BÁO CÁO ĐỒ ÁN

MÔN : NHẬP MÔN THỊ GIÁC MÁY TÍNH

FACIAL EXPRESSION IDENTIFICATION SYSTEM WITH EUCLIDEAN DISTANCE OF FACIAL EDGES

NHẬN DIỆN BIỂU CẢM KHUÔN MẶT NGƯỜI BẰNG PHƯƠNG PHÁP TINH KHOẢNG CÁC EUCLIDEAN CỦA CÁC CẠNH KHUÔN MẶT

Người thực hiện :

1512641 – Võ Anh Tuấn

1512647 – Nguyễn Sinh Tú

Nội dung

I. TỔNG QUAN

Trong bài báo cáo này, chúng em xin được trình bày một hệ thống nhận dạng biểu cảm khuôn mặt người. Việc xác định và phân loại dựa trên 7 biểu hiện cơ bản : Happy , Surprise, Fear, Sad, Anger và Neutral. Hệ thống gồm 3 phần chính. Phần thứ nhất là phát hiện khuôn mặt (face feature detection) để trích xuất vùng tâm khuôn mặt. Phần thứ hai là chuẩn hóa ảnh khuôn mặt sau khi trích xuất và trích xuất cạnh khuôn mặt. Ở bước này, chúng ta có một hình ảnh trích xuất cạnh khuôn mặt – được sử dụng để tinh khoảng cách Euclidian của tất cả các điểm ảnh tạo thành các cạnh. Bước thứ ba là phân loại trạng thái cảm xúc khác nhau dựa trên mô hình phân lớp SVM. Ở phần cuối cùng, chúng em sẽ đề cập đến phương pháp sử dụng mạng Convolutional Neural Network để giải quyết bài toán nhận dạng cảm xúc khuôn mặt.

Ngôn ngữ sử dụng : C++ (Hệ thống nhận diện khuôn mặt dùng SVM) và Python (Hệ thông nhận diện khuôn mặt dùng CNN).

II. BẢNG PHÂN CÔNG CÔNG VIỆC

MSSV	HỌ VÀ TÊN	CÁC CÔNG VIỆC PHÂN CÔNG
1512641	Võ Anh Tuấn	Tìm hiểu về hệ thống nhận diện khuôn mặt : - Trích xuất khuôn mặt sử dụng Haar Cascade. - Trích xuất các điểm đặc trưng bằng thuật toán ASM. - Phương pháp Canny tìm góc cạnh khuôn mặt. - Phương pháp tính khoảng cách Euclidian của các cạnh trên khuôn mặt. - CNN - Implement detect Khuôn mặt bằng c++ - Viết báo cáo
1512647	Nguyễn Sinh Tú	Tìm hiểu về hệ thống nhận diện khuôn mặt : - Trích xuất đặc trưng của khuôn mặt bằng cách tính khoảng cách Euclidian giữa những đặc trưng của khuôn mặt - Thuật toán phân lớp SVM để phân lớp 7 biểu cảm khuôn mặt - Sinh dữ liệu để traning - Viết báo cáo - Implement ý tưởng bằng C++ phần SVM, phát hiện góc cạnh, testing

III. GIỚI THIỆU

Việc xác định các trạng thái cảm xúc của một người hoặc xác định biểu hiện trên khuôn mặt (Facial Expression Recognition FER) là một nhiệm vụ khó khăn. Có rất nhiều trạng

trái biểu cảm khuôn mặt khác nhau, sự khác biệt giữa các biểu cảm trên khuôn mặt người thường rất nhỏ và rất khó phân biệt chính xác cảm xúc của khuôn mặt người. Trong thực tế, ngay cả con người cung khó có thể giải thích được nét mặt và cảm xúc của họ. Có ba nhiệm vụ chính trong việc xây dựng hệ thông nhận diện cảm xúc khuôn mặt (Facial Expression Recognition System) là phát hiện khuôn mặt, tách khuôn mặt, phân loại cảm xúc.

Một số nhà nghiên cứu đề cập đến việc phát hiện khuôn mặt thông qua một số phương pháp như neural networks và support vector machine hoặc cách khai thác những đặc tinh không thay đổi như màu da hoặc một số đặc điểm khuôn mặt cụ thể như mắt, mũi, miệng… Việc khai thác, trích xuất các đặc trưng trên khuôn mặt có thể đi đến việc nhận dạng biểu cảm khuôn mặt và giải thích được tâm trạng, cảm xúc của họ.

Chúng ta có thể phân loại các kỹ thuật nhận dạng biểu thức thành hai phương pháp chính là tổng thể (holistic) [2] và các phương pháp hình học (geometric approaches)[1]. Các phương pháp tổng thể bao gồm các kỹ thuật dựa trên mô hình phân tích toàn bộ khuôn mặt, sau đó được coi là một mô hình toàn cục mà không phân tách khuôn mặt thành các thành phần. Một vector đặc trưng, biểu thị thông tin biểu thức, thu được sau khi xử lý hình ảnh khuôn mặt. Nhiều phép biến đổi đã được áp dụng cho việc khai thác tính năng biểu hiện khuôn mặt, như Gabor wavelet [3] curvelets [4], và mô hình nhị phân cục bộ (local binary pattern) [5,6]. Phân tích thành phần chính (principal component analysis) thường được sử dụng để giảm kích thước và mạng nơron đa lớp hoặc máy hỗ trợ vector được sử dụng rộng rãi để phân loại [1, 7].

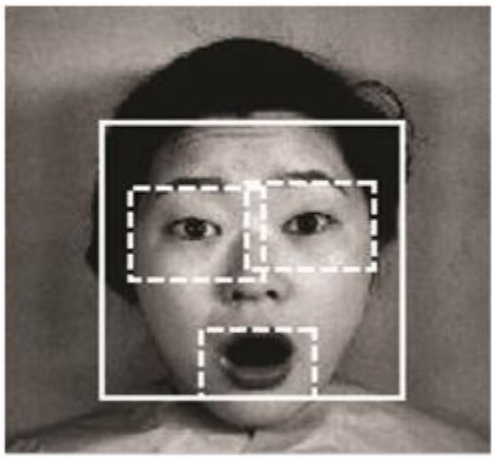
Phương pháp dựa trên tính năng phân tích hoặc hình học chia khuôn mặt thành các thành phần nhỏ hơn hoặc phần phụ mà từ đó các trạng thái biểu cảm có thể được xác định. Các điểm đặc trưng được phát hiện, và khoảng cách giữa các điểm đặc trưng được tính toán để tạo thành các vectơ đặc trưng hoặc xây dựng một mô hình appearance [7, 9, 10]. Hệ thống được trình bày ở đây có thể được xác định là một cách tiếp cận toàn diện. Mô hình được được tổ chức như sau: các bước xử lý liên quan đến hệ thống nhận dạng biểu thức của chúng tôi, như localization of the face và các tính năng của nó, trích xuất các cạnh khuôn mặt (face edge detection), tính toán bản đồ khoảng cách Euclidian của các pixel cạnh (computation of Euclidian distance map of edge pixels) và phân loại theo SVM. Trong phần cuối, chúng tôi trình bày dữ liệu được sử dụng để đánh giá hiệu suất và trình bày kết quả thử nghiệm. Cuối cùng, một kết luận được đưa ra trong phần cuối cùng.

IV. HỆ THỐNG NHẬN DIỆN KHUÔN MẶT - FACIAL EXPRESSION IDENTIFICATION SYSTEM

Trong phần này, ta sẽ chú trọng đến việc nhận diện khuôn mặt (face detection) cùng với bước Extracting the region of interest (Trích xuất vùng quan tâm).

Hệ thống được mô tả dưới đây được thiết kế để xác định bảy nét mặt cơ bản là: niềm vui, sợ hãi, bất ngờ, ghê tởm, buồn bã, giận dữ và biểu hiện trung lập. Tính năng phát hiện khuôn mặt và khuôn mặt được thực hiện theo phương pháp Viola và Jones [8,9].

