ĐẠI HỌC ĐÀ NẴNG
TRƯỜNG ĐẠI HỌC BÁCH KHOA
KHOA ĐIỆN TỬ – VIỄN THÔNG

BÁO CÁO

PBL4: TRÍ TUỆ NHÂN TẠO

ĐỀ TÀI:

DỰ ĐOÁN Ý ĐỊNH ĐIỀU KHIỂN THIẾT BỊ CHO BỆNH NHÂN HẠN CHẾ VẬN ĐỘNG DỰA TRÊN GIAO TIẾP BẰNG MẮT

Sinh viên thực hiện:
Trần Anh Toàn – 106220237 – 22KTMT1
Giảng viên hướng dẫn:
TS. Trần Thị Minh Hạnh

Đà Nẵng, 2026

DỰ ĐOÁN Ý ĐỊNH ĐIỀU KHIỂN THIẾT BỊ CHO BỆNH NHÂN HẠN CHẾ VẬN ĐỘNG DỰA TRÊN GIAO TIẾP BẰNG MẮT

Trần Anh Toàn
toantran1752004@gmail.com

Tóm tắt—
Từ khóa:

1 Giới thiệu

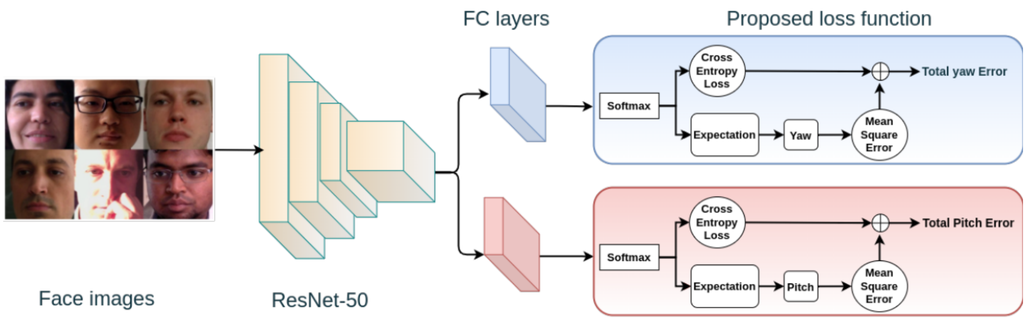
Hướng nhìn của mắt là một tín hiệu quan trọng, được khai thác trong nhiều lĩnh vực ứng dụng khác nhau. Thông tin này phản ánh mức độ tập trung và sự tương tác của người dùng trong các hệ thống tương tác người – robot cũng như các hệ thống hội thoại mở. Bên cạnh đó, trong lĩnh vực thực tế tăng cường, hướng nhìn của người dùng được sử dụng để dự đoán vùng chú ý, từ đó nâng cao khả năng nhận thức của thiết bị và giảm mức tiêu thụ năng lượng. Chính vì những lợi ích này, nhiều nghiên cứu đã được đề xuất nhằm ước lượng chính xác hướng nhìn của con người. Các phương pháp hiện có có thể được phân thành hai nhóm chính: phương pháp dựa trên mô hình và phương pháp dựa trên đặc trưng hình ảnh. Trong đó, các phương pháp dựa trên mô hình thường yêu cầu phần cứng chuyên dụng, khiến chúng khó triển khai trong các môi trường không bị ràng buộc. Ngược lại, các phương pháp dựa trên đặc trưng hình ảnh ước lượng trực tiếp hướng nhìn từ hình ảnh thu được bởi các camera thông dụng, nhờ đó dễ dàng áp dụng trong nhiều bối cảnh và điều kiện khác nhau.

Dựa vào những yếu tố trên, dự án này sẽ sử dụng phương pháp dựa trên đặc trưng hình ảnh để đưa ra các dự đoán về hướng nhìn của mắt nhằm nhận biết được ý định điều khiển thiết bị cho bệnh nhân hạn chế vận động. Mô hình sẽ được thực thi trên phần cứng Raspberry Pi (hoặc Jetson Nano) để ứng dụng cho xe lăn.

