# Hệ Thống Nhận Diện Biểu Cảm Khuôn Mặt Kết Hợp Phản Hồi Âm Thanh

**Vũ Đức Anh**, **Nguyễn Quang Hiệp**, **Nguyễn Xuân Thuận**, **Lê Đức Mạnh**
*Nhóm 7,Lớp 1601 Khoa Công Nghệ Thông Tin*
*Trường Đại Học Đại Nam*, Việt Nam
**ThS. Nguyễn Văn Nhân**, **ThS. Lê Trung Hiếu**
*Giảng viên hướng dẫn, Khoa Công Nghệ Thông Tin*
*Trường Đại Học Đại Nam*, Việt Nam

***Abstract*—- Ngày nay, nhận dạng khuôn mặt đã trở thành một trong những công nghệ tiên tiến nhất trong lĩnh vực thị giác máy tính. Hệ thống nhận diện khuôn mặt tự động giúp xác định danh tính con người từ hình ảnh tĩnh hoặc video theo thời gian thực, vượt qua những thách thức về điều kiện ánh sáng, góc nhìn và biểu cảm khuôn mặt. Trong nghiên cứu này, chúng tôi đề xuất một hệ thống nhận diện khuôn mặt sử dụng kết hợp các công nghệ hiện đại. Chúng tôi áp dụng TensorFlow và Keras để xây dựng mô hình deep learning dựa trên CNN, kết hợp với OpenCV để phát hiện khuôn mặt từ webcam bằng thuật toán Haar cascade. Để tối ưu và trực quan hóa kết quả, chúng tôi sử dụng NumPy để xử lý dữ liệu số học, Matplotlib và Seaborn để hiển thị biểu đồ và confusion matrix. Hệ thống đánh giá hiệu suất bằng Scikit-learn với các chỉ số như accuracy, precision và recall. Ngoài ra, chúng tôi tích hợp gTTS để chuyển văn bản thành giọng nói và Pygame để phát âm thanh phản hồi. Ứng dụng này có thể được triển khai trên giao diện Tkinter, với sự hỗ trợ của Pillow để hiển thị ảnh, tạo ra một hệ thống hoàn chỉnh giúp nhận diện khuôn mặt và cung cấp phản hồi âm thanh theo thời gian thực.**

***Index Terms*—Nhận diện biểu cảm, Học sâu, CNN, Landmark, Âm thanh tương tác.**

## I. Giới thiệu

Nhận diện biểu cảm khuôn mặt là một trong những lĩnh vực quan trọng trong thị giác máy tính và trí tuệ nhân tạo, với nhiều ứng dụng thực tế trong giám sát an ninh, hỗ trợ tâm lý, chăm sóc sức khỏe và cải thiện tương tác giữa con người với máy móc. Từ lâu, các nhà khoa học đã không ngừng nghiên cứu các phương pháp nhận diện cảm xúc thông qua biểu cảm khuôn mặt nhằm giúp máy tính có thể hiểu và phản hồi lại trạng thái cảm xúc của con người một cách tự nhiên nhất [1, 2].

Nhìn chung, lĩnh vực này đã có những bước tiến đáng kể nhờ vào sự phát triển mạnh mẽ của học sâu (Deep Learning) và các mô hình mạng nơ-ron tích chập (CNN). Những phương pháp truyền thống trước đây chủ yếu dựa vào trích xuất đặc trưng cục bộ như HOG, SIFT hoặc các phương pháp thống kê hình dạng khuôn mặt [2, 3]. Tuy nhiên, các phương pháp này thường bị hạn chế bởi khả năng tổng quát kém, khó xử lý trong các môi trường phức tạp khi điều kiện ánh sáng, góc nhìn hoặc sự đa dạng của biểu cảm thay đổi đáng kể [4].

Trong những năm gần đây, trí tuệ nhân tạo (AI) đã mở ra nhiều hướng nghiên cứu mới cho bài toán nhận diện biểu cảm khuôn mặt. Các mô hình học sâu hiện đại như CNN, ResNet hay các kiến trúc Transformer đã cho thấy tiềm năng vượt trội trong việc học đặc trưng biểu cảm một cách tự động mà không cần trích xuất thủ công như trước đây [5]. Đặc biệt, việc kết hợp giữa phân tích đặc trưng khuôn mặt bằng landmark detection với CNN đã giúp cải thiện đáng kể độ chính xác trong nhận diện cảm xúc, cho phép mô hình nhận diện khuôn mặt tốt ngay cả khi có sự biến đổi về tư thế hoặc môi trường xung quanh [3].

Bên cạnh những thành tựu đạt được, bài toán nhận diện biểu cảm khuôn mặt vẫn còn tồn tại một số thách thức nhất định. Thứ nhất, dữ liệu huấn luyện cho các mô hình học sâu cần phải đủ lớn và đa dạng để đảm bảo tính tổng quát cao, trong khi dữ liệu cảm xúc thực tế thường bị mất cân bằng, với một số biểu cảm hiếm gặp hơn các biểu cảm phổ biến [2, 4]. Thứ hai, các mô hình hiện tại vẫn chưa đạt độ chính xác tuyệt đối trong môi trường thực tế, đặc biệt là khi biểu cảm khuôn mặt không rõ ràng hoặc bị che khuất bởi khẩu trang, kính mắt, hoặc tóc [1, 5].

