TRƯỜNG ĐẠI HỌC BÁCH KHOA
KHOA ĐIỆN TỬ VIỄN THÔNG

BÁO CÁO CUỐI KỲ
PBL4: Trí tuệ nhân tạo

Đề tài:

Nhận diện đối tượng bằng ảnh năng lượng dáng đi

Sinh viên thực hiện : Nguyễn Văn Huy
Phạm Tiến Sơn

Giảng viên hướng dẫn : **Hồ Phước Tiến**

Ngày 19 Tháng 4 năm 2025

Mục lục

I GIỚI THIỆU

1 Bài báo tham khảo gốc

Báo cáo được lấy ý tưởng và thực hiện dựa trên tham khảo bài báo gốc là ***Individual Recognition Using Gait Energy Image*** được đăng vào tháng 2 năm 2006 bởi Ju Han và Bir Bhanu[1].

Trong đó ứng dụng hình ảnh năng lượng dáng đi (Gait Energy Image) và phương pháp PCA vào việc trích xuất các đặc trưng dáng đi của con người để nhận dạng đối tượng dựa trên dáng đi. Bên cạnh đó giải quyết vấn đề thiếu hụt số lượng ảnh do việc tạo mới dữ liệu bằng ảnh GEI để huấn luyện bằng cách tạo ra các hình ảnh tổng hợp từ các dữ liệu huấn luyện gốc.

2 Thiết bị phần cứng

Tất cả các bước và code trong cáo được thực hiện trên IDE VisualStudio Code với Python phiên bản 3.12.6 trên laptop Lenovo Ideapad S-145 với 20GB RAM và CPU Intel(R) Core i3-8130 @2.20GHz.

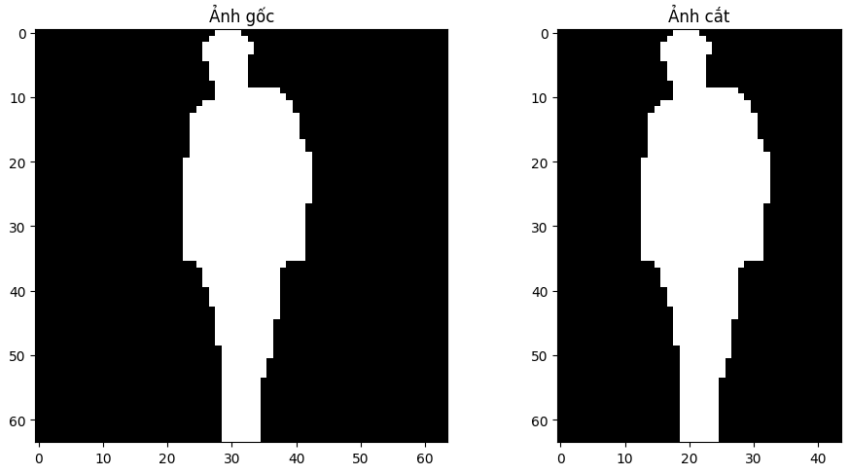
3 Tập dữ liệu

Trong đó bài báo tham khảo gốc ban đầu[1] của nhóm sử dụng tập dữ liệu USF HumanID, nhưng vì mục đích của môn học và hạn chế về quyền truy cập cũng như kích thước quá lớn nên tập dữ liệu được sử dụng trong dự án cho học phần này là Casia-B[2]. Với số lượng đối tượng thử nghiệm ban đầu là 10 người và góc quay là 90^o. Trước khi kết hợp với cả 11 góc quay và tăng số đối tượng để xem rằng liệu có thể nhận dạng với nhiều hướng đi khác nhau chỉ với ảnh GEI không và độ chính xác tương ứng với các góc khác nhau.

