

CLASSIFYING FACE SENSITIVES USING SUPPORT VECTOR MACHINE AND CONVOLUTIONAL NEURAL NETWORK

Le Thu Trang*, Nguyen Thu Huong

TNU - University of Information and Communication Technology

ARTICLE INFO	ABSTRACT
Received: 29/4/2021 Revised: 23/11/2021 Published: 24/11/2021	Distinguishing emotions through facial recognition is one of the problems that many researchers are interested in. For the purpose of assisting the user in recognizing facial emotions to use research in fields such as clinical or behavioral science. The challenge with this problem is that human facial expressions have similarities and overlaps in different emotional expressions. In this paper, the authors propose to use the SVM method combined with the transitional neural network to classify facial emotions on the FER dataset and build 3 strategic models to conduct experiments. Determining the exact emotion of the face is always difficult. The experimental results have shown that the CNN model is accurate when compared to the HOG-SVM model. The CNN model takes the real image as input, and the CNN model imports the combined image that has approximately the same results and is more stable.
KEYWORDS Facial expression Object classification Machine learning SVM – Support Vector Machine CNN – Convolutional Neural Network	

SỬ DỤNG PHƯƠNG PHÁP SUPPORT VECTOR MACHINE VÀ CONVOLUTIONAL NEURAL NETWORK ĐỂ PHÂN LOẠI CẢM XÚC CỦA KHUÔN MẶT

Lê Thu Trang*, Nguyễn Thu Hương

Trường Đại học Công nghệ thông tin và Truyền thông – ĐH Thái Nguyên

THÔNG TIN BÀI BÁO	TÓM TẮT
Ngày nhận bài: 29/4/2021 Ngày hoàn thiện: 23/11/2021 Ngày đăng: 24/11/2021	Phân loại cảm xúc khuôn mặt thông qua việc nhận diện nét mặt hiện là một trong những bài toán được nhiều nhà nghiên cứu quan tâm. Với mục đích hỗ trợ được người dùng trong việc nhận diện được cảm xúc khuôn mặt để sử dụng nghiên cứu trong các lĩnh vực như khoa học lâm sàng hay khoa học hành vi. Thách thức với bài toán này là nét mặt của con người có sự tương đồng, trùng lặp trong các biểu thị cảm xúc khác nhau. Trong bài báo này, nhóm tác giả đề xuất sử dụng phương pháp Support Vector Machine – SVM kết hợp với mạng thần kinh chuyển đổi để phân loại cảm xúc trên khuôn mặt trên bộ dữ liệu FER và xây dựng 3 mô hình chiến lược để tiến hành các thí nghiệm. Việc xác định cảm xúc chính xác của khuôn mặt luôn khó khăn. Các kết quả của thực nghiệm đã cho thấy rằng mô hình Convolutional Neural Network - CNN chính xác hơn khi so sánh với mô hình Histogram Of Orientation Gradient + Support Vector Machine - HOG + SVM. Mô hình CNN lấy hình ảnh thực làm đầu vào và mô hình CNN nhập hình ảnh kết hợp có kết quả gần đúng với nhau và có tính chất ổn định hơn.
TỪ KHÓA Biểu cảm khuôn mặt Phân lớp đối tượng Máy học SVM – Support Vector Machine Mạng nơ-ron tích chập	

