XÂY DỰNG CÔNG CỤ TỰ ĐỘNG NHẬN DẠNG CẢM XÚC DỰA TRÊN MẶT NGƯỜI

Trầm Vũ Kiệt1

1Khoa Công nghệ Thông tin & Truyền thông, Trường Đại học Cần Thơ

# TÓM TẮT

Trong bài báo này, tôi trình bày thuật toán thị giác máy tính và máy học để tự động phát hiện và nhận dạng cảm xúc con người dựa trên biểu cảm gương mặt, sử dụng đặc trưng cục bộ HOG, mạng nơ-ron nhân tạo và máy học vec-tơ hỗ trợ. Hệ thống có khả năng phát hiện gương mặt và nhận dạng cảm xúc con người với sáu cảm xúc cơ bản: vui vẻ, buồn bực, ngạc nhiên, giận dữ, sợ hãi và kinh tởm. Qua thực nghiệm với độ chính xác của tập huấn luyện là khoảng *99%* và độ chính xác của tập kiểm thử là *94%*.

**Từ khóa:** *Nhận dạng cảm xúc gương mặt, phân tích thành phần chính, đặc trưng HOG, mạng Nơ-ron nhân tạo, máy học véc-tơ hỗ trợ.*

# 1 GIỚI THIỆU

Ứng dụng công nghệ thông tin, xử lý ảnh và thị giác máy tính trong việc giải quyết các vấn đề thực tế đang là một lĩnh vực hấp dẫn ở thời điểm hiện tại. Các ứng dụng thông minh giúp tăng hiệu quả và độ chính xác của công việc, giảm thiểu thời gian, sức lực và nhân lực. Phát hiện và nhận dạng cảm xúc con người dựa trên gương mặt là một trong những giai đoạn chính hỗ trợ trong xây dựng các hệ thống nhân tạo thông minh. Các thành quả nghiên cứu được phát triển và ứng dụng thực tế trong hầu hết các lĩnh vực của đời sống xã hội. Thị trường phần mềm phân tích cảm xúc được dự báo có thể đạt tới 10 tỉ USD vào năm 2020 so với 100 triệu USD ở năm 2016 (theo Liên hiệp các Hội Khoa học và Kĩ thuật Sóc Trăng). Trên thế giới đã có nhiều công trình nghiên cứu của các nhà khoa học như Pascal Ackermann et al., ; Nishant Hariprasad Pandey et al., 2013; N Dharmesh et al., 2015; Monika Dubey et al. 2016. Một số đề tài nghiên cứu trong nước như của Nguyễn Thanh Bình et al., 2017; Lê Hoàng Thái và Võ Nhật Trường, Tạ Tiến Thành.

Bài báo này trình bày phương pháp phát hiện và nhận dạng cảm xúc bằng cách sử dụng nhiều phương pháp kết hợp như phân đoạn ảnh, phát hiện biên, phân tích hình dáng đối tượng để phát hiện vùng ứng biên có phải là gương mặt người hay không. Sau đó sử dụng các loại đặc trưng như đặc trưng HOG để tìm ra các đặc điểm nổi trội của từng lớp cần nhận dạng. Cuối cùng là các đặc trưng này được dùng vào việc huấn luyện mạng nơ-ron nhân tạo hay máy học véc-tơ hỗ trợ để tạo ra bộ nhận dạng cảm xúc. Tiến trình tổng quan của hệ thống được trình bày như hình 1. Đầu tiên, ảnh đầu vào sẽ được tiền xử lý, chuyển về mức xám để làm nổi bật các đối tượng gương mặt trong ảnh. Sau đó, các vùng ứng viên sẽ được rút trích đặc trung HOG, các đặc trưng rút ra sẽ được dùng làm đầu vào để phân lớp với mô hình mạng nơ-ron nhân tạo hay máy học véc-tơ hỗ trợ và tạo ra tập huấn luyện, đây là quá trình huấn luyện tập bộ dữ liệu nhận dạng. Quá trình nhận dạng cảm xúc của hình ảnh mới cũng được thực hiện tương tự, sau khi các đặc trưng được rút trích thì sẽ được sử dụng để làm cơ sở phân lớp cảm xúc trên gương mặt được phát hiện. Kế tiếp, các loại cảm xúc phổ biến nhất của con người được công nhận trên toàn cầu sẽ được giới thiệu. Phần tiếp theo của bài báo trình bày phương pháp nghiên cứu và quá trình thực hiện với các kỹ thuật rút trích đặc trưng HOG của các vùng ứng viên và huấn luyện bộ nhận dạng. Các kết quả thực nghiệm và kết luận, đánh giá cũng như định hướng sẽ được giới thiệu ở phần cuối của bài báo.