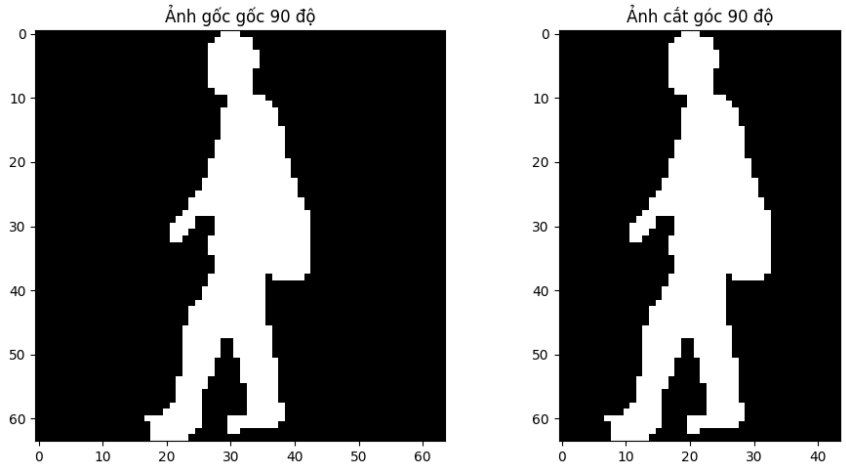
CASIA-B là một trong các tập dữ liệu được cung cấp bởi Viện tự động hóa thuộc Học viện khoa học Trung Quốc (CASIA) nhằm mục đích thúc đẩy các nghiên cứu về liên quan đến nhận dạng dáng đi. Tập dữ liệu CASIA-B bao gồm dữ liệu của 124 đối tượng với 11 góc đi khác nhau với 3 biến thể lần lượt là dáng đi bình thường (nm), dáng đi khi mặc áo khoác (cl) và dáng đi khi mang theo túi (bg). Dưới định dạng video tập dữ liệu được đánh dấu 'xxx-mm-nn-ttt.avi' với xxx là số thứ tự của đối tượng từ 001 đến 124, mm là kiểu dáng đi (nm: bình thường, cl: mặc áo khoác, bg: mang túi xách), nn là số thứ tự của khung ảnh trong video, ttt là góc của dáng đi (000, 018, ..., 180).

Tập dữ liệu CASIA-B được tạo ra bằng cách quay 1 đoạn video khi 1 người di chuyển với các 11 góc quay khác nhau từ 0^o đến 180^o rồi sau đó cắt mỗi khung hình của video thành nhiều ảnh trắng đen để có thể trích xuất được hình ảnh dáng đi của người đó tại mỗi thời điểm khi di chuyển. Từ đó có thể tạo ra một chuỗi các ảnh dáng đi của mỗi người theo trình tự thời gian cũng như với mỗi góc độ khác nhau.

Mỗi ảnh dáng đi có kích thước 64x64 pixels, với những các điểm ảnh màu trắng là điểm mà tại có đó có hiện diện dáng người, còn ngược lại là điểm đen tương ứng với môi trường. Nhưng nhóm nhận thấy rằng khoảng $\frac{2}{3}$ bức ảnh đã là phần nền nên đã quyết định cắt bớt 10 pixels 2 bên cạnh trái và cạnh phải để giảm đi những phần thông tin không quan trọng từ nền ảnh. Hiển thị ảnh trước và sau khi cắt, các phần điểm ảnh trắng vẫn được đảm bảo giữ lại toàn vẹn. Trong hình 1.3.2 kể cả với góc đi 90 độ, bức ảnh vẫn giữ lại được toàn bộ thông tin dáng đi.



Hình 1.3.1: Bên trái là ảnh gốc, bên phải là ảnh sau khi cắt với góc 0 độ



Hình 1.3.2: Bên trái là ảnh gốc, bên phải là ảnh sau khi cắt với góc 90 độ

Lưu ý, tập dữ liệu đối tượng thứ 5 bị thiếu dáng đi góc chính diện 0 độ. Nên trong quá trình thực hiện nhóm sẽ bỏ qua mà ko sử dụng đối tượng thứ 5.

4 Cấu trúc báo cáo

Trong phần I của bài báo đã giới thiệu về bài báo tham khảo gốc và tập dữ liệu được sử dụng trong dự án lần này cũng như thiết bị phần cứng được sử dụng để thực hiện nó. Sau đó phần II sẽ tập trung chính vào cơ sở lý thuyết của các phương pháp được ứng dụng trong mô hình phân loại nhận dạng dáng đi dựa trên ảnh năng lượng. Phần III sẽ đưa ra kết quả của phương pháp và nhận xét kết luận và cũng như điểm mạnh và hạn chế của phương pháp này.