DOI: <https://doi.org/10.34238/tnu-jst.4443>

* Corresponding author. Email: trangtip@gmail.com

1. Giới thiệu

Nhận dạng cảm xúc trên khuôn mặt là quá trình xác định các cảm xúc trên khuôn mặt của con người thông qua khuôn mặt. Mọi người rất khác nhau với độ chính xác khác nhau trong việc nhận biết cảm xúc của người khác. Trong sự lan rộng của các độ chính xác khác nhau này, công nghệ có thể chứng minh kênh nhận dạng cảm xúc cũng như phát triển một số mô hình để xác định cảm xúc khuôn mặt chính xác của con người. Các phương tiện giao tiếp bao gồm cử chỉ cơ thể, nét mặt,... ngoài lời nói. Cử chỉ cơ thể giúp điều chỉnh giao tiếp trong lời nói, trong khi nét mặt thể hiện cảm xúc của con người theo phản xạ. Nhận dạng cảm xúc khuôn mặt có ứng dụng trong nhiều lĩnh vực khác nhau bao gồm khoa học lâm sàng, khoa học hành vi,... Phát hiện nét mặt cũng là điều kiện tiên quyết ngày nay đối với giao diện người - máy. Trong vài năm qua, một số tiến bộ đã được thực hiện trong lĩnh vực phát hiện nét mặt, như kỹ thuật trích xuất các đặc điểm trên khuôn mặt và phân loại biểu cảm. Nhưng vẫn phải phát triển một hệ thống phát hiện biểu cảm khuôn mặt tự động. Dự án này thực hiện nhận dạng khuôn mặt theo nhiều cách, bằng cách sử dụng bộ mô tả (Histogram Of Orientation Gradient) và SVM (Máy vector hỗ trợ) cho mô hình đầu tiên trong khi với các chiến lược đầu vào khác nhau cho Mạng nơ-ron tích chập Convolutions Neural Network (CNN) trong các mô hình khác. Sau đó, CNN và SVM dự đoán nhãn theo một trong các cảm xúc trên khuôn mặt sau: bình thường, hạnh phúc, sợ hãi, buồn bã, chán ghét, tức giận, ngạc nhiên. Động cơ của việc kết hợp hai hoặc nhiều kỹ thuật và tiền xử lý để đạt được kết quả có thể so sánh được thành công.

Trước đây, các nhà nghiên cứu đã tiến hành nghiên cứu nhận biết các đơn vị hành động cho khuôn mặt thông qua việc phân tích các biểu thức tự động [1]. Tiếp đến, đã có nhiều kết quả nghiên cứu đề cập và thực hiện giải quyết các bài toán về nhận dạng như xác định người trong ảnh [2] hay nhận dạng các biểu hiện khuôn mặt dựa trên các đặc điểm hình ảnh bằng các công cụ khác nhau có thể được kết hợp để cải thiện độ chính xác của hệ thống [3]. Ngoài ra cũng đã có nhiều nghiên cứu về nhận dạng liên quan đến mạng độ sâu như Karen Simonyan và cộng sự trong [4], hệ thống tự động nhận biết các cử động trên khuôn mặt [5]. Hay hệ thống nhận dạng cảm xúc trên khuôn mặt thể hiện việc phân loại khuôn mặt thành một số bộ cảm xúc ban đầu như vui vẻ, buồn bã và tức giận [6].

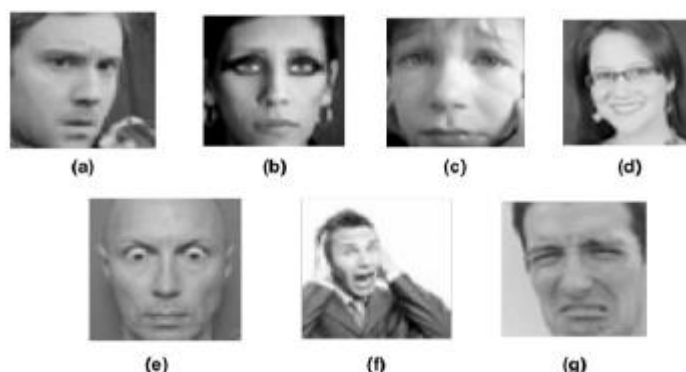
Khuôn mặt tạo ra các chuyển động cơ riêng lẻ để tạo ra một khuôn mặt khách quan và hệ thống mã hóa hành động trên khuôn mặt (FACS) là khung tâm lý - logic được sử dụng để mô tả các chuyển động trên khuôn mặt. Đây là một phương pháp được sử dụng để phân loại các chuyển động trên khuôn mặt của con người theo ngoại hình của họ bằng đơn vị hành động (AU). Chúng là sự thư giãn hoặc co lại của một hoặc nhiều cơ. Có nhiều kỹ thuật được sử dụng để nhận dạng cảm xúc trên khuôn mặt như mạng nơ-ron tích chập để nhận dạng biểu hiện trên khuôn mặt [7].

Với mục đích tiếp tục cải thiện và kết hợp so sánh với các phương pháp đã có và sử dụng trước đây, bài báo đề xuất một phương pháp sử dụng để phân lớp cảm xúc khuôn mặt dựa vào SVM và CNN. Các kết quả thực nghiệm dựa trên tập dữ liệu FER2013 từ trang web Kaggle đã chỉ ra hiệu quả của phương pháp đề xuất.