Các tính năng trên khuôn mặt được coi là mắt và miệng. Các vùng tương ứng của chúng trong khuôn mặt được bản địa hóa theo thuật toán phát hiện Viola và Jones [8,9]. Thuật toán này cung cấp tỷ lệ phát hiện cạnh tranh trong thời gian thực. Mặc dù nó có thể được đào tạo để phát hiện một loạt các lớp đối tượng, nó được thúc đẩy chủ yếu bởi vấn đề phát hiện khuôn mặt.Viola và Jones sử dụng các tính năng giống Haar giống với chức năng cơ sở Haar, trước đây được sử dụng trong lĩnh vực phát hiện đối tượng dựa trên hình ảnh.



Trích xuất khu vực quan tâm (Extracting the region of interest)

Mục đích của phần này là giới hạn vùng quan tâm trên khuôn mặt được bản địa hóa, chăm sóc bao gồm cả các đặc điểm khuôn mặt được phát hiện. Chúng tôi đã bản địa hoá mắt và vùng miệng; sau đó chúng tôi có thể dễ dàng xác định ranh giới của khu vực quan tâm bằng cách xem xét các giới hạn của vùng mắt và miệng như sau:

$$Upper\ Lim = (lim\ left\ eye\ up + lim\ right\ eye\ up)/2$$

$$-(witdh\ left\ eye + witdh\ right\ eye)/2 - cst$$

Trong đó cst là giá trị xác định dựa trên thực nghiệm.

$$Left\ lim = left\ lim\ left\ eye.$$

$$Right\ lim = right\ lim\ right\ eye.$$

$$Lower\ lim = lower\ lim\ mouth + (witdh\ mouth/2)$$

Chuẩn hóa khu vực quan tâm (Normalization of the region of interest)

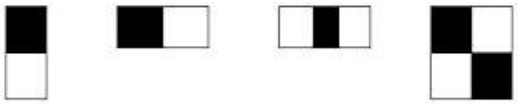
Vùng quan tâm được xác định được chuẩn hóa thành kích thước 60 x 60 pixel. Phương pháp nội suy lân cận gần nhất (nearest neighbor interpolation) được sử dụng để thay đổi kích thước này.

Một số phương pháp phát hiện khuôn mặt :

A. ***Phát hiện khuôn mặt và khuôn mặt được thực hiện theo phương pháp Viola và Jones**:*

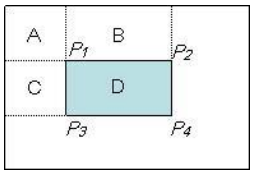
Đây là phương pháp phát hiện khuôn mặt của Paul Viola và Michael Jones đề xuất vào năm 2001. Phương pháp sử dụng đặc trưng Haar-Like kết hợp với máy phân lớp Ada Boost giúp tăng tốc độ của chương trình [10].

Các đặc trưng Haar-like là các hình chữ nhật đen trắng để xác định khuôn mặt người. Gồm 4 đặc trưng cơ bản:



Đặc trưng Haar-Like.

Để tăng tốc độ tính toán và xử lí, Viola-Jones đề xuất một khái niệm mới là Intergral Image (tích phân ảnh). Intergral Image là một mảng hai chiều có kích thước bằng kích thước của ảnh đang xét. Khi đó, tổng mức xám của 1 vùng được tính như sau:

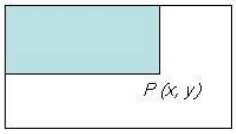


Tính tích phân ảnh

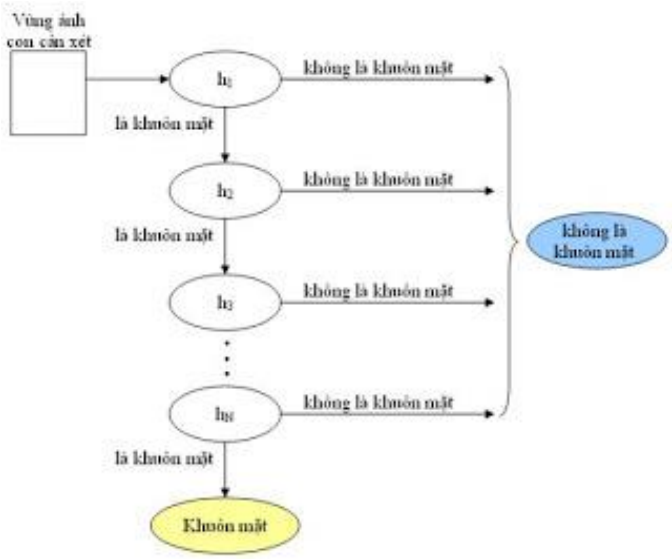
$$D = A + B + C + D - (A + B) - (A + C) + A$$

$$D = P_1(x_1, y_1) + P_2(x_2, y_2) + P_3(x_3, y_3) + P_4(x_4, y_4) - \left(P_1(x_1, y_1) + P_2(x_2, y_2)\right) - \left(P_1(x_1, y_1) + P_3(x_3, y_3)\right) + P_1(x_1, y_1)$$

$$P(x, y) = \sum_{x' \le x, y' \le y} i(x', y')$$



Sau khi tính được tích phân ảnh từ vùng ảnh cần xét, thuật toán Viola-Jones sử dụng bộ phân lớp AdaBoost như Hình 6 để loại bỏ các đặc trưng không cần thiết.



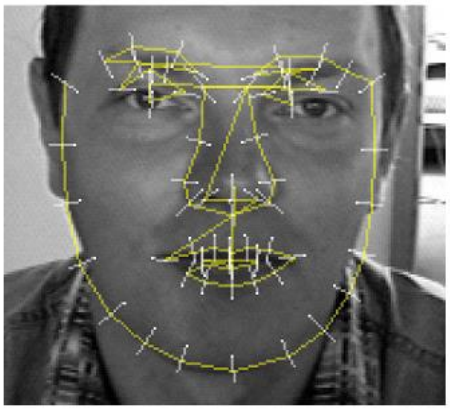
Máy phân lớp AdaBoost [11].

B. *Chiết xuất các điểm đặt trưng bằng thuật toán ASM*

Thuật toán ASM (Active Shape Model) là thuật toán chiết xuất các điểm đặc trưng của một đối tượng cụ thể, do Cootes và Taylor đề xuất[6]. Ý tưởng cơ bản của thuật toán là xác định các điểm đặc trưng của khuôn mặt bằng cách mô tả vùng ảnh xung quanh nó và chỉ ra mối quan hệ của nó với các điểm khác. Để làm được điều này, ASM cung cấp 2 mô hình: Profile model và shape model.

Profile model: Có nhiệm vụ mô tả vùng ảnh xung quanh mỗi điểm đánh dấu (landmark). Nói cách khác, mô hình này chỉ ra rằng, vùng ảnh sẽ "trông như thế nào?" tương ứng với mỗi điểm landmark. Trong quá trình huấn luyện, ta tạo thành các profile training cho mỗi điểm landmark bằng cách lấy mẫu ảnh xung quanh mỗi điểm landmark trong tập ảnh huấn luyện. Trong quá trình tìm kiếm, ta lấy mẫu vùng ảnh của các điểm xung quanh landmark khởi tạo. Điểm được chọn là điểm khớp với các profile training nhất. Tập hợp các điểm tìm được dựa trên landmark khởi tạo, thu được tập các điểm kiểm thử (suggested point) tạo thành shape kiểm thử (suggested shape) [12].

Profile được lấy mẫu dựa trên cường độ mức xám nằm trên các whisker đi qua điểm đặc trưng [12]. Hình dưới biểu diễn các whisker đi qua các điểm đặc trưng của khuôn mặt.



