NGHIÊN CỨU NHẬN DẠNG BIỂU CẢM KHUÔN MẶT BẰNG PHƯƠNG PHÁP HỌC SÂU SỬ DỤNG KIẾN TRÚC RESNET

RESEARCH OF FACIAL EXPRESSION RECOGNITION BY DEEP LEARNING USING RESNET ARCHITECTURE

HỒ THỊ HƯƠNG THƠM*, NGUYỄN KIM ANH

Khoa Công nghệ Thông tin, Trường Đại học Hàng hải Việt Nam

**Email liên hệ: thomhth@vimaru.edu.vn*

Tóm tắt

Nhận dạng biểu cảm khuôn mặt là phương pháp chính cho các ý định xử lý phi ngôn ngữ. Nghiên cứu nhận dạng biểu cảm khuôn mặt đã và đang được quan tâm nghiên cứu và ứng dụng ở nhiều nơi trên thế giới. Do đó trong bài báo này tập trung vào bài toán nhận dạng biểu cảm khuôn mặt bằng phương pháp học sâu sử dụng kiến trúc mạng ResNet101. Độ tin cậy của mô hình được đánh giá dựa trên tập dữ liệu mẫu có sẵn FER2013 cho tỷ lệ nhận dạng cao nhất là 71,22%. Từ phân tích chi tiết độ chính xác từng loại biểu cảm nhóm tác giả đưa ra giải pháp đề xuất ba nhóm biểu cảm chính để xây dựng chương trình đánh giá chất lượng dịch vụ với ba mức độ: hài lòng, bình thường và không hài lòng.

Từ khóa: *CNN, FER, ResNet.*

Abstract

Facial recognition is the main method for nonverbal processing intentions. Research on facial expression recognition has been interested in research and application in many parts of the world. Therefore, this paper focuses on the problem of facial expression recognition by deep learning method using ResNet101 network architecture. The reliability of the model was assessed based on the sample data set available FER2013 for the highest recognition rate of 71.22%. From the detailed analysis of the accuracy of each type of expression, the author offers the solution to propose three main expressive groups to develop a service quality assessment program with three levels: satisfaction, normal and unsatisfactory.

Keywords: *CNN, FER, ResNet.*

1. Giới thiệu

Biểu cảm khuôn mặt là một phương pháp phi ngôn ngữ chính thể hiện cảm xúc giao tiếp của con người. Theo các nghiên cứu trong [15] cho thấy 55% thông điệp liên quan đến cảm xúc và thái độ là ở nét mặt, 7% trong đó có thể nói ra, phần còn lại là biểu đạt ngôn ngữ (cách mà các từ được nói). Biểu cảm trên khuôn mặt đóng một vai trò quan trọng trong toàn bộ quá trình trao đổi thông tin. Với sự phát triển nhanh chóng của trí tuệ nhân tạo, tự động nhận dạng biểu cảm khuôn mặt đã được nghiên cứu mạnh mẽ trong những năm gần đây. Nghiên cứu về nhận dạng biểu cảm khuôn mặt (Facial Expression Recognition - FER) đang rất được chú ý quan tâm trong các lĩnh vực tâm lý học, thị giác máy tính và nhận dạng mẫu. FER có các ứng dụng rộng rãi trong nhiều lĩnh vực, bao gồm tương tác máy tính và con người [11,14], thực tế ảo [2], thực tế tăng cường [3], hệ thống hỗ trợ người lái tiên tiến [1], giáo dục [7] và giải trí [9].

Có nhiều phương pháp nhận dạng biểu cảm có thể nhóm theo bốn hướng chính: Hướng tiếp cận dựa trên tri thức, hướng tiếp cận dựa trên đặc trưng không gian thay đổi, hướng tiếp cận dựa trên đặc trưng so khớp mẫu, hướng tiếp cận dựa trên diện mạo (hướng tiếp cận theo phương pháp học). Đặc biệt hướng tiếp cận theo phương pháp học là hướng tiếp cận rất được quan tâm vì khả năng nhận dạng cho tỷ lệ chính xác cao với sai số có thể chấp nhận được.

