**NHẬN DIỆN CẢM XÚC CỦA SINH VIÊN DỰA TRÊN BIỂU CẢM KHUÔN MẶT TRONG GIÁO DỤC TRỰC TUYẾN**

Phạm Quang Huy

Khoa Công Nghệ Thông Tin, Trường Đại học Sư Phạm Hà Nội

**Tóm Tắt**

Giáo dục trực tuyến đã phát triển nhanh chóng do sự tiện lợi không thể thay thế của nó, những tiến bộ về công nghệ trong giáo dục trực tuyến mang đến cơ hội phát triển học tập cho sinh viên trong hoàn cảnh đặc biệt nghiêm trọng do dịch bệnh COVID-19 gây nên.

Hầu hết cách trường học trên khắp thế giới đã trì hoãn việc mở cửa và áp dụng giáo dục trực tuyến như một trong những hình thức phương pháp giảng dạy chính để đảm bảo sự an toàn cho sinh viên trước sự nguy hiểm mà dịch bệnh mang lại (4.777.503 trường hợp tử vong , báo cáo của WHO. Tính đến ngày 3 tháng 10 năm 2021). Tuy nhiên, hiệu quả của các lớp học trực tuyến từ lâu đã đặt ra nhiều dấu chấm hỏi. So với lớp học truyền thống, là sự thiếu phản hồi trực tiếp, kịp thời và hiệu quả giữa giáo viên và học sinh trong các lớp học trực tuyến.

Trước các nghiên cứu đã chỉ ra rằng, có một mối quan hệ chặt chẽ giữa nét mặt và cảm xúc của một người nói chung. Từ góc độ mô phỏng máy tính, kết hợp thuật toán nhận dạng biểu cảm khuôn mặt, từ đó đề xuất giải pháp nhận diện cảm xúc của sinh viên dựa trên biểu cảm khuôn mặt trong giáo dục trực tuyến. Máy ảnh điện tử trong các thiết bị là thuận tiện cho việc sử dụng để thu thập hình ảnh khuôn mặt sinh viên. Các biểu cảm được phân tích và phân loại thành tám loại cảm xúc khác nhau bằng thuật toán nhận diện cảm xúc.

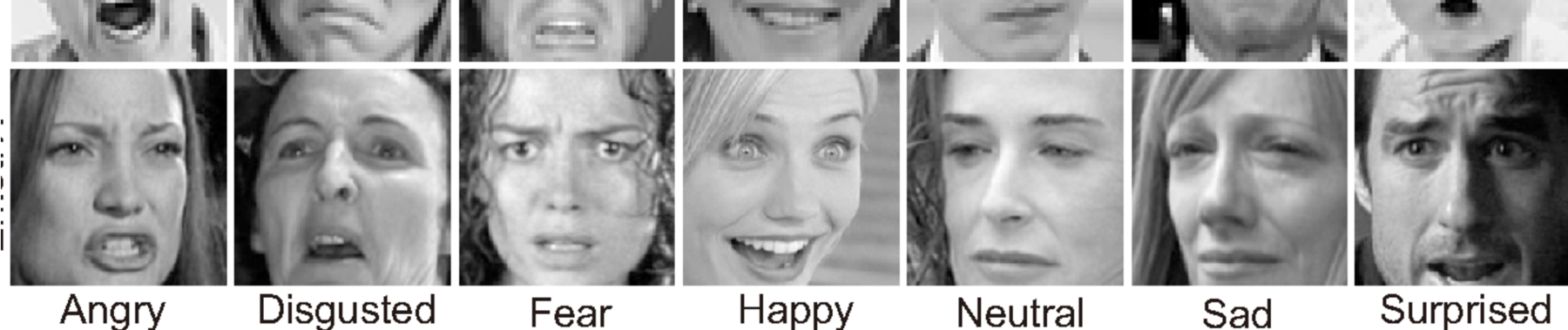
Phương pháp được đề xuất hoạt động hiệu quả trong những cuộc thử nghiệm trong các lớp học trực tuyến tại khoa công nghệ thông tin trường đại học sư phạm Hà Nội. Đây có thể coi như là một bộ khung có thể áp dụng linh hoạt trong các tình huống tương tự tại các lớp học trực tuyến.

