PHƯƠNG PHÁP ƯỚC LƯỢNG GÓC NHÌN DỰA TRÊN ĐIỂM 3D ĐẶC TRƯNG KHUÔN MẶT VÀ ỨNG DỤNG GIÁM SÁT THI TRỰC TUYẾN

Dương Thăng Long*, Trần Tiến Dũng*, Vương Thu Trang*, Dương Chí Bằng

Ngày tòa soạn nhận được bài báo: 02/12/2022
Ngày nhận kết quả phản biện đánh giá: 02/06/2023
Ngày bài báo được duyệt đăng: 28/06/2023

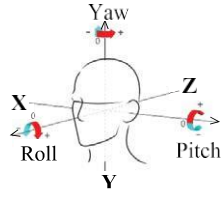
DOI: 10.59266/houjs.2023.270

Tóm tắt: *Ước lượng góc nhìn khuôn mặt (HPE) là một bài toán phức tạp đòi hỏi sự kết hợp giữa xử lý hình ảnh, thị giác máy tính và kỹ thuật học máy với các phương pháp hiện nay dựa trên mạng nơron tích chập (CNN) để xác định ánh xạ giữa không gian ảnh 2D và mô hình 3D khuôn mặt và xác định các góc nhìn. HPE được ứng dụng trong nhiều vấn đề thực tiễn và có ý nghĩa cao như các giám sát an ninh, phát hiện sự tập trung của lái xe, giám sát người học và thi trực tuyến,... Nghiên cứu này sử dụng mô hình CNN hiện đại để phát hiện các điểm đặc trưng khuôn mặt và đề xuất một phương pháp ước lượng góc nhìn khuôn mặt sử dụng thuật toán rừng ngẫu nhiên dựa trên các điểm đặc trưng 3D của khuôn mặt từ ảnh 2D để xác định góc nhìn của khuôn mặt trên ảnh đó. Kết quả thử nghiệm của phương pháp đề xuất trên bốn tập dữ liệu phổ biến đạt chất lượng tốt, cho sai số thấp nhất ở hai trong số 4 tập dữ khi so sánh các phương pháp. Chúng tôi đưa ra một thiết kế tích hợp giữa phương pháp đề xuất với hệ thống quản lý học tập trực tuyến nhằm hỗ trợ giám sát và đánh giá sự tập trung tham gia học tập và làm bài thi của người học.*

Từ khoá: *Giám sát thi trực tuyến, thị giác máy tính, mạng nơron tích chập, hồi quy rừng ngẫu nhiên.*

1. Giới thiệu

Ước lượng góc nhìn khuôn mặt (HPE) là việc xác định các góc quay 3D gồm Pitch, Yaw, Roll đối với hệ toạ độ tham chiếu của tư thế đầu, Hình 1 thể hiện minh hoạ chi tiết các góc này. Đây là một bài toán quan trọng trong thị giác máy tính và có ứng dụng trong tương tác người-máy tính, giám sát video và hệ thống hỗ trợ người lái xe. Ước lượng góc nhìn có thể giúp tạo ra tương tác người-máy tính một cách tự nhiên và trực quan hơn. Bài toán này có thể được ứng dụng để phát hiện hành vi đáng ngờ thông qua giám sát video, phát hiện sự mệt mỏi hoặc mất tập trung của người lái xe và cảnh báo phù hợp.



Hình 1. Các tham số thể hiện quay ngang, dọc hay nghiêng đầu

*Trường Đại học Mở Hà Nội

HPE là một bài toán phức tạp yêu cầu kết hợp xử lý hình ảnh, thị giác máy tính và kỹ thuật học máy. Có nhiều phương pháp được sử dụng để ước lượng góc nhìn khuôn mặt, bao gồm: (i) Sử dụng các đặc trưng cụ thể trên khuôn mặt để tính toán tư thế đầu, (ii) Sử dụng mô hình 3D của đầu để ước tính tư thế, (iii) Sử dụng toàn bộ hình ảnh để ước tính tư thế đầu, và (iv) Kết hợp các phương pháp trên để cải thiện độ chính xác. Các nghiên cứu gần đây tập trung vào sử dụng mạng nơron tích chập (CNN) để trích xuất đặc trưng và kết hợp với hình ảnh có chiều sâu để cải thiện độ chính xác trong ước tính tư thế đầu.

Tingting Liu và cộng sự [1] giới thiệu một mô hình CNN có tên là NGDNet để tăng hiệu quả ước lượng góc nhìn khuôn mặt trên ảnh hồng ngoại chất lượng thấp. Họ cũng nhận thấy rằng góc nhìn khuôn mặt thay đổi rất lớn khi góc Pitch và Yaw thay đổi. Phương pháp này cho kết quả tốt và hiệu quả đối với các khuôn mặt bị che khuất và đáp ứng tốt với các tư thế đầu đa dạng. Zhongxu Hu và cộng sự [2] đề xuất một bản đồ nhiệt Bernoulli và mạng CNN để ước lượng góc nhìn khuôn mặt. Bản đồ nhiệt Bernoulli không chỉ hồi quy góc tư thế đầu mà còn phân biệt được tiền cảnh và hậu cảnh, từ đó cải thiện độ chắc chắn của dự đoán. Các tác giả trong [3] sử dụng đám mây điểm 3D và mạng CNN để ước lượng góc nhìn khuôn mặt. Đám mây điểm 3D là một tập hợp các điểm 3D trên bề mặt có thể nhìn thấy của đối tượng, được tạo ra bằng hình ảnh độ sâu chụp từ camera 3D. Zhiwen Cao và cộng sự [4] sử dụng độ đo chuẩn Frobenius của hai ma trận xoay 3D để đánh giá độ tin cậy của ước tính tư thế đầu. Họ sử dụng mạng ResNet50 để trích xuất đặc trưng từ ảnh đầu vào và tính toán sai số tuyệt đối trung bình giữa các véctơ trực giao của ma trận xoay ước tính và ma trận xoay thực tế của ảnh.