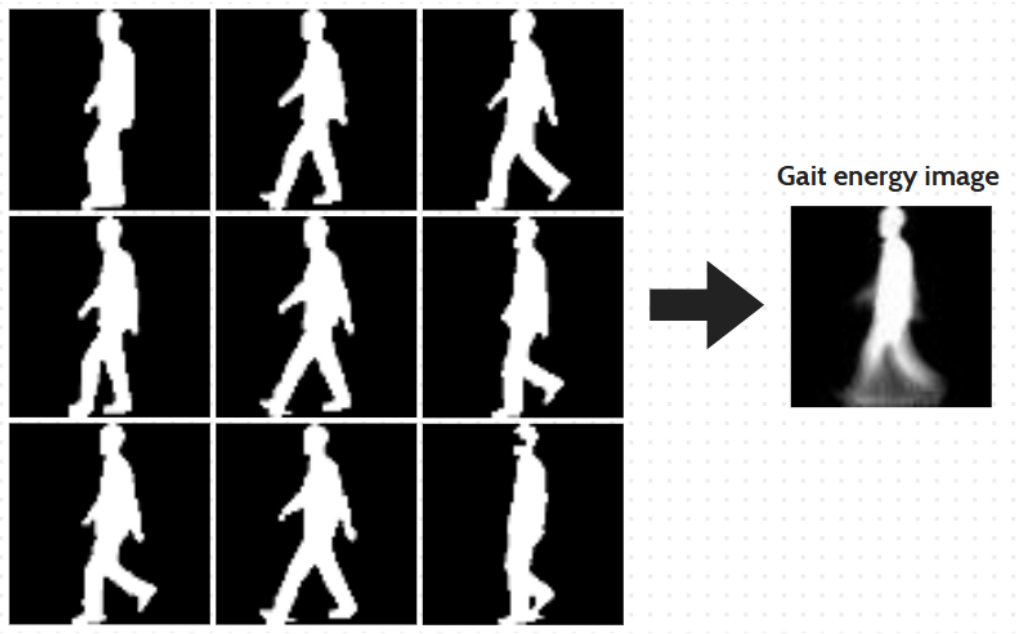
II CƠ SỞ LÝ THUYẾT

1 Ảnh năng lượng dáng đi (Gait Energy Image - GEI)

Việc dataset được tạo nên từ chuỗi các ảnh dáng đi trắng đen lại làm cho việc lưu trữ và tính toán trở nên phức tạp và tốn kém hơn. Nếu không xét đến việc dáng đi của mỗi đối tượng thường có tính chu kì lặp đi lặp lại một cách riêng biệt cũng như độ dài sải tay và chân khi di chuyển. Từ đó, bằng cách sử dụng ảnh năng lượng dáng đi, có thể giảm được tất cả ảnh trong một chuỗi về một bức ảnh duy nhất đại diện cho dáng đi mỗi đối tượng để dễ dàng xử lý. Sử dụng công thức dưới đây:

$$G(x,y) = \frac{1}{N}\sum_{t=1}^{N} B_t(x,y),$$

Với (x,y) là tọa độ của mỗi điểm ảnh, B_t và t lần lượt là ảnh tương ứng với mốc thời gian trong chuỗi, N là số ảnh trắng đen dáng đi trong chuỗi và G là giá trị pixel tương ứng trong ảnh GEI. Bằng cách tính lấy trung bình cộng mỗi điểm ảnh tương ứng ta có thể tạo ra được một ảnh năng lượng dáng đi, trong đó những điểm sáng nhất đại diện cho vị trí cơ thể ít cử động nhất, và giảm dần màu xám cho những phần mà cơ thể đối tượng thường xuyên cử động khi đi. Từ đó vẫn giúp lưu giữ lại các đặc trưng về tốc độ cũng như độ dài sải tay và bước chân của các bước ảnh.