Do đó, nghiên cứu này của chúng tôi tập trung vào việc xây dựng một hệ thống nhận diện biểu cảm khuôn mặt tối ưu, kết hợp giữa phương pháp landmark facial detection và mô hình CNN để cải thiện khả năng nhận diện cảm xúc. Hệ thống này của chúng tôi không chỉ giúp nâng cao độ chính xác mà còn có thể phản hồi bằng âm thanh để tạo ra một trải nghiệm tương tác chân thực hơn. Mô hình đề xuất sẽ được đánh giá trên nhiều bộ dữ liệu khác nhau để kiểm tra khả năng nhận diện chính xác trong nhiều điều kiện khác nhau. Kết quả nghiên cứu không chỉ đóng góp vào việc nâng cao hiệu suất của các hệ thống nhận diện cảm xúc mà còn mở ra nhiều ứng dụng tiềm năng trong các lĩnh vực như giáo dục, y tế, giám sát an ninh và các hệ thống AI hỗ trợ con người trong tương lai. Phần còn lại của bài báo trên chúng tôi sẽ dược trình bày như sau. Phần II là tổng quan lại các tài liệu nghiêm cứu liên quan về nhận dạng biểu cảm khuôn mặt, phần III là quy trình thu thập dữ liệu và chú thích. Phần IV là phương pháp về nhận dạng biểu cảm khuôn mặt và đánh giá thử nghiệm của mô hình trong phần V. Bài báo kết thúc bằng phần VI, VII, phần kết thúc và đưa ra hướng phát triển.

## II. Nghiên cứu liên quan

Nhận diện biểu cảm khuôn mặt là một lĩnh vực quan trọng trong thị giác máy tính và trí tuệ nhân tạo, với nhiều ứng dụng thực tế từ giám sát an ninh, giáo dục, y tế đến cải thiện tương tác giữa con người và máy móc [1, 2]. Việc hiểu và phân tích cảm xúc không chỉ giúp cải thiện trải nghiệm người dùng mà còn mở ra nhiều tiềm năng trong các lĩnh vực như hỗ trợ điều trị tâm lý, phát hiện mệt mỏi khi lái xe hay giao tiếp người-máy trong môi trường thông minh.

Có nhiều hướng tiếp cận đã được đề xuất nhằm giải quyết bài toán nhận diện biểu cảm khuôn mặt. Trước đây, các phương pháp truyền thống chủ yếu dựa trên đặc trưng hình học và thống kê, trong đó các điểm đặc trưng như mắt, mũi, miệng được trích xuất và phân tích [3]. Một số nghiên cứu sử dụng các kỹ thuật như Phân tích thành phần chính (PCA), Phân tích biệt số tuyến tính (LDA) hoặc máy vector hỗ trợ (SVM) để nhận diện biểu cảm [4]. Dù đạt được những kết quả đáng khích lệ, các phương pháp này vẫn gặp khó khăn khi đối mặt với sự thay đổi về ánh sáng, góc nhìn hoặc khi khuôn mặt bị che khuất một phần.

Sự phát triển của học sâu, đặc biệt là mạng nơ-ron tích chập (CNN), đã mang lại bước tiến lớn trong lĩnh vực này. Các nghiên cứu gần đây đã cho thấy CNN có khả năng học đặc trưng tốt hơn, giúp nhận diện biểu cảm một cách chính xác ngay cả khi dữ liệu đầu vào có nhiều biến đổi [5, 12]. Các mô hình hiện đại như ResNet [12], EfficientNet [17] hay những kiến trúc tích hợp cơ chế tập trung (attention mechanism) [17] đã được đề xuất để tăng cường hiệu suất của hệ thống. Ngoài ra, một số nghiên cứu cũng khai thác Generative Adversarial Networks (GAN) nhằm tăng cường dữ liệu và tạo ra các biểu cảm đa dạng để cải thiện độ chính xác [21].

Không chỉ dừng lại ở nghiên cứu thuật toán, nhiều ứng dụng thực tế đã tận dụng công nghệ nhận diện biểu cảm khuôn mặt để giải quyết các vấn đề cụ thể. Trong lĩnh vực an toàn giao thông, các hệ thống giám sát có thể phát hiện tình trạng buồn ngủ hoặc mất tập trung của tài xế, từ đó đưa ra cảnh báo kịp thời [8]. Trong giáo dục, các hệ thống hỗ trợ giảng dạy có thể phân tích biểu cảm của học sinh để điều chỉnh phương pháp giảng bài [14]. Bên cạnh đó, trong y tế, công nghệ này đã được ứng dụng để hỗ trợ điều trị cho bệnh nhân mắc chứng tự kỷ hoặc rối loạn cảm xúc thông qua phân tích biểu cảm trong môi trường thực tế ảo [9, 10].

Mặc dù đạt được nhiều tiến bộ, việc nhận diện biểu cảm trong điều kiện thực tế vẫn còn nhiều thách thức. Các yếu tố như điều kiện ánh sáng kém, góc quay đa dạng, hoặc sự pha trộn giữa các cảm xúc có thể làm giảm độ chính xác của hệ thống. Để khắc phục những hạn chế này, nghiên cứu của chúng tôi đề xuất một phương pháp kết hợp giữa CNN và phân tích đặc trưng khuôn mặt dựa trên landmark, giúp cải thiện khả năng nhận diện biểu cảm ngay cả trong những điều kiện không lý tưởng. Ngoài ra, chúng tôi cũng tập trung vào tối ưu hóa mô hình trên phần cứng nhúng như Jetson TX2 để có thể triển khai hệ thống nhận diện cảm xúc theo thời gian thực [7].