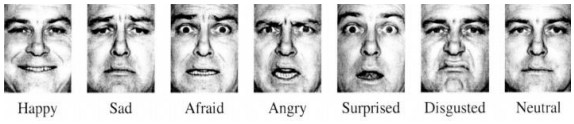
Trong nghiên cứu của bài báo này quan tâm đến nhận dạng biểu cảm khuôn mặt bằng phương pháp học sâu sử dụng kiến trúc Residual Network (ResNet) [5], đây là kỹ thuật đã cho ra kết quả rất khả quan trong thời gian gần đây đối với các bài toán nhận dạng đối tượng. Nội dung của bài báo được trình bày cụ thể như sau: Mục 2 giới thiệu tổng quan các loại biểu cảm khuôn mặt; Mục 3 trình bày mô hình học sâu sử dụng để nhận dạng biểu cảm khuôn mặt; Mục 4 đề xuất giải pháp ứng dụng nhận dạng biểu cảm để đánh giá chất lượng phục vụ dịch vụ và đánh giá kết quả thử nghiệm; Mục 5 kết luận.

TẠP CHÍ ISSN: 1859-316X
KHOA HỌC CÔNG NGHỆ HÀNG HẢI
JOURNAL OF MARINE SCIENCE AND TECHNOLOGY

2. Biểu cảm khuôn mặt

Cảm xúc của con người được thể hiện qua các biểu cảm khuôn mặt, nhận diện được biểu cảm của người đối diện là một trong các bản năng tự nhiên của con người. Vậy làm thế nào để "dạy" cho máy tính biết cách phân biệt các loại cảm xúc này? Câu trả lời là khi con người thể hiện cảm xúc, luôn tồn tại một số đặc trưng chung trên khuôn mặt của tất cả mọi người bất kể độ tuổi, vị trí địa lý hay điều kiện sống,… Dựa vào đặc trưng này, ta có thể rút ra các đặc điểm quan trọng của cảm xúc, mô hình hóa và "dạy" cho máy tính hiểu được cảm xúc đó.

Nhận dạng chính xác biểu cảm khuôn mặt là một bài toán khó khăn vì con người có rất nhiều "cung bậc cảm xúc" khác nhau. Để bài toán không quá phức tạp có thể chia biểu cảm khuôn mặt vào bảy loại sắc thái chính sau: hạnh phúc (happy), đau khổ (Sad), sợ hãi (Afraid/fear), tức giận (angry), ngạc nhiên (surprised), căm phẫn (disgusted) và trung lập (neutral) [1, 2, 7, 8,9,13] - như minh họa trong Hình 1.



***Hình 1. Bảy cảm xúc chính của khuôn mặt: hạnh phúc, buồn, sợ hãi, tức giận, ngạc nhiên, căm phẫn, trung lập** [13]*

Nhiệm vụ của một hệ thống nhận diện cảm xúc là phải phân loại được một trạng thái mặt người vào nhóm một trong bảy biểu cảm trên.

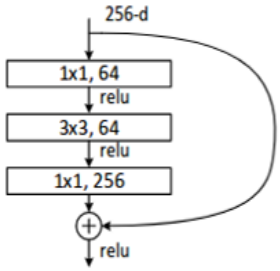
3. Mô hình học sâu sử dụng cho bài toán nhận dạng biểu cảm

Hiện nay có nhiều mô hình mạng học sâu nhân chập CNN (Convolutional neural networks) được vận dụng trong các bài toán nhận dạng như: LeNet, AlexNet, VGG, GoogLeNet, ResNet,… [1, 3, 6, 9, 10, 15], trong nghiên cứu này lựa chọn mạng ResNet cho mô hình nhận dạng biểu cảm vì một số lý do được trình bày chi tiết sau đây.