Ảnh đầu vào

Phát hiện vùng mặt

Xác định landmarks

Trích chọn đặc trưng

Huấn luyện SVM

Nhận dạng

Huấn luyện ANN

Hiển thị kết quả

# 2 PHƯƠNG PHÁP NGHIÊN CỨU

## 2.1 Các loại cảm xúc của con người

Cảm xúc có tác động rất lớn đến cuộc sống của con người. Cảm xúc rất phức tạp, có thể thay đổi nhanh chóng, một người có thể có hơn một cảm xúc tại một thời điểm. Vì vậy, cảm xúc là gì và con người có tất cả bao nhiêu cảm xúc là các câu hỏi chưa có một đáp án thuyết phục. Theo Paul Ekman và Wallace V. Freisen, có sáu cảm xúc cơ bản thuộc tất cả các nền văn hóa khác nhau trên thế giới, bao gồm vui vẻ, buồn bã, ngạc nhiên, giận dữ, sợ hãi và kinh tởm.



*Hình 1: Sáu loại cảm xúc cơ bản của con người theo Paul Ekman*

TS. Rachael Jack của đại học Glasgow, Vương quốc Anh cho rằng con người chỉ có bốn loại cảm xúc cơ bản khi mà những biểu hiện của các cơ trên gương mặt trong hai cảm xúc sợ hãi và ngạc nhiên là gần như nhau, tương tự với giận dữ và ghê tởm. Thật sự là hai loại cảm xúc trong từng cặp ở trên có sự giống nhau trong quá trình vận động các nhóm cơ mặt để hình thành cảm xúc, chúng chỉ khác nhau khi được thể hiện đầy đủ và xong xuôi, vì thực chất chúng là những cảm xúc khác nhau.

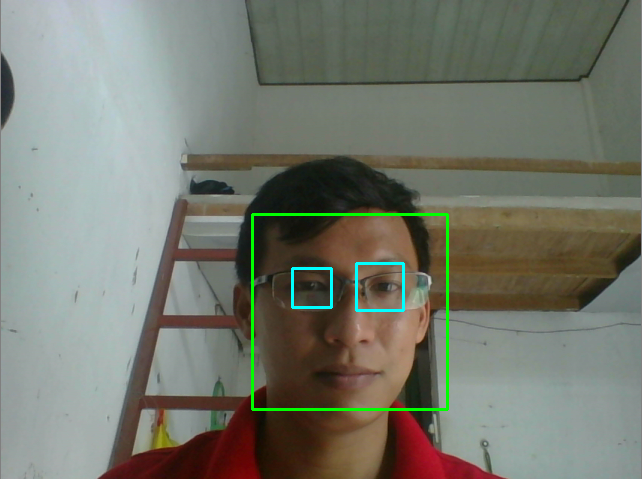
Đề tài được thực hiện để nhận dạng sáu loại cảm xúc cơ bản nhất của con người (vui vẻ, buồn bã, ngạc nhiên, sợ hãi, ghê tởm, giận dữ), bên cạnh đó còn có cảm xúc bình thường, có nghĩa là không cảm xúc.

## 2.2 Tiền xử lý ảnh

Tiền xử lý ảnh là một bước quan trọng và không thể thiếu trong các ứng dụng liên quan đến xử lý ảnh. Để hỗ trợ cho quá trình huấn luyện bộ nhận dạng cảm xúc được hiệu quả hơn, ảnh đầu vào được thay đổi hình dạng về dạng hình vuông với kích thước tối thiểu là 256x256 pixel, nhưng vẫn giữ nguyên được tỉ lệ hình ảnh ban đầu, bằng phương pháp điền thêm phần nền với màu đen. Sau đó, ảnh được chuyển về ảnh xám và được sử dụng như dữ liệu đầu vào.