1. **Giới Thiệu**

Trên bất kỳ quốc gia nào, không kể chủng tộc và giới tính, thì biểu cảm trên khuôn mặt là một trong những tín hiệu mạnh mẽ, tự nhiên và phổ biến nhất để con người truyền tải trạng thái cảm xúc và ý nghĩ của họ­ [1], [2], có rất nhiều ứng dụng liên quan đến vấn đề này như: quản lý sức khỏe [3], hỗ trợ lái xe và nhiều vấn đề khác [4].

Ekam và Friese đã chỉ ra rằng: con người nhận thức được một số cảm xúc cơ bản theo cùng một cách trên bất kể nền tảng văn hóa hay quốc gia nào và họ đã xác định các biểu hiện cảm xúc điển hình trên khuôn mặt gồm sáu loại: tức giận, ghê tởm, sợ hãi, hạnh phúc, buồn bã và ngạc nhiên. Dựa trên các nghiên cứu khác của Ekman và Heider, đã chứng minh rằng một biểu hiện cảm xúc phổ biến khác đó là: khinh bỉ [5].

Ngoài ra, FER 2013: một bộ cơ sở dữ liệu quy mô lớn được giới thiệu trong IMCL 2013, các hình ảnh được gắn nhãn biểu cảm trên khuôn mặt bao gồm: giận dữ, ghê tởm, sợ hãi, hạnh phúc, buồn bã, ngạc nhiên và bình thường. Trong các nghiên cứu tiếp theo, Ekam đã giới thiệu nhiều loại mô hình khác nhau để cung cấp nhiều loại cảm xúc hơn vì trên thực tế, những cảm xúc cơ bản chiếm một phần khá nhỏ trong các biểu hiện cảm xúc hàng ngày của chúng ta.[6] Trong đó, phân loại cảm xúc cơ bản một cách riêng biệt vẫn là phương pháp được sử dụng phổ biến nhất trong nhận diện cảm xúc do tính tổng quát và định nghĩa trực quan của biểu cảm khuôn mặt [7] . Hình 1 là mô tả bảy kiểu biểu cảm khuôn mặt cơ bản từ bộ cơ sở dữ liệu FER2013.



Hình 1: Mô tả 7 loại cảm xúc phổ biến từ FER2013

Với sự phát triển của trí tuệ nhân tạo và học sâu, nhiều thuật toán nhận diện cảm xúc được đề xuất để xử lý thông tin trong các biểu cảm được biểu diễn trên khuôn mặt, điều này đã cải thiện dần độ chính xác của nhận dạng và đạt được hiệu xuất tốt hơn so với các phương pháp truyền thống. Các hình ảnh trong bài toán nhận diện nói chung được chia ra là hai lại: hình ảnh tĩnh(biểu thị bằng hình ảnh) và hình ảnh động (biểu thị bằng video)[8] có tính chất thay đổi liên tục giữa các hình ảnh, do đó đặt ra nhiều thách thức hơn so với hình ảnh trước đây. Ngoài ra, các phương pháp dựa trên thị giác và sinh trắc học khác cũng có thể được áp dụng trong việc nhận diện cảm xúc khuân mặt.

Với cơ sở dữ liệu đào tạo được dán nhãn đầy đủ bao gồm nhiều biểu cảm khuân mặt có thể sẽ là điều quan trọng đối với các nhà nghiên cứu để thiết kế và thử nghiệm mô hình hoặc hệ thống FER. Các cơ sở dữ liệu hiện có chủ yếu được kiểm soát, tiêu biểu như bộ dữ liệu CK Plus 48[9], được tạo ra trong phòng thí nghiệm với đầy đủ ánh sáng và nền. Ngày nay, bởi vì hầu hết các ảnh thực thế đều phức tạp hơn và có thể thay đổi do các yếu tố như ánh sáng, FER trong phòng thí nghiệm hoặc môi trường được kiểm soát thường được coi là ít có ý nghĩa thực tế và được sử dụng chủ yếu để chứng minh khái niệm về các phương pháp khai thác và phân loại tính năng. Mặt khác, các cơ sở dữ liệu không được kiểm soát, chẳng hạn như FER2013[10] được thu thập từ các môi trường phức tạp với nền tảng, suy biến và độ chiếu sáng rất khác nhau, những ảnh này giống với tình huống thực tế hơn và ngày càng được sử dụng nhiều hơn trong các nghiên cứu.