Bằng cách kết hợp các thuật toán tiên tiến với khả năng triển khai linh hoạt, trong công trình nghiên cứu này chúng tôi không chỉ nâng cao độ chính xác của việc nhận diện biểu cảm mà còn mở ra nhiều ứng dụng tiềm năng trong tương lai. Từ hỗ trợ y tế, giáo dục đến các hệ thống giao tiếp thông minh, công nghệ này có thể giúp xây dựng những nền tảng AI thấu hiểu con người hơn, mang lại trải nghiệm tương tác tự nhiên và hiệu quả hơn trong nhiều lĩnh vực khác nhau.

## III. Tập dữ liệu

### 1. Giới thiệu tập dữ liệu FER2013

FER2013 (Facial Expression Recognition 2013) là một trong những tập dữ liệu tiêu chuẩn được sử dụng rộng rãi trong bài toán nhận diện biểu cảm khuôn mặt. Tập dữ liệu này được công bố lần đầu tại cuộc thi "Facial Expression Recognition Challenge" trong hội nghị ICML 2013 và hiện nay vẫn được nhiều nghiên cứu sử dụng làm cơ sở đánh giá hiệu suất của các mô hình [11].

FER2013 bao gồm tổng cộng 35.887 hình ảnh khuôn mặt có độ phân giải thấp (48×48 pixel) kết hợp với thu thập dữ liệu ngoài vá augment thì được thêm 2500 ảnh cho mỗi cảm xúc nhằm đảm bảo tính đa dạng về đặc điểm nhân khẩu học, điều kiện môi trường và góc nhìn. Mỗi ảnh trong tập dữ liệu được gán nhãn theo 7 loại cảm xúc chính như hình 1 sau:


Hình 1. Đây là 7 loại cảm xúc khuôn mặt người trong tập dữ liệu

### 2. Quy trình thu thập dữ liệu

Tập dữ liệu FER2013 được xây dựng thông qua các bước chính sau đây:

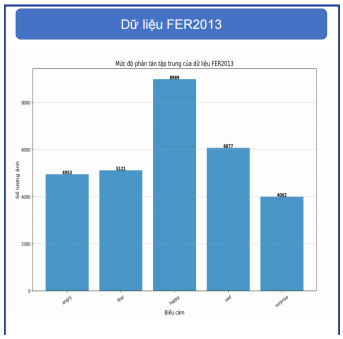
#### 2.1 Thu thập hình ảnh

Các hình ảnh trong FER2013 được lấy từ nhiều nguồn ảnh công khai trên Internet, bao gồm:

Cơ sở dữ liệu hình ảnh khuôn mặt trong các nghiên cứu trước.

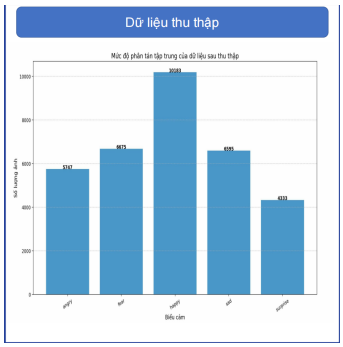
Ảnh từ mạng xã hội, trang web công cộng.

Video trích xuất từ các kho lưu trữ đa phương tiện. Chúng tôi đã kiểm tra mức độ phân tán của tập dữ liệu [11] qua hình 2 như sau




Hình 2. Đây là số lượng ảnh của mỗi loại cảm xúc trong dữ liệu FER2013

Để đảm bảo sự đa dạng, dữ liệu được lấy từ nhiều nhóm người với độ tuổi, giới tính, sắc tộc khác nhau và trong các điều kiện ánh sáng khác nhau. Chúng tôi đã thêm dữ liệu thu thập thêm từ bên ngoài nhằm tăng thêm độ chính xác cho mô hình dữ liệu được thể hiện thông qua hình 3




Hình 3. Đây là số lượng ảnh thêm vào dữ liệu thu thập

### 2.2 Xử lý tiền xử lý hình ảnh

Sau khi thu thập, tất cả hình ảnh đều trải qua quá trình tiền xử lý gồm:

**Chuyển về ảnh xám:** Toàn bộ hình ảnh được chuyển về ảnh đơn kênh (grayscale) nhằm giảm bớt độ phức tạp của mô hình mà vẫn giữ được thông tin biểu cảm.

**Chuẩn hóa kích thước:** Mọi ảnh đều được resize về kích thước 48×48 pixel để đảm bảo tính đồng nhất.

**Căn chỉnh khuôn mặt:** Một số hình ảnh có thể bị nghiêng hoặc lệch góc, do đó các thuật toán nhận diện khuôn mặt (như Haar Cascade, MTCNN) được sử dụng để căn chỉnh lại khuôn mặt về chính giữa khung hình.