2. Dữ liệu để nghiên cứu

Dự án sử dụng Tập dữ liệu FER2013 từ trang web Kaggle có chứa 35.887 (48 pixel * 48 pixel) hình ảnh xám mô tả các biểu cảm trên khuôn mặt. Tất cả các hình ảnh đều có các khuôn mặt gần như chính giữa và chiếm một lượng không gian tương tự. Tập dữ liệu có bảy lớp khác nhau và mỗi hình ảnh thuộc một lớp cụ thể của cảm xúc trên khuôn mặt. Ví dụ, Hình 1 cho thấy bảy cảm xúc khác nhau trên khuôn mặt.

Các ảnh từ (a) đến (g) mô tả số lớp của hình ảnh và do đó cảm xúc trên khuôn mặt lần lượt là tức giận, bình thường, buồn bã, vui vẻ, ngạc nhiên, sợ hãi và chán ghét. Kích thước của tập dữ liệu là (35887, 48, 48, 1). Mô tả phân loại của tập dữ liệu (số lượng hình ảnh và cảm xúc khuôn mặt tương ứng. Các hình ảnh của biểu cảm khuôn mặt có số lượng như sau trong cơ sở dữ liệu:



Hình 1. Mô tả các cảm xúc có trong tập dữ liệu ảnh cảm xúc khuôn mặt FER

- Tức giận có 4953 hình ảnh;
- Chán ghét có 547 hình ảnh;
- Sợ hãi có 5120 hình ảnh;
- Vui vẻ có 8988 hình ảnh;
- Buồn bã có 6077 hình ảnh;
- Ngạc nhiên có 4002 hình ảnh;
- Bình thường có 6198 hình ảnh.

Tập dữ liệu đã được chia thành 3 tập riêng biệt, đó là tập huấn luyện (để huấn luyện mô hình), tập xác nhận (để điều chỉnh siêu tham số) và tập thử nghiệm (để kiểm tra mô hình). Có 28196 hình ảnh trong bộ đào tạo, 3546 hình ảnh trong bộ xác nhận và 3545 hình ảnh.

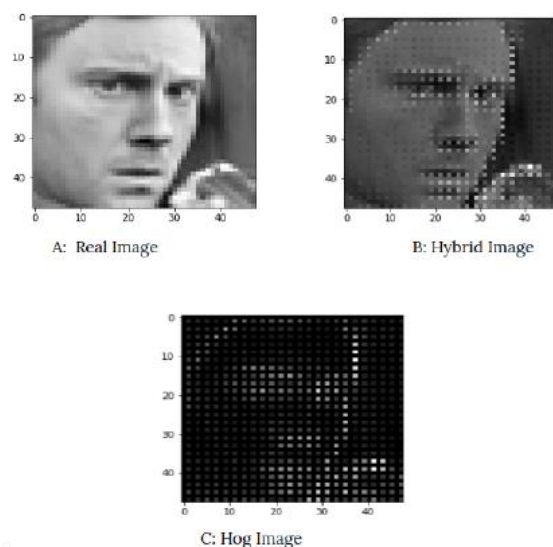
3. Xây dựng mô hình để phân lớp cảm xúc khuôn mặt dựa vào SVM và CNN

Ba mô hình để phân lớp các biểu cảm khuôn mặt được xây dựng như sau:

- *Mô hình 1 (HOG + SVM)*: Mô hình này sử dụng biểu đồ của các gradient có định hướng (HOG) để trích xuất đối tượng và máy vector hỗ trợ (SVM có RBF kernel) để phân loại cảm xúc trên khuôn mặt của ảnh [2]. HOG là một trong những bộ mô tả khuôn mặt trong học máy và thị giác máy tính nhưng cũng có thể được sử dụng để định lượng và thể hiện cả hình dạng và kết cấu. Các sự phổ biến của HOG có thể được ủng hộ với khả năng mô tả đặc điểm của diện mạo và hình dạng đối tượng cục bộ sử dụng phân bố cường độ cục bộ gradient (gradient ngang và dọc). Bộ mô tả HOG được mô hình hóa để cung cấp các chiều của vector đặc trưng có giá trị thực, tùy thuộc vào các tham số sau: hướng, pixel trên mỗi ô và các ô trên mỗi khối. Hàm cơ sở hướng tâm (RBF) được sử dụng trong SVM làm phương pháp hạt nhân.