Bị hạn chế bởi phần cứng và khả năng xử lý không đủ, phần lớn các phương pháp truyền thống cho FER sử dụng các tính năng thủ công hoặc học nông, chẳng hạn như các mẫu nhị phân cục bộ (LBP)[8] và phân tích nhân tử ma trận không âm (NMF)[11]. Với sự phát triển của khả năng xử lý và mô phỏng máy tính, tất cả các loại thuật toán học máy, chẳng hạn như Mạng thần kinh nhân tạo (ANN), Máy vectơ hỗ trợ (SVM) và bộ phân loại Bayes, đã được áp dụng cho FER với độ chính xác cao đã được chứng minh trong môi trường được kiểm soát để có thể phát hiện khuôn mặt một cách hiệu quả. Tuy nhiên, các phương pháp này yếu về khả năng khái quát hóa trong khi đây là chìa khóa để đánh giá tính thực tiễn của một mô hình[12]. Các thuật toán học sâu có thể giải quyết vấn đề này và nó cũng mạnh mẽ trong các môi trường không được kiểm soát. Các công trình gần đây đã chỉ ra tính hiệu quả của mạng nơ-ron tích tụ (CNN), vì tính hiệu quả của chúng trong các bài toán phân loại và trích xuất đối tượng, hoạt động tốt trong việc giải quyết các vấn đề về thị giác máy tính, đặc biệt là trong FER[13], và nhiều mô hình dựa trên cấu trúc CNN được đề xuất liên tục và đã đạt được kết quả tốt hơn các phương pháp trước đây. Simonyan và Zisserman[14] đã thông qua kiến ​​trúc của các bộ lọc tích chập rất nhỏ (3 × 3) để tiến hành đánh giá toàn diện các mạng với độ sâu ngày càng tăng và hai mô hình ConvNet hoạt động tốt nhất đã được công bố công khai để tạo điều kiện cho các nghiên cứu sâu hơn trong lĩnh vực này. Bằng cách tăng chiều sâu và chiều rộng của mạng trong khi vẫn giữ ngân sách tính toán không đổi, Szegedy et al.[15] đã giới thiệu một kiến ​​trúc mạng nơ-ron phức hợp sâu có tên là “Inception” trong đó việc sử dụng tài nguyên máy tính có thể được cải thiện đáng kể, và Jahandad et al.[16] đã làm việc trên 2 kiến ​​trúc mạng nơ-ron phức hợp (Inception-v1 và Inception-v3) dựa trên “Inception” và chứng minh rằng 2 mô hình này hoạt động tốt hơn các mô hình khác và Inception-v1 với mạng học sâu 22 lớp hoạt động tốt hơn mạng Inception-v3 với 42 lớp sau khi thực nghiệm với hình ảnh đầu vào có độ phân giải thấp và hình ảnh chữ ký 2D; tuy nhiên, Inception-v3 hoạt động tốt hơn với ImageNet. Xu hướng chung của mạng nơ-ron là tăng độ sâu của mạng và độ rộng của lớp. Về lý thuyết, các mô hình mạng nơ-ron càng sâu thì khả năng học càng mạnh nhưng càng khó đào tạo. Ông và cộng sự[17] đã đề xuất một mô hình mạng nơ-ron còn lại (Residual neural network) nhằm làm giảm độ khó đào tạo của các mạng sâu hơn và đã chứng minh kỹ lưỡng rằng các mạng RNN này dễ tối ưu hóa hơn trong khi tăng độ chính xác từ độ sâu tăng lên đáng kể. Ngoài ra, một bộ phận các nhà nghiên cứu đề xuất rằng độ chính xác của nhận dạng có thể được cải thiện hơn nữa bằng cách kết hợp CNN với RNN trong đó CNN được sử dụng làm đầu vào cho RNN.