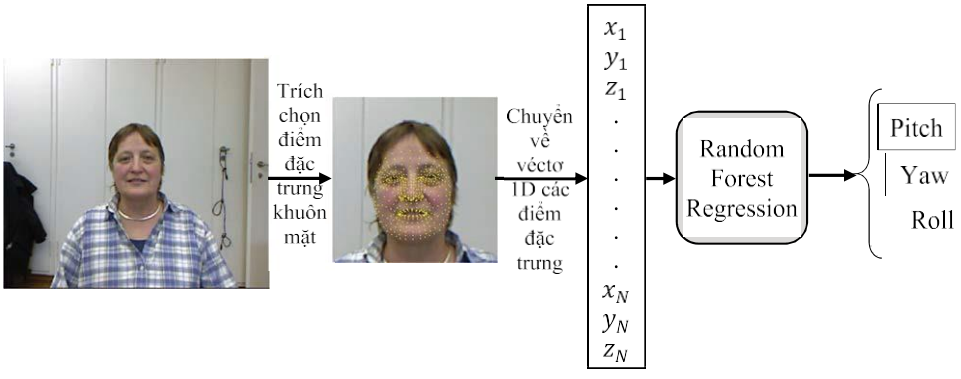
Ngày này, sự phát triển và ứng dụng mạnh mẽ của hệ thống học tập trực tuyến (LMS) nhằm cung cấp môi trường thuận lợi, dễ dàng trong giáo dục và mang lại hiệu quả cao trong thực tiễn. Theo đó, hệ thống LMS được áp dụng để giám sát, quản lý và đánh giá các hoạt động học tập và làm bài thi của người học là không thể thiếu. Đã có một số nghiên cứu [5, 6, 7, 8] sử dụng phương pháp phát hiện và nhận dạng khuôn mặt hoặc biểu cảm khuôn mặt để giám sát và đánh giá sự tập trung tham gia học tập của người học với kết quả khả quan, đạt độ chính xác nhận dạng trên 99% của các tập dữ liệu. Tuy nhiên, hiện nay chưa có nhiều nghiên cứu phương pháp ước lượng góc nhìn khuôn mặt và ứng dụng trong LMS để hỗ trợ giám sát quá trình học tập và thi của người học trực tuyến, qua đó góp phần nâng cao chất lượng học tập của người học.

Nghiên cứu này tận dụng lợi thế của các mô hình CNN hiện đại phát hiện các điểm đặc trưng khuôn mặt (facial landmarks) để đề xuất một phương pháp ước lượng góc nhìn khuôn mặt dựa trên các điểm đặc trưng 3D của khuôn mặt từ ảnh 2D đầu vào để xác định góc nhìn Pitch, Yaw, Roll của khuôn mặt trên ảnh đó. Chúng tôi cũng thiết kế ứng dụng tích hợp phương pháp này vào hệ thống LMS để hỗ trợ giám sát, đánh giá sự tập trung của quá trình học tập và thi của người học trực tuyến. Tiếp theo, Phần 2 sẽ xây dựng phương pháp ước lượng góc nhìn khuôn mặt từ ảnh đầu vào, đồng thời thiết kế tích hợp vào hệ thống LMS. Phần 3 trình bày kịch bản thử nghiệm và phân tích kết quả trên một số tập dữ liệu được công bố. Phần 4 là kết luận và một số hướng nghiên cứu tiếp theo.

2. Phương pháp ước lượng góc nhìn khuôn mặt

Trong nghiên cứu này, chúng tôi sử dụng phương pháp ước lượng góc nhìn khuôn mặt bằng cách sử dụng các điểm 3D đặc trưng khuôn mặt từ hình ảnh 2D

đầu vào. Theo đó, trước hết chúng ta phải trích xuất và chuyển đổi từ hình ảnh 2D đầu vào thành tập các điểm 3D đặc trưng khuôn mặt có kích thước , trong đó là số lượng điểm, là số chiều của mỗi điểm và trong nghiên cứu này, là các bất kỳ điểm đặc trưng bổ sung nào cần dùng và chúng tôi đặt , tức là không có điểm đặc trưng bổ sung nào được sử dụng. Tiếp theo, chúng tôi áp dụng phương pháp RandomForest để hồi quy các góc quay Pitch, Yaw, Roll cho việc xác định tư thế góc nhìn của khuôn mặt trên ảnh. Phương pháp tiếp cận này được tóm tắt minh hoạ trong quy trình ở Hình 2.



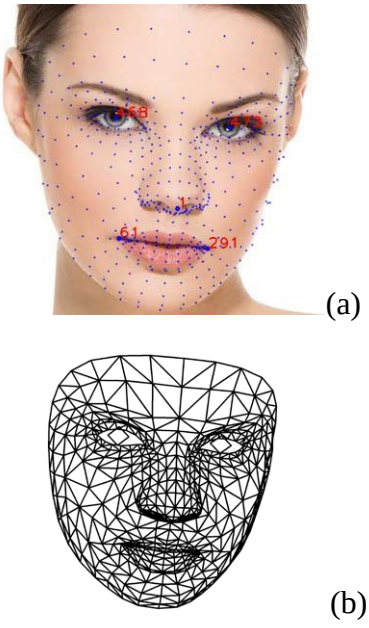
Hình 2. Sơ đồ tổng thể quá trình thực hiện của phương pháp tiếp cận

