HỆ THỐNG CẢNH BÁO PHÁT HIỆN TÉ NGÃ TRONG PHÒNG BẰNG CAMERA

FALL DETECTION SYSTEM IN THE ROOM BY CAMERA

SVTH: Nguyễn Phước Đại Toàn, Nguyễn Trần Mỹ Duyên, Trần Thị Ngà

Lớp 20T1, Khoa Công nghệ thông tin, Trường Đại học Bách Khoa - Đại học Đà Nẵng;
Email: nguyentoan102002@gmail.com, myduyen14125@gmail.com, tranngadn02@gmail.com

GVHD: KS. Nguyễn Bá Hoàng, TS. Ninh Khánh Duy

Khoa Công nghệ thông tin, Trường Đại học Bách Khoa - Đại học Đà Nẵng;
Email: nbhoang@dut.udn.vn, nkduy@dut.udn.vn

Tóm tắt

Té ngã là một hành động bất thường tiềm ẩn nhiều nguy cơ ảnh hưởng đến sức khỏe như gãy xương, vỡ đốt sống, chấn thương não,... và có thể gây ra tàn tật, tử vong nếu không phát hiện kịp thời. Đặc biệt trong bối cảnh tỷ lệ già hóa ngày càng tăng, té ngã càng trở nên phổ biến dẫn đến nhu cầu phát triển các hệ thống phát hiện té ngã là vô cùng cấp thiết. Do đó, mục đích của nghiên cứu này nhằm đề xuất một hệ thống giúp phát hiện té ngã bằng camera với khả năng cảnh báo trong thời gian thực. Bài báo nghiên cứu dựa trên đặc trưng tư thế con người nhằm phát hiện té ngã dựa trên hai hướng tiếp cận phổ biến là phương pháp Top-down và phương pháp Bottom-up để so sánh hiệu suất của chúng từ các khía cạnh như độ chính xác và thời gian xử lý. Mô hình được sử dụng trong nghiên cứu là LSTM với độ chính xác tốt nhất là 99.82% với phương pháp Top-down và 99.63% với phương pháp Bottom-up.

Từ khóa: Phát hiện té ngã; Học máy; Học sâu; Phân đoạn hình ảnh; Ước tính tư thế cơ thể con người; Thời gian thực.

Abstract

Falling is an abnormal action that carries various potential risks to health, such as bone fractures, spinal injuries, brain trauma,.. and can cause disability or even death if not detected in time. Especially in the context of an increasing aging rate, falls are becoming more and more common, emphasizing the urgent need for developing fall detection systems. Therefore, the purpose of this research is to propose a camera-based fall detection system with real-time alert capabilities.The research paper focuses on human pose estimation features for fall detection using two popular approaches, namely the Top-down method and the Bottom-up method., to compare their performance in terms of accuracy and processing time. The employed model in the research is LSTM, achieving the best accuracy of 99.82% using the Top-down approach and 99.63% using the Bottom-up approach.

Key words: Fall detection; Machine learning; Deep learning; Image segmentation; Human Body Pose Estimation; Realtime.

1. Đặt vấn đề

Ngày nay, tỉ lệ người bị đột quỵ và tỉ lệ già hóa đang ngày càng tăng cao đã làm tăng nguy cơ té ngã. Theo thống kê của Hội đồng Lão khoa quốc gia Hoa Kỳ (NCOA), trong vòng một năm, cứ 4 người trên 65 tuổi thì có 1 người bị té ngã. Một thống kê khác cho thấy tỉ lệ người cao tuổi bị té ngã mỗi năm là khoảng 28% - 35% đối với những người có tuổi từ 65 tuổi trở lên và 32% - 42% đối với những người trên 75 tuổi. Trong đó, có hơn 15% người cao tuổi bị té ngã trên 2 lần trong năm [1].

Theo nghiên cứu được thực hiện bởi Trung tâm Bệnh tật Kiểm soát và Phòng ngừa (CDCP), khoảng 36 triệu ca té ngã được báo cáo mỗi năm dẫn đến 32.000 ca tử vong [2]. Tại Việt Nam, ước tính có khoảng 1,5 - 1,9 triệu người cao tuổi bị té ngã mỗi năm, 5% trong số đó phải nhập viện vì các chấn thương [1]. Do đó, xây dựng hệ thống phát hiện té ngã là cần thiết giúp giảm thiểu những rủi ro về sức khỏe.

Hệ thống phát hiện té ngã thường tiếp cận dựa trên cảm biến, camera hoặc có thể kết hợp cả cảm biến và camera. Có nhiều ứng dụng phát hiện té ngã dựa trên cảm biến như: Thiết bị đeo tay Fall Detector của ELDERCARE dựa trên cảm biến đo chuyển động và gia tốc; Hệ thống Bitcare sử dụng các cảm biến áp suất,... Tuy nhiên, hệ thống sử dụng các cảm biến cố định chỉ dành riêng cho một người và vị trí duy nhất, mặc dù đem lại hiệu quả cao nhưng chi phí lại rất đắt đỏ. Trong khi đó, sử dụng hệ thống phát hiện ngã bằng camera cho phép phát hiện được nhiều người khác nhau mà không cần sử dụng các cảm biến.