## 2.3 Phát hiện vùng ứng viên

Sử dụng bộ phát hiện gương mặt với đặc trưng Haar-like và bộ phân lớp Cascade để phát hiện gương mặt trong ảnh hoặc video. Hình 2 trình bày kết quả dò tìm và phát hiện gương mặt trong ảnh. Gương mặt được phát hiện và được bao quanh bởi hình vuông màu xanh lá. Phần màu xanh lục là vùng mắt được tìm thấy.



*Hình 2: Vùng ứng viên gương mặt và vùng mắt được phát hiện*

Hình ảnh vùng ứng viên là một hình vuông với kích thước cố định cho tất cả các đối tượng khác nhau. Vấn đề này đảm bảo cho việc huấn luyện được thực hiện hiệu quả hơn.

## 2.4 Đặc trưng HOG

**HOG** (Histogram of Oriented Gradients) là một bộ mô tả tính năng được sử dụng để phát hiện đối tượng trong thị giác máy tính và xử lý ảnh, được tính toán trên một lưới dày đặc các ô và chuẩn hóa sự tương phản giữa các khối để nâng cao độ chính xác. Đặc trưng HOG được đề xuất bơi N Dalal et al., 2005. Ý tưởng chính của thuật toán này là dựa trên việc đếm số lần xuất hiện của các hướng gradient (*gradient orientation*) trong các vùng cục bộ của ảnh.

Các thông số về hình dáng và bề ngoài của các đối tượng cục bộ trong ảnh được mô tả bằng cách sử dụng thông tin phân bố của các gradients cường độ (intensity gradients) và các hướng của cạnh (edge directions). Thuật toán **HOG** được tiến hành bằng cách chia nhỏ một bức ảnh thành các vùng con nhỏ hơn, gọi là các ô (cell), và thực hiện tính từng biểu đồ về hướng (histogram of gradients) cho từng điểm trong ô. Khi tổng hợp các biểu đồ này lại sẽ được biểu đồ biểu diễn hình ảnh ban đầu.

Để tăng cường hiệu quả nhận dạng, thay vì làm việc trên từng ô thì tiến hành làm việc trên từng khối (block) chứa các ô. Các biểu đồ cục bộ được chuẩn hóa về độ tương phản bằng cách tính một ngưỡng cường độ của khối. Giá trị ngưỡng này được sử dụng để chuẩn hóa tất cả các ô trong khối. Việc này cho kết quả là các vector đặc trưng có tính bất biến cao hơn với các ảnh hưởng của điều kiện ánh sáng. Phần tiếp theo sẽ trình bày các bước áp dụng đặc trưng HOG trong đề tài.

### 2.4.1 Các bước trích đặc trưng HOG trên ảnh đầu vào

**Bước 1**: Chuẩn hóa hình ảnh trước khi xử lý.

**Bước 2**: Tính gradient theo *x* và *y*.

Nhận chập ảnh gốc với hai nhân một chiều *Dx = [-1 0 1]* và *Dy = [1 0 -1]T*. Việc này tương ứng với việc lấy đạo hàm của ảnh theo hai chiều Ox và Oy. Với ảnh I đầu vào, ta tính được hai ảnh đạo hàm riêng theo hai hướng với công thức: *Ix = I\*Dx* và *Iy = I\*Dy*

Tính cường độ ảnh: *G = sqrt(Ix2 + Iy2)* và hướng của ảnh: . Dựa vào G và sẽ tính được một biểu đồ cường độ gradient, với các cột dựa trên và trọng số dựa trên G

**Bước 3**: Thống kê thành phần véc-tor có cùng trọng số trong mỗi ô và vẽ một histogram cho mỗi ô.

**Bước 4**: Chuẩn hóa các khối.

Chia hình ảnh theo các khối, mỗi khối chứa các ô. Các khối này thường có kích thước là 2x2 hoặc 3x3 để dễ tính toán. Các khối này sẽ chồng lên nhau. Tiếp theo, tiến hành thu thập và ghép các biểu đồ của từng ô trong khối.

