MÔ HÌNH MẠNG NƠRON TÍCH CHẬP ĐA NHIỆM NHẬN DẠNG KHUÔN MẶT VÀ BIỂU CẢM CHO ỨNG DỤNG HỖ TRỢ GIÁM SÁT HỌC TRỰC TUYẾN

MULTI-TASK CNN MODEL FOR FACE AND FACIAL EXPRESSION RECOGNITION AND APPLICATION FOR MONITORING ONLINE LEARNING

Dương Thăng Long, Chu Minh*, Phí Quốc Chính†

Ngày tòa soạn nhận được bài báo: 02/11/2021
Ngày nhận kết quả phản biện đánh giá: 04/05/2022
Ngày bài báo được duyệt đăng: 26/05/2022

Tóm tắt: *Hệ thống quản lý học tập trực tuyến (LMS) đang được phát triển mạnh, góp phần nâng cao chất lượng đào tạo. Tuy nhiên, việc tăng cường giám sát và hỗ trợ người học, theo dõi và quản lý học tập dựa trên các công nghệ hiện đại chưa được nghiên cứu sâu rộng. Đặc biệt là ứng dụng của công nghệ nhận dạng khuôn mặt và biểu cảm khuôn mặt giúp cho việc theo dõi, giám sát người học được tự động hoá cao độ và hỗ trợ kịp thời. Bằng việc ứng dụng công nghệ mạng nơron tích chập đa nhiệm (MTCNN), nghiên cứu này đề xuất một mô hình MTCNN nhằm thực hiện hai nhiệm vụ là nhận dạng khuôn mặt và nhận dạng biểu cảm khuôn mặt. Mô hình được thử nghiệm trên các tập dữ liệu công bố gồm CK+, OuluCASIA và dữ liệu người học được thu thập cho kết quả khả quan khi so sánh với một số kiến trúc hiện đại trong khi kích thước mô hình đơn giản hơn. Chúng tôi cũng thiết kế tích hợp mô hình được đề xuất với hệ thống quản lý học tập trực tuyến (LMS) theo hướng kết nối mở để gia tăng thêm tính năng giám sát và theo dõi quá trình học tập, chủ động cảnh báo cho giáo viên, người học biết để điều chỉnh hoạt động dạy và học nhằm nâng cao chất lượng đào tạo.*

Từ khoá: *Mạng nơron tích chập đa nhiệm, nhận dạng khuôn mặt, nhận dạng biểu cảm khuôn mặt, hệ thống quản lý học tập trực tuyến.*

Abstract: *The online learning management system (LMS) is being more and more widely developed and contributes to improving the quality of training at educational institutions. However, at present, there are few systems with enhanced monitoring and support for learners based on modern technologies. Especially, the application of this facial recognition and facial expression technology makes the tracking and monitoring of learners highly automated and timely supported. By using multi-tasking convolutional neural networks, this study proposes such a network model to perform two tasks of face recognition and facial expression recognition. The model is tested on published data sets including CK+,*

---

* Trường Đại học Mở Hà Nội

† VNPT Hà Nội

OuluCASIA and our collected data. The experimental results are significant in comparison with some modern architectures while the model size is simpler. Based on the proposed model, we design an integrated proposed model with the online LMS in the direction of open connection to increase the monitoring and tracking learning activities, therefore, it can give warnings as well as notify teachers and learners to adjust teaching and learning activities to improve training quality.

Keywords: *Multi-task convolutional neural network, face recognition, facial expressions recognition, online learning management systems.*

I. Giới thiệu

Trong những năm gần đây, sự phát triển mạnh mẽ của e-learning đang thu hút ngày càng nhiều người lựa chọn cách học và tiếp thu kiến thức bằng trực tuyến thông qua hệ thống học tập trực tuyến (LMS). Trong E-learning, mọi người có thể học nhiều thứ họ cần ở bất cứ lúc nào và bất cứ nơi đâu. E-Learning khá linh hoạt và có thể mở rộng dễ dàng, sử dụng phương pháp học cá nhân hoá cao độ, ít tốn kém và đã được chứng minh là hiệu quả hơn so với giáo dục truyền thống. Vì vậy, e-learning ngày càng trở nên phổ biến. Tuy nhiên, giám sát và đánh giá chất lượng hoạt động học tập trực tuyến chắc chắn là điều cần được quan tâm đặc biệt. Chúng ta phải hạn chế đến mức tối thiểu các tình trạng gian lận trong học tập, thi kiểm tra trên các hệ thống trực tuyến và tốt nhất là không để xảy ra tình trạng đó, sẽ ảnh hưởng rất lớn đến kết quả học tập của người học và chất lượng của hệ thống giáo dục. Do đó, các hệ thống quản lý học tập trực tuyến cần phải cung cấp khả năng xác định và giám sát các hoạt động của người học [1]. Một số nghiên cứu tìm kiếm những cách tốt hơn để sử dụng phương pháp sinh trắc học giúp xác định và giám sát trong quá trình học tập và thi trực tuyến [2], [3]. Tuy nhiên, một hệ thống nhận dạng khuôn mặt (FR) và nhận dạng biểu cảm khuôn mặt (FER) sẽ rất thân thiện với con người vì chúng không cần tiếp xúc và không cần phần cứng bổ sung khi hiện nay hầu hết các máy tính hoặc thiết bị người dùng đều có camera tích hợp. Quan trọng hơn, hệ thống FR/FER có thể được sử dụng để xác thực liên tục người học trong toàn bộ quá trình học tập hoặc kiểm tra theo thời gian thực và giám sát, đo đếm các thể hiện quá trình học tập của người học trên biểu cảm khuôn mặt để dựa vào đó, các nhà sư phạm và quản lý có thể điều chỉnh các hoạt động của mình nhằm đáp ứng tốt hơn cho quá trình đào tạo đối với từng người học.

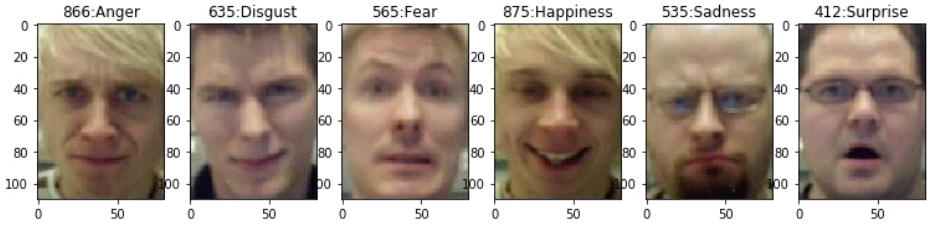
Bài toán FR/FER là những bài toán thú vị và thu hút nhiều nghiên cứu với kết quả tích cực trong lĩnh vực thị giác máy tính, ứng dụng rộng rãi của các bài toán này như giám sát trạng thái người lái xe [4], giám sát người dùng điện thoại, phát hiện biểu cảm không thật, nhận dạng trầm cảm [5], hệ thống giám sát tại các cơ sở y tế và trong giáo dục [3], [2]. Tuy nhiên, bài toán FR/FER vẫn còn nhiều thách thức do sự đa dạng của những người có nét mặt giống nhau và sự thể hiệu biểu cảm trên khuôn mặt của mỗi người có thể thay đổi theo thời gian. Hiện nay, các tác giả chủ yếu tiếp cận vấn đề này dựa trên mạng nơron tích chập (CNN) với các mô hình hiện đại như VGGNet, GoogleNet, ResNet, SENet và chúng đều cho kết quả khả quan. Mặc dù kết quả