### 2.3 Gán nhãn cảm xúc

Quá trình gán nhãn cảm xúc được thực hiện bằng hai phương pháp chính:

**Gán nhãn tự động**: Sử dụng các thuật toán phân tích cảm xúc để gán nhãn ban đầu cho các hình ảnh.

**Kiểm duyệt thủ công**: Một nhóm chuyên gia đã kiểm tra và hiệu chỉnh lại nhãn để đảm bảo độ chính xác.

Để cải thiện tính nhất quán, mỗi hình ảnh được gán nhãn bởi nhiều người khác nhau, sau đó sử dụng phương pháp bỏ phiếu đa số để xác định nhãn cuối cùng.

### 3. Phân chia tập dữ liệu

FER2013 được chia thành ba phần chính: tập huấn luyện (28.709 ảnh, 80%), tập kiểm tra (3.589 ảnh, 10%) và tập kiểm thử (3.589 ảnh, 10%). Cách chia này giúp mô hình có đủ dữ liệu để học và đánh giá khả năng tổng quát hóa trên dữ liệu chưa thấy trước.

Mặc dù FER2013 là một tập dữ liệu lớn và đa dạng, nó cũng có một số hạn chế. Độ phân giải thấp (48×48 pixel) có thể làm mất một số chi tiết quan trọng của biểu cảm khuôn mặt. Ngoài ra, nhãn dữ liệu không đồng đều, một số loại cảm xúc như Ghê tởm có rất ít ảnh, có thể dẫn đến mất cân bằng dữ liệu và ảnh hưởng đến độ chính xác của mô hình.

Để khắc phục những hạn chế này, chúng tôi đã đề xuất áp dụng các phương pháp như:

**Tăng cường dữ liệu (Data Augmentation):** Sử dụng các kỹ thuật như xoay, lật, thay đổi độ sáng để tăng sự đa dạng của hình ảnh.

**Tạo dữ liệu bổ sung bằng GAN:** Mô hình Generative Adversarial Network (GAN) có thể tạo ra các ảnh giả lập nhằm cân bằng số lượng ảnh giữa các nhãn cảm xúc.

**Kết hợp với các tập dữ liệu khác:** Việc sử dụng thêm các bộ dữ liệu như AffectNet, RAF-DB có thể giúp cải thiện chất lượng và độ phong phú của dữ liệu huấn luyện.

Tập dữ liệu FER2013 là một tập dữ liệu nền tảng quan trọng trong nghiên cứu nhận diện cảm xúc khuôn mặt. Dù có những hạn chế nhất định, tập dữ liệu này vẫn đóng vai trò quan trọng trong việc huấn luyện mô hình và phát triển các ứng dụng thực tế như giám sát cảm xúc, hỗ trợ trị liệu tâm lý và giao tiếp người-máy.

## IV. Phương Pháp

### 1. Nhận diện biểu cảm khuôn mặt

Nhận diện biểu cảm khuôn mặt là quá trình phân tích các đặc điểm khuôn mặt để xác định trạng thái cảm xúc của một người. Công nghệ này kết hợp thị giác máy tính và trí tuệ nhân tạo để trích xuất và phân loại cảm xúc dựa trên dữ liệu hình ảnh. Trong hệ thống, các cảm xúc chính bao gồm: hạnh phúc, tức giận, buồn bã, bất ngờ, ghê tởm và sợ hãi.

### 2. Các phương pháp phát hiện biểu cảm khuôn mặt

Các phương pháp nhận diện biểu cảm khuôn mặt có thể được chia thành hai nhóm chính: **Phương pháp truyền thống:** Sử dụng các kỹ thuật như phân tích hình dạng khuôn mặt (Facial Action Coding System - FACS) và trích xuất đặc trưng dựa trên các thuật toán như HOG (Histogram of Oriented Gradients), SIFT (Scale-Invariant Feature Transform), LBP (Local Binary Patterns), PCA (Principal Component Analysis).

**Phương pháp học sâu:** Sử dụng các mô hình học sâu như CNN (Convolutional Neural Networks), RNN (Recurrent Neural Networks) hoặc kết hợp với Attention Mechanism để tự động trích xuất đặc trưng và phân loại cảm xúc.

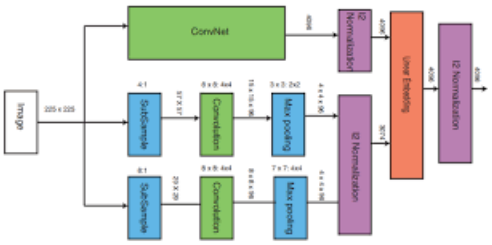
### 3. Thuật toán

Hệ thống nhận diện biểu cảm khuôn mặt được xây dựng dựa trên sự kết hợp của hai phương pháp chính:

**Landmark Detection:** Trích xuất các điểm đặc trưng trên khuôn mặt, bao gồm mắt, mũi, miệng để phân tích các thay đổi về hình dạng.

**Deep Learning với CNN:** Sử dụng mạng nơ-ron tích chập (CNN) để phân tích ảnh khuôn mặt và xác định trạng thái cảm xúc.

**Kết hợp CNN và MLP:** CNN xử lý dữ liệu ảnh, trong khi MLP (Multilayer Perceptron) xử lý dữ liệu landmark để tăng cường độ chính xác của hệ thống được thể hiện qua hình 4 sau




Hình 4. Sơ đồ hoạt động về thuật toán Landmark Detection

Hệ thống sử dụng thư viện `pyttsx3` hoặc `gTTS` để phát giọng nói.