2.1. Sinh tập các điểm 3D đặc trưng khuôn mặt

Các công cụ như Dlib hay MTCNN được sử dụng khá rộng rãi để phát hiện các vùng ảnh và điểm đặc trưng khuôn mặt. Trong đó, mô hình của nhóm nghiên cứu Google [9] được sử dụng để phát hiện điểm đặc trưng khuôn mặt 3D có cấu trúc nhẹ và dễ dàng sử dụng dưới dạng thư viện chuẩn hoá MediaPipe. Mô hình này là một biến thể của kiến trúc mạng residual, tập trung vào việc lấy mẫu thô trong các lớp đầu tiên và dành nhiều tính toán cho bước đầu. Điều này giúp các nơron trong mô hình có khả năng xác định biên của các đối tượng trên ảnh, phân biệt giữa các đặc trưng như miệng và mắt. Kết quả đầu ra của mô hình bao gồm N=468 điểm 3D đặc trưng trên khuôn mặt và được biểu diễn bằng công thức (1) sau:

$$\mathcal{M}(a) = \{(x_i, y_i, z_i) : i = \overline{1,N}\} \quad (1)$$

trong đó, các toạ độ được chuẩn hoá theo kích thước ảnh đầu vào x_i, y_i, $z_i \in [0,1]$, x_i tỷ lệ theo kích thước chiều ngang, y_i tỷ lệ theo chiều cao và z_i là tỷ lệ theo kích thước chiều ngang của ảnh. Hình 2 minh hoạ các điểm đặc trưng trên khuôn mặt và mô hình lưới 3D mô tả bề mặt khuôn mặt tương ứng, 5 điểm đặc quan trọng như mắt, mũi, miệng được thể hiện ở các vị trí tương ứng trên hình (a).



(a)

(b)

Hình 2. 468 điểm đặc trưng trên khuôn mặt (a) và mô hình tương ứng (b)

2.2. Biểu diễn véctơ các góc quay tư thế nhìn của khuôn mặt

Để hiển thị các góc quay Pitch (p), Yaw (y), Roll (r) của một tư thế nhìn khuôn mặt trong chế độ xem 3D, chúng ta chuyển ba góc quay này (dưới dạng Euler) thành một ma trận xoay như sau:

$$R_x = \begin{Bmatrix} 1 & 0 & 0 \\ 0 & \cos p & -\sin p \\ 0 & \sin p & \cos p \end{Bmatrix},\ R_y = \begin{Bmatrix} \cos y & 0 & \sin y \\ 0 & 1 & 0 \\ -\sin y & 0 & \cos y \end{Bmatrix},\ R_z = \begin{Bmatrix} \cos r & -\sin r & 0 \\ \sin r & \cos r & 0 \\ 0 & 0 & 1 \end{Bmatrix} \quad (2)$$

$$R = R_x \cdot R_y \cdot R_z$$

Sau đó, ta đặt một điểm làm điểm gốc toạ độ (thường là điểm giữa khuôn mặt và sử dụng điểm ở vị trí cao nhất của mũi) và sử dụng một giá trị tỷ lệ của véctơ cần vẽ cùng với ma trận xoay ở trên để xác định được vị trí của ba điểm đầu véctơ như sau:

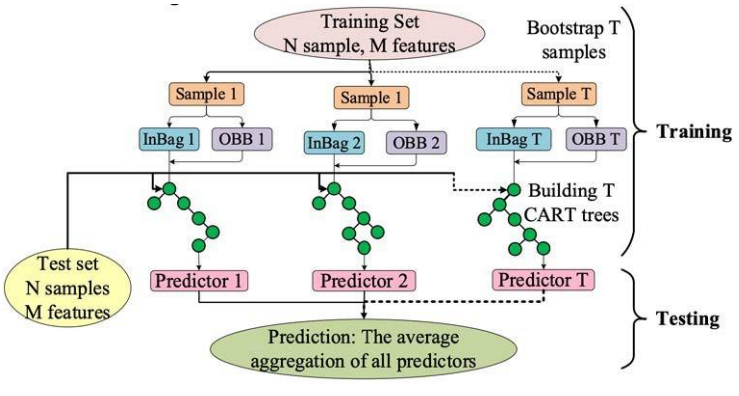
$$p_x = [s \quad 0 \quad 0] \cdot R,\ p_y = [0 \quad s \quad 0] \cdot R\ ,\ p_z = [0 \quad 0 \quad s] \cdot R \quad (3)$$

Ba đường kết nối giữa điểm gốc γ với các điểm , và hiển thị các góc quay của tư thế đầu hay thể hiện cho góc nhìn khuôn mặt ở chế độ 3D. Nghiên cứu này sử dụng công cụ tính toán Rodrigues được cung cấp trong thư viện OpenCV. Hình 3 minh hoạ các góc này trên một khuôn mặt cụ thể.

Hình 3. Minh hoạ các góc quay () tư thế đầu trên hình ảnh

2.3. Phương pháp hồi quy góc quay tư thế nhìn của khuôn mặt

Kỹ thuật hồi quy để ước lượng tư thế góc nhìn của khuôn mặt đã được đa số nghiên cứu đề cập sử dụng. Mạng nơron nhân tạo cũng cho thấy hiệu suất tốt hơn so với các thuật toán học máy khác trong nhiều lĩnh vực, nhưng nó có hạn chế khi xử lý dữ liệu đầu vào dưới dạng chuỗi số. Rừng ngẫu nhiên là một kỹ thuật học máy mà mỗi cây quyết định đưa ra dự đoán riêng cho bài toán phân loại hoặc hồi quy. Rừng ngẫu nhiên ít tốn kém tính toán hơn so với mạng nơron và có khả năng đưa ra hiểu biết về sự khác biệt của các cây quyết định. Nó cũng giúp tránh hiện tượng quá khớp dữ liệu và yêu cầu ít dữ liệu huấn luyện hơn so với mạng nơron.