nhận dạng trong các mô hình CNN ngày càng tốt hơn khi các phiên bản kiến trúc mạng được điều chỉnh và cải tiến, nhưng vẫn còn một số vấn đề cần được cải thiện, đặc biệt là trong các ứng dụng thực tế. Hơn nữa, các mô hình CNN này thường được thiết kế độc lập cho từng bài toán và có độ phức tạp lớn đối với một số ứng dụng trong thực tế khi có giới hạn về tài nguyên tính toán của máy tính, có những mô hình lên đến hàng trăm triệu tham số [6]. Nghiên cứu này tập trung thiết kế một mô hình CNN đa nhiệm (Multi-Task CNN) cho hai bài toán FR/FER đồng thời với độ phức tạp vừa phải nhưng vẫn đảm bảo chất lượng và hiệu quả cho bài toán. Mô hình sẽ được chạy thử nghiệm để đánh giá trên một số bộ dữ liệu phổ biến như OuluCASIA [7] và được thiết kế để tích hợp với hệ thống LMS để hỗ trợ giám sát và đánh giá quá trình học tập trực tuyến của người học.

II. Một số nghiên cứu liên quan

2.1. Nhận dạng khuôn mặt và biểu cảm

Trong bài toán nhận dạng biểu cảm khuôn mặt, Paul Ekman và cộng sự [5] đã xác định sáu cảm xúc cơ bản được biểu cảm trên khuôn mặt của con người dựa trên nghiên cứu sự giao thoa giữa các nền văn hóa. Theo đó, mọi người cùng thể hiện và cảm nhận được những cảm xúc cơ bản bằng biểu cảm trên khuôn mặt theo cùng một cách bất kể họ thuộc dân tộc hay nền văn hóa nào. Nói cách khác, các nét biểu cảm trên khuôn mặt cơ bản độc lập với nền văn hoá mà con người đang trải nghiệm, sinh sống. Những biểu cảm cơ bản trên khuôn mặt (Hình 2.1) bao gồm tức giận (An-anger), ghê tởm (Di-disgust), sợ hãi (Fe-fear), hạnh phúc (Ha-happiness), buồn bã (Sa-sadness) và ngạc nhiên (Su-surprise). Một biểu cảm khác cũng có thể được sử dụng đó là sự khinh bỉ (Co-contempt). Một số nghiên cứu sử dụng thêm biểu cảm trung tính (Ne-neutral) như một trong số các biểu cảm cơ bản.



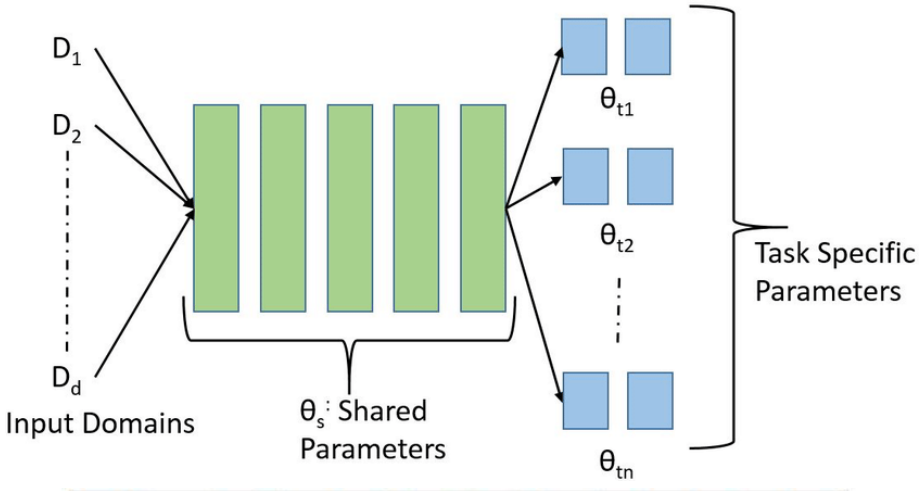
Hình 2.1. Các biểu cảm khuôn mặt cơ bản

Hệ thống FR/FER nói chung có thể được chia thành hai giai đoạn chính, giai đoạn 1 thực hiện trích xuất các đặc trưng của hình ảnh khuôn mặt đại diện cho định danh khuôn mặt và biểu cảm tương ứng và giai đoạn 2 là phân loại các đặc trưng đó vào các định danh và biểu cảm. Việc trích xuất các đặc trưng khuôn mặt cho bài toán FR/FER là rất quan trọng và nó ảnh hưởng đến độ chính xác của việc nhận dạng. Một số phương pháp truyền thống được đề cập trong [8] như kỹ thuật HOG (biểu đồ của gradient có định hướng), kỹ thuật LBP (mẫu nhị phân cục bộ), kỹ thuật Gabor và các đặc trưng kiểu Haar. Các phương pháp này có thể hoạt động tốt trên các tập dữ liệu đơn giản và thuần nhất, nhưng trên thực tế, các tập dữ liệu rất phức tạp và đa dạng, trong đó có nhiều biến thể đặc biệt thể hiện sự đa dạng của biểu cảm khuôn mặt trong hình ảnh, chẳng hạn dạng điệu, tư thế góc nhìn, độ sáng tối,.... Đây là

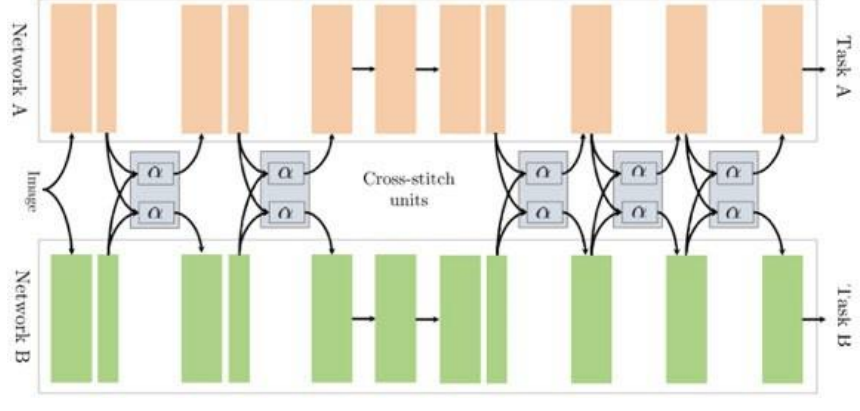
những thách thức lớn đối với các phương pháp truyền thống, vì vậy các phương pháp hiện đại dựa trên mô hình CNN được thiết kế trong các công trình nghiên cứu với độ chính xác cao về khả năng nhận dạng và có nhiều tiềm năng ứng dụng hơn. Gần đây, các mô hình CNN được thiết kế nhận dạng hình ảnh với các kiểu kiến trúc phức tạp như VGG, ResNet, SENet hay MobileNet [6], [9] và có xu hướng ngày càng sâu hơn.