Gọi *v* là véc-tơ cần chuẩn hóa chứa tất cả các histogram của một khối, *||vk||*là giá trị chuẩn của nó theo các chuẩn *k = 1, k = 2* và hằng số nhỏ *e*. Khi đó, các giá trị chuẩn hóa có thể được tính bằng một trong ba công thức:

L2-norm:

L1-norm:

L1-sqrt:

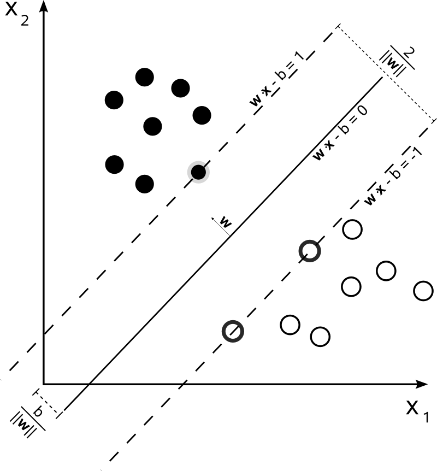
**Bước 5**: Thu thập tất cả các biểu đồ cường độ gradient định hướng của các khối để tạo ra vector tính năng cuối cùng.

### 2.4.2 Áp dụng đặc trưng HOG cho vùng ứng viên

Sử dụng đặc trung HOG được huấn luyện trong thư viện dlib để xác định các landmark của vùng ứng viên, tìm ra đặc trưng cơ bản nhất của các thành phần trên gương mặt. Với từng vị trí của các landmark sẽ tìm được cảm xúc biểu lộ tương ứng.

## 2.5 Quá trình phân lớp

Phân lớp là một giai đoạn quan trọng tiên quyết trong các bài toán nhận dạng. Quá trình phân lớp thật sự là việc sử dụng các thuật toán máy học để gán dữ liệu đầu vào vào lớp mong muốn. Với bài báo này, mô hình phân lớp Support Vector Machine được sử dụng để chia các cảm xúc cơ bản thành các lớp, với mục đích nhận dạng cảm xúc của gương mặt vừa được phát hiện như vùng ứng viên ở bước trên. Tổng quát về Máy học véc-tơ hỗ trợ được trình bày như một phương pháp học có giám sát dùng cho phân lớp và hồi quy.



*Hình 3: Hình ảnh minh họa giải thuật phân lớp của Support Vector Machine*

Hai lớp đại diện bởi hình tròn màu đen và trắng được phân ra thành hai lớp nhờ siêu phẳng wx – b = 0 và hai siêu phẳng hỗ trợ song song.

Cũng trong bài báo này, mạng nơ-ron nhân tạo được dùng để huấn luyện để tạo ra bộ phân lớp để nhận dạng các loại cảm xúc như một giải thuật so sánh với độ hiệu quả của SVM. Bài báo sử dụng mạng nơ-ron tích chập Convolutional Neutral Network – là một mạng nơ-ron Multilayer Perception đa tầng với cấu trúc đặc biệt, là một trường hợp đặc trưng và phổ biến của mạng nơ-ron nhân tạo.

MLP là mạng nơ-ron truyền thẳng, bao gồm một tầng đầu vào, một tầng đầu ra và không giới hạn số tầng ẩn (các tầng không phải là tầng đầu vào và tầng đầu ra). Số lượng tầng ẩn được điều chỉnh để độ chính xác mô hình đạt được ở mức tốt nhất có thể, thông thường chỉ dao động từ 1 đến 5. CNN hay còn gọi là mạng nơ-ron tích chập là một mô hình học sâu tiên tiến, là một cấu trúc đặc biệt của mạng MLP ở chỗ Convolution và Pooling. Hình 4 minh họa cấu trúc của một mạng CNN.