Hình 4. Minh hoạ xây dựng mô hình rừng ngẫu nhiên từ dữ liệu [11]

Trong nghiên cứu này, phương pháp rừng ngẫu nhiên được sử dụng để hồi quy các góc quay tư thế nhìn của khuôn mặt từ các điểm đặc trưng trên khuôn mặt. Mô hình rừng ngẫu nhiên được xây dựng từ nhiều cây quyết định, mỗi cây được huấn luyện trên một tập dữ liệu con ngẫu nhiên và sử dụng một tập đặc trưng con ngẫu nhiên. Dự đoán cuối cùng được tính bằng cách lấy trung bình các dự đoán của các cây và có thể có trọng số. Hình 4 mô tả quy trình tổng thể của mô hình này. Dữ liệu để huấn luyện rừng ngẫu nhiên là tập các điểm đặc trưng khuôn mặt được trích xuất từ mô hình mạng nơron tích chập và đã được chuyển thành véctơ một chiều các toạ độ điểm, như trong công thức (1) và Hình 2. Với mỗi ảnh 2D đầu vào kèm theo đầu ra mong muốn là các góc nhìn tư thế khuôn mặt được xác định, tập dữ liệu huấn luyện rừng ngẫu nhiên như sau:

$$\{(\mathcal{X}_i, \mathcal{Y}_i): \mathcal{X}_i = \mathcal{M}(a_i),\ \mathcal{Y}_i = pyr_i = (p_i, y_i, r_i),\ i = 1, \dots, |D|\} \quad (4)$$

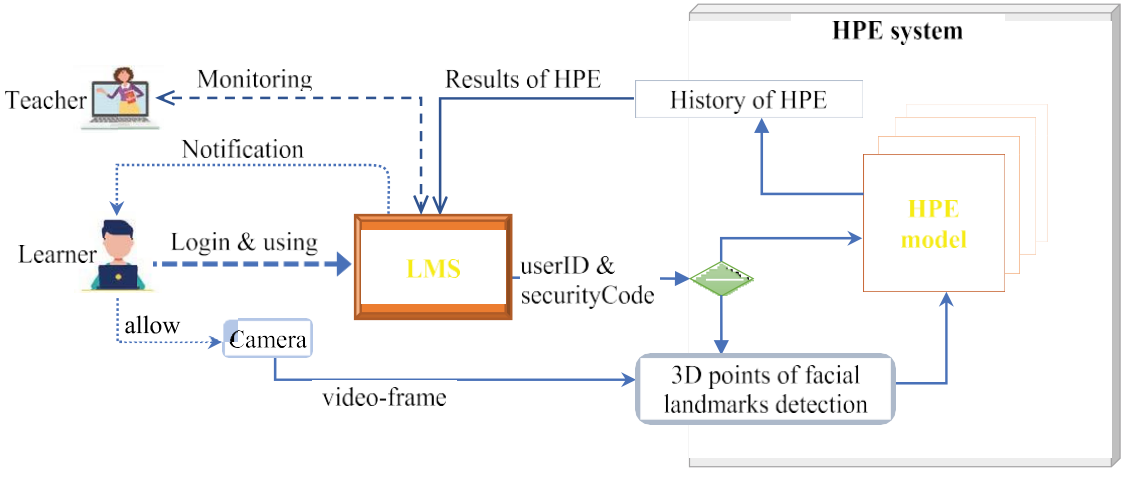
trong đó, $\mathcal{X}_i \in [0,1]^{468\times3}$ là dữ liệu đầu vào của mô hình gồm các toạ độ điểm đặc trưng trên khuôn mặt, $\mathcal{Y}_i \in R^3$ là dữ liệu đầu ra mong muốn (các góc quay Pitch, Yaw, Roll tư thế nhìn khuôn mặt), $D = \{(a_i, pyr_i)\}$ là tập dữ liệu. Sử dụng đánh giá sai số của mô hình theo MAE ở công thức (5) trên cả ba góc quay của tư thế nhìn khuôn mặt gồm Pitch, Yaw, Roll.

$$MAE = \frac{1}{|D|}\sum_{i=0}^{|D|}|t_i - \hat{t}_i| \quad (5)$$

trong đó, $t_i \in \{p_i, y_i, r_i\}$ là góc quay thực tế của tư thế nhìn khuôn mặt cần đánh giá sai số và $\hat{t}_i = \mathcal{P}^{RFR}(\mathcal{M}(a_i))$ là giá trị dự đoán đầu ra tương ứng của mô hình đề xuất ($\mathcal{P}^{RFR}$).

2.4. Thiết kế ứng dụng mô hình ước lượng góc nhìn khuôn mặt cho giám sát thi trực tuyến

Trong phần này, chúng tôi đề xuất một hệ thống tích hợp để giám sát quá trình học tập và thi trực tuyến bằng cách tích hợp mô hình ước lượng góc nhìn khuôn mặt (HPE) vào hệ thống LMS thông qua kết nối API. Khi người học đăng nhập vào hệ thống và bắt đầu làm bài thi, LMS gửi yêu cầu kích hoạt HPE để thực hiện giám sát trên thiết bị người học. Quá trình giám sát diễn ra trong suốt phiên làm bài thi, với việc chụp ảnh, trích xuất đặc trưng và ước lượng góc nhìn khuôn mặt. Kết quả giám sát được hiển thị trên thiết bị người học và gửi đến LMS theo chu kỳ thời gian. Mô hình kết nối giữa HPE và LMS được mô tả trong Hình 5.



Hình 5. Sơ đồ kết nối hai hệ thống LMS và HPE

Mục tiêu của hệ thống này là cung cấp thông tin chi tiết về quá trình học tập và thi trực tuyến, từ việc nhận dạng khuôn mặt và biểu cảm của người học đến việc đo và ghi nhận các hoạt động học tập. Điều này giúp hỗ trợ giảng viên và cơ sở đào tạo trong việc đánh giá và đưa ra phương pháp phù hợp với từng cá nhân người học.