2.2. Mạng nơron tích chập đa nhiệm

Mạng nơron tích chập đa nhiệm (Multi-Task CNN - MTCNN) là kiểu mô hình CNN học sâu hiệu quả trong việc cải thiện chất lượng cho mục tiêu của một nhiệm vụ với sự trợ giúp của một số nhiệm vụ có liên quan. Mô hình MTCNN thực hiện chia sẻ tham số để tìm kiếm các biểu diễn đặc điểm chung của các bài toán cần giải quyết trong các lớp tích chập ở mức sâu. Có hai kiểu chia sẻ tham số mô hình trong MTCNN gồm chia sẻ cứng (hard-sharing) và chia sẻ mềm (soft-sharing). Chia sẻ cứng trong MTCNN là việc sử dụng một kiến trúc mạng xương sống chung để trích chọn đặc trưng cho các bài toán và phân lớp độc lập theo từng nhiệm vụ (Hình 2.2a). Chia sẻ mềm là sử dụng mỗi khối kiến trúc trích chọn đặc trưng cho riêng từng bài toán nhưng có liên kết chéo các lớp nơron giữa các khối này (Hình 2.2b). Các mô hình MTCNN được nghiên cứu và xây dựng đã thực nghiệm cho thấy có hiệu quả trong các nhiệm vụ thị giác máy tính khác nhau [9].



(a)



(b)

Hình 2.2. Hai kiểu chia sẻ tham số MTCNN

Ban và cộng sự [10] thiết kế MTCNN kiểu phân tầng với tầng 1 cho hai bài toán phân loại học (taxonomic assigment) và tầng thứ hai có sử dụng kết quả tầng 1 cho bài toán phân vùng gen (genomic region assigment). Mô hình này dựa trên kiến trúc VGG với độ sâu 11 lớp CONV. Kiểu mô hình MTCNN dạng phân tầng và có liên kết chéo giữa các lớp nơron (soft-sharing) cũng được phát triển cho bài toán phát hiện các loại phương tiện hàng hải [11]. Mô hình này sử dụng các lớp tích chập lõi chung để trích xuất đặc trưng dựa trên kiến trúc mạng VGG với độ sâu 16 lớp CONV.

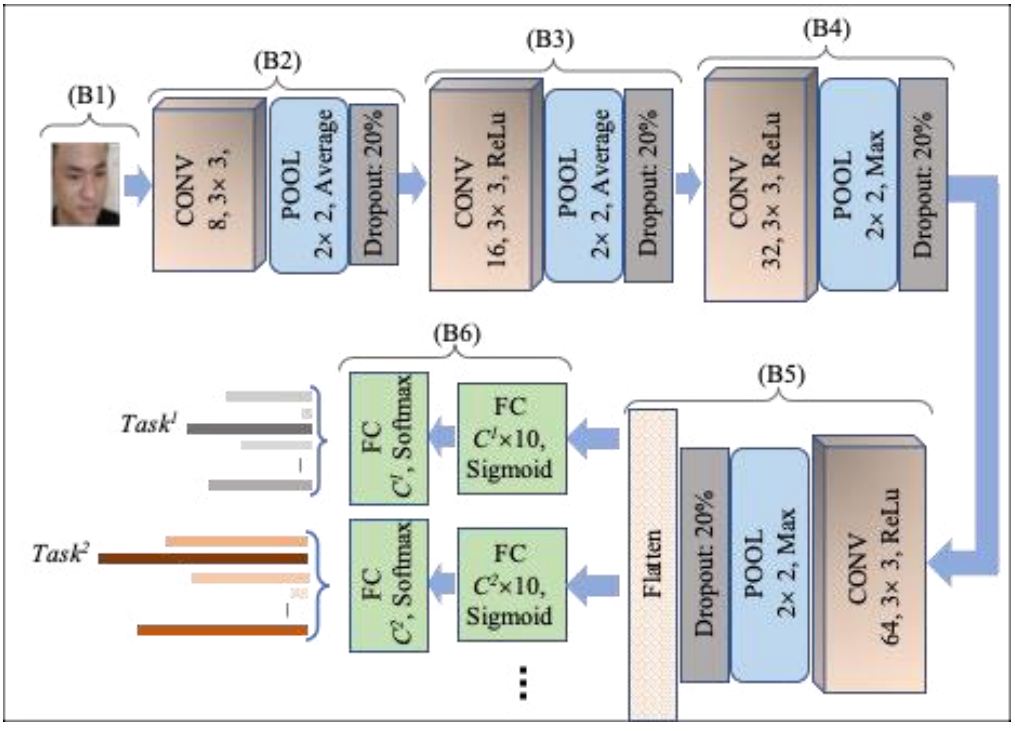
Cuong và cộng sự [12] thiết kế mô hình MTCNN có 9 lớp tích chập (CONV) và 3 lớp phân loại (FC) theo kiểu chia sẻ tham số và đặc trưng dạng "hard-sharing" để thực hiện phát hiện giới tính, trạng thái cười và biểu cảm trên khuôn mặt. Wang và công sự [9] đã thiết kế mô hình CNN đa nhiệm và đa nhãn cũng theo kiểu "hard-sharing" dựa trên kiến trúc ResNet50 cho bài toán nhận dạng các thuộc tính trên ảnh khuôn mặt như trạng thái đeo kính, đội mũ, hay để tóc mái hoặc mỉn cười, mũi nhọn hoặc môi to.

III. Mô hình MTCNN nhận dạng khuôn mặt và biểu cảm

3.1. Kiến trúc mô hình CNN đa nhiệm

Trong phần này, chúng tôi thiết kế mô hình MTCNN (gọi tắt là mô hình MFER) để thực hiện 2 nhiệm vụ cùng lúc gồm nhận dạng định danh khuôn mặt (FR) và nhận dạng biểu cảm khuôn mặt (FER). Mô hình MFER này được chia thành hai giai đoạn chính (Hình 3.1) bao gồm: (1) các đặc trưng hình ảnh được trích xuất biểu thị cho các định danh bằng khuôn mặt và biểu cảm trên khuôn mặt; và (2) phân loại các đặc trưng thành các nhãn phân lớp tương ứng với mỗi bài toán thực hiện. Số lớp và độ lớn (số lượng nơron) của mỗi lớp ảnh hưởng đến chất lượng của mô hình và độ phức tạp trong tính toán. Các nghiên cứu thường điều chỉnh hai yếu tố này theo từng bài toán ứng dụng để đạt được chất lượng mong đợi và độ phức tạp tính toán có thể chấp nhận được cùng một lúc. Vì vậy, chúng tôi thiết kế mô hình này với số lượng lớp vừa phải để phù hợp với hệ thống tính toán của chúng tôi.