Shape model: Nhiệm vụ của shape model là biến shape kiểm thử được sinh ra từ profile model thành khuôn mặt (face shape) chính thức. Shape model định nghĩa mọi shape bằng công thức sau:

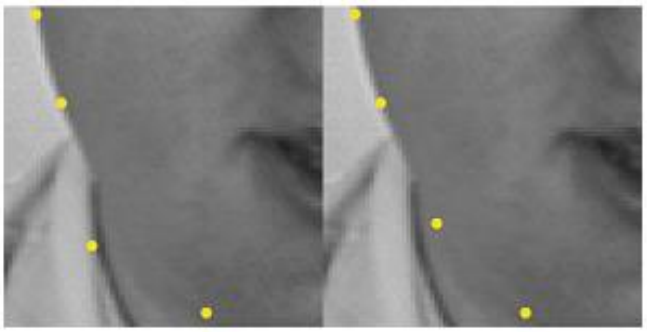
$$\hat{x} = \bar{x} + \Phi b \qquad (1)$$

$\hat{x}$ là shape được sinh ra. $\bar{x}$ là shape trung bình được tính từ các shape huấn luyện đã được align. Φ là ma trận vector riêng (matrix of eigenvectors) của ma trận hiệp phương sai (convariance matrix) tính từ tập huấn luyện.

$\hat{x}$Như vậy, bằng việc biến đổi tham số b, ta có thể sinh ra bất kì shape nào với hình dạng mong muốn. Bằng việc tìm b thích hợp, ta có thể đưa shape kiểm thử trở thành khuôn mặt chính thức. Việc tìm b thỏa mãn khoảng cách sau là nhỏ nhất thu được kết quả như Hình 8. Chi tiết cách tìm b được trình bày trong [12]. Trong đó, T là một phép biến đổi tương đương để đưa về vị trí của khuôn mặt trên ảnh.

$$distance(\boldsymbol{x},\ \boldsymbol{T}(\bar{\boldsymbol{x}} + \boldsymbol{\Phi b}))$$

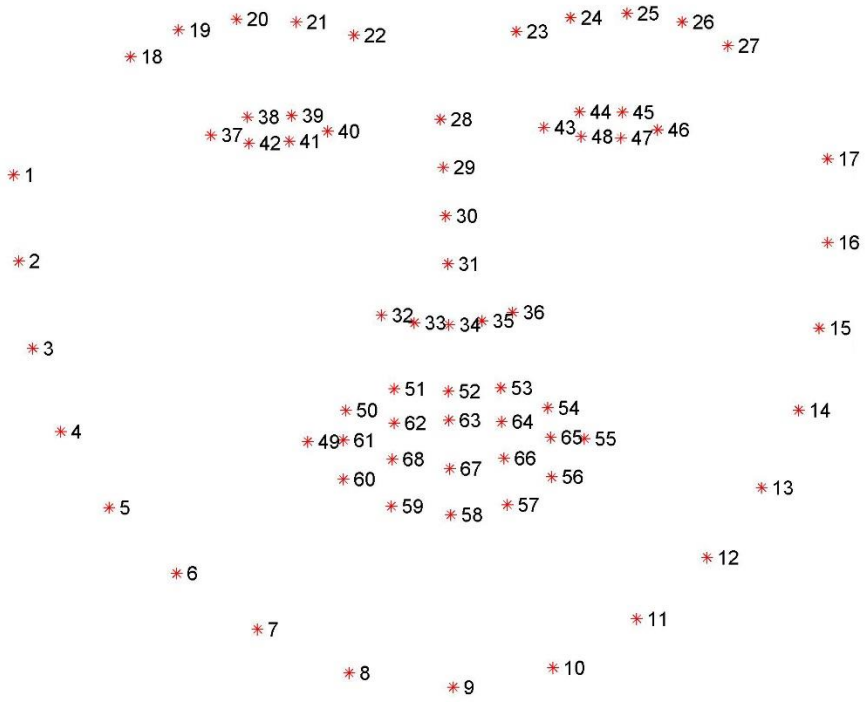
$$\boldsymbol{T}\begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} x_{translate} \\ y_{translate} \end{pmatrix} + \begin{pmatrix} s\cos\theta & s\sin\theta \\ -s\sin\theta & s\cos\theta \end{pmatrix}\begin{pmatrix} x \\ y \end{pmatrix}$$



C. Facial landmarks with dlib, OpenCV

Thiết bị dò tìm khuôn mặt được đào tạo bên trong thư viện dlib được sử dụng để ước tính vị trí của 68 (x, y) - các tọa độ có bản đồ đối với các cấu trúc khuôn mặt trên mặt.

Các chỉ số của 68 toạ độ có thể được hiển thị trên hình ảnh dưới đây:



TRÍCH XUẤT CẠNH KHUÔN MẶT

Biểu hiện khuôn mặt thay đổi với cảm xúc, khuôn mặt dịch những biến thể này bằng các nếp nhăn biểu hiện; bằng cách sửa đổi các đường nét đặc trưng và sự xuất hiện của các nếp nhăn đặc biệt xung quanh miệng và mắt trong khuôn mặt. Đối với điều này, chúng tôi áp dụng các máy dò cạnh Canny [20] trên khu vực mặt của hình ảnh quan tâm. Ngưỡng được chọn để địa phương hóa các đường nét chính trên khuôn mặt và hầu hết các cạnh nhăn biểu hiện.

D. Canny edge detector - Phương phát phát hiện cạnh Canny :

Là một toán tử phát hiện cạnh sử dụng thuật toán nhiều giai đoạn để phát hiện một loạt các cạnh trong hình ảnh [20]. Canny cũng sản xuất một lý thuyết tính toán về phát hiện cạnh. Thuật toán của nó chạy trong năm bước riêng biệt.

- Smoothing - Làm mịn: Làm mờ hình ảnh để giảm tiếng ồn.
- Finding gradients - Tìm gradient: Các cạnh phải được đánh dấu nơi các gradient của hình ảnh có độ lớn lớn.
- Non-maximum suppression – Chặn không cực đại: Chỉ ra các cực đại địa phương mới được đánh dấu là cạnh.

- Double thresholding – Ngưỡng kép: Các cạnh tiềm năng được xác định bằng các ngưỡng.
- Edge tracking by hysteresis: Các cạnh cuối cùng được xác định bằng cách triệt tiêu tất cả các cạnh không được kết nối với cạnh mạnh.

E. Euclidean distance of facial edges image – Tính khoảng cách Euclidian của các cạnh trên khuôn mặt:

Đối với mỗi pixel trong hình ảnh cạnh nhị phân trên khuôn mặt, phép biến đổi khoảng cách gán một số là khoảng cách giữa điểm ảnh đó và điểm ảnh không viền màu trắng (gần) nhất của cạnh trên. Chúng tôi sử dụng chỉ số khoảng cách Euclide, các cạnh trên khuôn mặt. Ma trận khoảng cách có cùng kích thước với hình ảnh cạnh trên khuôn mặt. Nó được chuyển đổi thành một vector bằng cách nối các hàng cho nhiệm vụ phân loại.