- *Mô hình 2 (Hình ảnh thực + CNN)*: Mô hình này sử dụng mạng nơ-ron tích chập (mô hình CNN) để nhấn chìm cả hai mục đích trích xuất và phân loại đặc điểm của cảm xúc trên khuôn mặt và đầu vào hình ảnh thực (hình ảnh thang độ xám) (Hình 2). Mô hình được cấu trúc với 9 lớp tích chập, chức năng kích hoạt phi tuyến, ReLU, sau đó dẫn đến 4 lớp được kết nối đầy đủ [7]. Các hoạt động bỏ qua, chuẩn hóa hàng loạt (BN) và gộp tối đa được sử dụng sau mỗi lớp. Cuối cùng, mạng có một lớp dày đặc tính toán điểm số.

- *Mô hình 3 (CNN Balanced Dataset)*: Các bộ dữ liệu không phải lúc nào cũng cân bằng về số lượng hình ảnh nằm trong mỗi lớp phân loại cảm xúc khuôn mặt, không thể hoàn toàn dựa vào ergo để xác định độ chính xác. Tương tự là trường hợp với tập dữ liệu FER. Nó có tổng cộng 35887 hình ảnh thuộc 7 cảm xúc khác nhau. Số lượng danh mục chính xác như sau: Tức giận có 4953 hình ảnh, Chán ghét có 547 hình ảnh, Sợ hãi có 5120 hình ảnh, Vui vẻ có 8988 hình ảnh, Buồn bã có 6077 hình ảnh, Ngạc nhiên có 4002 hình ảnh, Bình thường có 6198 hình ảnh. Mô hình này khai thác khái niệm lấy mẫu xuống để cân bằng tập dữ liệu. 547 hình ảnh từ mỗi danh mục nêu trên được thu thập là "Chán ghét" có số lượng hình ảnh ít nhất, là 547, trong số tất cả các danh mục. Bộ dữ liệu mới hiện có 3829 hình ảnh.



Hình 2. Dữ liệu đầu vào cho các mô hình

4. Phân tích và đánh giá thực nghiệm

4.1. Phân tích và đánh giá thực nghiệm

Trong Model-1, tập dữ liệu (35887 hình ảnh) được chia thành 2 tập là tập huấn luyện chứa 28196 hình ảnh và tập thử nghiệm với 7691 hình ảnh. Kích thước của mỗi hình ảnh là 48 pixel * 48 pixel và được chia thành các khối có kích thước tương đương khác nhau có chứa pixel.

Trong Model-2 và Model-3, tập dữ liệu được chia thành 3 bộ, cụ thể là bộ huấn luyện, bộ xác nhận và bộ thử nghiệm. Tập dữ liệu có 35887 hình ảnh, mỗi hình ảnh có kích thước 48 pixel * 48 pixel. Tập huấn luyện được sử dụng để huấn luyện mô hình chứa 28196 hình ảnh, mỗi hình ảnh 48 pixel * 48 pixel. Bộ kiểm tra được sử dụng để kiểm tra mô hình và chứa 3845 hình ảnh, mỗi hình ảnh 48 pixel * 48 pixel. Bộ xác thực được sử dụng cho điều chỉnh các siêu tham số và chứa 3846 ảnh còn lại, mỗi ảnh 48 pixel * 48 pixel.

4.2. Tiến hành thí nghiệm

Thuật toán thực hiện của mô hình 1:

Bước 1. Tính toán độ lớn/ hướng của gradient tại mỗi pixel bằng cách sử dụng diện tích ô 8 * 8 ô pixel.

Bước 2. Tạo biểu đồ gồm 64 vector gradient đã tạo (8 * 8).

Bước 3. Chia từng ô của hình ảnh thành các vùng góc cạnh (Hình 3), trong đó mỗi vùng tương ứng với một hướng gradient từ 0 độ đến 180 độ (20 độ mỗi vùng).

Bước 4. Lặp lại tất cả ba bước cho đến khi toàn bộ hình ảnh được bao phủ bởi tổ hợp chập, để có được hình ảnh gradient.