2 Phương pháp

Hiện tại có 2 phương pháp đang được xem xét:

2.1 L2CS-NET



Hình 1: Kiến trúc của phương pháp L2CS-NET

Hình 1 là kiến trúc của phương pháp L2CS-NET [1] với phần chính là lớp trích xuất đặc trưng ResNet-50, lớp Fully-Connected và cuối cùng là hàm mất mát được trình bày

Ảnh đầu vào được ResNet-50 thực hiện trích xuất đặc trưng hướng nhìn theo không gian. Tiếp theo đặc trưng sẽ được chia thành hai lớp Fully-Connected, và cũng áp dụng hai hàm mất mát khác nhau để tính góc của hướng nhìn theo phương ngang và phương dọc. Với mỗi đầu của lớp Fully-Connected, sử dụng softmax để chuyển từ mạng đầu ra của nơ-ron thành

phân bố xác suất. Sau đó, hàm mất mát cross-entropy được áp dụng để tính toán mất mát phân loại nhị phân giữa xác suất đầu ra và nhãn nhị phân mong đợi.
Mất mát cross-entropy được tính như sau:

$$H(\mathbf{y}, \mathbf{p}) = -\sum_i y_i \log p_i$$

Tiếp theo, tính toán độ mong đợi của phân bố xác suất để dự đoán được hướng nhìn chi tiết. Cuối cùng tính sai số bình phương trung bình cho dự đoán và cộng nó với mất mát phân loại. Sai số bình phương trung bình được tính như sau:

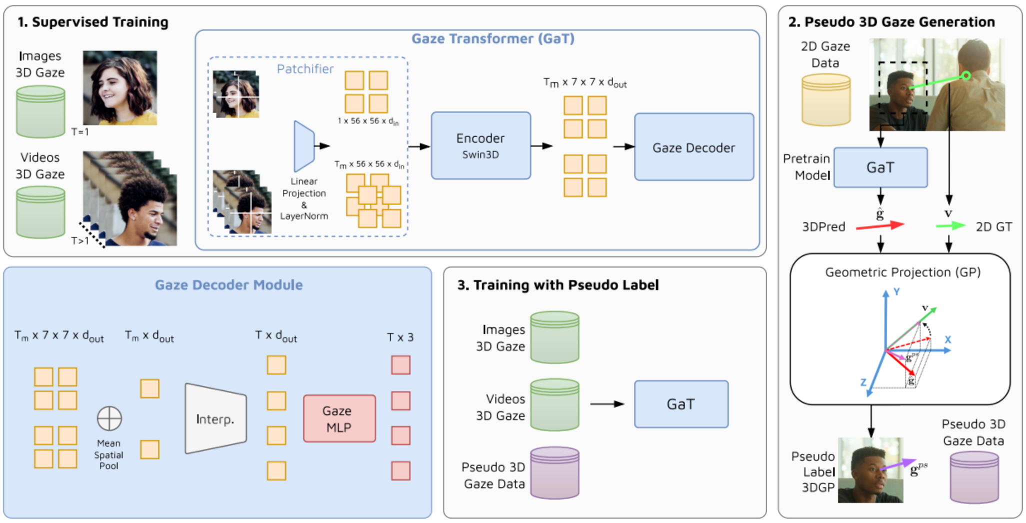
$$\mathrm{MSE}(\mathbf{y}, \mathbf{p}) = \frac{1}{N}\sum_{0}^{N}(y-p)^2$$

Hàm mất mát được trình bày cho mỗi góc hướng mát là một hàm kết hợp tuyến tính của mất mát cross-entropy và sai số bình phương trung bình, được tính như sau:

$$\mathrm{CLS}(\mathbf{y}, \mathbf{p}) = H(\mathbf{y}, \mathbf{p}) + \beta \cdot \mathrm{MSE}(\mathbf{y}, \mathbf{p})$$

Trong đó, CLS là hàm mất mát kết hợp, **p** là giá trị dự đoán, **y** là giá trị nhãn thực, và β là hệ số hồi quy.