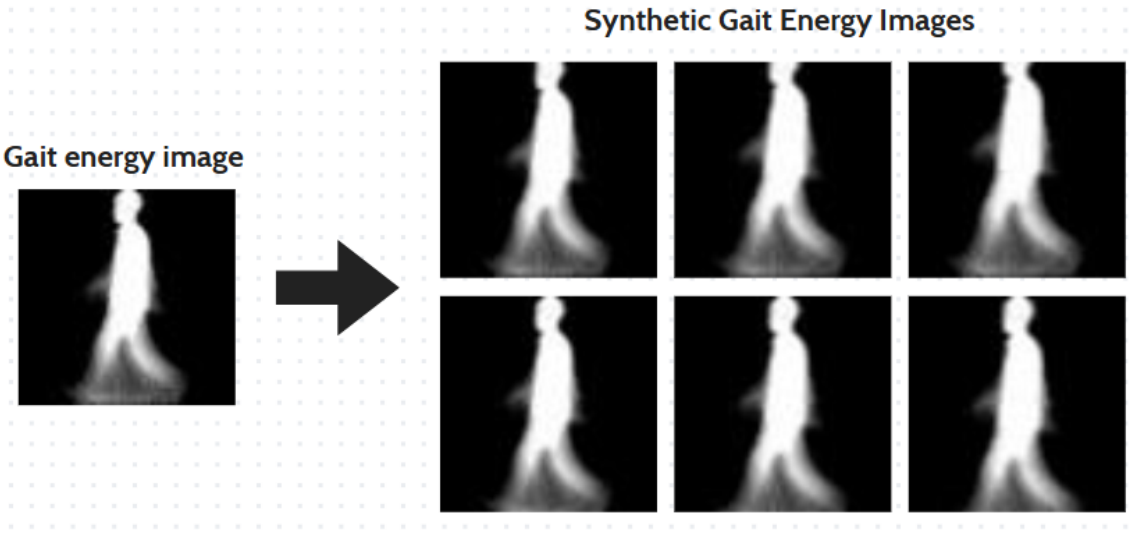


Hình 2.1.1: Ảnh năng lượng dáng đi sau khi tổng hợp từ các ảnh nhị phân dáng đi

2 Ảnh tổng hợp (Synthetic template)

Tuy vậy, việc gộp tất cả ảnh dáng đi về thành làm 1 ảnh năng lượng dáng đi lại khiến số lượng dữ liệu lại giảm đi đáng kể, khiến mỗi đối tượng chỉ còn lại 1 ảnh GEI cho mỗi tập dữ liệu dáng đi. Do đó, bài báo còn đề xuất thêm cách làm để tăng dữ liệu huấn luyện lên bằng cách sử dụng các ảnh GEI và biến đổi thành các ảnh tổng hợp (synthetic). Bằng cách lần lượt cắt đi 1 phần phía dưới của bước ảnh đi đến 1 tỉ lệ nhất định là $\frac{3}{24}$ tỉ lệ chiều cao sau đó kéo dãn lại về kích thước ban đầu. Lý do của việc cắt một phần bức ảnh này là để mô phỏng sự biến dạng dáng đi của đối tượng khi di chuyển trên các bề mặt (trên nền đường bê tông hoặc trên cỏ, làm chân bị che đi mất 1 phần) cũng như môi trường khác nhau. Bên cạnh đó, xét tới cả việc người đi đường thường hay đội thêm mũ cũng như ít có cử động phần đầu đáng kể khi đang đi, nhóm em quyết định cũng tạo thêm các ảnh tổng hợp khác tương ứng cho phần thân trên (đầu) của ảnh.

Chưa tính đến việc dữ liệu thực tế có thể có chút sai lệch với 11 góc quay mặc định mà dữ liệu không bao phủ hết tất cả góc độ. Nên cần thực hiện thêm 1 bước nữa là lật ảnh, để toàn bộ dữ liệu có thể thể chiếm hết 360^o khi thực hiện nhận diện với nhiều hướng khác nhau. Dưới đây là đoạn mã giả cho thuật toán tạo ảnh tổng hợp GEI của bài báo.



Hình 2.2.1: Một số bức ảnh sau khi thực hiện thuật toán trên bao gồm cắt dưới và trên.

1. Cho một tấm ảnh GEI kích thước X x Y
2. Cho h là hàng cao nhất từ dưới ảnh tính lên ứng với mức độ biến dạng cho phép tối đa.
3. Cho k = 2
4. Khởi tạo i = i.
5. Loại bỏ $r = k * i$ hàng từ dưới bức ảnh gốc.
6. Kéo dãn phần còn lại từ $(X - r)$ x Y về X x $\frac{XY}{X-r}$ bằng phép nội suy.
7. Cắt bỏ đều 2 bên rìa trái và phải bức ảnh để tạo ra ảnh tổng hợp kích thước X x Y.
8. Cho $i = i + 1$.
9. Nếu $k * i \leq h$ thì quay lại bước 5, không thì dừng lại.