**4. Mô hình hoạt động và cách vận hành mô hình**

**Tiền xử lý dữ liệu:** Ảnh đầu vào được chuyển thành ảnh xám, áp dụng Bilateral Filter để giảm nhiễu nhưng vẫn giữ được biên cạnh.

**Trích xuất đặc trưng:** Hệ thống sử dụng Landmark Detection để xác định vị trí các điểm quan trọng trên khuôn mặt, đồng thời CNN xử lý ảnh để trích xuất đặc trưng học sâu.

**Huấn luyện mô hình:** Dữ liệu được chia thành tập huấn luyện và kiểm tra. Mô hình được tối ưu hóa để tăng độ chính xác và giảm độ nhiễu.

**Nhận diện và phản hồi:** Khi một khuôn mặt được phát hiện, hệ thống dự đoán cảm xúc và đưa ra phản hồi bằng âm thanh hoặc hiển thị thông báo phù hợp.

**5 .Công thức vận hành mô hình**

Mô hình kết hợp CNN (để xử lý ảnh khuôn mặt) và MLP (để xử lý dữ liệu landmark), với hàm mất mát và tối ưu hóa phù hợp. Công thức chính bao gồm:

*1) Biểu diễn đầu vào:* Ảnh khuôn mặt được chuẩn hóa:

$$I' = \frac{I - \mu}{\sigma} \tag{1}$$

Trong đó, $I$ là ảnh đầu vào, $\mu$ là giá trị trung bình của tập dữ liệu, và $\sigma$ là độ lệch chuẩn.

Landmark vector $L$:

$$L = (x_1, y_1, x_2, y_2, ..., x_n, y_n) \tag{2}$$

với $(x_i, y_i)$ là tọa độ của điểm landmark thứ $i$.

*2) Mạng CNN trích xuất đặc trưng:* Đặc trưng từ ảnh được trích xuất qua các lớp CNN:

$$F_{CNN} = f(W_{CNN} * I' + b_{CNN}) \tag{3}$$

Trong đó, $W_{CNN}$ và $b_{CNN}$ là trọng số và bias của CNN, $*$ là toán tử tích chập, và $f$ là hàm kích hoạt ReLU.

*3) Mạng MLP xử lý landmark:* Đặc trưng từ landmark được trích xuất qua MLP:

$$F_{MLP} = g(W_{MLP} \cdot L + b_{MLP}) \tag{4}$$

Trong đó, $W_{MLP}$ và $b_{MLP}$ là trọng số và bias của MLP, $g$ là hàm kích hoạt ReLU.

*4) Kết hợp đặc trưng và phân loại cảm xúc:* Kết hợp hai đặc trưng từ CNN và MLP:

$$F = h(W \cdot [F_{CNN}, F_{MLP}] + b) \tag{5}$$

Trong đó, $h$ là hàm kích hoạt softmax để phân loại cảm xúc.

Hàm softmax:

$$P(y_i) = \frac{e^{z_i}}{\sum_{j=1}^{N} e^{z_j}} \tag{6}$$

với $z_i$ là đầu ra của lớp fully connected, và $N$ là số lớp cảm xúc.

*5) Hàm mất mát và tối ưu hóa:* Sử dụng Categorical Cross-Entropy để đo sai số giữa nhãn thực tế $y$ và dự đoán $\hat{y}$:

$$L = -\sum_{i=1}^{N} y_i \log(\hat{y}_i) \tag{7}$$

Tối ưu hóa bằng thuật toán Adam:

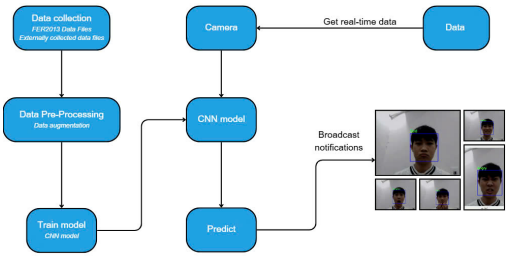
$$\theta_{t+1} = \theta_t - \eta \cdot \frac{m_t}{\sqrt{v_t} + \epsilon} \tag{8}$$

với $m_t$ và $v_t$ là các ước lượng động lượng, và $\eta$ là tốc độ học.

## V. Phân Tích Hoạt Động Của Mô Hình

**1. Luồng Xử Lý Hệ Thống**

Sơ đồ dưới đây minh họa quy trình hoạt động của hệ thống qua hình 5




Hình 5. Sơ đồ hoạt động của mô hình

**2.Các Bước Xử Lý Chi Tiết**

Bước đầu tiên trong quá trình xử lý là thu thập dữ liệu từ các nguồn như FER2013 và các tập dữ liệu thu thập từ bên ngoài. Sau đó, dữ liệu ảnh sẽ được tiền xử lý, bao gồm chuẩn hóa, cân bằng số lượng ảnh giữa các lớp và thực hiện các kỹ thuật tăng cường dữ liệu như xoay, lật, thay đổi độ sáng nhằm cải thiện khả năng tổng quát của mô hình.