Các hình vuông thể hiện cho các tầng, hai quá trình tích chập và tổng hợp được thực hiện xen kẽ nhau. Tầng đầu vào không chứa các nơ-ron mà chứa các giá trị ban đầu được cung cấp bởi người dùng. Giá trị này được xử lý trong mạng và tổng hợp thành một đặc trưng đại diện cho một lớp. Các lớp này tương ứng với các loại cảm xúc cần phải phân biệt của yêu cầu hệ thống.

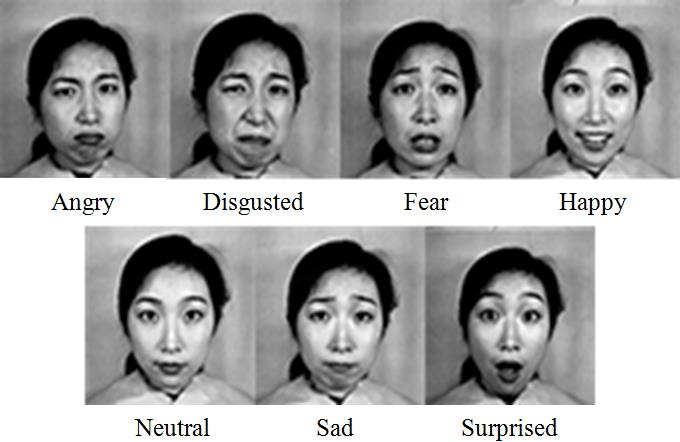


*Hình 4: Mô hình mạng nơ-ron CNN đa tầng*

# 3 KẾT QUẢ VÀ THẢO LUẬN

## 3.1 Các tập cơ sở dữ liệu

Các tập dữ liệu ảnh được sử dụng là JAFFE, Cohn-Kanade và các hình ảnh tự thu thập từ cá nhân tác giả. **JAFFE** là một cơ sở dữ liệu chứ 213 hình ảnh gương mặt của 10 người mẫu nữ Nhật Bản thể hiện 7 loại cảm xúc cơ bản. JAFFE được chuẩn bị bởi Michael Lyons, Miyuki Kamachi và Jiro Gyoba.

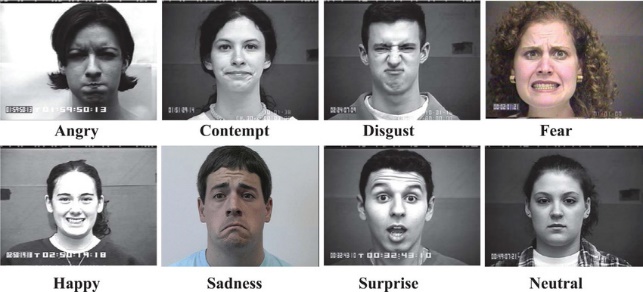


*Hình 5: Một số hình ảnh đại diện với cảm các xúc tương ứng trong tập dữ liệu JAFFE*

**Cohn-Kanade** (CK và CK+): Là cơ sở dữ liệu biểu cảm gương mặt dựa trên các đơn vị vận động gương mặt, được dùng cho nghiên cứu trong tự động phân tích và tổng hợp hình ảnh gương mặt, và cho các lĩnh vực học về cảm xúc. Cohn-Kanade hiện tại có hai phiên bản:

*Phiên bản 1:* **CK**. Là phiên bản đầu tiên, bao gồm 486 chuỗi hình ảnh cảm xúc gương mặt của 97 đối tượng. Mỗi đối tượng thể hiện một chuỗi cảm xúc bắt đầu với không cảm xúc và tăng dần để đạt đến cảm xúc mạnh mẽ nhất. Từng hình ảnh cảm xúc trong mỗi chuỗi được mã hóa và gán nhãn.

*Phiên bản 2:* **CK+**. Chứa 593 chuỗi hình ảnh, mỗi hình ảnh thể hiện một trong bảy loại cảm xúc: vui vẻ, buồn bã, ngạc nhiên, giận dữ, sợ hãi, kinh tởm và không cảm xúc. Các hình ảnh hầu hết là ảnh mức xám với độ phân giải quy định là 640x490 pixels.



*Hình 6: Các hình ảnh đại diện với các cảm xúc tương ứng trong tập Conh-Kanade*

Bên cạnh đó, hệ thống còn sử dụng một số hình ảnh tự sưu tầm để huấn luyện và nhận dạng.