Những đóng góp chính của nghiên cứu này là:

(1) Phân tích, so sánh hai hướng tiếp cận phổ biến là Top-down và Bottom-up trong bài toán phát hiện té ngã.

(2) Xây dựng một mô hình học sâu giúp phát hiện té ngã bằng camera.

(3) Xây dựng hệ thống cảnh báo đến người dùng khi phát hiện té ngã.

Bài nghiên cứu này được trình bày thành năm phần, trong đó **Phần 2** trình bày các nghiên cứu liên quan; phương pháp của nghiên cứu ở **Phần 3**; **Phần 4** thảo luận về các thực nghiệm và kết quả tương ứng; và cuối cùng là phần kết luận của bài báo ở **Phần 5**.

2. Nghiên cứu liên quan

Có nhiều kỹ thuật hoặc phương pháp xử lý khác nhau được sử dụng trong các hệ thống cảnh báo té ngã. Trong hầu hết các hệ thống phát hiện té ngã, có hai loại phương pháp chính đó là phân tích thống kê và học máy.

Để dự đoán hành động tẽ ngã, phương pháp phân tích áp dụng kỹ thuật thống kê như: phân ngưỡng [3], lọc Bayes [4], mô hình Markov [4]. Trong phương pháp này, sự ngã được dự đoán khi đặc trưng của tín hiệu nhận vào có sự thay đổi. Các cách tiếp cận với phương pháp này thường sử dụng các cảm biến như thiết bị đeo, áp suất, trọng lực,.. để theo dõi dữ liệu.

Trong phương pháp học máy, đối với việc dự đoán hành động té ngã dựa trên video, đã có nhiều nghiên cứu được tiến hành để phân tích các đặc trưng liên quan đến hình dáng và tư thế của con người, luồng quang, và ước tính tư thế,... Các phương pháp máy học như SVM, Random Forest, CNN, RNN, LSTM đã được áp dụng trong các nghiên cứu này [8]. Ví dụ, trong nghiên cứu của tác giả [7], họ đã sử dụng đặc trưng chuyển động và hình dáng cơ thể để phát hiện hành động té ngã. Trong một bài báo khác [20], tác giả đã sử dụng đặc trưng của các khu vực cơ thể người dựa trên biểu đồ giám sát MEWMA kết hợp với mô hình học máy SVM, và kết quả đạt được đạt tỷ lệ chính xác 96.66%. Các tác giả khác [21] cũng đã sử dụng đặc trưng ước tính tư thế cùng với mô hình LSTM và đạt kết quả cao lên đến 97.23%.

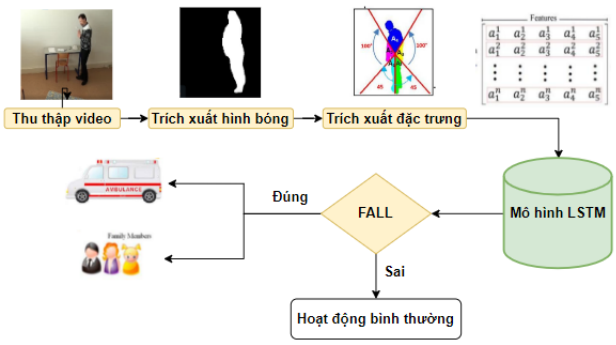
Nhìn chung, trong các nghiên cứu về phát hiện té ngã dựa trên video, phần lớn đều tập trung vào sử dụng các đặc trưng liên quan đến tư thế cơ thể con người như hình bóng và khung xương [8]. Đối với ước tính tư thế con người, hai phương pháp phổ biến được sử dụng là phương pháp Top-down và phương pháp Bottom-up được trình bày trong bài báo [9]. Trong bài báo này, chúng tôi sẽ thử nghiệm cả hai phương pháp này kết hợp với mô hình học máy LSTM để so sánh và đánh giá kết quả.

3. Phương pháp

Trong một số nghiên cứu, người ta đã sử dụng các đặc trưng chuyển động, hình dáng cơ thể, ước lượng tư thế [7, 8],... để phát hiện té ngã dựa trên hai hướng tiếp cận là phương pháp Top-down và phương pháp Bottom-up [9]. Trong bài báo này, chúng tôi sử dụng đặc trưng ước lượng tư thế cùng với hai phương pháp trên để so sánh kết quả.

3.1. Phương pháp Top-down

Phương pháp Top-down sử dụng quy trình hai bước, trước tiên phát hiện các hộp giới hạn của con người (quá trình này có thể kết hợp phân vùng hình ảnh) và sau đó thực hiện trích xuất đặc trưng của một người cho từng hộp giới hạn.