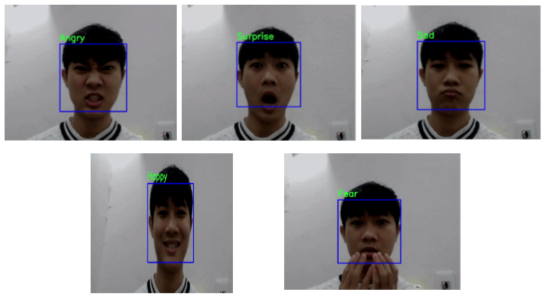
Tiếp theo, mô hình CNN sẽ được sử dụng để trích xuất đặc trưng và phân loại cảm xúc từ ảnh khuôn mặt. Để đảm bảo hệ thống có thể hoạt động hiệu quả trong thời gian thực, quá trình tối ưu hóa mô hình sẽ được thực hiện, tập trung vào giảm độ trễ và tăng tốc độ xử lý. Sau khi huấn luyện, mô hình sẽ được đánh giá dựa trên các tiêu chí như độ chính xác (accuracy), tốc độ xử lý (FPS) và thời gian suy luận (inference time).

Khi hệ thống nhận diện được cảm xúc, phản hồi tương ứng sẽ được xác định. Đồng thời, công nghệ Text-to-Speech (TTS)

sẽ được sử dụng để phát thông báo âm thanh phù hợp với cảm xúc của người dùng. Cuối cùng, mô hình được triển khai thực nghiệm trên luồng video trực tiếp, với camera được đặt ở vị trí phù hợp để đảm bảo nhận diện chính xác và ổn định.

## VI. Kết quả thực nghiệm

Hệ thống nhận diện được 5/7 cảm xúc trong tập dữ liệu kiểm thử. Các cảm xúc nhận diện được bao gồm: Hạnh phúc, Buồn, Giận dữ, Ngạc nhiên, và Sợ hãi. Hệ thống phát thông báo âm thanh tương ứng với từng cảm xúc nhận diện được thể hiện qua hình 6

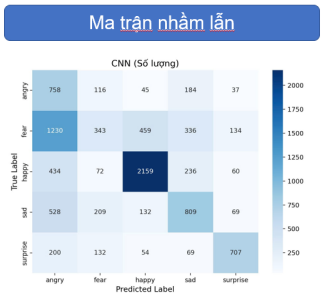



Hình 6. kết quả thử nhiệm khi chạy mô hình

Độ chính xác trung bình đạt 50%, cho thấy cần cải thiện mô hình để nâng cao hiệu suất qua hình 7

| Đánh giá mô hình | |
|---|---|
| Accuracy | 0.50 |
| Precision | 0.54 |
| Recall | 0.50 |
| F1-score | 0.49 |

Hình 7. Độ chính xác, tốc độ xử lý của mô hình




Hình 8. Ma trận nhầm lẫn về số lượng của mô hình

```
Báo cáo chi tiết:
              precision    recall  f1-score   support

       angry       0.24      0.66      0.35      1140
        fear       0.39      0.14      0.20      2502
       happy       0.76      0.73      0.74      2961
         sad       0.50      0.46      0.48      1747
    surprise       0.70      0.61      0.65      1162

    accuracy                           0.50      9512
   macro avg       0.52      0.52      0.49      9512
weighted avg       0.54      0.50      0.49      9512
```

Hình 9. Báo cáo chi tiết về tỉ lệ % train mô hình

Hình 8 và Hình 9 minh họa ma trận nhầm lẫn của mô hình CNN trong bài toán nhận diện biểu cảm khuôn mặt, lần lượt theo số lượng mẫu và tỷ lệ phần trăm. Quan sát Hình 8, ta thấy rằng mô hình hoạt động tốt nhất với biểu cảm "happy" khi có 2.159 mẫu được nhận diện chính xác, chiếm phần lớn trong tập dữ liệu thử nghiệm. Tuy nhiên, với các biểu cảm như "fear" và "sad", mô hình có xu hướng nhầm lẫn với các biểu cảm khác, đặc biệt là giữa "fear" với "angry" và "sad" với "happy".

Hình 9 thể hiện các chỉ số đánh giá mô hình, bao gồm Precision, Recall và F1-score. Kết quả cho thấy mô hình hoạt động tốt nhất với nhãn "happy" (F1-score = 0.74), trong khi "fear" có độ chính xác thấp nhất (F1-score = 0.20). Độ chính xác tổng thể của mô hình đạt 50%, cho thấy vẫn còn dư địa để cải thiện, đặc biệt đối với các nhãn có tỷ lệ nhầm lẫn cao.

Một điểm đáng chú ý là sự chồng chéo giữa các biểu cảm có thể xuất phát từ đặc điểm sinh lý tự nhiên của con người. Ví dụ, một số người khi sợ hãi có thể mở to mắt giống như biểu cảm ngạc nhiên, trong khi một số trạng thái buồn có thể dễ nhầm lẫn với trạng thái bình thường hoặc hạnh phúc nhẹ. Điều này gợi ý rằng việc chỉ dựa vào thông tin hình ảnh có thể chưa đủ để đạt độ chính xác cao. Trong nghiên cứu tương lai, có thể cân nhắc kết hợp thêm thông tin từ chuyển động khuôn mặt theo thời gian hoặc đặc trưng sinh trắc học khác để cải thiện độ chính xác của mô hình.

## VII. Kết Luận

Trong nghiên cứu này, chúng tôi đã xây dựng một hệ thống nhận diện cảm xúc dựa trên mô hình học sâu, có khả năng phân loại cảm xúc từ hình ảnh khuôn mặt. Hệ thống đã nhận diện được 5 trên 7 cảm xúc trong tập kiểm thử, bao gồm Hạnh phúc, Buồn, Giận dữ, Ngạc nhiên, và Sợ hãi. Tuy nhiên, độ chính xác trung bình chỉ đạt 50%, điều này cho thấy hệ thống vẫn còn nhiều hạn chế và cần tiếp tục được cải thiện.