Kiến trúc của mô hình MFER này sử dụng phương pháp chia sẻ tham số dạng "hard-sharing" nhằm giảm kích thước và độ phức tạp của mô hình cho việc tích hợp vào các ứng dụng có điều kiện tính toán hạn chế. Khối lõi của mô hình MFER dựa trên kiến trúc VGG để thực hiện trích chọn các đặc trưng cho các bài toán cần thực hiện, tuy nhiên, để giảm kích thước mô hình chúng tôi thiết kế số lớp tích chập (CONV) là 4, sau mỗi hai lớp tích chập đầu sử dụng lớp kết gộp tín hiệu đặc trưng bằng phép trung bình (Average POOL) và sau mỗi hai lớp tích chập cuối sử dụng phép gộp tín hiệu ở dạng lớn nhất (Max POOL). Chia các lớp nơron này thành 4 khối gồm (B2), (B3), (B4) và (B5) có cấu trúc cơ bản như nhau, mỗi khối này có một lớp tích chập (CONV) theo sau là lớp gộp tín hiệu (POOL). Khối (B1) là ảnh đầu vào, để giảm kích thước tham số mô hình và phù hợp với ảnh thu thập từ camera của thiết bị đầu cuối thông dụng có độ phân giải ở mức vừa phải chúng tôi đặt kích thước ảnh đầu vào là H(cao) × W(rộng) × D(sâu) = 80 × 60 × 3.



Hình 3.1. Mô hình MFER

Các bộ lọc của nơron ở lớp CONV có kích thước là 3×3, ở lớp POOL có kích thước là 2×2. Các nơron tích chập sử dụng hàm kích hoạt dạng "ReLu" thông dụng nhằm cho phép kích hoạt thưa ở mức khoảng 50% được kích hoạt ở đầu ra khi tổng tín hiệu đầu là dương, giảm thiểu khả năng suy biến gradient trong quá trình học, tính toán đơn giản và tăng tốc độ huấn luyện cho mô hình. Để giảm thiểu hiện tượng quá khớp (overfitting) trong học máy chúng tôi sử dụng kỹ thuật loại bỏ ngẫu nhiên kết nối của các nơron (tức là đầu ra của nơron được loại bỏ là bằng 0) theo tỷ lệ 20% (Dropout = 0.2). Số lượng các bộ lọc (filter) trong mỗi lớp nơron CONV được tăng dần theo chiều sâu từ 8, 16, 32 và 64 nhằm tăng thêm cơ hội trích chọn được nhiều hơn các đặc trưng ẩn sâu bên trong hình ảnh ở các lớp nơron tích chập ở mức sâu hơn.

Khối (B6) dùng để phân loại ảnh đầu vào đến các lớp theo của bài toán. Khối này có 2 lớp nơron kết nối đầy đủ (FC) cho mỗi bài toán cần thực hiện, lớp FC ẩn sử dụng hàm kích hoạt phi tuyến dạng "sigmoid" và lớp FC ra có kích hoạt bằng hàm "softmax" (công thức (3.2)) để tính xác suất thuộc từng lớp cho mỗi hình ảnh đầu vào. Khối (B5) có thêm cơ chế trải tín hiệu đặc trưng về dạng phẳng để truyền tín hiệu đặc trưng theo kết nối đầy đủ đến khối phân loại (B6). Chúng tôi áp dụng mô hình MFER này cho hai bài toán nhận dạng định danh khuôn mặt (FR) và nhận dạng biểu cảm khuôn mặt (FER). Để tăng khả năng phân loại và nhận dạng, số nơron của lớp FC ẩn được tăng thêm 10 lần so với số nơron ở lớp FC ra, tức là bằng 10 lần số lớp cần nhận dạng của bài toán. Như vậy, số nơron lớp FC ra và lớp FC ẩn của khối (B6) cho bài toán FR tương ứng là số người cần định danh (C^1) và $C^1 \times 10$, cho bài toán FER tương ứng là số loại biểu cảm trên khuôn mặt (C^2) và $C^1 \times 10$. Công thức tính đầu

ra nơron phân lớp theo hàm kích hoạt 'softmax' có dạng:

$$O_j^t = softmax(y_j^t) = \frac{e^{y_j^t}}{\sum_{k=1}^{M^t} e^{y_k^t}} \quad (3.2)$$

trong đó, O_j^t là đầu ra của nơron thứ j^{th} của lớp ra tương ứng với nhiệm vụ $t \in \{1,2,\dots,T\}$, y_j^t, là tổng tín hiệu đầu vào của nơron thứ j^{th} trong lớp phân loại tương ứng nhiệm vụ t, $M^t \in \{C^1, C^2\}$ là số nơron lớp ra của nhiệm vụ t. Ở đây, rõ ràng tổng các giá trị đầu ra của các nơron lớp ra thuộc khối phân lớp (B6) cho mỗi bài toán bằng 1, $\sum_j O_j^t = 1$.

Trên cơ sở xác suất được tính ở mỗi nơron lớp ra của nhiệm vụ t, chúng ta chọn lớp có xác suất cao nhất tương ứng để phân loại $\Im^t$ cho nhiệm vụ t tương ứng theo công thức (3.3).

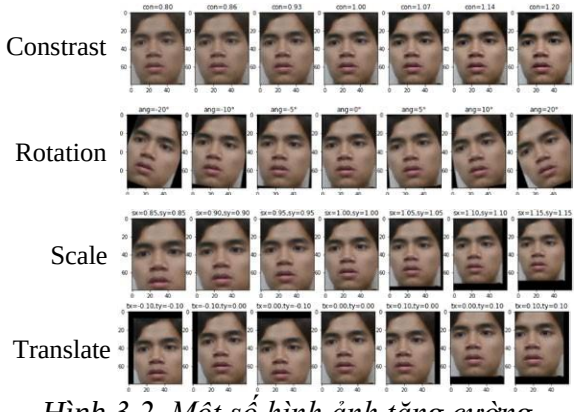
$$\Im^t = argmax_{C_j^t}\{O_j^t : j = 1,\dots,\mathrm{M}^t\} \quad (3.3)$$