2.2 ST-WSGE và GaT



Hình 2: Kiến trúc của phương pháp ST-WSGE

Hình 2 thể hiện kiến trúc của phương pháp ST-WSGE [2] với hai giai đoạn chính

Giai đoạn 1 là huấn luyện mô hình có giám sát với Gaze Transformer (GaT), dữ liệu đầu vào là dạng ảnh và video được chia thành các spatio-temporal patches bởi Patchifier. Sau đó áp dụng Encoder với bộ mã hóa phân cấp không gian - thời gian dựa trên Swin Transformer. Kiến trúc này giúp bắt chước các chi tiết nhỏ ở vùng mắt (local features) đồng thời hiểu được ngữ cảnh lớn hơn từ tư thế đầu (global context). Và cuối cùng là áp dụng Gaze Decoder: đi qua các lớp gom cụm (spatial pooling) và nội suy thời gian, một lớp MLP chia sẻ sẽ dự đoán vector ánh nhìn 3D chuẩn hóa cho từng token.

Giai đoạn 2 là Pseudo-label Generation, sử dụng mô hình đã huấn luyện ở giai đoạn 1 để dự đoán ánh nhìn 3D trên các bộ dữ liệu 2D. Do nhãn 2D chỉ thiếu thành phần chiều sâu (z), phương pháp áp dụng Phép chiếu Hình học (Geometric Projection) để căn chỉnh dự đoán 3D với nhãn 2D gốc, từ đó tạo ra nhãn giả (pseudo-labels) 3D chất lượng cao.

Giai đoạn 3 là huấn luyện lại mô hình bằng cách kết hợp cả dữ liệu 3D chuẩn và dữ liệu 2D đã được gán nhãn giả 3D.

Công thức sử dụng cho phép chiếu hình học:

$$\mathbf{g}^{ps} = \left(v_x \left\|(\hat{g}_x, \hat{g}_y)\right\|_2,\ v_y \left\|(\hat{g}_x, \hat{g}_y)\right\|_2,\ \hat{g}_z\right)$$

Hàm mất mát được áp dụng cho phương pháp:

$$\mathcal{L}_{\text{gaze}}(\hat{\mathbf{g}}, \mathbf{g}) = \frac{1}{T} \sum_{t=1}^{T} \frac{180}{\pi} \arccos\left(\frac{\hat{\mathbf{g}}_t^\top \mathbf{g}_t}{\|\hat{\mathbf{g}}_t\| \, \|\mathbf{g}_t\|} \right)$$

2.3 So sánh hai phương pháp

Bảng 1: So sánh phương pháp L2CS-Net và phương pháp ST-WSGE và GaT

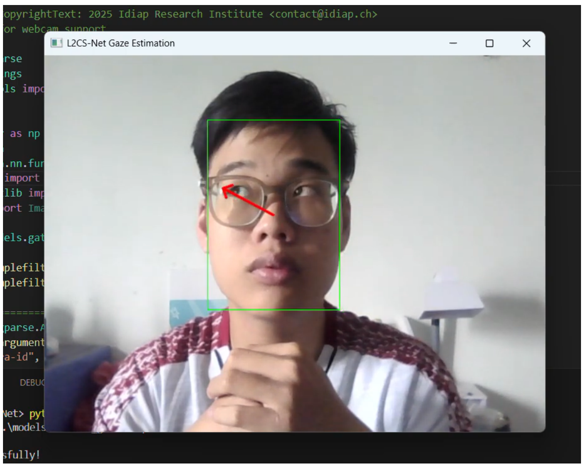
Tiêu chí	**L2CS-Net**	**ST-WSGE và GaT**
Bài toán chính	Ước lượng gaze 3D (yaw, pitch)	Ước lượng gaze 3D vector
Dạng giám sát	Giám sát đầy đủ (Fully supervised)	Giám sát yếu + Self-training
Nhãn huấn luyện	Gaze 3D chính xác	Gaze 2D + pseudo 3D gaze
Kiến trúc chính	CNN (ResNet-based)	Transformer (GaT, Swin3D)
Hàm mất mát	Classification + Regression (L2CS loss)	Angular loss + consistency loss
Độ phức tạp mô hình	Thấp – Trung bình	Cao