## 3.2 Kết quả huấn luyện

Với tập dữ liệu Cohn-Kanade, mô hình mạng nơ-ron tích chập CNN được sử dụng để huấn luyện mô hình phân lớp cảm xúc. Huấn luyện 4,234 ảnh và kiếm thử 1058 ảnh. Sử dụng mạng nơ-ron tích chập 5 tầng. Tầng 1 nhận vào véc-tơ 34,944 chiều. Tầng 2 xử lý 614,656 nơ-ron. Tầng 3 huấn luyện 885,120 nơ-ron. Tầng 4 xử lý 1,327,488 nơ-ron. Tầng 5 tính toán trên 884,992 nơ-ron. Kết quả huấn luyện và kiểm tra được trình bày ở bảng 1.

*Bảng 1: Kết quả huấn luyện mạng CNN*

|  |  |  |  |
| --- | --- | --- | --- |
| **Tập dữ liệu** |  | **Số mẫu phân lớp đúng** | **Số mẫu phân lớp sai** |
| Cohn-Kanade | Tập huấn luyện (4234 mẫu)  Tập kiểm tra (1058 mẫu) | 4,209/4,234  (chiếm 99.41%) | 25/4,234  (chiếm 0.59%) |
| JAFFE | Tập huấn luyện (171 mẫu)  Tập kiếm tra (42 mẫu) | 150/171  (chiếm 87.72 %) | 21/171  (chiếm 12.28%) |

Bên cạnh đó, giải thuật phân lớp SVM được sử dụng để huấn luyện bộ nhận dạng cảm xúc trên các tập dữ liệu. Đây là mô hình máy học véc-tơ hỗ trợ cho một tập các phương pháp học có giám sát dùng để phân loại và hồi quy. Các tham số dùng cho hàm train\_svm là: *C=0.01* (), *kernel=poly* (), *decision function shape=ovo* (one vs one), *probability=True* (). Kết quả huấn luyện được trình bày trong bảng 2.

*Bảng 2: Kết quả huấn luyện SVM*

|  |  |  |  |
| --- | --- | --- | --- |
| **Tập dữ liệu** |  | **Số mẫu phân lớp đúng** | **Số mẫu phân lớp sai** |
| Cohn-Kanade | Tập huấn luyện (4234 mẫu)  Tập kiểm tra (1057 mẫu) | 3,748/4,234  (chiếm 88.54%) | 486/4,234  (chiếm 11.46%) |
| JAFFE | Tập huấn luyện (169 mẫu)  Tập kiếm tra (42 mẫu) | 97/169  (chiếm 57.14%) | 72/169  (chiếm 42.86%) |

Bài báo cũng đã tiến hành thực nghiệm với video được quay trực tiếp từ webcam trong điều kiện ánh sáng ban ngày bình thường và ban đêm có đèn. Sau đó trích từng frame ảnh có độ phân giải 640x480 để xử lý. Mô hình nhận dạng sử dụng mạng nơ-ron nhận tạo cho kết quả tốt hơn (độ chính xác khoảng 99%). Mô hình nhận dạng sử dụng máy học véc-tơ hỗ trợ cho kết quả chấp nhận được, độ chính xác sẽ cải thiện hơi với số lượng ảnh huấn luyện đủ lớn. Về thời gian, việc huấn luyện với SVM thực hiện nhanh hơn, nhưng cho kết quả không tốt hơn so với CNN.

Kết quả nhận dạng cảm xúc là khá tốt đối với các loại cảm xúc cơ bản, đáp ứng yêu cầu hệ thống nhận dạng cảm xúc thông qua gương mặt. Tuy nhiên, do sự sai khác không nhiều của các loại cảm xúc tương có biểu cảm gương mặt gần tương tự nhau, nên sai số xảy ra và số lượng ảnh chưa đủ lớn đề có thể tạo ra tập huấn luyện tốt hơn. Bên dưới là một số hình ảnh minh họa cho việc nhận dạng cảm xúc thông qua gương mặt trên video của webcam máy tính.