Bước 5. Sau khi kết thúc các tính năng HOG từ tập huấn luyện (kích thước của các tính năng Hog = (kích thước trong tổng số hình ảnh, 900))

Xtrain = (28196,900),

Xtest = (7691,900),

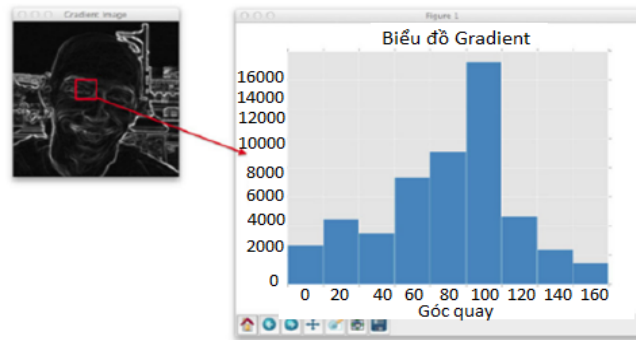
ytrain = (28196,1),

Ytest = (7691,1).

Bước 6. Huấn luyện SVM (nhân rbf, gamma = 0,1) trên tập huấn luyện.

Bước 7. Kiểm tra mô hình SVM trên dữ liệu thử nghiệm.

Bước 8. Tính toán độ chính xác cũng như ma trận nhầm lẫn.

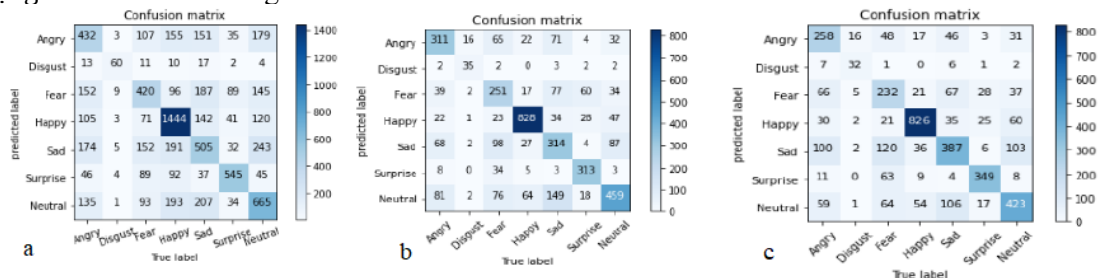


Hình 3. Chuyển đổi Gradient

Trong mô hình 2, mô hình 3, kiến trúc bao gồm tám lớp phức hợp và bốn lớp được kết nối đầy đủ (để tính toán tổn thất và điểm số), tiếp theo là chuẩn hóa hàng loạt và bỏ học. Lớp CNN đầu tiên (Hình 4) sử dụng 64 bộ lọc, mỗi bộ lọc có kích thước 3×3 , sải chân (kích thước 1) và ReLU làm chức năng kích hoạt. Lớp thứ hai sử dụng 64 bộ lọc, mỗi bộ lọc có kích thước 3×3 , bước (kích thước 1), chuẩn hóa hàng loạt, tổng hợp tối đa với các bộ lọc có kích thước 2×2 , loại bỏ là 0,5 (Hình 2) và ReLU làm chức năng kích hoạt. Lớp CNN thứ ba tương tự như lớp đầu tiên nhưng chứa 128 bộ lọc, mỗi bộ lọc có kích thước 3×3 . Lớp thứ tư tương tự như lớp thứ hai nhưng có 128 bộ lọc, mỗi bộ lọc có kích thước 3×3 . Lớp CNN thứ năm tương tự như lớp thứ ba nhưng sử dụng 256 bộ lọc. Lớp thứ sáu đồng nhất với lớp thứ tư nhưng với 256 bộ lọc, mỗi bộ có kích thước 3×3 . Lớp CNN thứ bảy tương tự như lớp CNN thứ năm nhưng sử dụng bộ lọc 512. Lớp thứ tám tương tự như lớp thứ sáu nhưng có bộ lọc 512. Lớp được kết nối đầy đủ đầu tiên có 512 tế bào thần kinh trong một lớp ẩn và ReLU làm chức năng kích hoạt, tiếp theo là sự bỏ qua 0,4. Lớp được kết nối đầy đủ thứ hai có 256 tế bào thần kinh trong một lớp ẩn và Relu dưới dạng chức năng kích hoạt theo sau là bỏ qua 0,4. Lớp được kết nối đầy đủ thứ ba có 128 tế bào thần kinh trong một lớp ẩn, ReLU làm chức năng kích hoạt, tiếp theo là 0,4 người bỏ học. Lớp kết nối đầy đủ cuối cùng, có 7 tế bào thần kinh trong một lớp ẩn và chức năng kích hoạt Softmax. GPU đã được khai thác để tăng thời gian xử lý của các mô hình.