F. Phân loại biểu cảm khuôn mặt sử dụng SVM

Chúng ta sử dụng support vector machine SVM để thực hện phân loại các biểu cảm cảm xúc của khuôn mặt

Cho tập dữ liệu có nhãn $T = (x_i, y_i), i = 1 \ldots . l$ where $x_i \in R^n$ và $y_i \in 1, -1$ dữ liệu sẽ input sẽ đươc phân loại theo công thức sau :

$$f(x) = sgn\left(\sum_{i=1}^{l} a_i\, y_i k(x_i, x) + b\right) (2)$$

Trong đó $k(x_i, x)$ là kernel function, a_i là số Lagrange của bài toán tối ưu kép, b là tham số của hypelane tối ưu. Cho một ánh xạ phi tuyến ∅ ánh xạ input data vào không gian features. Kernels có dạng $k(x_i, x) =< \emptyset(x_i).\emptyset(y_j) >$ SVM tìm kiếm một siêu phẵng phân chia tuyến tính với maximal margin để phân chia training data

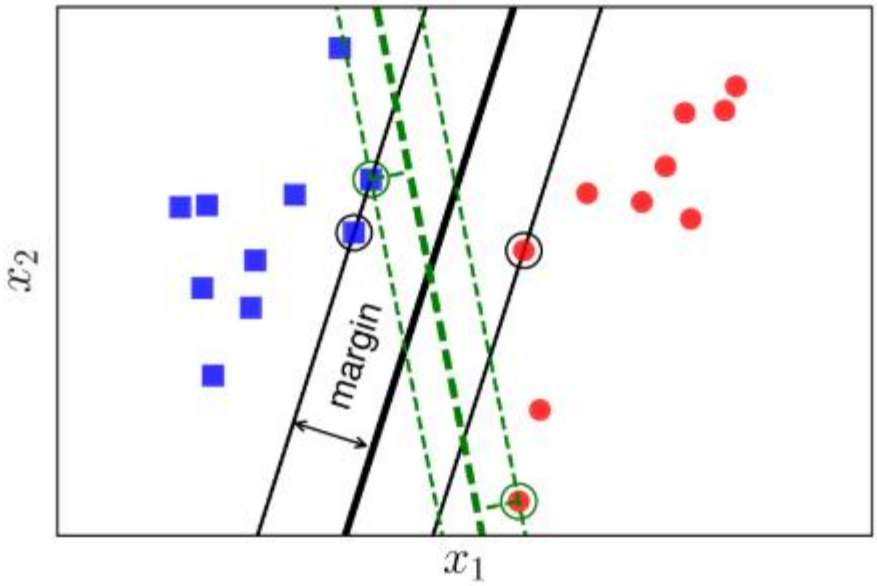


Figure 1: SVM – margin

SVM cho phép chúng ta lựa chon kernel function. Những kernel thường được hay sử dụng linear, polynomial, RBF kernel từ đó SVM tạo ra những binary decision.

Phân loại nhiều lớp được thực hiện bởi chuỗi các phân loại nhị phân cùng nhau. Muticlass SVM được thực hiện dựa trên hai phương pháp phổ biến là “one against all” bằng cách xây dựng SVM cho mỗi class. Thực hiện training phân loại những mẫu thuộc một class từ tất cả các mẫu thuộc các class còn lại. Một phương pháp phổ biến khác lài “one against one”, được thực hiện bằng cách xây dựng SVM cho từng cặp classes. Vì vậy nến ta có c class thì sẽ có $c(c-1)/2$ SVMs được train dể thực hiện phân lớp. Và trong dồ án này chúng tôi sử dụng chiến lược “one against one” nghĩa là chúng ta phải train 21 SVM cho 7 class

G. Kết quả bước đầu

Tập dữ liệu để train : JAFFE (Japanese Female Facial Expression)

Mô tả tập dữ liệu

+ Số lượng : 213 ảnh khuôn mặt với 7 biểu cảm

+ Tên ảnh : ví dụ KA.**SU**1.36.tiff là ảnh có biểu cảm bất ngờ (surprise)

KA.**NE**1.26.tiff là ảnh có biểu cảm trung lập(neutral)

Sử dụng 139 ảnh để train và 74 ảnh để Test

Biểu cảm	Tức giận(Ang)	Ghê tởm(Disg.)	Sợ hãi (Fear)	Hạnh phúc (Hap.)	Trung lập(Neut.)	Buồn (Sad.)	Bất ngờ(Surp.)

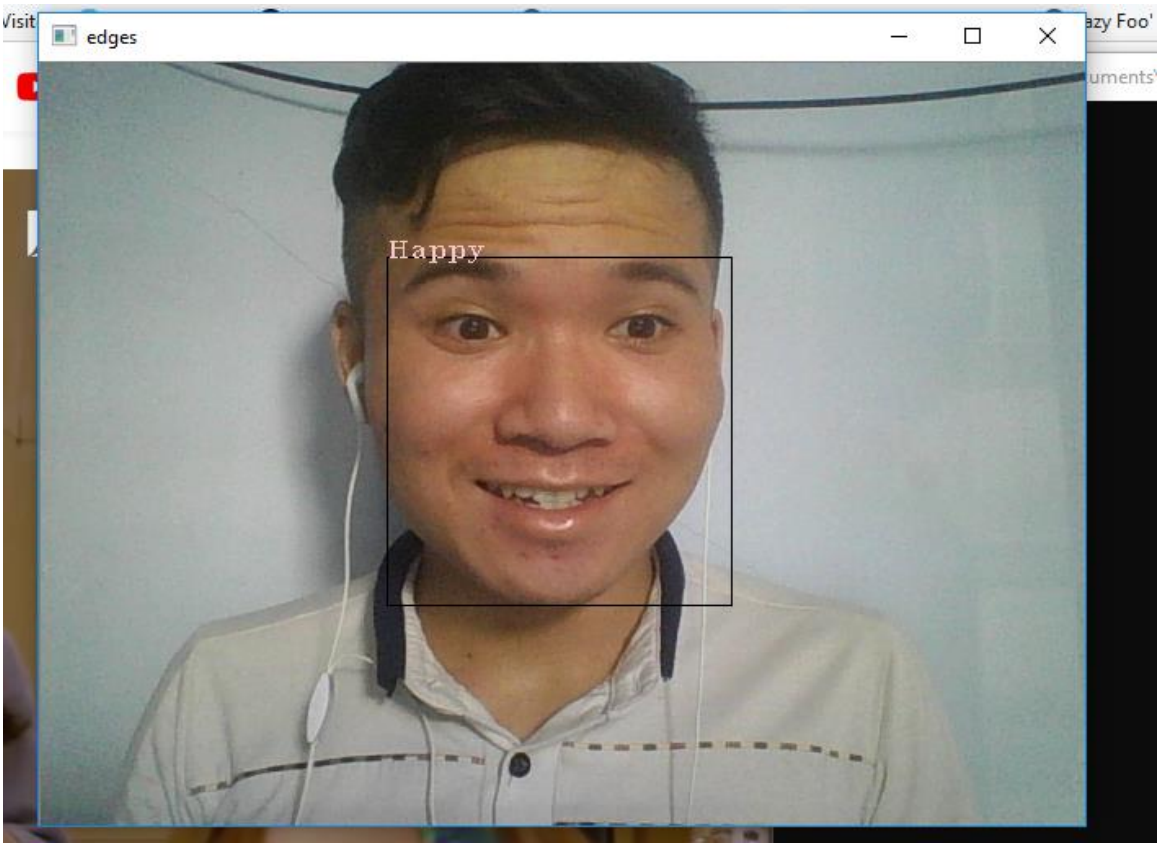
Tỉ lệ đúng(trên ảnh test)	85%	70%	83.2%	100%	91.66%	100%	90%