3.2. Tăng cường dữ liệu và *huấn luyện mô hình MTCNN*

Phần này áp dụng một số phương pháp tiền xử lý hình ảnh đầu vào gồm dò tìm và cắt ảnh để lấy vùng ảnh chỉ chứa khuôn mặt, sau đó thực hiện một số kỹ thuật nâng cao chất lượng ảnh. Trong các ứng dụng thực tế, hình ảnh đầu vào thường được chụp từ máy ảnh, chúng bao gồm nền với bất kỳ vật thể nào bên trong ảnh. Vì vậy, chúng ta phải thực hiện phương pháp phát hiện khuôn mặt (Face detection - FD) để xác định vùng ảnh có chứa khuôn mặt và sau đó cắt bỏ phần nền của ảnh và chỉ giữ lại vùng ảnh chứa khuôn mặt. Để thực hiện điều này, chúng tôi sử dụng một mô hình dựa trên CNN nổi tiếng được gọi là MTCNN như trong [1]. Để tránh hiện tượng quá khớp trong huấn luyện mô hình và giúp cho mô hình có khả năng nhận dạng cao hơn, chúng tôi tăng cường hình ảnh huấn luyện bằng cách sử dụng một số kỹ thuật xử lý hình ảnh 2D như thêm nhiễu, xoay, cắt và dịch chuyển, tăng cường độ sáng hoặc làm tối hình ảnh. Với hình ảnh đầu vào , chúng ta nhận được danh sách các hình ảnh sau khi tiền xử lý như sau:

$$\{\Im^\alpha(f^D(a), p^\alpha)\} \quad (3.4)$$

trong đó, f^D là bộ dò tìm và phát hiện khuôn mặt trên ảnh, chẳng hạn MTCNN, p^α là các tham số cho hoạt động tăng cường hình ảnh với một phép xử lý α = *{nhiễu, xoay, co giãn, dịch chuyển, độ tương phản,...}*, $\Im^\alpha$ biểu thị sự biến đổi của hình ảnh đối với phép xử lý tăng cường α. Chẳng hạn, bằng cách áp dụng các phép xử lý gồm tăng giảm độ tương phản (Constrast), xoay ảnh (Rotation) theo góc sang trái (dương) và sang phải (âm), co giãn (Scale) và dịch chuyển (Translate) theo một tỷ lệ so với kích thước ảnh ta có kết quả như Hình 3.2.



Hình 3.2. Một số hình ảnh tăng cường

Trong Hình 3.2, ảnh trái cùng của dòng thứ nhất và ảnh chính giữa của các dòng sau là ảnh gốc ban đầu. Các hình ảnh còn lại là kết quả xử lý biến đổi tương ứng với từng dòng và giá trị tham số biến đổi

được ghi trên tiêu đề mỗi hình ảnh. Các tham số biến đổi này được lựa chọn ở mức độ vừa phải để đảm bảo những thông tin chính trên ảnh được duy trì cho việc trích chọn đặc trưng cho bài toán. Chẳng hạn ảnh phải cùng ở dòng đầu có nhiễu lớn nên rất khó để trích chọn đặc trưng và nhận dạng, kể cả bằng mắt thường. Một hình ảnh tăng cường có thể áp dụng cùng lúc đồng thời các phép xử lý và trong nghiên cứu này chúng tôi áp dụng ngẫu nhiên các giá trị tham số điều chỉnh của các phép xử lý.

Trong mô hình MFER này, chúng tôi áp dụng hàm đánh giá sai số dạng cross-entropy. Sai số theo mỗi nhiệm vụ của MFER được đánh giá riêng biệt, sau đó kết hợp chúng lại thông qua một hệ số để đưa ra sai số cuối cùng của mô hình. Hàm sai số của nhiệm vụ được biểu diễn bằng công thức sau:

$$\mathcal{L}_t = -\frac{1}{N}\sum_{i=1}^{N}\left(\alpha_i^t \sum_{j=1}^{M^t} y_i^t(j)\log(\tilde{y}_i^t(j))\right) \quad (3.5)$$

trong đó, N là số mẫu dữ liệu huấn luyện, $\alpha_i^t \in \{0,1\}$ cho biết mẫu dữ liệu thứ i^{th} xác định cho công việc t nếu bằng 1 và ngược lại là bằng 0, M^t là số lớp (class) của nhiệm vụ t, $y_i^t(j) \in \{0,1\}$ cho biết mẫu dữ liệu thứ i^{th} có thuộc nhãn phân lớp thứ j^{th} nếu bằng 1 và ngược lại bằng 0, $\tilde{y}_i^t(j) \in [0,1]$ là xác suất nhận dạng vào lớp thứ j^{th} của mô hình đối với mẫu dữ liệu thứ i^{th} ở nhiệm vụ T. Trong nghiên cứu này, chúng tôi áp dụng các tập dữ liệu với mẫu dữ liệu được xác định đủ cho cả đồng thời hai nhiệm vụ FR và FER, tức $\alpha_i^t = 1$. Như vậy, hàm sai số chung của mô hình đối với các nhiệm vụ được xác định như sau:

$$\mathcal{L}^* = \sum_{t=1}^{T} w_t \mathcal{L}_t \quad (3.6)$$

trong đó, T là số lượng các nhiệm vụ cần thực thi của mô hình, w_t là hệ số đánh giá vào hàm sai số chung $\mathcal{L}^*$ đối với nhiệm vụ t. Nghiên cứu này áp dụng với $T=2$ gồm bài toán FR và bài toán FER.

Mô hình MFER được huấn luyện theo phương pháp tối ưu hoá Adam [13], đây là một kỹ thuật tối ưu hóa được sử dụng rộng rãi trên cơ sở kết hợp những điểm mạnh của phương pháp Momentum và RMSprop bằng cách sử dụng các giá trị bình phương gradient để chia tỷ lệ học mạng theo kỹ thuật RMSprop và sử dụng trung bình động của các bước thay đổi gradient. Chi phí bộ nhớ hiệu quả hơn và giảm thiểu tính toán là hai lợi thế của phương pháp Adam. Cơ chế điều chỉnh trọng số mạng để tìm điểm tối ưu của Adam được thể hiện trong công thức (3.7) như sau:

$$w_{ij}{}^t = w_{ij}{}^{t-1} - \frac{\eta}{\sqrt{\hat{v}^t}+\epsilon} * \hat{m}^t \quad (3.7)$$

trong đó, $\hat{m}^t$ và $\hat{v}^t$ tương ứng là giá trị trung bình suy giảm theo cấp số nhân của các gradient và của các gradient bình phương tại thời điểm học thứ t, η là hệ số học (tốc độ huấn luyện, thường sử dụng là 10^{-3}), hệ số $\epsilon = 10^{-8}$. Chúng tôi sử dụng kỹ thuật học Adam cho MFER và tham số của mô hình được khởi tạo ngẫu nhiên theo phân phối đều trong khoảng giới hạn (công thức (3.8)) [14] nhằm đem lại các ưu điểm của quá trình huấn luyện mạng và đạt kết quả cao.

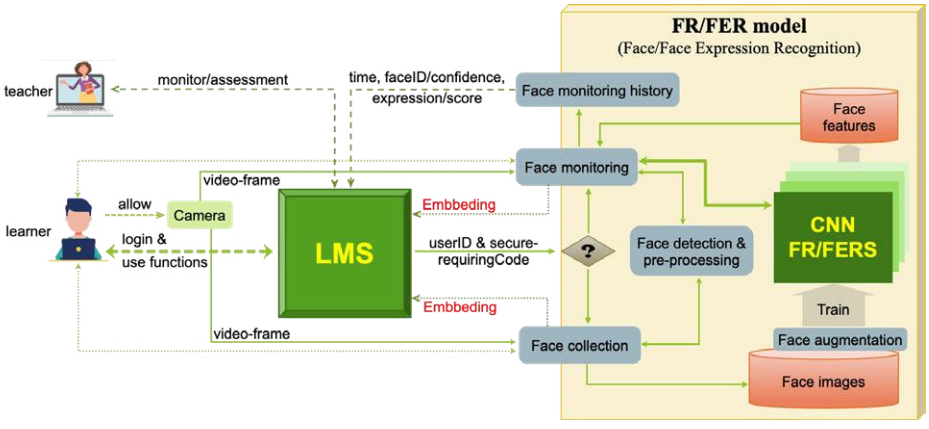
$$W \sim U\left[-\frac{\sqrt{6}}{\sqrt{n_j+n_{j+1}}}, \frac{\sqrt{6}}{\sqrt{n_j+n_{j+1}}}\right] \quad (3.8)$$

trong đó, n_j và n_{j+1} là số các tham số vào và số các tham số ra của lớp nơron.