3.1. Mạng học sâu ResNet101

ResNet (Residual Network) được phát triển bởi Microsoft vào năm 2015 công bố trên bài báo "Deep residual learning for image recognition" [5]. ResNet đã chiến thắng với vị trí số một trong cuộc thi ILSVRC 2015 với tỷ lệ lỗi đứng trong top 5 chỉ 3,57%, thậm chí đứng vị trí đầu tiên trong cuộc thi ILSVRC và COCO 2015 với ImageNet Detection, ImageNet localization, Coco detection và Coco segmentation. ResNet có cấu trúc gần giống VGG với nhiều lớp ngăn xếp làm cho mô hình sâu hơn. Có nhiều biến thể của kiến trúc ResNet với số lớp khác nhau như ResNet-18, ResNet-34, ResNet-50, ResNet-101, ResNet-152,... Với tên là ResNet theo sau là một số chỉ kiến trúc ResNet với số lớp nhất định. Resnet giải quyết được vấn đề của học sâu truyền thống, nó có thể dễ dàng học với hàng trăm lớp.

Mạng ResNet (R) là một mạng CNN được thiết kế để làm việc với hàng trăm hoặc hàng nghìn lớp chập. Một vấn đề xảy ra khi xây dựng mạng CNN với nhiều lớp chập sẽ xảy ra hiện tượng Vanishing Gradient dẫn tới quá trình học tập không tốt. Chính vì vậy giải pháp mà ResNet đưa ra là sử dụng kết nối tắt đồng nhất để xuyên qua một hay nhiều lớp. Một khối như vậy được gọi là một Residual Block, như trong Hình 2.



Hình 2. Một khối Residual của ResNet

ResNet gần như tương tự với các mạng CNN khác gồm có: nhân chập (convolution), tổng hợp (pooling), kích hoạt (activation) và kết nối đầy đủ (fully-connected layer). Hình 3 hiển thị khối dư được sử dụng trong mạng. Xuất hiện một mũi tên cong xuất phát từ đầu và kết thúc tại cuối khối dư hay ResNet sử dụng các kết nối tắt (kết nối trực tiếp đầu vào của lớp (n) với (n+x) được hiển thị dạng mũi tên cong. Qua mô hình nó chứng minh được có thể cải thiện hiệu suất trong quá trình huấn luyện khi mô hình có hơn 20 lớp. Như vậy có thể hiểu việc tăng số lượng các lớp trong mạng làm giảm độ chính xác, nhưng muốn có một kiến trúc mạng sâu hơn có thể hoạt động tốt.

Do đó trong nghiên cứu này, sử dụng mạng CNN với mô hình ResNet101 [5] để xây dựng cho bài toán nhận dạng biểu cảm khuôn mặt.

3.2. Cấu hình ResNet101

Cấu trúc mạng ResNet101 cho bài toán nhận dạng biểu cảm khuôn mặt được thiết lập như Hình 3 gồm năm phân đoạn (stage), chi tiết mỗi stage được miêu tả dưới đây.

Ký hiệu "ID BLOCK" trong Hình 4 là viết tắt của từ Identity block, ID BLOCKx3 nghĩa là có 3 khối Identity block chồng lên nhau. Cụ thể như sau:

Zero-padding: Input với (3,3).

Stage 1: Tích chập (Conv1) với 64 filters với

shape(7,7), sử dụng stride(2,2). BatchNorm (epsilon = 1.1e-5, axit = 1|3), MaxPooling (3,3).

Stage 2: Convolutiontal block (a) sử dụng 3 bộ lọc filter với size 64x64x256, f=3, s=1, strides(1,1). Có 2 Identity blocks (b, c) với filter size 64x64x256, f=3.

Stage 3: Convolutional block (a) sử dụng 3 bộ lọc filter size 128x128x512, f=3, s=2. Có 3 Identity blocks (b_1, b_2, b_3) với filter size 128x128x512, f=3.

