loại cảm xúc trong môi trường lớp học. Hệ thống được huấn luyện trên tập dữ liệu bao gồm các hình ảnh lớp học thực tế, nâng cao khả năng áp dụng vào các bối cảnh thực tiễn. Ngoài ra, chúng tôi đã đề xuất một công thức mới, giúp cải thiện đáng kể khả năng của hệ thống trong việc đánh giá chất lượng lớp học. Kết quả thực nghiệm cho thấy hệ thống đạt hiệu quả cao, đặc biệt trong bối cảnh lớp học thực tế.

Hướng nghiên cứu trong tương lai sẽ tập trung vào việc mở rộng và làm phong phú thêm tập dữ liệu tổng hợp nhằm tiếp tục cải thiện hiệu suất của hệ thống trong các ứng dụng thực tế. Chúng tôi cũng dự kiến tiến hành các nghiên cứu thực nghiệm toàn diện, triển khai hệ thống tại các lớp học ở Việt Nam để tiếp tục đánh giá và nâng cao chất lượng lớp học.

LỜI CẢM ƠN

Nghiên cứu này được hỗ trợ bởi Dự án số CT07/03-2022-3, được cấp bởi Sở Khoa học và Công nghệ thành phố Hà Nội

TÀI LIỆU THAM KHẢO

[1] C. M. Tyng, H. U. Amin, M. N. M. Saad, and A. S. Malik, "The influences of emotion on learning and memory," *Frontiers in Psychology*, vol. 8, p. 1454, 2017.

[2] K. Stockinger, R. Rinas, and M. Daumiller, "Student adaptability, emotions, and achievement: Navigating new academic terrains in a global crisis," *Learning and Individual Differences*, vol. 90, p. 102046, 2021.

[3] J. Zhang, Z. Yin, P. Chen, and S. Nichele, "Emotion recognition using multi-modal data and machine learning techniques: A tutorial and review," *Information Fusion*, vol. 59, pp. 103–126, 2020.

[4] M. Jin, Q. Wen, Y. Liang, C. Zhang, S. Xue, X. Wang, J. Zhang, Y. Wang, H. Chen, and X. Li, "Large models for time series and spatio-temporal data: A survey and outlook," *arXiv preprint*, 2023.

[5] M. Bouhlal, K. Aarika, R. Ait Abdelouahid, S. Elfilali, and E. Benlahmar, "Emotions recognition as innovative tool for improving students' performance and learning approaches," in *Procedia Computer Science*, vol. 175. Elsevier B.V., 2020, pp. 597–602.

[6] Z. Cheng, Z.-Q. Cheng, J.-Y. He, J. Sun, K. Wang, Y. Lin, Z. Lian, X. Peng, and A. Hauptmann, "Emotion-llama: Multimodal emotion recognition and reasoning with instruction tuning," *arXiv preprint*, 2024, accepted at NeurIPS 2024. [Online]. Available: https://doi.org/10.48550/arXiv.2406.11161

[7] Z. Sedaghatjoo, H. Hosseinzadeh, and A. Shirzadi, "Efficient optimal dispersed haar-like filters for face detection," *arXiv preprint*, 2024.

[8] Y. Pang, Y. Yuan, X. Li, and J. Pan, "Efficient hog human detection," *arXiv preprint*.

[9] K. O'Shea and R. Nash, "An introduction to convolutional neural networks," *arXiv preprint*, 2015.

[10] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao, "Joint face detection and alignment using multi-task cascaded convolutional networks," *arXiv preprint*, 2016.

[11] J. Deng, J. Guo, Y. Zhou, J. Yu, I. Kotsia, and S. Zafeiriou, "Retinaface: Single-stage dense face localisation in the wild," *arXiv preprint*, 2019.

[12] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, "Mobilenets: Efficient convolutional neural networks for mobile vision applications," *arXiv preprint*, 2017.

[13] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *arXiv preprint*, 2015.