Tại mỗi phiên làm bài thi ký hiệu *k*, $\mathcal{A}^k = \{a_{t_i}^k \mid i = 1,2, \dots\}$ là chuỗi hình ảnh thu được từ thiết bị ghi hình của người học, với t_i là thứ tự thời điểm ảnh trong chuỗi. Các ảnh $a_{t_i}^k$ có thể được tiền xử lý và áp dụng phương pháp tìm kiếm và phát hiện điểm đặc trưng khuôn mặt để thu được tập điểm đặc trưng $\mathcal{X}_{t_i}$. Mô hình ước lượng góc nhìn khuôn mặt (HPE) đưa ra kết quả góc quay tư thế nhìn của khuôn mặt $\hat{\mathcal{Y}}_{t_i}^k = \mathcal{P}^{RFR}\left(\mathcal{M}(a_{t_i}^k)\right)$ và gửi kết quả này cho LMS để hỗ trợ cảnh báo và quản lý.

Để giảm thiểu truyền hình ảnh qua mạng và bảo vệ quyền riêng tư của người dùng, ta có thể thực hiện mô hình $\mathcal{M}$ trên thiết bị học tập sử dụng môi trường JavaScript. Ở đây sử dụng công cụ TensorflowJSLite để triển khai mô hình ở thể nhẹ. Vì LMS là môi trường học tập cá nhân, thiết bị ghi hình cá nhân có một khuôn mặt. Trong trường hợp có nhiều khuôn mặt, có thể cảnh báo và lựa chọn khuôn mặt lớn nhất để đảm bảo rằng người học gần nhất với thiết bị đang sử dụng.

3. Thử nghiệm và thảo luận kết quả

3.1. Dữ liệu và kịch bản thử nghiệm

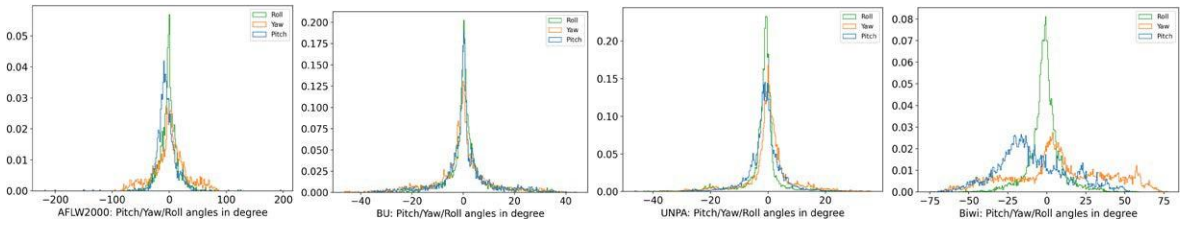
Để thử nghiệm và đánh giá kết quả phương pháp đề xuất trong nghiên cứu này, chúng tôi sử dụng bốn bộ dữ liệu thử nghiệm được công bố gồm AFLW2000, BU Head Tracking database (BU) và UPNA Head Pose database (UPNA) và Biwi Kinect Head Pose database (BIWI) như đề cập trong [12]. Hình ảnh trong các tập dữ liệu được trích chọn đặc trưng tạo thành tập dữ liệu các điểm $\mathcal{X}_i$ mô hình 3D của khuôn mặt, ký hiệu $D^{tr} = \{\mathcal{X}_i^{tr}, \mathcal{Y}_i^{tr}\}$ và $D^{te} = \{\mathcal{X}_i^{te}, \mathcal{Y}_i^{te}\}$. Một số hình ảnh có thể không được trích chọn điểm đặc trưng do khuôn mặt bị che khuất quá lớn. Kết quả số lượng hình ảnh được trích chọn điểm đặc trưng từ các tập dữ liệu và giới hạn phân bố các góc nhìn Pitch, Yaw, Roll được thể hiện ở Bảng 1.

Bảng 1. Thông tin mô tả về các tập dữ liệu thử nghiệm

Tập dữ liệu	Số lượng ảnh được trích chọn đặc trưng (#Img)	Độ phân giải khung hình	Giới hạn (min, max) của các góc quay tư thế khuôn mặt (tính bằng độ)		
			Pitch	Yaw	Roll
AFLW2000	1691	450x450	[-151.04, 195.87]	[-95.54, 85.32]	[-218.90, 124.11]
BU	8990	320x240	[-38.78, 44.03]	[-46.36, 39.73]	[-29.73, 35.66]
UPNA	36000	1280x720	[-30.02, 25.58]	[-42.49, 35.80]	-46.36, 35.84]
BIWI	14831	640x480	[-74.94, 53.55]	[-66.95, 76.89]	[-56.62, 58.46]

Hình 6 thể hiện phân bố các góc quay Pitch, Yaw, Roll của tư thế nhìn khuôn mặt trong các tập dữ liệu (biểu đồ histogram). Ba tập dữ liệu AFLW2000, BU và UPNA có phân bố tập trung chủ yếu quanh 0^0, trong đó AFLW2000 có mức phân bố trải rộng và tỷ lệ rất thấp, BU và UPNA có mức độ tập trung cao trong đoạn [-10^0, 10^0]. Dữ liệu BIWI có phân bố khá phức tạp, ba góc có mức độ phân bố khác nhau khá lớn.

Để đánh giá mô hình ước lượng góc nhìn khuôn mặt $\mathcal{P}^{RFR}$, chúng tôi áp dụng phương án thử nghiệm 5-folds, tức là chia tập dữ liệu *D* thành 5 phần bằng nhau, sử dụng 4 phần để huấn luyện mô hình (D^{tr}) và phần còn lại để kiểm tra (D^{te}). Quá trình này chạy thử nghiệm lặp lại với lần lượt các phần dữ liệu được dùng để kiểm tra, kết quả lấy trung bình cuối cùng của 5 lần chạy.