3.3. Thiết kế hệ thống tích hợp hỗ trợ giám sát và đánh giá học trực tuyến

Hệ thống tích hợp cần phải thực hiện được hai chức năng cơ bản gồm: (1) chụp ảnh khuôn mặt của người dùng, tiền xử lý hình ảnh gồm phát hiện khu vực chứa khuôn mặt trên ảnh và nâng cao chất lượng của hình ảnh

nếu cần; và (2) nhận dạng hình ảnh khuôn mặt để định danh và xác định biểu cảm biểu thị trên khuôn mặt. Tính năng chụp ảnh có thể được áp dụng cho việc thu thập hình ảnh khuôn mặt trong giai đoạn huấn luyện xây dựng mô hình hoặc áp dụng khi mô hình được ứng dụng tích hợp. Sơ đồ kết nối tổng thể về mặt logic của việc tích hợp mô hình nhận dạng khuôn mặt và biểu cảm (FR/FER) với hệ thống LMS được thể hiện trong Hình 3.4.



Hình 3.4. Kết nối giữa mô hình FR/FER và hệ thống LMS

Mô hình FR/FER tích hợp với LMS thực hiện trong điều kiện kết nối mở và hoạt động độc lập tương đối giữa hai hệ thống. Khi đó, hai hệ thống có thực thi một chức năng độc lập, chẳng hạn LMS thực hiện các nghiệp vụ và tính năng quản lý học tập của người học như bản thân có của nó, hệ thống FR/FER thực hiện thu thập hình ảnh hoặc thậm chí nhận dạng cũng độc lập với LMS. Do đó, việc tích hợp ở cơ chế lỏng và đơn giản là nhúng vào LMS các tính năng có liên quan của hệ thống nhận dạng để kích hoạt khi cần thiết và trả lại kết quả thực hiện tương ứng. Các giao tiếp giữa hai hệ thống có thể thực hiện qua môi trường Internet với kênh truyền bảo mật có xác thực. Điều này nhằm tận dụng tối đa hệ thống sau này khi cần kết nối với các hệ thống khác cũng như bố trí các tài nguyên máy chủ chuyên dụng phục vụ cho các chương trình học máy để xây dựng mô hình nhận dạng (như sử dụng máy chủ có GPU). Cơ chế kết nối này cho phép chúng ta đạt được một số đặc điểm sau đây của hệ thống tích hợp:

- Kết nối hệ thống FR/FER với hệ thống LMS thông qua web services hoặc nhúng vào giao diện của LMS nhưng đảm bảo không làm thay đổi nhiều đối với LMS đang có;

- Hệ thống LMS có thể vận hành mà không cần kết nối với hệ thống nhận dạng trong trường hợp người học không có thiết bị thu nhận hình ảnh hoặc chủ động không bật chế độ nhận dạng;

- Đảm bảo tính an ninh, bảo mật của cả hai hệ thống và của cả dữ liệu hình ảnh của người dùng;

- Tối thiểu hoá quá trình trao đổi tương tác giữa hai hệ thống, đặc biệt là dữ liệu hình ảnh giữa máy trạm của người học và máy chủ hệ thống LMS hoặc hệ thống nhận dạng.

Cơ chế bảo mật được sử dụng trong tích hợp này nhằm đảm bảo an toàn và

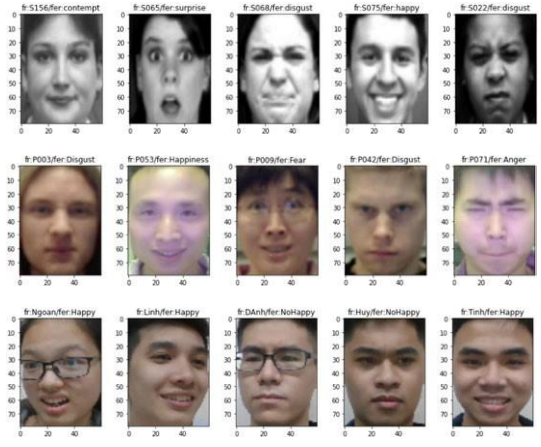
chính danh cho kết nối giữa hai hệ thống, theo đó, hệ thống FR/FER chỉ thực hiện các nhiệm vụ khi nhận được thông điệp kèm theo mã bảo mật đã được thiết lập. Trong trường hợp này, chúng tôi sử dụng mã bảo mật dựa trên sơ đồ mã hoá khoá công khai, tức là sử dụng cặp khoá bất đối xứng của RSA cho hệ thống, phần khoá công khai sử dụng ở hệ thống FR/FER và dĩ nhiên có thể được biết bởi bất kỳ ai (do tính công khai), phần khoá bí mật được sử dụng tại LMS. Do đó, hệ thống FR/FER thực hiện kiểm chứng dựa trên nguyên tắc chữ ký số, thông tin định danh người học đã ký số bởi phần khoá bí mật tại LMS và được xác thực số bởi phần khoá công khai tại FR/FER cho việc thực hiện chức năng nhận dạng và trả về kết quả cho LMS.

IV. Thử nghiệm

4.1. Dữ liệu và kịch bản thử nghiệm

Chúng tôi sử dụng ba bộ dữ liệu để thử nghiệm đánh giá mô hình gồm CK+ (Extended Cohn-Kanade) [15], Oulu-CASIA [7] và các hình ảnh được thu thập từ người học của chúng tôi, ký hiệu là FERS21 [16].

Tập dữ liệu CK+ gồm 327 video được gắn nhãn được thu thập từ 118 người khác nhau. Chúng tôi sử dụng video với bảy biểu cảm cơ bản đại diện cho sự tức giận (Anger), ghê tởm (Disgust), sợ hãi (Fear), hạnh phúc (Happy), buồn bã (Sadness), ngạc nhiên (Surprise) và sự khinh thường (Contempt). Các khung hình trong video biểu thị sự biến đổi trạng thái biểu cảm, tuy nhiên, chúng tôi chọn ba khung hình cuối cùng trong mỗi video để làm dữ liệu thử nghiệm, kết quả có tổng cộng 981 hình ảnh. Tập dữ liệu này là hình ảnh có màu đa cấp xám (Hình 4.1, dòng đầu), tiêu đề của hình ảnh được hiển thị là nhãn tương ứng của người trong ảnh (fr) và biểu cảm trên khuôn mặt (fer) của hình ảnh trong tập dữ liệu.