[14] Linzaer, "Ultra-light-fast-generic-face-detector-1mb," https://github.com/Linzaer/Ultra-Light-Fast-Generic-Face-Detector-1MB, accessed on 6 February 2022.

[15] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint*, 2014.

[16] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "Mobilenetv2: Inverted residuals and linear bottlenecks," *arXiv preprint*, 2018.

[17] P. Henderson and V. Ferrari, "End-to-end training of object class detectors for mean average precision," *arXiv preprint arXiv:1607.03476*, 2016.

IOT-BASED EMOTION RECOGNITION SYSTEM FOR ENHANCING STUDENT ENGAGEMENT IN CLASSROOM

Abstract — This paper explores the role of emotion recognition in improving student learning by monitoring psychological states in real-time. We introduce a multi-face automated emotion recognition system that analyzes the emotions of multiple students simultaneously, evaluated on the Emotion PTIT dataset, which consists of students in a classroom, manually labeled with 1,500 images and 300 videos for face recognition and classroom detection tasks, respectively. To enhance classroom dynamics, we propose a new formula for evaluating engagement based on emotional data. The system, designed for low-power IoT devices, addresses challenges like load balancing and latency, processing live video feeds to assess classroom interactions. Our prototype can detect up to 50 faces at 25 FPS with an accuracy of 88.18%.

***Keywords* —**Emotion recognition system, Iot device, Learning quality, Real-time system



Vương Thị Uyên sinh năm 2002 tại Hà Nội, Việt Nam. Cô hiện là sinh viên năm thứ tư tại Học viện Công nghệ Bưu chính Viễn thông, Hà Nội. Chuyên ngành của cô là hệ thống thông tin và các hướng nghiên cứu của cô bao gồm xử lý ảnh, thị giác máy tính, xử lý ngôn ngữ tự nhiên và học máy/học sâu.



Nguyễn Nam Hải sinh năm 2003 tại Hà Nội, Việt Nam. Anh ấy hiện là sinh viên năm thứ tư tại Học viện Công nghệ Bưu chính Viễn thông, Hà Nội. Chuyên ngành của anh ấy là hệ thống thông tin và các hướng nghiên cứu của anh ấy bao gồm xử lý ảnh, thị giác máy tính, xử lý ngôn ngữ tự nhiên và học máy/học sâu.



Đào Trọng Hiếu sinh năm 2002 tại Hải Phòng, Việt Nam. Anh ấy hiện là sinh viên năm thứ tư tại Học viện Công nghệ Bưu chính Viễn thông, Hà Nội. Chuyên ngành của anh ấy là hệ thống thông tin và các hướng nghiên cứu của anh ấy bao gồm xử lý ảnh, thị giác máy tính, xử lý ngôn ngữ tự nhiên và học máy/học sâu.



Nguyễn Việt Hưng nhận bằng Thạc sĩ vào năm 2009 từ Viện Công nghệ Grenoble và bằng Tiến sĩ về Xử lý Tín hiệu và Viễn thông từ Đại học Rennes 1, Pháp. Hiện nay, thầy là giảng viên tại Học viện Công nghệ Bưu chính Viễn thông, Việt Nam. Hoạt động nghiên cứu của thầy tập trung vào thiết kế ăng-ten và vi mạch sóng cho các hệ thống truyền thông không dây thế hệ tiếp theo.



Trần Tiến Công nhận bằng Tiến sĩ ngành Khoa học Máy tính từ Đại học Dankook, Yongin, Hàn Quốc, vào năm 2021. Trước đó, Thầy nhận bằng Thạc sĩ Khoa học Máy tính năm 2014 và bằng Cử nhân ngành Mạng và Truyền thông năm 2009 tại Đại học Quốc gia Hà Nội, Việt Nam. Từ tháng 9 năm 2021, anh công tác tại Khoa Công nghệ Thông tin, Học viện Công nghệ Bưu chính Viễn thông, Hà Nội, Việt Nam, với vai trò giảng viên. Lĩnh vực nghiên cứu của anh bao gồm phân tích mạng xã hội, khai phá dữ liệu và học máy.