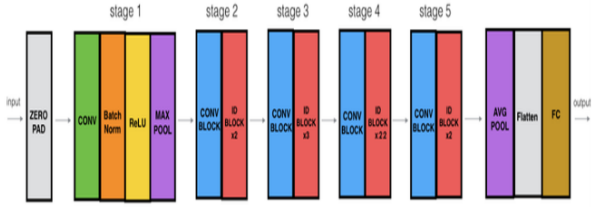
Stage 4: Convolutional block (a) sử dụng 3 filter size 256x256x1024, f=3, s=2. Có 22 Identity blocks (b_1, b_2,…b_{22}) với filter size 256x256x1024, f=3.

Stage 5: Convolutional block (a) sử dụng 3 filter size 512x512x2048, f=3, s=2. Có 2 Identity blocks (b,c) với filter size 512x512x2048, f=3.

The 2D Average Pooling: Sử dụng với kích thước (7,7).

The Flatten.

Fully Connected (Dense): sử dụng softmax activation.



Hình 3. Cấu trúc ResNet101 nhận dạng biểu cảm

3.3. Tập dữ liệu, cài đặt và thử nghiệm

Tập ảnh dùng đánh giá độ tin cậy của mô hình là tập ảnh Fer2013 được tải về từ [4] trên Kaggle gồm 35.887 ảnh cấp xám kích cỡ 48x48 trong đó: 28.709 ảnh dùng để huấn luyện (training), 3.589 ảnh kiểm tra thẩm định (public test) và 3.589 ảnh kiểm tra riêng (private test) với 7 lớp biểu cảm (Angry, Disgust, Fear, Happy, Sad, Surprise, Neutral), Hình 4 minh họa một phần bộ ảnh.

Mô hình nhận dạng trên ResNet101 được cài đặt trên ngôn ngữ Python Ver 3.7 và thư viện Keras/Tensorflow được sử dụng để cài đặt, trên máy tính PC i7- 4600U CPU@ 2.10Hz. Quá trình xử lý qua 5 bước sau:

Bước 1: Nhập ảnh đầu vào (có thể là ảnh màu hoặc ảnh xám).

Bước 2: Phát hiện vùng ảnh mặt người bằng hàm haar cascade (của thư viện OpenCV).

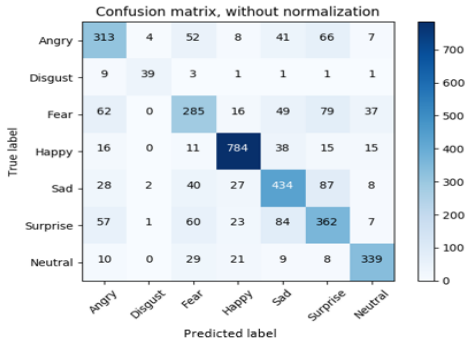
Bước 3: Vùng ảnh mặt người được chuyển về kích thước 48x48.

Bước 4: Ảnh vùng mặt 48x48 (sử dụng cả 3 kênh màu) đưa vào mạng học sâu sử dụng cấu trúc ResNet101.

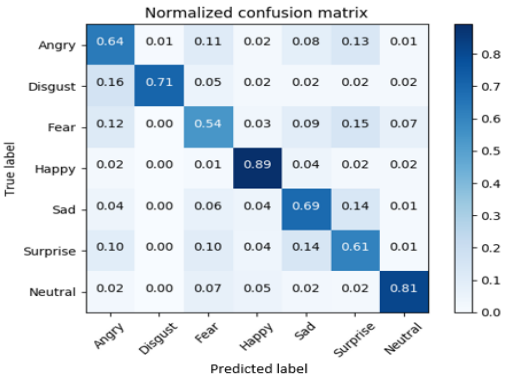
Bước 5: Đầu ra của ResNet101 là xác xuất của bảy cảm xúc chính.

***Hình 4. Minh họa một phần tập ảnh Fer2013** [6]*

Số ảnh kiểm tra độ chính xác từ tập dữ liệu Fer2013 là 3.589 ảnh cho kết quả cao nhất trên Fer2013 là 71,22%. Hình 5 và Hình 6 thống kê kết quả phân loại của từng loại biểu cảm trong đó Hình 5 thống kê theo số lượng ảnh, Hình 6 thống kê theo tỷ lệ tương ứng. Hình 7 đề mô giao diện hệ thống nhận dạng biểu cảm.