Hình 6. Biểu đồ phân bố (histogram) các góc quay tư thế nhìn khuôn mặt

3.2. Kết quả thử nghiệm và thảo luận

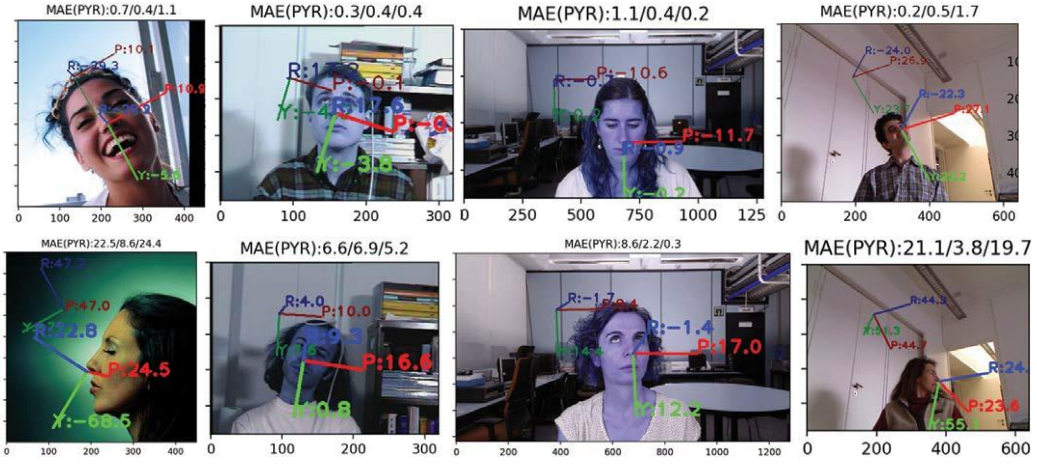
Bảng 3 thể hiện chi tiết kết quả đánh giá ước lượng các góc nhìn khuôn mặt trong các lần chạy. Sai số góc quay (MAE) có dấu * bên cạnh thể hiện cho bé nhất. Dòng cuối cùng là trung bình của các góc và trong các lần chạy. Dữ liệu AFLW2000 có sai số lớn nhất bởi vì phạm vi các góc quay khuôn mặt trong dữ liệu là rất rộng và dữ liệu lại có ít hình ảnh nhất. Sai số thấp nhất là 4.21° ở góc quay Roll, trong khi góc Pitch và Yaw có sai số thấp nhất tương ứng là 5.63° và 4.68°. Đối với 3 tập dữ liệu còn lại có sai số khá thấp bởi vì phạm vi các góc quay khuôn mặt là nhỏ và có nhiều dữ liệu hình ảnh để huấn luyện mô hình. Dữ liệu BIWI có sai số lớn thứ hai là 1.60° bởi vì phân bố phạm vi các góc quay khuôn mặt cũng lớn thứ hai tương ứng. Dữ liệu UPNA có sai số thấp nhất là 0.35° bởi vì phân bố phạm vi các góc quay khuôn mặt nhỏ nhất và đồng thời có nhiều dữ liệu hình ảnh nhất để huấn luyện mô hình.

Xét hai dòng cuối cùng của Bảng 3 thể hiện tỷ số giữa MAE và trung bình của độ lớn phạm vi các góc quay khuôn mặt (pyr) là TS(pyr), giữa sai số trung bình các góc của các lần chạy (MAE) và số lượng hình ảnh (#Img) là TS(Img), sử dụng hệ số nhân vì quá nhỏ. Các giá trị này cũng phản ánh tập dữ liệu AFLW2000 đạt cao nhất và UPNA đạt thấp nhất. Điều này một lần nữa cho thấy dữ liệu có mức độ phân tán các góc quay càng lớn và số lượng hình ảnh càng ít thì càng cho sai số lớn đối với mô hình ước lượng đề xuất .

Bảng 3. Sai số MAE° ước lượng góc nhìn khuôn mặt trong các lần chạy

Dữ liệu \ Lần chạy		AFLW2000	BU	UPNA	BIWI
#1	Pitch	6.12	0.97*	0.45	1.65
	Yaw	5.34	1.35	0.25*	1.31
	Roll	4.21*	1.00	0.35*	1.96
#2	Pitch	6.25	0.98	0.46	1.54*
	Yaw	5.75	1.33	0.26	1.22
	Roll	4.73	1.00	0.36	2.00
#3	Pitch	5.63*	1.02	0.44*	1.64
	Yaw	5.21	1.29*	0.25*	1.27
	Roll	4.22	0.96	0.35*	1.90*
#4	Pitch	6.82	0.97*	0.45	1.54*
	Yaw	5.58	1.32	0.27	1.20*
	Roll	5.10	0.95*	0.35*	1.96
#5	Pitch	5.69	1.01	0.45	1.57
	Yaw	4.68*	1.37	0.25*	1.26
	Roll	4.50	0.97	0.35*	1.99
Trung bình		**5.32**	**1.10**	**0.35**	**1.60**
TS(pyr)		0,018	0,014	0,007	0,012
TS(Img)x10^3		3,146	0,122	0,010	0,108

Hình 7 minh họa trực quan các góc quay khuôn mặt được ước lượng so sánh với góc quay thực tế từ các tập dữ liệu. Các góc thực tế được thể hiện ở vị trí phía trái trên so với vị trí góc quay ước lượng của mô hình và có màu nhạt hơn. Trên tiêu đề của mỗi hình ảnh là sai số (MAE) theo thứ tự của các góc Pitch, Yaw và Roll. Các hình ảnh ở dòng cuối được chọn từ các kết quả HPE có sai số lớn. Tập dữ liệu của AFLW2000, BIWI có phạm vi góc quay khuôn mặt lớn nên có sai số lơn hơn hai tập dữ liệu còn lại, hơn nữa dữ liệu AFLW2000 có ít dữ liệu hình ảnh nên càng khó khăn cho huấn luyện mô hình được chất lượng cao và dẫn đến có sai số lớn nhất.