5. Kết quả thực nghiệm

Hiệu suất của các mô hình được xác định bởi độ chính xác của từng mô hình (Hình 4). Biểu đồ của kỷ nguyên chính xác và kỷ nguyên mất mát đã được vẽ cho tất cả các mô hình để đánh giá hiệu suất của chúng. Hình 4 cho thấy ma trận nhầm lẫn của tất cả các mô hình và hỗ trợ đánh giá hiệu suất của chúng tương ứng. Ma trận này cũng hỗ trợ trong việc xác định sự hiện diện có thể xảy ra của / các hư cấu phân loại không chính xác, là những khả năng xảy ra với các hư cấu nhận dạng cảm xúc của con người.



Hình 4. Ma trận nhập nhằng của lần lượt các mô hình: a) mô hình 1, b) mô hình 2, c) mô hình 3

Việc xác định cảm xúc chính xác của khuôn mặt luôn khó khăn. Có thể thấy rõ qua kết quả rằng, mô hình CNN chính xác khi so sánh với mô hình HOG-SVM. Mô hình CNN lấy hình ảnh thực làm đầu vào và mô hình CNN nhập hình ảnh kết hợp có kết quả gần đúng với nhau.

Bảng 1. Độ chính xác của các mô hình theo từng lớp cảm xúc

Biểu cảm	Mô hình 1	Mô hình 2	Mô hình 3
Tức giận	90,67	88,56	88,78
Chán ghét	91,2	91,02	90,43
Sợ hãi	93,04	93,85	85,98
Vui vẻ	91,7	85,04	81,35
Ngạc nhiên	89,43	89,76	90,7
Bình thường	83,1	83,01	89,90
Buồn bã	80,6	94,87	93,6

6. Kết luận

Ba mô hình được phát triển để nhận dạng cảm xúc trên khuôn mặt và màn trình diễn của chúng được đánh giá bằng các kỹ thuật khác nhau. Kết quả ở Bảng 1 cho thấy rằng, các mô hình CNN tốt hơn mô hình SVM. Trái ngược với lần quan sát đầu tiên, việc lấy mẫu xuống không giúp cải thiện hiệu suất của mô hình mặc dù được thoải mái trong việc đạt được độ tin cậy tốt hơn. Mô hình CNN đưa hình ảnh gốc vào ngang bằng với mô hình CNN lấy hình ảnh lai về độ chính xác. Kết quả đáng chú ý về thời gian xem xét là thời gian cần thiết để đạt được độ chính xác thực tế là ít hơn trong trường hợp mô hình CNN nhập hình ảnh thực.

TÀI LIỆU THAM KHẢO/ REFERENCES

- [1] Y. Tiana and T. Kanade, "Recognizing action units for facial expression analysis," *IEEE transactions on pattern analysis and machine intelligence*, vol. 23, no. 2, pp. 97-115, 2001.
- [2] D. Navneet and T. Bill, "Histograms of Oriented Gradients for Human Detection," *Comput. Vision Pattern Recognit.*, vol. 1, pp. 886-893, 2005, doi: 10.1109/CVPR.2005.177.
- [3] A. Durmusoglu and Y. Kahraman, "Facial expression recognition using geometric features," *International Conference on Systems, Signals and Image Processing (IWSSIP)*, 2016, pp. 1-15.
- [4] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large- scale image recognition," *arXiv preprint arXiv:pp. 1409-1556v6 [cs.CV]*, 2015.
- [5] G. Littlewort, M. Frank, and C. Lainscsek, "Automatic recognition of facial actions in spontaneous expressions," *Journal of Multimedia*, vol. 1, pp. 1-9, 2006.
- [6] S. Du and Y. Tao, "Compound facial expressions of emotion," *Proceedings of the National Academy of Sciences*, vol. 15, pp. 111-129, 2014.
- [7] A. Shima and F. Azar, "Convolutional Neural Networks for Facial Expression Recognition," *arXiv:1704.06756v1 [cs.CV]*, 2017.