Trong suốt những thập kỷ qua, giáo dục trực tuyến đã phát triển nhanh chóng dù là tại các trường đại học hay cơ sở đào tạo[18], mang lại cơ hội ứng dụng tiềm năng cho FER. Vấn đề khó khăn lớn giữa lớp học trực tuyến học trực tiếp truyền thống đó là các lớp học trực tuyến thường được coi là ít ràng buộc hơn và giao tiếp kém hiệu quả, chắc chắn sẽ dẫn đến sự nghi ngờ của giảng viên cũng như học sinh, sinh viên đối với phương pháp giáo dục mới lạ này trong khi có một số nghiên cứu tranh luận rằng kết quả học tập của sinh viên đạt được bằng giáo dục trực tuyến có thể tương đương với các lớp học truyền thống, ngoại trừ các kỹ năng đòi hỏi độ chính xác tối ưu và mức độ nhận thức xúc giác cao hơn[19]. Không thể phủ nhận rằng tốc độ phát triển nhanh chóng của giáo dục trực tuyến có thể mang lại sự thuận tiện và linh hoạt cho nhiều học sinh hơn, vì vậy nó cũng có không gian phát triển rộng rãi trong tương lai. Do đó, làm thế nào để đảm bảo rằng học sinh giữ được mức độ tập trung và hiệu quả học tập như các lớp học truyền thống trong quá trình giáo dục trực tuyến là rất quan trọng để thúc đẩy sự phát triển hơn nữa của giáo dục trực tuyến. Tóm lại, đóng góp chính của bài báo này như sau. Bằng cách kết hợp các nền tảng giáo dục trực tuyến hiện có với mô hình nhận dạng nét mặt dựa trên kiến ​​trúc của mạng nơ-ron phức hợp, công trình này đã đề xuất một phương pháp cho phép theo dõi thời gian thực cảm xúc của học sinh trong các khóa học trực tuyến và đảm bảo rằng phản hồi được thể hiện bằng nét mặt có thể cung cấp cho giáo viên kịp thời để họ có thể linh hoạt điều chỉnh chương trình dạy học và cuối cùng là nâng cao chất lượng và hiệu quả của giáo dục trực tuyến.

1. **Phương pháp đề xuất**

Khung chủ yếu bao gồm hai phần: nền tảng các lớp học trực tuyến, trong bài báo này, tôi sẽ sử dụng các lớp học trực tuyến trên nền tảng học trực tuyến zoom và google meet của Khoa công nghệ thông tin trường đại học sư phạm Hà Nội làm ví dụ cho cuộc thử nghiệm và mô hình học sâu dựa trên CNN, trước đó có lưu ý rằng các hình ảnh gốc được thu thập từ các lớp học trực tuyến cần phải được xử lý trước, bao gồm nhận diện khuôn mặt, căn chỉnh, xoay và thay đổi kích thước, tùy theo các yếu tố khác nhau trong hình ảnh gốc. Hình số 2 trình bày quy trình của FER, và các bước chi tiết của khung đề xuất như sau: đầu tiên, các máy ảnh tích hợp trong các thiết bị điện tử được sử dụng để chụp ảnh khuôn mặt của các sinh viên đang theo học. Thứ hai, thuật toán nhận dạng nét mặt được đào tạo bởi cơ sở dữ liệu biểu cảm khuôn mặt được sử dụng để phát hiện các khuôn mặt và phân loại các biểu hiện trên khuôn mặt như tức giận, ghê tởm, sợ hãi, hạnh phúc, buồn bã, ngạc nhiên, khinh thường và trung tính. Thứ ba, biểu đồ phân bố xác suất của biểu thức được vẽ và cung cấp cho giáo viên để có thể điều chỉnh kế hoạch dạy học kịp thời.

1. Nền tảng giáo dục trực tuyến

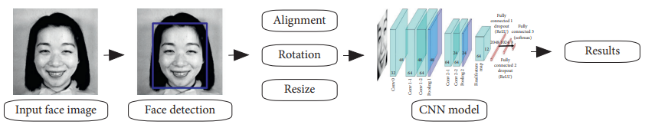
Những tiến bộ trong công nghệ đã tạo ra một số lượng lớn các nền tảng giáo dục trực tuyến và cải thiện đáng kể tính linh hoạt của giáo dục, cho phép giáo viên áp dụng các phương pháp kỹ thuật đa dạng để hỗ trợ giảng dạy mà không phải lo lắng về giới hạn số lượng học sinh trong lớp học dựa trên lớp học truyền thống và học sinh ở các khu vực khác nhau có thể giao tiếp trong thời gian thực mà không cần phải xem xét đến lưu lượng truy cập và các vấn đề khác. Các tài liệu giảng dạy tương tự như các lớp học truyền thống có thể được tải lên các nền tảng này để sinh viên tham khảo. Hiện tại, trong các nền tảng có chức năng giảng dạy trực tuyến, chẳng hạn như Zoom, Google meet, MS Team, giáo viên có thể áp dụng phương pháp họp video và tận dụng camera tích hợp trong thiết bị để chụp và nhận dạng nét mặt của học sinh trong thời gian thực. Các hình ảnh được chụp sẽ được xử lý trước và sau đó được sử dụng làm đầu vào của CNN.