Một trong những thách thức chính của hệ thống là sự nhầm lẫn giữa các cảm xúc có biểu cảm tương tự, điều này có thể bắt nguồn từ sự đa dạng về khuôn mặt, điều kiện ánh sáng, và góc nhìn trong dữ liệu. Ngoài ra, thời gian suy luận cũng là một yếu tố cần được tối ưu để đảm bảo hệ thống hoạt động hiệu quả trong thời gian thực.

Dù còn tồn tại những hạn chế, kết quả thử nghiệm ban đầu cho thấy hệ thống có tiềm năng ứng dụng trong nhiều lĩnh vực khác nhau, từ trợ lý ảo đến theo dõi tâm lý và hỗ trợ giao tiếp người-máy. Việc cải thiện mô hình và tối ưu hóa hiệu suất sẽ giúp hệ thống trở nên hữu ích hơn trong thực tế.

## VIII. Hướng Phát Triển

Một trong những hướng phát triển quan trọng của hệ thống nhận diện cảm xúc là mở rộng tập dữ liệu huấn luyện. Việc thu thập thêm dữ liệu từ nhiều nguồn khác nhau sẽ giúp cải thiện khả năng tổng quát của mô hình, giảm bớt tình trạng nhầm lẫn giữa các cảm xúc tương đồng. Bên cạnh đó, thử nghiệm các mô hình tiên tiến hơn như Vision Transformer (ViT) hoặc hybrid CNN-LSTM có thể nâng cao độ chính xác của hệ thống.

Để cải thiện khả năng phân biệt cảm xúc, việc sử dụng Attention Mechanism có thể giúp mô hình tập trung vào các đặc trưng quan trọng của khuôn mặt, như mắt, miệng và chân mày – những bộ phận thể hiện cảm xúc mạnh nhất. Nhờ cơ chế này, mô hình có thể học được các đặc điểm chi tiết hơn, từ đó tăng độ chính xác trong việc nhận diện cảm xúc.

Bên cạnh độ chính xác, tốc độ xử lý cũng là yếu tố quan trọng để đảm bảo hệ thống hoạt động hiệu quả trong thời gian thực. Việc triển khai mô hình trên GPU hoặc TPU có thể giúp tăng tốc độ suy luận đáng kể. Đồng thời, phát triển các mô hình nhẹ hơn như MobileNetV3 hoặc phiên bản nhỏ của EfficientNet có thể giúp hệ thống hoạt động tốt trên các thiết bị có phần cứng hạn chế như điện thoại di động hoặc hệ thống nhúng.

Ngoài ra, quá trình tiền xử lý ảnh cũng cần được cải thiện để nâng cao chất lượng đầu vào. Các yếu tố như điều kiện ánh sáng, góc chụp và độ phân giải có thể ảnh hưởng đến kết quả dự đoán. Do đó, sử dụng các phương pháp nâng cao chất lượng hình ảnh, như cân bằng sáng hoặc tăng cường dữ liệu với nhiều biến thể hơn, sẽ giúp mô hình hoạt động ổn định hơn.

Nhận diện cảm xúc không chỉ dựa vào khuôn mặt mà còn có thể kết hợp với giọng nói và tư thế cơ thể để nâng cao độ chính xác. Việc sử dụng các mô hình đa modal (multi-modal) kết hợp dữ liệu từ nhiều nguồn khác nhau có thể giúp hệ thống đưa ra dự đoán chính xác hơn. Điều này đặc biệt hữu ích trong các ứng dụng như theo dõi tâm lý, hỗ trợ chẩn đoán sớm các vấn đề tâm lý hoặc nâng cao khả năng tương tác giữa con người và máy móc.

Với những tiềm năng ứng dụng rộng rãi, hệ thống nhận diện cảm xúc có thể được triển khai trong nhiều lĩnh vực thực tế. Trong lĩnh vực trợ lý ảo thông minh, hệ thống có thể hỗ trợ người dùng bằng cách phản hồi phù hợp với cảm xúc của họ. Trong chăm sóc sức khỏe tâm lý, công nghệ này có thể giúp phát hiện các dấu hiệu bất thường về mặt cảm xúc, hỗ trợ chẩn đoán sớm các vấn đề tâm lý. Ngoài ra, trong lĩnh vực robot giao tiếp, việc nâng cao khả năng nhận diện cảm xúc sẽ giúp robot hiểu được cảm xúc của người đối diện và phản hồi một cách tự nhiên, tạo ra trải nghiệm giao tiếp chân thực hơn. Với những cải tiến trong tương lai, hệ thống này có thể trở thành một công cụ hữu ích trong nhiều lĩnh vực, từ công nghệ cá nhân đến y tế và giáo dục.

## IX. Tài liệu tham khảo


[1] Đinh Xuân Nhất. *Nghiên cứu các thuật toán nhận dạng cảm xúc khuôn mặt trên ảnh 2D*. Khóa luận tốt nghiệp, Trường Đại học Bách khoa Hà Nội, 2018.