Nhưng trong quá trình thực hiện code khi nhóm muốn tùy chỉnh số lượng ảnh tổng hợp được tạo ra với mỗi ảnh thì lại khiến kích thước các ảnh ra ko không được đồng đều với nhau. Nên nhóm em đã tối ưu cũng như đơn giản hóa lại thuật toán trên bài báo để có thể ra được ảnh tổng hợp ổn định hơn. Cũng như nhận thấy rằng việc loại bỏ một số hàng pixel phía thân dưới ảnh cũng có thể được áp dụng lên phía thân trên mà không làm tổng quan bức ảnh thay đổi quá nhiều về mặt cảm quan sau khi kéo dãn về kích thước ban đầu.

1. Cho một tấm ảnh GEI kích thước X * Y
2. Cho k = 2 và iter = 3
3. Cho i = 1
4. Cắt k * i pixels hàng dưới cùng của ảnh gốc
5. Kéo dãn bức ảnh sau khi cắt ở bước 4. về kích thước X * Y
6. Cắt k * i pixels hàng trên cùng của ảnh gốc
7. Kéo dãn bức ảnh sau khi cắt ở bước 6. về kích thước X * Y
8. Cho i = i +1
9. Quay lại bước 4., dừng lại khi i > iter

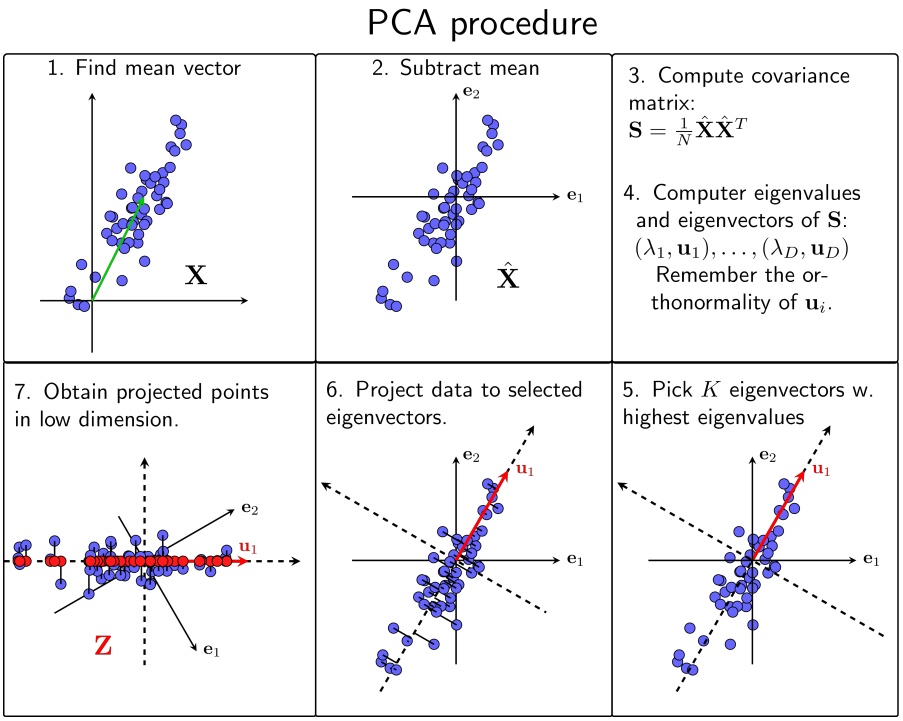
3 Phương pháp Phân tích thành phần chính (Principle Components Analysis -PCA)

Đến đây, chúng ta đã có một chuỗi các ảnh GEI (gốc và tổng hợp) cho mỗi đối tượng, thì một vấn đề nữa xảy ra đó là kích thước quá lớn của bức ảnh với bức ảnh tập dữ liệu mà nhóm đang sử dụng có kích thước là 64 x 64. Khi làm phẳng ra để đưa vào các bộ phân loại sẽ là 4096 đặc trưng ảnh, thì cần đến 4096 neuron cho

lớp đầu vào, làm tăng khối lượng tính toán không cần thiết. Trong khi xét đến ảnh dữ liệu GEI thì phần nền đen phía sau đã chiếm khoảng $\frac{2}{3}$ diện tích bức ảnh. Khi đó, chúng ta cần phải thực hiện giảm chiều dữ lại nhưng đồng thời hạn chế làm mất đi những đặc trưng cơ bản của dữ liệu bằng phương pháp *Phân tích thành phần chính* PCA - Principal Component Analysis. PCA về mặt cơ sở lý thuyết có chút hơi phức tạp nhưng có thể trình bày thành các bước tính toán đơn giản như sau:

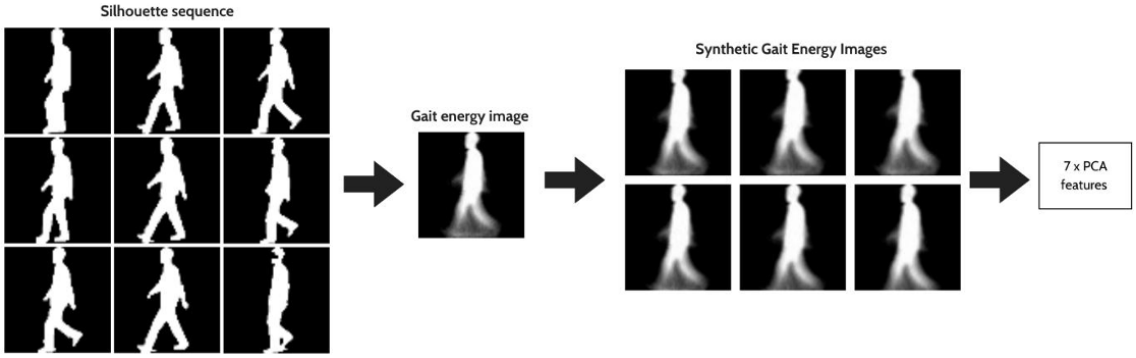
1. Tính vector trung bình của tập dữ liệu (train).
2. Trừ vector trung bình đó cho tập dữ liệu (train).
3. Tính ma trận phương sai của tập dữ liệu (train).
4. Tính các giá trị riêng và vector riêng của ma trận phương sai đó.
5. Chọn K vector riêng tương ứng với giá trị riêng lớn nhất.
6. Chiếu tập dữ liệu khác (test) lên những vector riêng đó.
7. Thu được tập dữ liệu (test) mới với K-chiều

Trong đó, K được chọn sao cho tổng K giá trị riêng chiếm khoảng 95% tổng tất cả giá trị riêng đó. Mà thường chỉ chiếm một phần rất nhỏ trong 4096 đặc trưng đó, từ đó giúp chúng ta có thể giảm được số chiều của dữ liệu một cách đáng kể mà không làm thất thoát quá nhiều thông tin quan trọng. Bên dưới là hình ảnh trực quan hóa các bước thực hiện phương pháp PCA.



Hình 2.3.1: Các bước thực hiện phương pháp PCA trực quan hóa

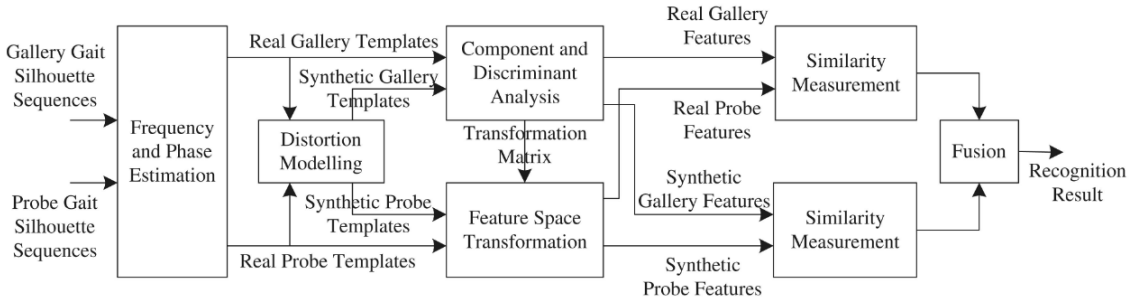
Và sau khi đã thực hiện hết tất cả các bước trên, ta sẽ được được 7 đặc trưng PCA cho với mỗi dáng đi - mỗi góc đi - mỗi người. Có thể được mô tả bằng hình phía dưới.