1. Tiền xử lý hình ảnh

Xử lý hình ảnh trước khi dự đoán có thể làm giảm sự can thiệp của các vật thể giống như khuôn mặt trong nền khi phát hiện khuôn mặt trong một hình ảnh và sau đó chuẩn hóa các hình ảnh về ảnh đa mức xám, điều này sẽ giúp nâng cao hiệu quả của mô hình học sâu. Trong bài nghiên cứu này, tôi sử dụng phương pháp haar-cascade[20], một gói phần mềm có sẵn công khai tích hợp các thuật toán để theo dõi đặc điểm khuôn mặt, ước tính tư thế đầu, phát hiện thuộc tính khuôn mặt, v.v., làm công cụ xử lý trước. Haar-cascade cũng có thể được sử dụng để phát hiện nhiều khuôn mặt cùng một lúc. Các đặc điểm chính của từng khuôn mặt bao gồm lông mày, mắt, đầu mũi và miệng có thể được nhận ra một cách hiệu quả, và biểu hiện trên khuôn mặt có thể được phát hiện bằng các đường viền hình chữ nhật cho phù hợp, những đường viền này được xây dựng bởi các điểm đặc trưng ở cạnh của mọi mặt, bao gồm cả mặt trên và mặt dưới, xác định chiều rộng dọc, ngoài cùng bên phải và ngoài cùng bên trái, xác định chiều ngang của hình ảnh khuôn mặt. Để tránh bỏ sót thông tin trên khuôn mặt đồng thời giảm nhiễu nền, tôi đã chỉnh kích thước đường viền hình chữ nhật lên 3px lần để che được nhiều nội dung trên khuôn mặt hơn. Hơn nữa, xem xét kích thước hình ảnh đầu vào mô hình học tập được đặt trước là 48 × 48, hình ảnh được phát hiện sẽ được xoay với đầu mũi làm trung tâm và được thay đổi kích thước phù hợp để phù hợp với kích thước đầu vào.

1. Mô hình học máy dựa trên mạng nơ-ron tích chập

Kiến trúc của mô hình học sâu ứng dụng dựa trên CNN được minh họa trong Hình 3, dựa trên kết quả nghiên cứu do Kuo đề xuất[21] và hiệu suất trước đây của mô hình này trong FER so với mô hình tương tự khác cũng đã được chứng minh. Sau lớp tích hợp gồm 64 bộ lọc (channel) mỗi bộ lọc có kích thước cỡ 3 x 3, lớp đầu vào được theo sau bởi 2 khối, bao gồm 2 lớp tích tụ và 1 lớp tổng hợp tối đa với 64 bản đồ đối tượng riêng biệt. Và kích thước của các hạt nhân trong lớp tích chập đầu tiên được đặt thành 3 × 3, lớp thứ hai là 5 × 5, các lớp gộp tối đa đều bao gồm một hạt nhân có kích thước 2 × 2 và bước 2, và do đó, đầu vào hình ảnh sẽ được nén thành một phần tư. Và có 2 lớp được kết nối đầy đủ sau đây gồm 2048 và 1024 nơ-ron, trong đó Đơn vị tuyến tính chỉnh lưu (ReLU) [56–59] được sử dụng làm chức năng kích hoạt. Để ngăn chặn việc trang bị quá mức, một Dropout được thêm vào sau mỗi lớp trong số 2 lớp được kết nối đầy đủ, lớp này sẽ giải phóng một phần tế bào thần kinh theo xác suất rơi cài đặt trước; trong bài báo này, cả 2 giá trị đều được đặt thành 0,5. , e lớp đầu ra sau bao gồm 8 đơn vị và softmax [60] được sử dụng làm hàm kích hoạt để phân loại các biểu hiện được kiểm tra theo các khía cạnh giận dữ, ghê tởm, sợ hãi, hạnh phúc, buồn bã, ngạc nhiên, khinh thường và trung tính.

1. Huấn luyện mô hình dựa trên mạng nơ-ron tích chập
2. **Thực nghiệm và Đánh giá**
3. **Kết luận**

**Tài liệu tham khảo**

[1]. C. Darwin and P. Prodger. The Expression of the Emotions in Man and Animals. John Murray, 1998.