[2] Nguyễn Thị Thanh Tâm. *Nâng cao độ chính xác nhận dạng khuôn mặt dựa trên mô hình CNN học sâu kết hợp với đặc trưng HOG và bộ phân lớp SVM*. Tạp chí Nghiên cứu KH&CN quân sự, Số 54, 2018.

[3] Nguyễn Thị Duyên, Trương Xuân Nam, Nguyễn Thanh Tùng. *Một mô hình học sâu cho phát hiện cảm xúc khuôn mặt*. Kỷ yếu Hội nghị KHCN Quốc gia lần thứ XII về Nghiên cứu cơ bản và ứng dụng Công nghệ thông tin (FAIR), Huế, 2019.

[4] Huỳnh Cao Tuấn, Đỗ Năng Toàn, Nguyễn Thanh Bình. *Một kỹ thuật nhận dạng biểu cảm khuôn mặt dựa trên mô hình chất liệu*. Tạp chí Khoa học và Công nghệ, 2018.

[5] Hồ Thị Hương Thơm, Nguyễn Kim Anh. *Nghiên cứu nhận dạng biểu cảm khuôn mặt bằng phương pháp học sâu sử dụng kiến trúc ResNet*. Tạp chí Khoa học Công nghệ Hàng hải, ISSN: 1859-316X, 2020.

[6] Rebecca Mobbs, Dimitrios Makris, Vasileios Argyriou. *Emotion Recognition and Generation: A Comprehensive Review of Face, Speech, and Text Modalities*. School of Computer Science and Mathematics, Kingston University London, 2025. License: CC BY-NC-ND 4.0. arXiv: 2502.06803v1 [cs.LG], 02 Feb 2025.

[7] Phạm Minh Quyền, Phùng Thanh Huy, Đỗ Duy Tân, Huỳnh Hoàng Hà, và Trương Quang Phúc. *Nhận diện cảm xúc khuôn mặt dùng mạng nơ-ron tích chập CNN trên phần cứng Jetson TX2*. Trường Đại học Sư phạm Kỹ thuật TP.HCM, Việt Nam, 2020.

[8] Assari, M.A.; Rahmati, M. *Driver drowsiness detection using face expression recognition*. In Proceedings of the IEEE International Conference on Signal and Image Processing Applications, Kuala Lumpur, Malaysia; pp. 337-341, 16-18 November 2011.

[9] Bekele, E.; Zheng, Z.; Swanson, A.; Crittendon, J.; Warren, Z.; Sarkar, N. *Understanding how adolescents with autism respond to facial expressions in virtual reality environments*. IEEE Trans. Vis. Comput. Graphics, Vol. 19, pp. 711-720, 2013.

[10] Chen, C.H.; Lee, I.J.; Lin, L.Y. *Augmented reality-based self-facial modeling to promote the emotional expression and social skills of adolescents with autism spectrum disorders*. Res. Dev. Disabil. Vol. 36, pp. 396-403, 2015.

[11] FER2013 Dataset, Available at: https://www.kaggle.com.

[12] K. He, X. Zhang, S. Ren, and J. Sun. *Deep residual learning for image recognition*. In IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 770-778, June 27-30, 2016.

[13] K. He, X. Zhang, S. Ren, and J. Sun. *Identity mappings in deep residual networks*. In B. Leibe, J. Matas, N. Sebe, and M. Welling, editors, Computer Vision ECCV, volume 9908 of Lecture Notes in Computer Science, Amsterdam, October 8-16, 2016.

[14] Kapoor, A.; Burleson, W.; Picard, R.W. *Automatic prediction of frustration*. Int. J. Hum.-Comput. Stud. Vol. 65, pp. 724-736, 2007.

[15] L. Wolf, T. Hassner, I. Maoz. *Face Recognition in Unconstrained Videos with Matched Background*

*Similarity*. Computer Vision and Pattern Recognition (CVPR), 2011.

[16] Lankes, M.; Riegler, S.; Weiss, A.; Mirlacher, T.; Pirker, M.; Tscheligi, M. *Facial expressions as game input with different emotional feedback conditions*. In Proceedings of the 2008 International Conference on Advances in Computer Entertainment Technology, Yokohama, Japan, December 3-5, pp. 253-256, 2008.

[17] Li, S.; Deng, W. *Reliable crowdsourcing and deep locality-preserving learning for unconstrained facial expression recognition*. IEEE Trans. Image Process, Vol. 28, pp. 356-370, 2019.

[18] Li, Y.; Zeng, J.; Shan, S.; Chen, X. *Occlusion Aware Facial Expression Recognition Using CNN With Attention Mechanism*. IEEE Trans. Image Process. Vol. 28, pp. 2439-2450, 2019.

[19] M. Mathias, R. Benenson, M. Pedersoli, L. Van Gool. *Face detection without bells and whistles*. European Conference on Computer Vision, 2014.

[20] Matthew N. Dailey, Garrison W. Cottrell, Curtis Padgett, and Ralph Adolphs. *EMPATH: A Neural Network that Categorizes Facial Expressions*. Journal of Cognitive Neuroscience 14:8, pp. 1158-1173, 2014.

[21] Yang, H.; Zhang, Z.; Yin, L. *Identity-adaptive facial expression recognition through expression regeneration using conditional generative adversarial networks*. In Proceedings of the 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018), Xi'an, China, 15-19, pp. 294-301, May 2018.