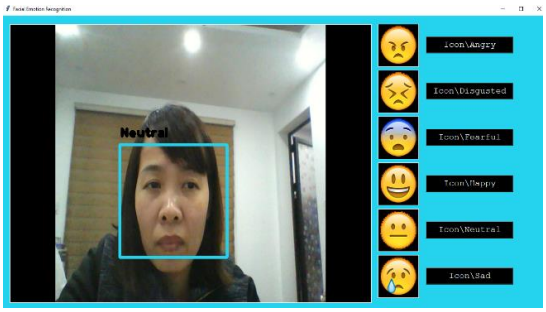


Hình 5. Kết quả phân loại biểu cảm theo số lượng ảnh



Hình 6. Kết quả nhận dạng biểu cảm theo tỷ lệ

TẠP CHÍ ISSN: 1859-316X
KHOA HỌC CÔNG NGHỆ HÀNG HẢI
-JOURNAL OF MARINE SCIENCE AND TECHNOLOGY-



Hình 7. Đề mô hệ thống nhận dạng

Từ kết quả nhận được theo Hình 6 và 7 của mô hình nhận dạng biểu cảm khuôn mặt sử dụng ResNet cho kết quả tốt đối với các cảm xúc: hạnh phúc - happy (89%), trung lập - Neutral (81%), kết quả mức trung đối với cảm xúc: ghê tởm - căm phẫn (71%) và buồn - sad (69%), kết quả mức thấp với cảm xúc: giận dữ - angry (64%), ngạc nhiên - surprise (61%) và sợ hãi - fear (54%).

4. Đề xuất giải pháp ứng dụng đánh giá chất lượng dịch vụ

Theo kết quả thử nghiệm trong mục 3, có thể thấy mô hình đánh giá tốt với cảm xúc hạnh phúc (89%) và trung lập (81%), đánh giá kém với cảm xúc sợ hãi (54%) và ngạc nhiên (61%). Dựa trên lợi thế nhận dạng tốt cảm xúc hạnh phúc và trung lập hay căm phẫn có thể đề xuất ứng dụng vào hệ thống đánh giá chất lượng mang tên "hành chính nụ cười" hay "dịch vụ hạnh phúc" theo ba nhóm cảm xúc với ba mức độ về chất lượng dịch vụ như sau:

+ Nhóm 1 (hài lòng về dịch vụ): Nhóm cảm xúc hạnh phúc;

+ Nhóm 2 (bình thường về dịch vụ): Nhóm cảm xúc trung lập;

+ Nhóm 3 (không hài lòng về dịch vụ): Nhóm cảm xúc còn lại (tức giận, căm phẫn, sợ hãi, buồn và ngạc nhiên).

Khi mỗi khách hàng (sinh viên hoặc công dân) được phục vụ ra về yêu cầu họ cho biết cảm xúc của họ qua hệ thống nếu họ cảm thấy hài lòng hãy nở nụ cười, nếu họ cảm thấy bình thường hãy giữ thái độ trung lập, họ không thỏa mãn họ có thể thể hiện cảm xúc trong năm biểu cảm (tức giận, căm phẫn, sợ hãi, buồn và ngạc nhiên). Hệ thống sẽ tự động đếm số mức độ (hài lòng, bình thường và không hài lòng) để tổng hợp đánh giá tình hình chất lượng phục vụ từ đó đưa ra giải pháp điều chỉnh phù hợp và nâng cao chất lượng phục vụ cần thiết.

Thực hiện thử nghiệm cho một nhóm sinh viên với 142 sinh viên (của 3 lớp THVP N17, N02 và N09) để đánh giá 3 nhóm biểu cảm đã đề xuất trên, Hình 8 minh họa một phần tập ảnh hệ thống lưu lại sau khi nhận dạng.