Hình 7. Các góc quay khuôn mặt ước lượng và thực tế

Bảng 4 thể hiện so sánh kết quả HPE của phương pháp đề xuất với các nghiên cứu gần đây. Ký hiệu "x" sau phương pháp là huấn luyện mô hình trên dữ liệu nhân tạo độc lập 300W-LP và kiểm tra trên toàn bộ tập dữ liệu sử dụng, "s24" là thử nghiệm kiểu k-fold với k=24 và chia tập dữ liệu thành các phần theo 24 đối tượng người trong ảnh, "t2" hoặc "t3" là thử nghiệm với việc sử dụng 20% hoặc 30% dữ liệu để kiểm tra, "u" là không có thông tin và "f5" là thử nghiệm kiểu 5-fold (k=5 và chia dữ liệu ngẫu nhiên). Kết quả đạt cao nhất được thể hiện bằng chữ in đậm, "-" là không có dữ liệu kết quả.

Bảng 4. So sánh kết quả ước lượng (MAE°) giữa các phương pháp trên các tập dữ liệu

Phương pháp	AFLW2000	BU	UPNA	BIWI
M.Shao et. al. [14]/x	5.48	-	-	5.99
H.Wang et. al. [15]/x	5.40	-	-	3.02
Z.Cao et. al. [4]/x	4.50	-	-	2.80
M.Ariz et. al. [16]/u	-	2.58	1.09	-
A.Asperti & D.Filippini [12]/x	1.46	-	-	3.36
Y.Xu et. al. [3]/s24	-	-	-	**1.42**
A.F.Abate et. al. [10]/t3	3.39	-	-	2.43
K.Khan et. al., [17]/py	-	2.20	-	-
H.Kawai et. al. [18]/t2	-	-	1.81	-
Z.Hu et. al. [2]/t3	-	-	-	3.31
K.Khan et. al. [19]/*,u	-	2.40	-	2.00
S.Malek et. al. [13]/f5	-	-	-	1.60
Our method/f5	5.32	**1.10**	**0.35**	1.60

Ở tập dữ liệu AFLW2000, kết quả tốt nhất với sai số 1.46° trong [12] nhưng họ thử nghiệm phương pháp "x", kết quả phương pháp đề xuất là 5.32° đứng thứ tư. Nếu kiểu thử nghiệm trong cùng bộ dữ liệu (chia tỷ lệ dữ liệu kiểm tra hoặc k-fold) thì kết quả phương pháp của chúng tôi đứng thứ 2 và chỉ đứng sau [10]. Đối với tập dữ liệu BIWI đạt sai số thấp nhất là 1.41° trong [3], tuy nhiên, kết quả này chạy thử nhiệm có dữ liệu kiểm tra ít hơn vì chia dữ liệu với 24 phần. Sai số của phương pháp đề xuất cùng với kết quả trong [13] đứng thứ hai với sai số là 1.60°, trong khi kích thước dữ liệu kiểm tra nhiều hơn 480% so với [3]. Đối với 2 tập dữ liệu BU và UPNA, kết quả của phương pháp đề xuất đạt sai số thấp nhất, tương ứng là 1.10° và 0.35°. Nhìn chung, phương pháp đề xuất cho kết quả ước lượng tốt và cao nhất ở các góc quay phạm vi vừa phải và điều này giúp cho việc giám sát góc nhìn của người dùng ở phạm vị hẹp là khả thi và hữu ích.

4. Kết luận

Trong nghiên cứu này, chúng tôi đề xuất phương pháp ước lượng góc nhìn khuôn mặt trong hình ảnh dựa trên các điểm đặc trưng khuôn mặt được trích chọn bằng mô hình mạng nơron học sâu. Phương pháp ước lượng góc nhìn khuôn mặt dựa trên một lượng lớn các điểm đặc trưng 3D của khuôn mặt từ ảnh 2D đầu vào để xác định góc nhìn Pitch, Yaw, Roll của khuôn mặt trên ảnh đó, sau đó sử dụng thuật toán rừng ngẫu nhiên để huấn luyện mô hình ước lượng các góc nhìn khuôn mặt. Kết quả ước lượng cho sai số thấp ở các bộ dữ liệu thử nghiệm, đạt chất lượng cao nhất ở hai trong số 4 tập dữ liệu thử nghiệm với phạm vi phân bố góc quay vừa phải, ở mức sai số chỉ là 1.10° và 0.35°. Điều này cho thấy phương pháp ước lượng đề xuất có thể áp dụng tốt cho bài toán đặt ra là giám sát quá trình học tập và thi của người học trên hệ thống trực tuyến. Hơn nữa, chúng tôi cũng thiết kế một kết nối ứng dụng và mô hình của phương pháp đề xuất với bất kỳ một hệ thống quản lý học tập và thi trực tuyến (LMS) nào một cách đơn giản, linh hoạt để dễ dàng áp dụng trong thực tiễn.

Trong những nghiên cứu tiếp theo, chúng tôi sẽ tích hợp việc giám sát định danh người học, nhận dạng các trạng thái học tập và giám sát góc nhìn khuôn mặt để hỗ trợ và giúp cho quá tình quản lý học tập trực tuyến được toàn diện và chất lượng hơn. Các mô hình mạng nơron theo kiến trúc hiện đại cũng được nghiên cứu áp dụng nhằm thực hiện đa tác vụ và cho chất lượng kết quả cao hơn.

Tài liệu tham khảo:

[1]. T. Liu, J. Wang, B. Yang và X. Wang, "NGDNet: Nonuniform Gaussian-label distribution learning for infrared head pose estimation and on-task behavior understanding in the classroom," Neurocomputing, tập 436, p. 210–220, 2021.

[2]. Z. Hu, Y. Xing, C. Lv, P. Hang và J. Liu, "Deep convolutional neural network-based Bernoulli heatmap for head pose estimation," Neurocomputing, tập 436, p. 198–209, 2021.