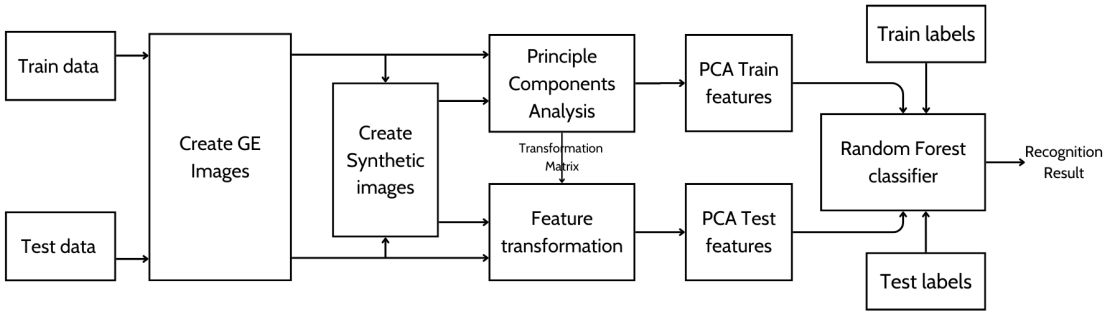
Hình 2.3.2: Luồng trích xuất đặc trưng của chương trình

4 Mô hình phân loại

Trong bài báo nhóm tham khảo sử dụng mô hình như hình 2.5.1, nhưng sau đó nhóm đã điều chỉnh lại một chút phần **Similarity measurement** thành **Random Forest classifier** để phù hợp và chính xác hơn với phương pháp mà nhóm muốn ứng dụng tại hình 2.5.2 để có thể dễ dàng theo dõi luồng thực hiện chương trình hơn.



Hình 2.4.1: Mô hình phân loại của bài báo



Hình 2.4.2: Mô hình phân loại của nhóm sau khi điều chỉnh

Và dưới đây là tổng quan dữ dữ liệu khi sử dụng mô hình huấn luyện với 10 đối tượng.

	Train	Test
Số đối tượng	10	
Góc đi	11 góc (0 → 180 độ)	
Kích thước	64 x 44 (cắt bớt background)	
Dáng đi	nm-1, nm-2 nm-3, nm-4, bg-1, cl-1	nn5, nm-6, bg-2, cl-2
Số ảnh GE/ng	66 x (64x44)	44 x (64x44)
Số ảnh/ng sau khi tổng hợp	462 x (64x44)	308 x (64x44)
Sau khi PCA	4620 x 18	3080 x 18

Hình 2.4.3: Tổng quan luồng dữ liệu

III KẾT QUẢ VÀ KẾT LUẬN

1 Kết quả

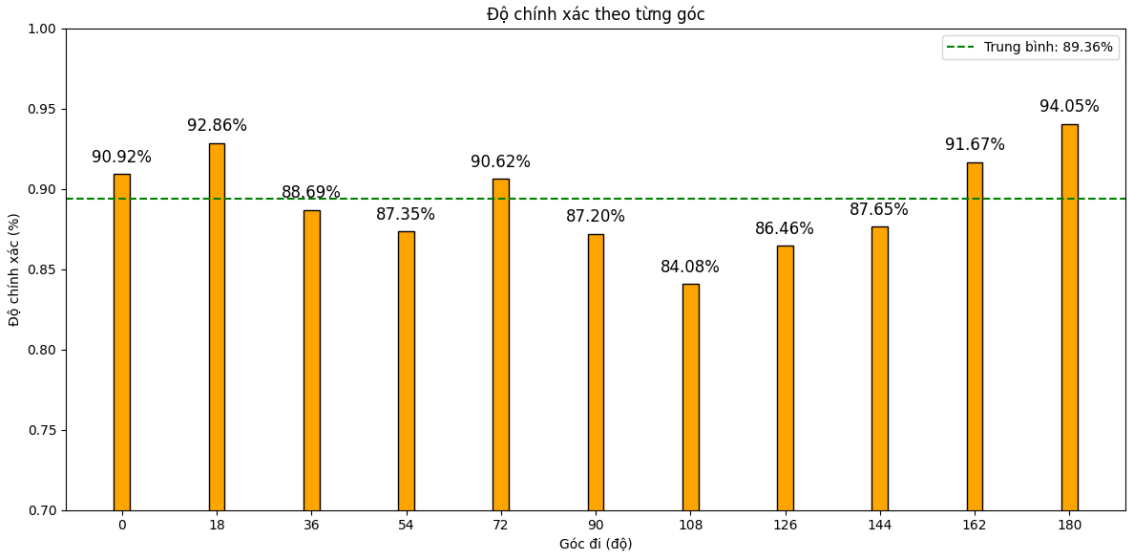
Bấm vào link *https://github.com/sillycroisant/GaitEnergy-Detection* để xem cách thực hiện code và kết quả cũng như trực quan hóa sau khi huấn luyện và dự đoán.