Bảng 1 thể hiện sự khác nhau giữa hai phương pháp, với việc ứng dụng vào phần cứng Raspberry Pi thì nên lựa chọn sử dụng phương pháp L2CS-Net. Tuy nhiên việc sử dụng backbone là ResNet-50 theo phương pháp đó cũng là một điểm gây khó khăn khi thực thi với phần cứng, nếu được thì nên sử dụng một backbone khác như MobileNet v2.

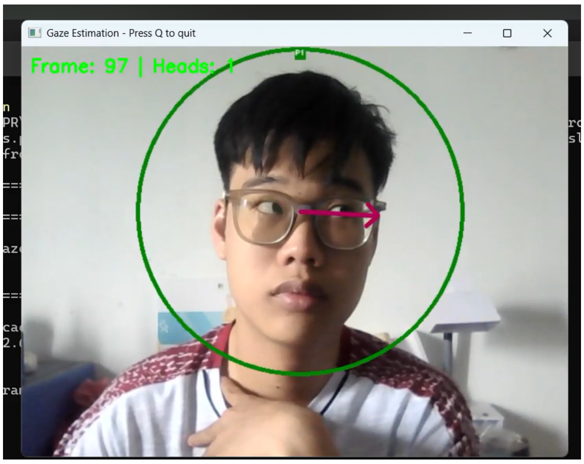
3 Kết quả thực hiện (đến thời điểm hiện tại)

3.1 Thực thi trên máy tính

Tính đến thời điểm hiện tại, đã thu thập được tập dữ liệu thô Gaze360 và thực thi mô hình pretrained của các phương pháp trên máy tính và thu được kết quả khá chính xác. Tập dữ liệu thô Gaze360 cung cấp rất nhiều chú thích hướng nhìn trong khoảng 360 độ. Bao gồm 238 đối tượng trải dài nhiều độ tuổi, giới tính và sắc tộc, có thể xem chi tiết ở liên kết dưới phần phụ lục. Hình 3 và Hình 4 thể hiện kết quả thử nghiệm cho mô hình của hai phương pháp được trình bày.



Hình 3: Kết quả của phương pháp L2CS-NET



Hình 4: Kết quả của phương pháp ST-WSGE

Để xem chi tiết kết quả chạy trên máy có thể truy cập vào liên kết dưới phần phụ lục

Bảng 2: So sánh thực thi L2CS-Net và ST-WSGE & GaT

Tiêu chí	**L2CS-Net**	**ST-WSGE và GaT**
Tham số	23.88 M	27.86 M
FPS	9.08	6.09
RAM	492.7 MB	829.6 MB

3.2 Thực thi trên phần cứng

Dựa vào bảng so sánh hai phương pháp để chọn thực thi trên Raspberry Pi thì thấy phương pháp L2CS phù hợp hơn. Tuy nhiên việc thực thi để lấy được kết quả trên Raspberry Pi vẫn chưa được do giới hạn phần cứng. Hướng xư lý là đổi backbone sang MobileNet v2.

A Phụ lục

A.1 Tập dữ liệu

Để xem chi tiết tập dữ liệu có thể truy cập vào liên kết sau:
https://drive.google.com/drive/folders/1eUz45L3aqO9PD9rvhOTsmGzdH6Nb7Jrh

A.2 Kết quả chạy mô hình pretrained

Để xem chi tiết kết quả có thể truy cập vào liên kết sau:
http://youtube.com/watch?v=2bAZR3xpuA8