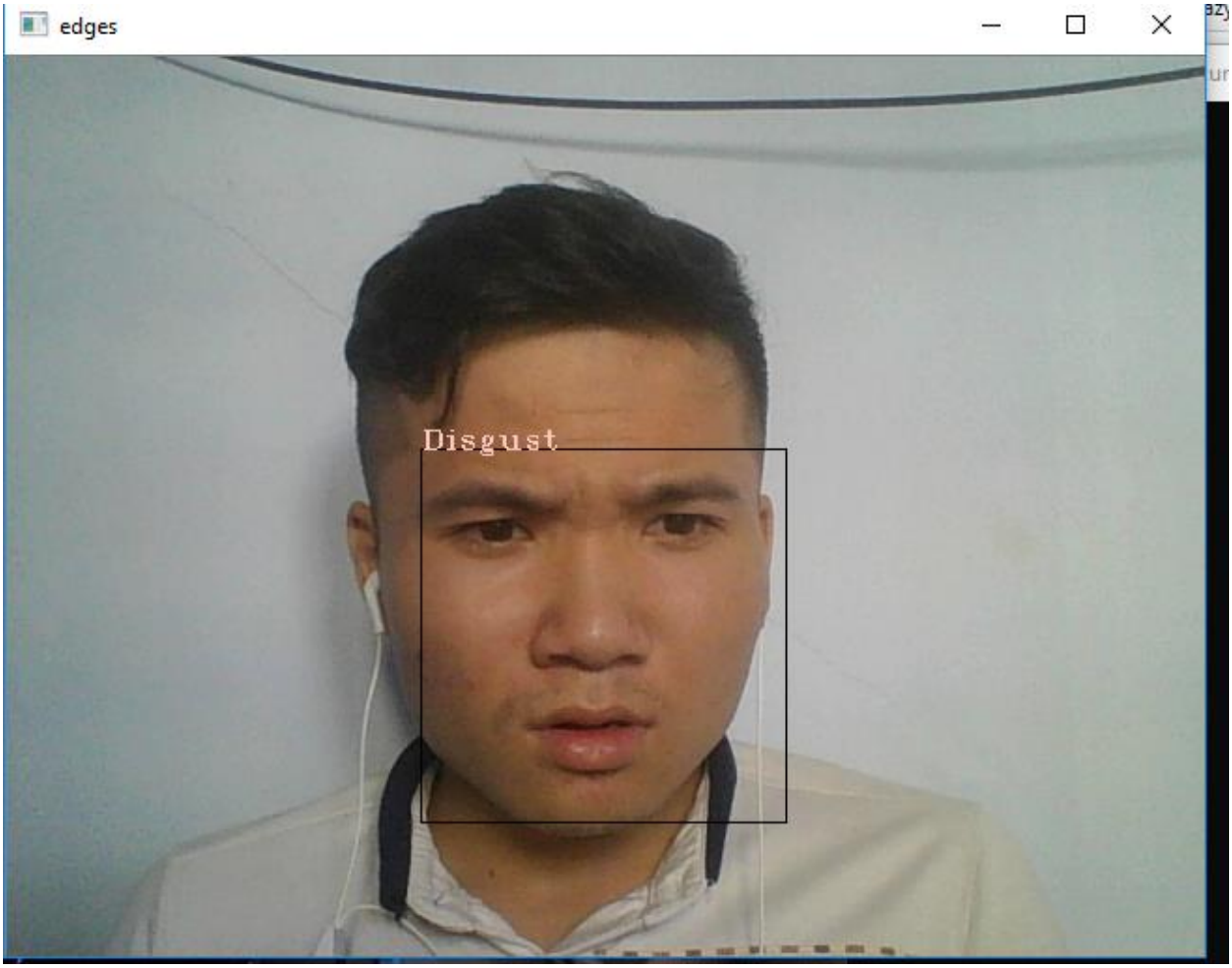
H. Vấn đề gặp phải

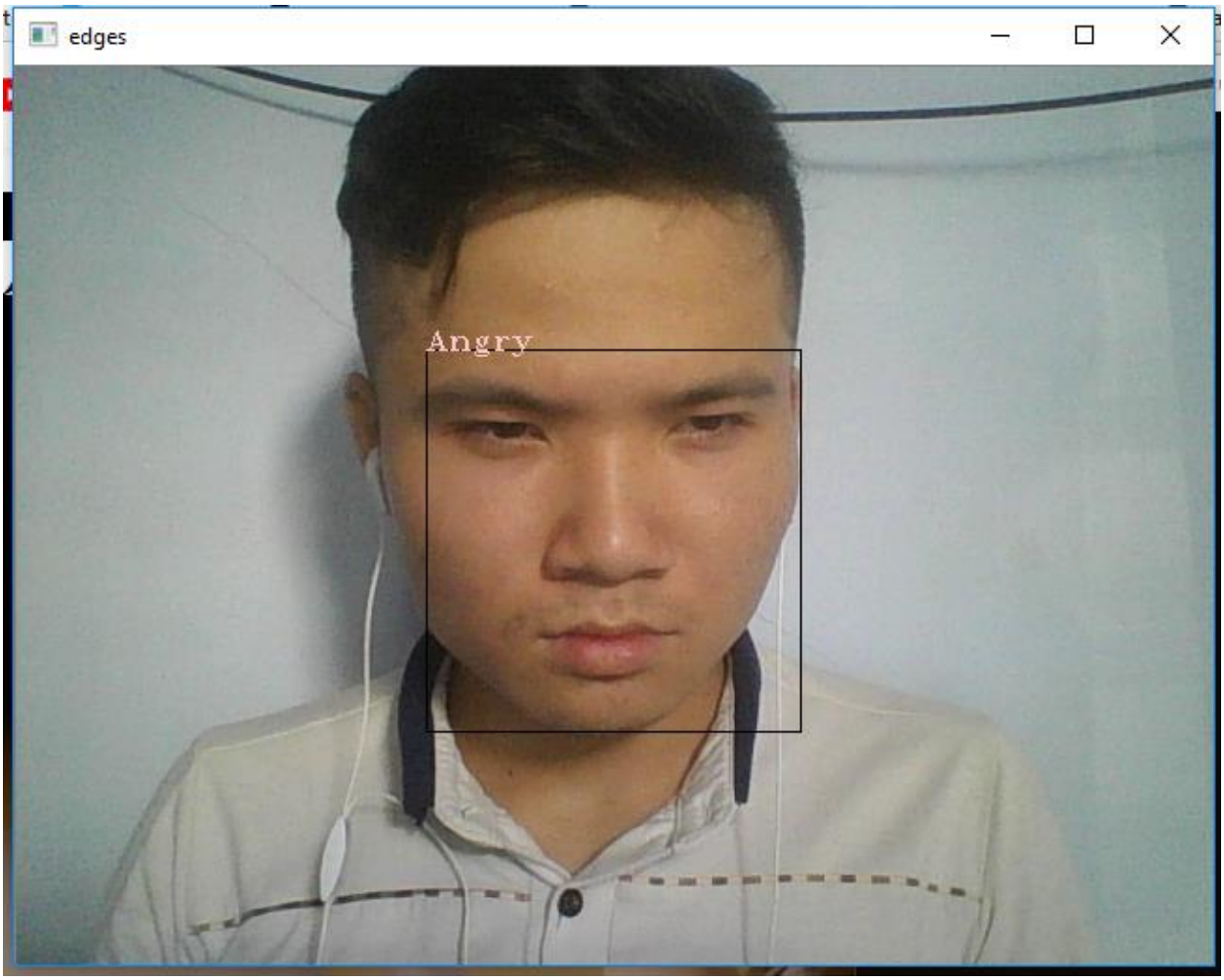
Mô hình SVM bị overfitting chỉ đúng với dữ liệu trong tập test nên khi test trên khuôn mặt thực tế kết quả phân loại sai hoàn toàn

Cách giải quyết : k- fold cross validation, thực hiện model selection để trọn model không bị overfitting và cho kết quả out-of-sample tốt hơn

Kết quả đạt được khi sử dụng SVM với k-fold cross validation




edges
Disgust



I. **Nhận xét :**

Kết quả thực nghiệm khi sử dụng phương pháp trên nhiều lúc vẫn cho kết quả không tốt, sau nhiều lần thực nghiệm chúng em rút ra được một số vấn đề sau

a. Cách trích xuất đặc trưng của khuôn mặt chưa tốt: Việc tính tìm góc cạnh rồi tính khoảng Euclidean rất dễ gây ra nhiễu
b. Dữ liệu tập **ferr** ảnh khá nhỏ, không rõ nét nên việc xác định góc cạnh không chính xác, nếu sữ dụng tập **jaffe** thì model hay bị overfitting dõ dữ liệu traning ít

Từ đó chúng em đã quyết định sử dụng mô hình CNN để training, sẽ được giới thiệu qua ở bên dưới, đấy chỉ làm mô hình em tìm hiểu thêm, mục đích của nhóm vẫn làm muôn cải tiến mô hình trên sao cho đạt kết quả tốt khi chạy trên thực tế, vì mô hình trên khá đơn giản, dể thực và dễ hiểu.