Bên cạnh mô hình chỉ sử dụng PCA và RandomForest, nhóm còn thực hiện nhận dạng đối tượng sử dụng một mô hình CNN đơn giản để làm đối chứng so sánh kết quả với nhau, như một cách so sánh cho phương pháp học máy truyền thống (PCA RandomForest) và hiện đại (mạng noron - mô hình CNN). Dưới đây là kết quả độ chính xác khi huấn luyện và dự đoán với 2 mô hình khác nhau đó.

Mô hình / số đối tượng	10 người	24 người
PCA RandomForest	93,80%	89,36%
CNN đơn giản	98,02%	96,32%
Độ chênh lệch	4.22%	6.96%

Bảng 1: Độ chính xác khi huấn luyện với các mô hình và số đối tượng khác nhau

Với phương pháp mô hình đầu tiên được sử dụng với các tham số là k = 18 cho số thành phần PCA, số lượng cây quyết định trong mô hình rừng cây là 300. Còn mô hình CNN sử dụng 3 lớp tích chập liên tiếp với nhau kết hợp với batch norm và max pooling, bộ tối ưu Adam, batch size là 32 với số lần huấn luyện 30 epochs.



Hình 3.1.1: Độ chính xác dựa trên dáng đi

Bên cạnh đó khi phân tích độ chính xác dựa trên góc, chúng ta cũng có thể thấy một sự bất đồng đều trong khả năng nhận diện của mô hình dựa trên dáng đi khi góc có độ chính xác cao nhất là góc 180 độ và thấp nhất là 108 độ.

2 Nhận xét và kết luận

Có thể thấy cả hai phương pháp học máy đều cho ra kết quả khá cao nhưng phương pháp hiện đại (CNN) sẽ luôn nhỉnh hơn phương pháp truyền thống về độ chính xác. Và khi số lượng đối tượng cần nhận dạng tăng lên thì độ chính xác chênh lệch giữa hai phương cũng dẫn tăng lên từ 4.22% đến 6.96%.

Từ đó có thể cho thấy được sự hiệu quả và linh hoạt hơn rõ ràng của phương pháp học máy hiện đại so với phương pháp truyền thống. Trong khi đó phươnng pháp truyền thống vẫn giữ một vai trò quan trọng khi ngày xưa các mô hình vẫn bị giới hạn do sự hạn chế về phần cứng và khả năng tính toán máy móc, thì ngày vẫn được sử dụng rộng rãi như một cách để hỗ trợ các mô hình học máy hiện đại phức tạp hơn trong việc trích xuất đặc trưng và giảm chiều dữ liệu để có thể thực hiện tính toán huấn luyện nhanh hơn nữa.

3 Bàn luận mở rộng

Dựa trên mô hình phương pháp truyền thống và hiện đại trên, chúng ta còn có thể mở rộng phát triển thêm một số bài toán với tập dữ liệu Casia-B như phát hiện người đi đường có mang áo khoác hay túi xách hay không, hoặc phân loại giới tính của đối tượng dựa trên dáng đi, hoặc dự đoán tốc độ di chuyển của đối tượng nếu chúng ta có thể thu thập thêm nhiều thông tin gán nhãn hơn cho quá trình huấn luyện mô hình.

Bên cạnh đó, thay vì chỉ sử dụng mỗi năng lượng dáng đi, còn nhiều loại ảnh khác cũng có thể cho biết thêm thông tin về dáng đi của đối tượng như ảnh gradient hoặc kết hợp cả 2 hoặc nhiều loại ảnh khác nhau để mô hình có thể học và dự đoán tốt hơn.

Tài liệu

[1] **Dataset Casia-B:** *https://www.kaggle.com/datasets/trnquanghuyn/casia-b/data*

[2] **Individual recognition using gait energy: image:** *https://ieeexplore.ieee.org/abstract/document/1561189*

[3] **Github:** *https://github.com/nikitaomare/Gait-Recognition*

[4] **Visualize filters and feature maps:** *https://www.kaggle.com/code/arpitjain007/guide-to-visualize-filters-and-feature-maps-in-cnn*