Hình 4.1. Một số ảnh trong các tập dữ liệu

Tập dữ liệu Oulu-CASIA bao gồm các video được thu thập trong các điều kiện ánh sáng khác nhau. Trong thử nghiệm này, chúng tôi sử dụng 480 video, được chụp từ 80 đối tượng trong điều kiện ánh sáng trung bình và cao. Có sáu nhãn biểu cảm trong dữ liệu Oulu-CASIA như trong tập dữ liệu CK+ trừ biểu cảm sự khinh thường (Contempt). Đối với mỗi video, chúng tôi chọn ba khung hình cuối cùng có khuôn mặt thể hiện biểu cảm cao nhất của loại tương ứng, Hình 4.1 (dòng giữa) cho thấy một số hình ảnh của tập dữ liệu Oulu-CASIA. Tập dữ liệu thử nghiệm có tổng cộng 1440 hình ảnh.

Bộ dữ liệu FERS21 được thu thập từ 20 người học của chúng tôi, cả nam và nữ. Để đơn giản trong việc thu thập và ứng dụng trong đánh giá quá trình học tập về mức độ hài lòng hoặc không hài lòng, chúng tôi sử dụng hai biểu cảm trên khuôn mặt đó là hài lòng/vui vẻ (Happy) và không hài lòng/buồn (NoHappy) được minh họa trong Hình 4.1 (dòng cuối).

Để chạy thử nghiệm, chúng tôi chia ngẫu nhiên mỗi tập dữ liệu thành 5 phần (fold) có kích thước tương đương nhau trong các lớp của mỗi bài toán FR và FER. Kịch bản thử nghiệm theo cơ chế kiểm tra chéo (cross-validation), mỗi lượt chạy sử dụng một phần dữ liệu để kiểm tra kết quả mô hình (), bốn phần còn lại để xây dựng mô hình, trong đó một phần để thẩm định và lựa chọn mô hình () và 3 phần còn lại được sử dụng để huấn luyện mô hình (). Kịch bản này được chạy lặp lại 5 lần theo thứ tự lần lượt các phần được chọn để kiểm tra mô hình, kết quả đánh giá cuối cùng là trung bình và độ lệch của 5 lần chạy.

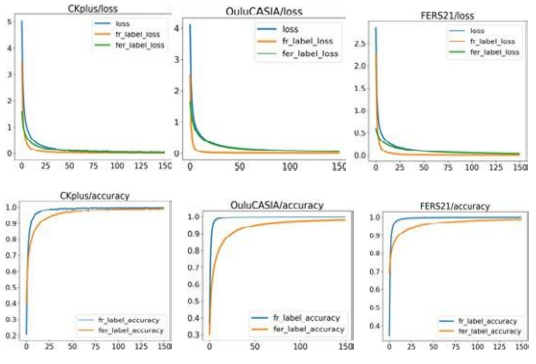
Trong mỗi lần chạy thử nghiệm, các phần dữ liệu huấn luyện mô hình () được tăng cường bằng cách áp dụng các phép biến đổi hình ảnh bao gồm . Các tham số cho mỗi phép biến đổi hình ảnh được chọn ngẫu nhiên trong khoảng giới hạn. Số lần được tăng cường là 20 cho mỗi hình ảnh tạo nên tập dữ liệu huấn luyện khá lớn nhằm đảm bảo độ đa dạng của dữ liệu, tránh bị hiện tượng quá khớp và kỳ vọng đạt được độ chính xác cao của mô hình. Chúng tôi sử dụng phương pháp tối ưu Adam [13] để huấn luyện mô hình với các tham số chi tiết trong Bảng 4.1.

Bảng 4.1. Các tham số chạy thử nghiệm

Stt	Tham số	Giá trị
1	Tốc độ huấn luyện ban đầu ()	
2	Kích thước gói dữ liệu (batch)	128
3	Số lần lặp huấn luyện	150

4.2. Kết quả thử nghiệm

Quá trình huấn luyện được tính trung bình trên 5 lần chạy thử nghiệm theo từng nhiệm vụ FR và FER thể hiện trong Hình 4.2. Các biểu đồ này cho thấy quá trình huấn luyện cho kết quả tốt hơn (sai số nhỏ hơn và độ chính xác cao hơn) của nhiệm vụ FR so với FER trên cả 3 tập dữ liệu. Mặc dù số lớp của nhiệm vụ FR (118, 80 và 20) lớn hơn nhiều so với FER (chỉ 6 hoặc 7) nhưng kết quả huấn luyện cao hơn, cho thấy mô hình MFER trích chọn đặc trưng cho phân biệt định danh khuôn mặt tốt hơn so với biểu cảm khuôn mặt. Hơn nữa, các hình ảnh khuôn mặt trong tập dữ liệu có sự phân biệt cao giữa những người được thu thập trong khi biểu cảm trên khuôn mặt là khó phân biệt hơn, thậm chí một số biểu cảm gần như giống nhau trên khuôn mặt.



Hình 4.2. Quá trình huấn luyện MTCNN

Kết quả nhận dạng trên tập dữ liệu kiểm tra của mô hình MFER sau khi đã được huấn luyện được tính toán trung bình và độ lệch chuẩn trên 5 lần chạy thử nghiệm thể hiện trong Bảng 4.2. Tương ứng với quá trình huấn luyện, kết quả nhận dạng đối với dữ liệu kiểm tra theo bài toán FER thấp hơn bài toán FR trên cả 3 tập dữ liệu, trong đó, tập dữ liệu CK+ không đáng kể nhưng tập dữ liệu OuluCASIA và FERS21 có chênh lệch tương ứng là 4.51% và 8.45%. Tương ứng độ lệch chuẩn của các kết quả nhận dạng theo bài toán FER cao hơn bài toán FR. Chứng tỏ bài toán FER là khó hơn so với bài toán FR, thực tế cũng cho thấy hình ảnh biểu cảm trên khuôn mặt có trường hợp rất khó phân biệt, đặc biệt là các loại biểu cảm như “Sadness”, “Anger” đã đề cập ở trên.

Bảng 4.2. Kết quả nhận dạng tập kiểm tra

Datasets	Trung bình (±độ lệch)	
	FR	FER
CK+	97.97 ±0.01	97.86 ±0.015
Oulu CASIA	99.16 ±0.0096	94.65 ±0.026
FERS21	99.66 ±0.004	±0.02

Để so sánh, chúng tôi sử dụng kiến trúc lõi trích chọn đặc trưng của mô hình MobileNetV2 [17] và mô hình ResNet50V2 [18], bổ sung thêm các lớp để phân loại cho hai bài toán FR và FER như mô hình MFER, sau đó chạy thử nghiệm trên cùng kịch bản và tham số. Kết quả thể hiện trong Bảng 4.3.