V. Phương pháp dùng mạng học sâu.

Sở dĩ nhóm chúng em đề cập thêm đến phương pháp này là vì trong quá trình nghiên cưu các phương pháp, chúng em thấy rằng phương pháp sử dụng CNN cho kết quá khá ấn tượng trên tập dữ liệu Fer2013 (sẽ đề cập bên dưới) , trong khi đó phương

pháp sử dụng SVM có sự phân lớp sai tương đối nhiều. Nên nhóm chúng em quyết định giới thiệu thêm và sử dụng mô hình CNN.

J. *Dataset*

Về tập dữ liệu huấn luyện , nhóm chúng em sử dụng tập dữ liệu fer2013 (https://www.kaggle.com/c/challenges-in-representation-learning-facial-expression-recognition-challenge/data) . Đây là tập dữ liệu gồm có 35887 ảnh xám có kích thước 48x48 pixels được gán nhãn với 7 loại cảm xúc khác nhau : anger, disgust, fear, happiness, sadness, surprise, and neutral.

Chúng ta phân chia tập dữ liệu ra thành 2 phần. Tập training gồm 28708 ảnh và tập testing gồm 3,589 ảnh.

K. *Mô hình one convolution layer*

Mỗi một neural trong mạng sẽ tính toán tổng các trọng số của tất cả các đầu vào của nó, cộng thêm một độ lệch “bias” để thông qua một số chức năng kích hoạt phi tuyến.

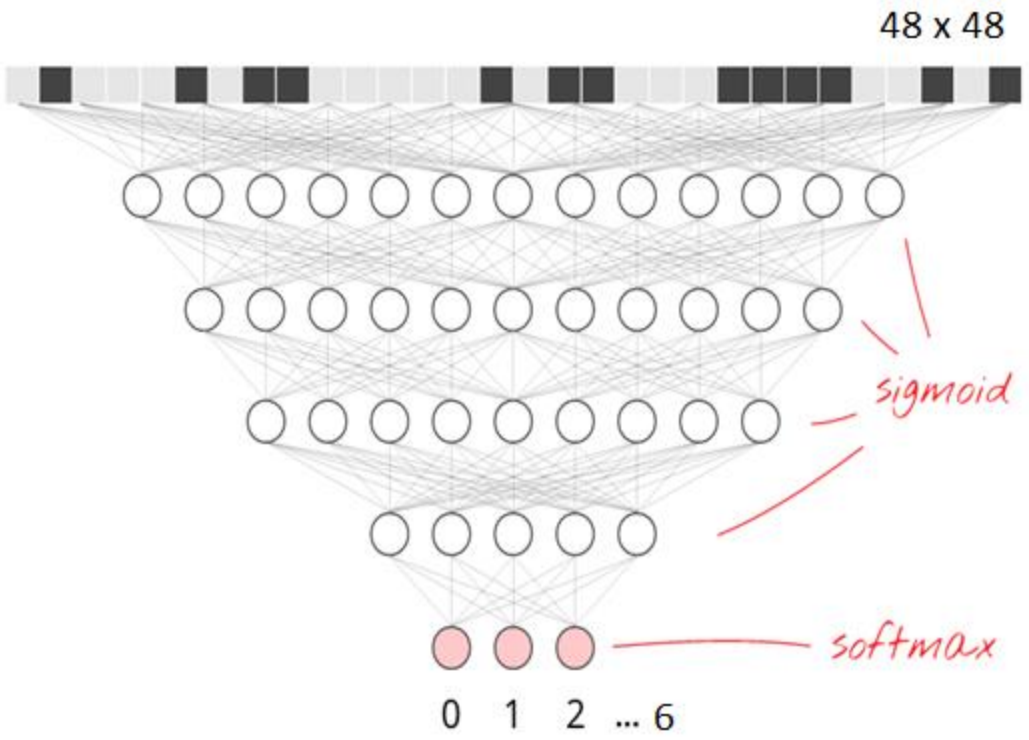
Ở đây chúng ta sẽ thiết kế một mạng nơ ron gồm 1 lớp có 7 neural ở output vì chúng ta muốn phân loại các chữ số thành 7 lớp (từ 0 đến 6). Output layer thường được biểu thị dưới dạng one-vs-rest, tức là tập output chỉ có 1 neural mang giá trị 1, tất cá các neural còn lại mang giá trị 0.

L. *Fully Connected Layer*

Là một lớp Perceptron Multi Layer sử dụng hàm kích hoạt SoftMax ở lớp output. Thuật ngữ "Kết nối hoàn toàn" ngụ ý rằng mỗi neuron trong lớp trước đó được nối với mỗi nơ-ron trên lớp tiếp theo. Output của các lớp Convolutional và Pooling đại diện cho những đặt điểm quang trọng của hình ảnh input. Mục đích của lớp Fully Connected là sử dụng các tính năng này để phân loại hình ảnh đầu vào thành các lớp khác nhau dựa trên bộ dữ liệu huấn luyện.

M. *Mô hình 5 – Layer fully connected neural network*

Để nâng cao độ chính xác trong việc nhận dạng, chúng ta sẽ thêm vào nhiều lớp vào mạng neural để có thể rút trích được nhiều đặc điểm (shape) để nhận dạng. Sau đây là mô hình ví dụ :

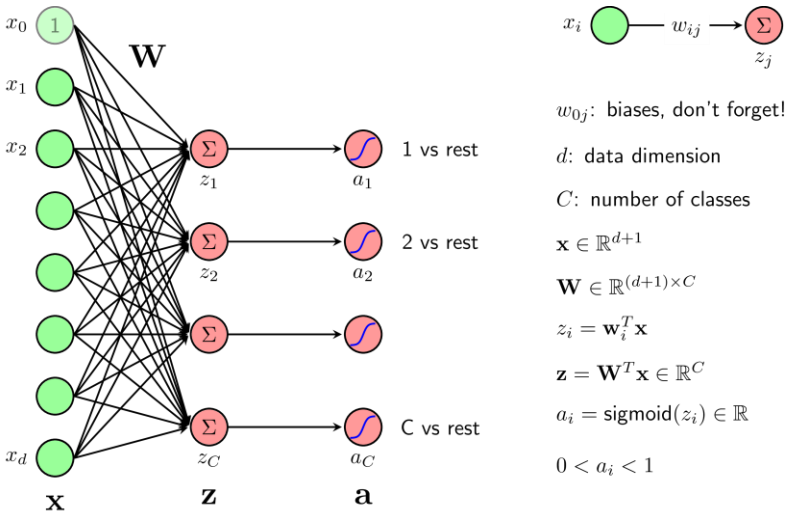


N. ***Sofmax***

Chúng ta cần một mô hình xác suất sao cho với mỗi input X (trong trường hợp này, số lượng của x là 48*48 + 1 = 2305 do có thêm một hệ số bias) , a_i thể hiện xác suất để input đó rơi vào class i. Vậy điều kiện cần là các a_i phải dương và tổng của chúng bằng 1. Để có thể thỏa mãn điều kiện này, chúng ta cần *nhìn vào* mọi giá trị Z_i và dựa trên quan hệ giữa các Z_i này để tính toán giá trị của a_i. Ngoài các điều kiện a_i lớn hơn 0 và có tổng bằng 1, chúng ta sẽ thêm một điều kiện cũng rất tự nhiên nữa, đó là: giá trị $Z_i = W_i^T X$ càng lớn thì xác suất dữ liệu rơi vào class i càng cao. Điều kiện cuối này chỉ ra rằng chúng ta cần một hàm đồng biến ở đây. Nhận thấy rằng Z_i có thể nhận các giá trị cả âm và dương. Một hàm số mượt đơn giản có thể chắc chắn rằng Z_i sẽ trở thành một giá trị dương, và đồng biến là hàm $\exp(z_i) = e^{z_i}$. Điệu kiện mượt để dễ tính đạo hàm hơn sau này. Điều kiện cuối cùng là tổng của các a_i bằng 1 có thể được đảm bảo nếu :