[2]. Y. Tian, T. Kanade, and J. F. Cohn. *Recognizing action units for facial expression analysis*. IEEE Transactions onPattern Analysis and Machine Intelligence, vol. 23, no. 2, 2001.

[3]. M. Bani, S. Russo, S. Ardenghi, G. Rampoldi, V. Wickline, S. Nowicki Jr, M. G. Strepparava . *Behind the Mask: Emotion Recognition in Healthcare Students*. Med.Sci.Educ. 2021.

[4]. M. Jeong, B. C. Ko. *Driver’s Facial Expression Recognition in Real-Time for Safe Driving*. Department of Computer Engineering, Keimyung University, Daegu 42601, Korea, 4 December 2018.

[5]. P. Ekman and W. V. Friesen. *Constants across cultures in the face and emotion*. Journal of Personality and Social Psychology*, vol. 17, no. 2, 124–129, 1971.*

[6]. Z. Zeng, M. Pantic, G. I. Roisman, and T. S. Huang. *A survey of affect recognition methods: audio, visual, and spontaneous expressions.* IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 31, no. 1, pp. 39–58, 2009.

[7]. S. Li and W. *Deng. Deep facial expression recognition: a survey*. IEEE Transactions on Affective Computing, In press.

[8]. C. Shan, S. Gong, and P. W. McOwan. *Facial expression recognition based on local binary patterns: a comprehensive study*. Image and Vision Computing, vol. 27, no. 6, pp. 803–816, 2009.

[9]. P. Lucey, J. F. Cohn, T. Kanade, J. Saragih, Z. Ambadar, and I. Matthews. *The extended Cohn-Kanade dataset (CK+): a complete dataset for action unit and emotion-specified expression*. In Proceedings of the 2010 IEEE Computer Society Conference on Computer Vision and Pattern RecognitionWorkshops, pp. 94–101, San Francisco, CA, USA, July 2010

[10]. D. Matsumoto. *More evidence for the universality of a contempt expression*. Motivation and Emotion, vol. 16, no. 4, pp. 363–368, 1992.

[11]. R. Zhi, M. Flierl, Q. Ruan, and W. B. Kleijn. *Graph-preserving sparse nonnegative matrix factorization with application to facial expression recognition*. IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics), vol. 41, no. 1, pp. 38–52, 2011.

[12]. A. Dhall, R. Goecke, J. Joshi, K. Sikka, and T. Gedeon. *Emotion recognition in the wild challenge 2014: baseline, data and protocol*. In Proceedings of the 16th International Conference on Multimodal Interaction, pp. 461–466, ACM, Istanbul Turkey, November 2014.

[13]. J. Li, K. Jin, D. Zhou, N. Kubota, and Z. Ju. *Attention mechanism-based CNN for facial expression recognition*. Neurocomputing, vol. 411, pp. 340–350, 2020.

[14]. K. Simonyan and A. Zisserman. *Very deep convolutional networks for large-scale image recognition*. 2014, https:// arxiv.org/abs/1409.1556.

[15]. C. Szegedy, W. Liu, Y. Jia et al. *Going deeper with convolutions*. in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1–9, Boston, MA, USA, June 2015.

[16]. A. Jahandad, S. M. Sam, K. Kamardin, N. N. Amir Sjarif, and N. Mohamed. *Offline signature verification using deep learning convolutional neural network (CNN) architectures GoogLeNet inception-v1 and inception-v3*. Procedia Computer Science, vol. 161, pp. 475–483, 2019.

[17]. K. He, X. Zhang, S. Ren, and J. Sun. *Deep residual learning for image recognition*. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778, Las Vegas, NV, USA, June 2016.

[18]. I. Allen and J. Seaman. *Digital compass learning: distance education enrollment report 2017*. Babson Survey Research Group, Babson Park, MA, USA, 2017.

[19]. ] E. Dolan, E. Hancock, and A. Wareing. *An evaluation of online learning to teach practical competencies in undergraduate health science students*. The Internet and Higher Education, vol. 24, pp. 21–25, 2015.

[20]. A.B.Shetty , Bhoomika , Deeksha , J.Rebeiro , Ramyashree. *Facial Recognition using Haar Cascade and LBP Classifiers*. Journal Pre-proof, 28 July 2021.

[21]. C. M. Kuo, S. H. Lai, and M. Sarkis. *A compact deep learning model for robust facial expression recognition*. in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pp. 2202–2208, Salt Lake City, UT, USA, June 2018.