Bảng 4.3. So sánh kết qủa các mô hình

Datasets **Models (on FR/FER)**	**CK+**	**Oulu CASIA**	**FERS21**
MobileNetV2 /FR	69.57	87.26	91.38
FER	81.69	85.17	88.07
ResNet50V2 / FR	96.45	98.32	99.19
FER	96.58	96.32	89.54
MFER / FR	97.97	99.16	99.66
FER	97.86	94.65	91.21

Mặc dù số lượng tham số của cả hai mô hình lõi MobileNetV2 (khoảng 2,5 triệu) và ResNet50V2 (khoảng 24 triệu) lớn hơn rất nhiều so với MFER (khoảng 0,24 triệu) nhưng kết quả nhận dạng của MFER cao hơn khi so sánh. Cụ thể, ở bài toán FR, mô hình MFER đạt kết quả nhận dạng tốt nhất, mô hình lõi MobileNetV2 có kết quả thấp nhất trong cả ba tập dữ liệu thử nghiệm. So với ResNet50V2, trường hợp cao hơn nhiều nhất là 1.52% tại dữ liệu CK+ và cao hơn ít nhất là 0.47% tại dữ liệu FERS21. Ở bài toán FER, mô hình kiến trúc lõi ResNet50V2 có độ sâu lớn hơn và số lượng tham số rất lớn nhưng kết quả nhận dạng của mô hình này đạt cao nhất tại tập dữ liệu OuluCASIA, trong khi mô hình MFER đạt cao nhất ở dữ liệu CK+ và FERS21.

V. Kết luận

Nghiên cứu này đã đề xuất mô hình mạng nơron tích chập đa nhiệm (MFER) cho bài toán nhận dạng khuôn mặt để định danh và nhận dạng biểu cảm khuôn mặt. Kiến trúc của mô hình theo thuần kiến trúc của mạng CNN kiểu VGG và có độ sâu không lớn chỉ ở mức 4 lớp tích chập, kèm theo kích thước tham số của mô hình ở mức thấp. Kết quả nhận dạng rất khả quan trên các tập dữ liệu thử nghiệm, đạt mức thấp nhất là 91.21% đối với bài toán FER và cao nhất là 99.66% đối với bài toán FR, cả hai trường hợp đều ở dữ liệu FERS21. Điều này cho thấy mô hình MFER có thể áp dụng được dễ dàng trên các hệ thống có năng lực tính toán không đòi hỏi quá cao và phù hợp đa dạng trong thực tế nhưng vẫn cho kết quả tốt đối với các bài toán ứng dụng.

Chúng tôi cũng đã thiết kế hệ thống tích hợp mô hình MFER vào hệ thống quản lý học tập trực tuyến (LMS) để hỗ trợ giám sát người học trên các hệ thống LMS. Qua đó, mỗi người học được giám sát chi tiết quá trình học tập, được đo đếm biểu cảm thể hiện trong suốt quá trình học tập, nếu có những bất thường hệ thống có thể tổng hợp báo cáo người dạy, người quản lý và hỗ trợ để nhắc nhở, giúp đỡ người học đạt kết quả học tập cao hơn. Việc tích hợp hệ thống này theo có chế mở, không gắn chặt với nhau, do đó, hệ thống hoạt động khá độc lập và có thiết kế đảm bảo tính an toàn, an ninh của dữ liệu và hệ thống kết nối tích hợp.

Trong những nghiên cứu tiếp theo, chúng tôi sẽ cải tiến kiến trúc lõi các lớp tích chập để trích chọn đặc trưng của mô hình MFER theo các kiến trúc hiện đại nhằm tăng cường chất lượng cho các bài

toán nhận dạng khác nhau với số lượng nhiệm vụ có thể thực hiện nhiều hơn.

Tài liệu tham khảo:

[1]. Duong Thang Long, A Lightweight Face Recognition Model Using Convolutional Neural Network for Monitoring Students in E-Learning, *I.J. Modern Education and Computer Science*, vol.6, pp.16-28, 2020.

[2]. Francisco D. Guillén-Gámez, Facial authentication software for the identification and verification of students who use virtual learning platform (LMS), *Advances in Educational Technology and Psychology*, 1: 1-8 Clausius Scientific Press, Canada, 2017.

[3]. Ayham Fayyoumi1 and Anis Zarrad, Novel Solution Based on Face Recognition to Address Identity Theft and Cheating in Online Examination Systems, *Advances in Internet of Things*, 2014, 4, 5-12.

[4]. Ekberjan Derman1 and Albert Ali Salah, *Continuous Real-Time Vehicle Driver Authentication Using Convolutional Neural Network Based Face Recognition*, 3th IEEE International Conference on Automatic Face & Gesture Recognition (FG), 2018.

[5]. Shan Li and Weihong Deng, *Deep Facial Expression Recognition - A Survey*, IEEE Transactions on Affective Computing, 2020.

[6]. Vijayan K. Asari and et al., *A State-of-the-Art Survey on Deep Learning Theory and Architectures*, Electronics, 8, 292, 2019.

[7]. G. Zhao, X. Huang, and et al., *Facial expression recognition from near-infrared videos*, Image and Vision Computing, 29(9):607–619, 2011.

[8]. I.Michael Revina and W.R. Sam Emmanuel, *A Survey on Human Face Expression Recognition Techniques*, https://doi.org/10.1016/ j.jksuci.2018.09.002, 2018.

[9]. Hanzi Wang et al., *Deep Multi-task Multi-label CNN for Effective Facial Attribute Classification*, https://arxiv.org/abs/2002.03683, 2020.

[10]. Haoran Ma et al., *A multi- task CNN learning model for taxonomic assignment of human viruses*, BMC Bioinformatics 22:194, 2021.

[11]. Zhaoying Liu et al., *A Multi-Task CNN for Maritime Target Detection*, IEEE Signal Processing Letters, Vol. 28, 2021.

[12]. Dinh Viet Sang and Le Tran Bao Cuong, *Effective Deep Multi-source Multi-task Learning Frameworks for Smile Detection*, Emotion Recognition and Gender Classification, Informatica, vol.42, pp.345–356, 2018.

[13]. Diederik P. Kingma and Jimmy Lei Ba, *Adam-A Method for Stochastic Optimization*, Published as a conference paper at ICLR 2015.

[14]. Xavier Glorot and Yoshua Bengio, *Understanding the difficulty of training deep feedforward neural networks*, Proceedings of the 13th International Conference on Artificial Intelligence and Statistics (AISTATS), 2010.

[15]. Patrick Lucey et al., *The Extended Cohn-Kanade Dataset (CK+ dataset)*, IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Workshops, 2010.

[16]. Duong Thang Long, *A Facial Expressions Recognition Method Using Residual Network Architecture for Online Learning Evaluation*, Journal of Advanced Computational Intelligence Informatics, Vol.25 No.6, 2021.

[17]. Chunrui Han and et al., *Face Recognition with Contrastive Convolution*, ECCV, DOI:10.1007/978-3-030-01240-3_8, 2018.

[18]. Kaiming He et al., *Identity Mappings in Deep Residual Networks*, arXiv:1603.05027v3 [cs.CV] 25 Jul 2016.

Địa chỉ tác giả: Trường Đại học Mở Hà Nội
Email: duongthanglong@hou.edu.vn