Hình 8. Minh họa một phần tập ảnh được lưu lại từ hệ thống đánh giá chất lượng phục vụ

Từ số lượng biểu cảm nhận được của hệ thống cho thấy kết quả nhận dạng tốt các nhóm cảm xúc đánh giá chất lượng dịch vụ: hài lòng, bình thường và không hài lòng. Hệ thống xác nhận thái độ khi biểu cảm được nhận dạng ổn định trong 5 giây, kết quả tỷ lệ nhận dạng trung bình của 3 mức độ trên 79%. Cụ thể theo Bảng 1.

Bảng 1. Bảng thống kê tỷ lệ nhận dạng của 3 mức thái độ

Thái độ	Biểu cảm dự định	Biểu cảm nhận dạng (dự đoán)	Tỷ lệ nhận dạng
Hài lòng	142	121	85,21%
Bình thường	142	115	80,97%
Không hài lòng	142	101	71,13%

5. Kết luận

Trong nghiên cứu này đã đưa ra mô hình nhận dạng biểu cảm khuôn mặt bằng mạng học sâu kiến trúc ResNet101. Tập dữ liệu ảnh Fer2013 [4] dùng để đánh giá độ chính xác của mô hình với tỷ lệ trên 70%. Rất nhiều nghiên cứu đã sử dụng tập dữ liệu Fer2013 để thử nghiệm nhưng tỷ lệ nhận dạng tốt nhất cũng chỉ trên dưới 70% kể cả với các công bố gần đây 2019 [1-3, 7, 9, 10], điều đó chứng tỏ tập dữ liệu này có nhiều mâu thuẫn hay có độ tương đồng giữa các biểu cảm. Về tổng thể các loại biểu cảm có độ chính.

Mô hình được huấn luyện trên tập dữ liệu Fer2013 nhưng vẫn làm nhận diện tốt trên các dữ liệu khác cho thấy mô hình đã học được các đặc trưng phù hợp của khuôn mặt người. Tuy nhiên hầu hết các dữ liệu học hiện nay thường sử dụng khuôn mặt người phương tây, nhóm tác giả sẽ xây dựng và bổ sung thêm tập dữ liệu

cảm xúc của người châu Á để phong phú dữ liệu huấn luyện và nâng cao chất lượng nhận dạng.

Ngoài ra cần tiến hành thử nghiệm thêm địa điểm thực tế phục vụ khách hàng hoặc sinh viên với số lượng mẫu nhiều hơn nữa (trên 1.000 khách hàng/sinh viên) để đưa ra tỷ lệ nhận dạng chính xác hơn đánh giá độ tin cậy của hệ thống nhận dạng trước khi đưa hệ thống vào ứng dụng thức tế.

Lời cảm ơn

Bài báo này là sản phẩm của đề tài nghiên cứu khoa học cấp Trường năm học 2019-2020, tên đề tài: "*Nhận dạng biểu cảm khuôn mặt bằng phương pháp học sâu*", được hỗ trợ kinh phí bởi Trường Đại học Hàng hải Việt Nam.

TÀI LIỆU THAM KHẢO

[1] 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.

[2] 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.

[3] 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.

[4] Fer2013, https://www.kaggle.com.

[5] 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.

[6] 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, Octoer 8-16 2016.

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

[8] L. Wolf, T. Hassner, I. Maoz, Face Recognition in Unconstrained Videos with Matched Background Similarity, Computer Vision and Pattern Recognition (CVPR), 2011.

[9] 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.

[10] 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.

[11] 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.

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

[13] Matthew N. Dailey, Garrison W. Cottrell1, Curtis Padgett, and Ralph Adolphs (2014), EMPATH: A Neural Network that Categorizes Facial Expressions, Journal of Cognitive Neuroscience 14:8, pp.1158-1173, 2014.

[14] 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.

[15] Yunxin Huang, Fei Chen, Shaohe Lv and Xiaodong Wang, Facial Expression Recognition: A Survey, Symmetry 2019, 11, 1189; doi:10.3390/sym11101189.

Ngày nhận bài:	14/04/2020
Ngày nhận bản sửa:	19/05/2020
Ngày duyệt đăng:	01/06/2020