$$a_i = \frac{\exp(z_i)}{\sum_{j=1}^{C} \exp(z_j)} , \forall i = 1,2, \dots, C$$

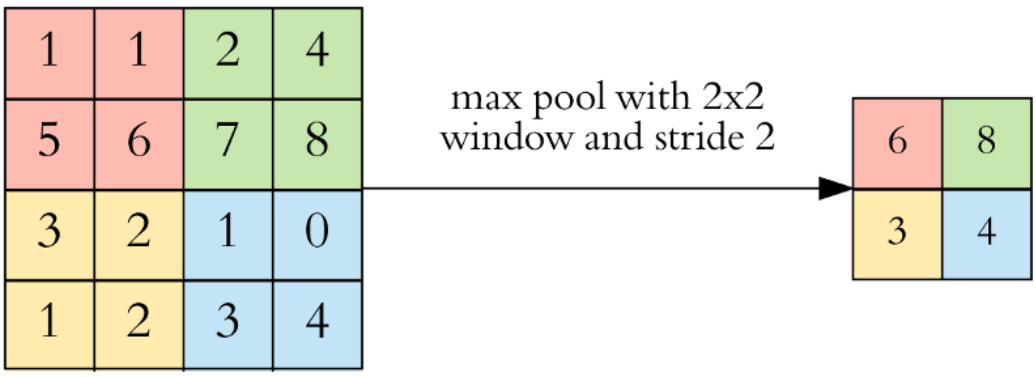
Ta gọi đây là hàm softmax.



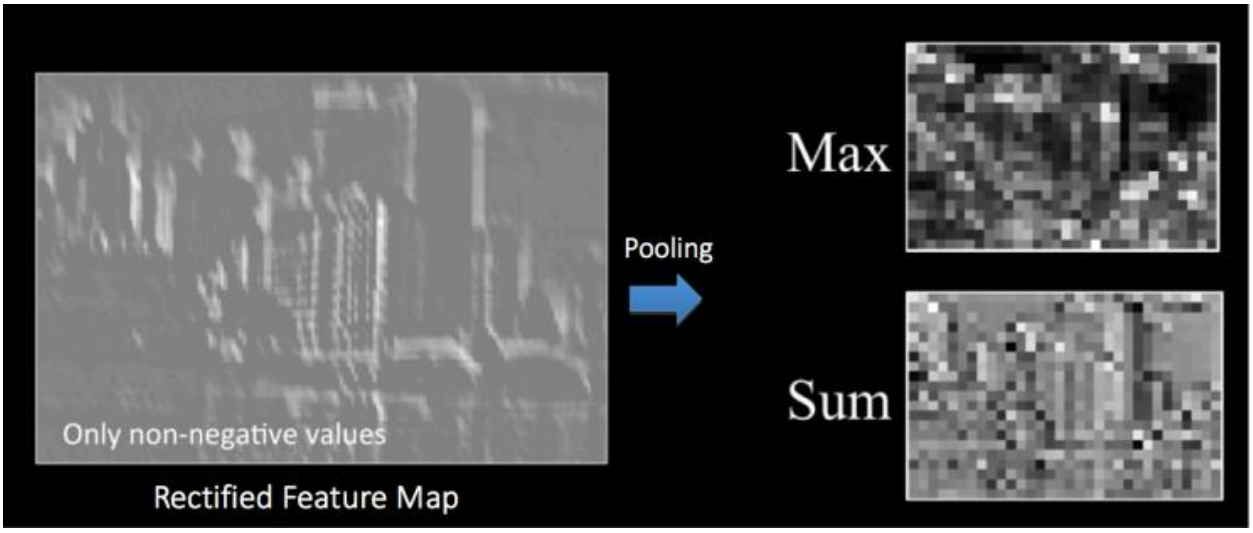
o. ***Max pooling***

Còn được gọi là *Subsampling* hay *Downsampling*, có chức năng làm giảm kích thước của bản đồ tính năng nhưng vẫn giữ lại được các thông tin quan trọng nhất. Có thể gồm nhiều loại khác nhau như Max, Average, Sum.

Mục đích của *pooling* rất đơn giản, nó làm giảm số hyperparameter mà ta cần phải tính toán, từ đó giảm thời gian tính toán, tránh overfitting. Loại pooling ta thường gặp nhất là max pooling, lấy giá trị lớn nhất trong một pooling window. Pooling hoạt động gần giống với convolution, nó cũng có 1 cửa sổ trượt gọi là pooling window, cửa sổ này trượt qua từng giá trị của ma trận dữ liệu đầu vào (thường là các feature map trong convolutional layer), chọn ra một giá trị từ các gía trị nằm trong cửa sổ trượt (với max pooling ta sẽ lấy giá trị lớn nhất).
Hãy cùng nhìn vào ví dụ sau, tôi chọn pooling window có kích thước là 2 * 2, stride = 2 để đảm bảo không trùng nhau, và áp dụng max pooling:



Cụ thể , với hình minh họa dưới đây, khi áp dụng Max Pooling và SumPooling sẽ cho ra 2 ma trận mới :



Mục đích của việc giảm kích thước ma trận là làm giảm số lượng phần tử và các thông số tính toán trong mạng, do đó có thể kiểm soát được tình trạng overfitting. Làm cho mạng không thay đổi với các biến đổi nhỏ.

Đề xuất mô hình mạng [13]

Five – Layer CNN
INPUT (48x48x1)
CONV3-64 BATCHNORM RELU MAX-POOL
CONV3-64 BATCHNORM RELU CONV3-64 BATCHNORM RELU AVG-POOL DROPOUT
CONV3-128 BATCHNORM RELU CONV3-128 BATCHNORM RELU AVG-POOL DROPOUT
FC-1024 BATCHNORM RELU

DROPOUT = 0.2
SOFTMAX

Kết quả :