[3]. Y. Xu, C. Jung và Y. Chang, "Head pose estimation using deep neural networks and 3D point clouds," Pattern Recognition, tập 121, số 108210, pp. 1-10, 2022.

[4]. Z. Cao, Z. Chu, D. Liu và Y. Chen, "A Vector-based Representation to Enhance Head Pose Estimation," Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pp. 1188-1197, 2021.

[5]. D. T. Long, "A Facial Expressions Recognition Method Using Residual Network Architecture for Online Learning Evaluation," Journal of Advanced Computational Intelligence and Intelligent Informatics, tập 25, số 6, pp. 1-10, 2021.

[6]. D. T. Long, "A Lightweight Face Recognition Model Using Convolutional Neural Network for Monitoring Students in E-Learning," International Journal of Modern Education and Computer Science, tập 6, pp. 16-28, 2020.

[7]. D. T. Long, T. T. Tung và T. T. Dung, "A Facial Expression Recognition Model using Lightweight Dense-Connectivity Neural Networks for Monitoring Online Learning Activities," International Journal of Modern Education and Computer Science, tập 6, pp. 53-64, 2022.

[8]. B. N. Anh, N. T. Son, P. T. Lam, L. P. Chi, N. H. Tuan, N. C. Dat, N. H. Trung, M. U. Aftab và T. V. Dinh, "A Computer-Vision Based Application for Student Behavior Monitoring in Classroom," Applied sciences , tập 9, số 4729, pp. 1-17, 2019.

[9]. Y. Kartynnik, A. Ablavatski, I. Grishchenko và M. Grundmann, "Real-time Facial Surface Geometry from Monocular Video on Mobile GPUs," CVPR Workshop on Computer Vision for Augmented and Virtual Reality, số https://arxiv.org/abs/1907.06724, pp.1-4, 2019.

[10]. A. F. Abate, C. Bisogni, A. Castiglione và M. Nappi, "Head pose estimation: An extensive survey on recent techniques and applications," Pattern Recognition, tập 127, số 108591, pp. 1-14, 2022.

[11]. M. H. Lipu, A. Ayob, M. H. M. Saad và M. Faisal, "State of Charge Estimation for Lithium-ion Battery Based on Random Forests Technique with Gravitational Search Algorithm," IEEE PES Asia-Pacific Power and Energy Engineering Conference (APPEEC), pp. 45-50, 2018.

[12]. A. Asperti và D. Filippini, "Deep Learning for Head Pose Estimation: A Survey," SN Computer Science, tập 4, số 349, pp. 1-41, 2023.

[13]. S. Malek và S. Rossi, "Head pose estimation using facial-landmarks classification for children rehabilitation games," Pattern Recognition Letters, tập 152, p. 406–412, 2021.

[14] M. Shao, Z. Sun, M. Ozay và T. Okatani, "Improving Head Pose Estimation with a Combined Loss and Bounding Box Margin Adjustment," IEEE International Conference on Automatic Face & Gesture Recognition, số doi: 10.1109/FG.2019.8756605, pp. 1-5, 2019.

[15] H. Wang, Z. Chen và Y. Zhou, "Hybrid coarse-fine classification for head pose estimation," ArXiv, tập abs/1901.06778, pp. 1-5, 2019.

[16]. M. Ariz, A. Villanueva và R. Cabeza, "Robustandaccurate 2D-tracking-based 3D positioning method: Application to head pose estimation," Computer Vision and Image Understanding, tập 180, pp. 13-22, 2019.

[17]. K. Khan, J. Ali, K. Ahmad, A. Gul, G. Sarwar, S. Khan, Q. T. H. Ta, T.-S. Chung và M. Attique, "3D Head Pose Estimation through Facial Features and Deep Convolutional Neural Networks," Computers, Materials & Continua, tập 66, số 2, pp. 1757-1770, 2021.

[18]. H. Kawai, J. Chen, P. Ishwar và J. Konrad, "VAE/WGAN-Based Image Representation Learning For Pose-Preserving Seamless Identity Replacement In Facial Images," IEEE 29th International Workshop on Machine Learning for Signal Processing (MLSP), số https://doi.org/10.1109%2Fmlsp.2019.8918926, 2019.

[19]. K. Khan, R. U. Khan, R. Leonardi, P. Migliorati và S. Benini, "Head pose estimation: A survey of the last ten years," Signal Processing: Image Communication, tập 99, số 116479, pp. 1-16, 2021.

A METHOD FOR HEAD POSE ESTIMATION BASED ON 3D POINTS OF FACIAL LANDMARKS AND APPLICATION TO MONITOR ONLINE EXAMINATION

Duong Thang Long[†], Tran Tien Dung[†], Vuong Thu Trang[†], Duong Chi Bang[†]
Email: duongthanglong@hou.edu.vn

Abstract: *Head pose estimating (HPE) is a complex problem requiring image processing, computer vision, and machine learning techniques. Current methods rely on convolutional neural networks (CNNs) to establish the mapping between 2D image space and the 3D face model, enabling the determination of head pose angles. HPE is applicated in various practical and highly significant areas such as security surveillance, driver attention monitoring, online learning and testing supervision, and more. This study utilizes a modern CNN model to detect facial landmarks. It proposes a method for estimating facial pose angles using a random forest algorithm based on 3D facial landmarks extracted from 2D images. This approach enables the determination of head pose angles from the given images. Experimental results on four popular datasets demonstrate the effectiveness of the proposed method, achieving low estimation errors, particularly outperforming other methods on two of the four datasets. We also present an integrated design of the proposed method with an online learning management system to facilitate monitoring and assessment of learners' engagement and performance during learning and testing activities.*

Keywords: *Monitor online examination, computer vision, convolutional neural network, random forest regression.*

[†] Hanoi Open University