|  |  |
| --- | --- |
|  |  |
|  |  |
|  |  |
| *Hình 7: Một số ví dụ thực tiễn về nhận dạng cảm xúc gương mặt* | |

# 4. KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN

## 4.1 Kết luận

Trong bài nghiên cứu này, tôi đã sử dụng các kỹ thuật máy học và xử lý ảnh để phát hiện và nhận dạng sáu loại cảm xúc cơ bản của con người thông qua các biểu hiện của gương mặt. Kết quả cho *99%* độ chính xác khi huấn luyện và *94%* độ chính xác khi kiểm thử. Kết quả này có thể được cải thiện bằng cách tăng cơ sở dữ liệu huấn luyện.

## 4.2 Hướng phát triển

Trong thời gian tới, tôi sẽ nghiên cứu, đánh giá và so sánh để tìm ra các đặc trưng phù hợp và hiệu quả hơn để huấn luyện. Đồng thời tăng số lượng dữ liệu huấn luyện và kiểm tra để nâng cao độ chính xác. Phát triển hệ thống nhận dạng thêm các loại cảm xúc phúc tạp hơn nữa.

# TÀI LIỆU THAM KHẢO

1. **Khánh, Đoan.** Liên hiệp các Hội Khoa học và Kĩ thuật tỉnh Sóc Trăng. *Công nghệ phân tích cảm xúc - Ứng dụng tiềm năng.* [Online] [Cited: June 22, 2018.] http://lienhiephoi.soctrang.gov.vn/index.php/khoa-h-c-va-cong-ngh/tin-khoa-h-c-va-cong-ngh/765-cong-ngh-phan-tich-c-m-xuc-ng-d-ng-ti-m-nang.

2. *EEG-based Automatic Emotion Recognition: Feature Extraction, Selection and Classification Methods.* **Pascal, Ackermann, et al., et al.** Aachen : s.n.

3. *Design and Inplementation of Real Time Facial Enmotion Recognition System.* **Nishant, Hariprasad Pandey, Preeti, Bajaj and Shubhangi, Giripunje.** Real time facial emotion recognition, Nagpur : s.n., 2013, Vol. 4. ISSN 2229-5518.

4. *Emotion Detection: A Feature Analysis.* **N Dharmesh and Mausmi , Kulshreshtha.** A feature analysis of emotion detection, Mumbai : IJASCSE, 2015, Vol. 4.

5. *Automatic Emotion Recognition Using Facial Expression: A Review.* **Monika, Dubey and Prof. Lokesh, Singh.** Facial expression and emotion relationship, Bhopal : IRJET , 2016, Vol. 3. 2395 - 0056.

6. *An Efficient Real-Time Emotion Detection Using Camera and Facial Landmarks.* **Binh, T. Nguyen, et al., et al.** Real time emotion detection , Ho Chi Minh : Seventh International Conference on Information Science and Technology, 2017.

7. *Face Alignment Using Active Shape Model and Support Vector Machine.* **Thái, Hoàng Lê and Trường , Nhật Võ.** Face alignment, Ho Chi Minh : International Journal of Biometrics and Bioinformatics (IJBB), Vol. 4.

8. *Một số phương pháp trích chọn đặc trưng khuôn mặt và ứng dụng.* **Thành, Tiến Tạ and Thảo, Thu Quách.** Hà Nội : s.n.

SUMMARY

DEVELOPING A TOOL THAT AUTOMATICALLY RECOGNIZES EMOTION BASED ON THE HUMAN FACE

Trầm Vũ Kiệt1

1Khoa Công nghệ Thông tin & Truyền thông, Trường Đại học Cần Thơ

In this journal, the computer vision and data science algorithms was shown in order to automatically detect human face and recognize human emotion base on facial expression, that intergrates Histogram of Gradients feature extraction, Artificial Neural Network and Support Vector Machine. The system abilities are detected and recognize human emotion with six universal emotions: happiness, sadness, surprise, anger, fear and disgust. An overall training accuracy of *95%* and test accuracy of 93*%* is archieved.

**Keywords:** *Facial emotion recognition, principal component analysis, HOG features, artifial neutral network, support vector machine.*