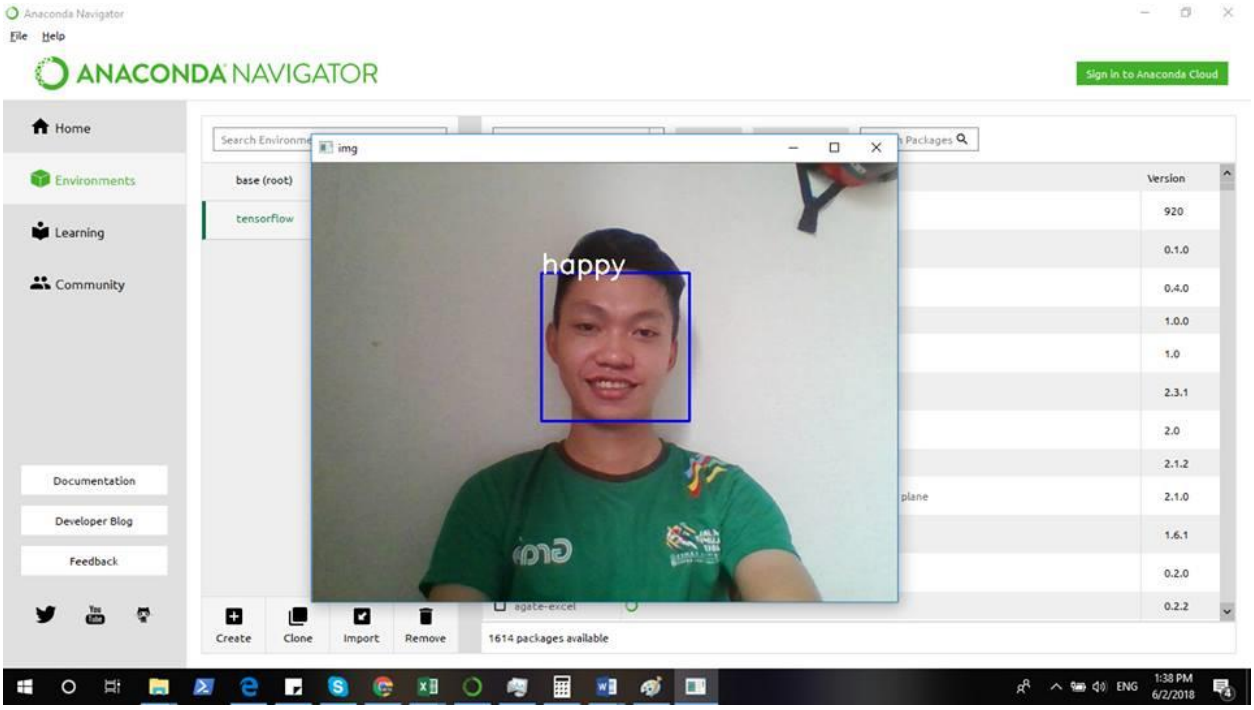
Train loss: 1.1242277171596173

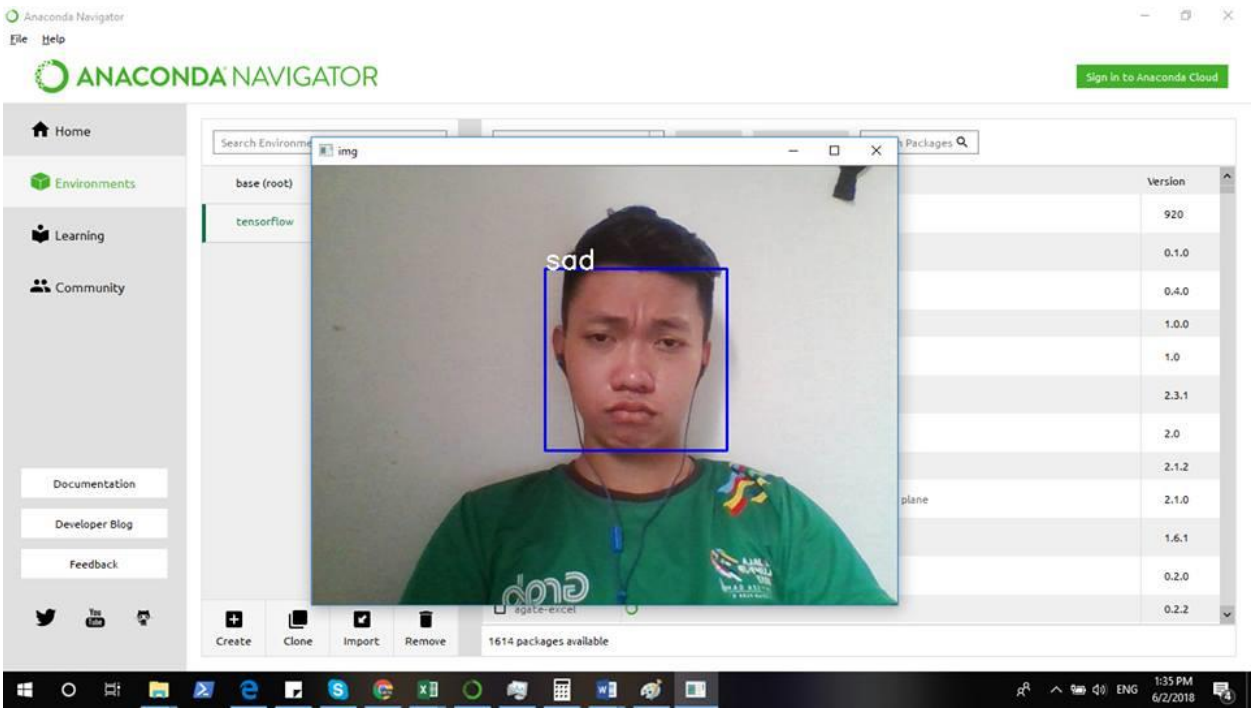
Train accuracy: 57.370767730249405

Test loss: 1.218308902311604

Test accuracy: 53.413207024775964

P. ***Hình ảnh thực nghiệm*** *:*





VI. Source code

Source code SVM : https://github.com/nguyensinhtu/Face-Expression.git

Source code CNN : https://github.com/voanhtuanvn12/Facial-Expression-Recognition-using-CNN

Link full đồ án : https://github.com/nguyensinhtu/1512647_1512641.git

II. TÀI LIỆU THAM KHẢO

[1] M.Pantic and L.J.M. Rothkrantz, “Expert Ststem For Automatic Analysis Of Facilal Expressions,” Image And Vision Computing,vol. 18, pp. 881-905,2000.

[2] P. Viola and M. Jones. “Robust Real Time Object Detection,”Second International Workshop on Statistical and Computational Theories of Vision Modelling Learning Computing and Sampling, Vancouver, Canada, July 13, 2001

[3] Juliano J. Bazzo Marcus V. Lamar, “Recognizing Facial Actions Using Gabor Wavelets with Neutral Face Average Difference,” Sixth IEEE International Conference on Automatic Face and Gesture Recognition (FGR’04), 2004.

[4] Xianxing Wu and Jieyu Zhao, "Curvelet Feature Extraction for Face Recognition and Facial Expression Recognition", Sixth International Conference on Natural Computation, pp.1212-1216, 2010.

[5] S. Singh and R. Maurya, A. Mittal,"Application of Complete Local Binary Pattern Method for Facial Expression Recognition",IEEE Proceedings of 4th International Conference on Intelligent Human Computer Interaction, Kharagpur, India, December 27-29, 2012.

[6] R.Hablani,N. Chaudhari andS.Tanwani "Recognition of Facial Expressions using Local Binary Patternsof Important Facial Parts," International Journal of Image Processing (IJIP), vol. 7 (2), 2013.

[7] D. Arumugan, S. Purushothaman.”Emotion Classification Using Facial Expression,” International Journal of advanced Computer Science and Application, vol. 2, n° 7,2011.

[8] P. Viola and M. Jones. “Robust Real Time Object Detection,”Second International Workshop on Statistical and Computational Theories of Vision Modelling Learning Computing and Sampling, Vancouver, Canada, July 13, 2001.

[9] P. Viola, and M. J. Jones, "Rapid Object Detection using a Boosted Cascade of Simple Features," Proceedings of the 2001 IEEE Computer Society Conf. on Computer Vision and Pattern Recognition, vol.1, pp.511–518, 2001.

[10]P. Viola, M. Jones, “Rapid Object Detection using a Boosted Cascade of Simple Features,” International Journal of Computer Vision 57(2), 137–154, 2004.

[11] Comvisap, “Phát hiện mặt người dựa trên các đặc trưng Haar-like,”

http://www.comvisap.com/2012/01/phat-hien-mat-nguoi-dua-tren-cac-ac.html.

[12] T.F. Cootes, C.J.Taylor, Statistical Models of Appearance for Computer Vision, Imaging Science and Biomedical Engineering, University of Manchester, March 8, 2004.

[13] Facial Expression Recognition with Convolutional Neural Networks, Arushi Raghuvanshi Stanford University, Vivek Choksi Stanford University.