Hình 1: *Sơ đồ thuật toán phương pháp Top-down*

3.1.1. Phân vùng hình ảnh

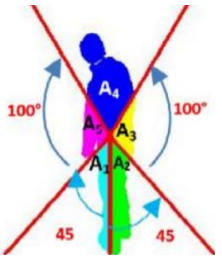
Phân vùng hình ảnh (Image Segmentation) là một nhánh nhỏ trong xử lý ảnh số và thị giác máy tính mục tiêu là gom nhóm các điểm ảnh hoặc các khu vực trong bức ảnh về một trong những các phân lớp mục tiêu nào đó, có thể hiểu như là một thuật toán phân loại các điểm ảnh thuộc các lớp khác nhau [22]. Có nhiều thuật toán nổi tiếng thường được áp dụng trong bài toán này như: Mask-RCNN, U-Net, YOLO, Deep Lap,... Trong nghiên cứu này sử dụng mô hình YOLO cho bài toán phân vùng hình ảnh bởi tốc độ xử lý nhanh, dễ triển khai và mở rộng.

Mục tiêu sau khi phân vùng được vùng chứa người, ta tiến hành khử nền để trích xuất hình bóng nhằm nắm bắt được những đặc trưng quan trọng liên quan đến hình dạng, tư thế của con người (Hình 1) mà không bị ảnh hưởng nhiễu của môi trường nền.

3.1.2. Trích xuất đặc trưng

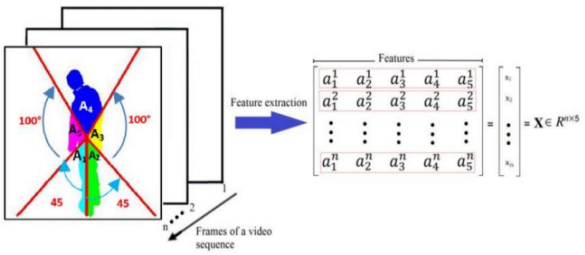
Sau khi khử nền và xác định vùng chứa con người, chúng ta tiến hành trích xuất đặc trưng của các frame. Trong việc trích xuất đặc trưng hình bóng cơ thể thì đã có nhiều đề xuất được sử dụng trước đây, tuy nhiên chúng phụ thuộc vào kích thước và vị trí của hình bóng so với camera. Do đó, chúng ta tiếp cận việc sử dụng năm khu vực cơ thể con người được xác định bằng cách xác định năm đường kẻ từ trọng tâm của hình bóng.

Việc phân vùng được thực hiện bằng cách thiết lập năm đường từ tâm của vật thể. Đường đầu tiên thẳng đứng, hai đường khác được đặt ở góc 45◦ ở hai bên của đường thẳng đứng và hai đường tiếp theo sau đó được đặt góc 100◦ ở cả hai phía của hai đường trước đó [20]. Kết quả thu được 5 vùng A1, A2, A3, A4, A5 như hình sau:



Hình 2: *Chia 5 khu vực hình bóng cơ thể [20]*

Sau đó, từng khu vực được bình thường hóa bằng cách chia giá trị của nó cho diện tích của khung hình. Sau khi trích xuất vector đặc trưng của n frame, ta thu được ma trận 2 chiều với số hàng là n và số cột là số chiều của vector đặc trưng, cụ thể số chiều ở đây là 5:



Hình 3: *Ma trận vector đặc trưng theo 5 khu vực*

3.2. Phương pháp Bottom-up

Phương pháp Bottom-up trong nhận diện té ngã áp dụng một quy trình trích xuất đặc trưng từ các điểm chính không có danh tính và sau đó nhóm chúng thành các phiên bản người dùng khác nhau [10]. Quá trình này không yêu cầu phát hiện hộp giới hạn của con người như phương pháp Top-down.

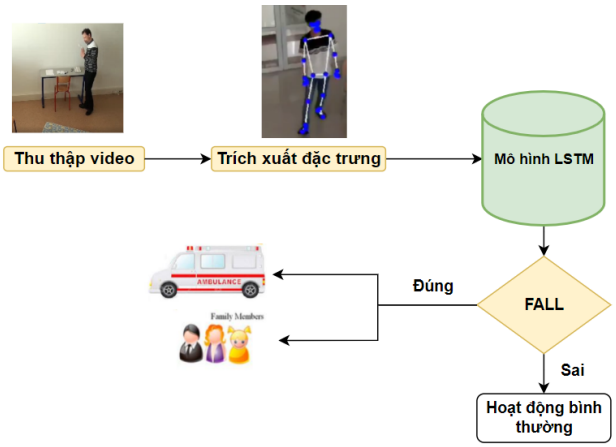
Các bước cơ bản của phương pháp Bottom-up để nhận diện té ngã gồm:

(1) Định vị các điểm chính không có danh tính trên

khung hình. Các điểm chính này có thể là các điểm mốc cơ thể như mắt, mũi, cổ tay, đầu gối, v.v.

(2) Xác định các liên kết giữa các điểm chính để xây dựng các mô hình người khác nhau. Các liên kết này thường được xác định dựa trên vị trí tương đối và khoảng cách giữa các điểm chính.

(3) Phân loại và phân đoạn các phiên bản người dùng để xác định xem có ngã hay không. Điều này có thể được thực hiện bằng cách áp dụng các thuật toán nhận diện hành động hoặc dựa trên các đặc trưng và thông số của các phiên bản người dùng.



Hình 4: *Sơ đồ thuật toán phương pháp Bottom-up*

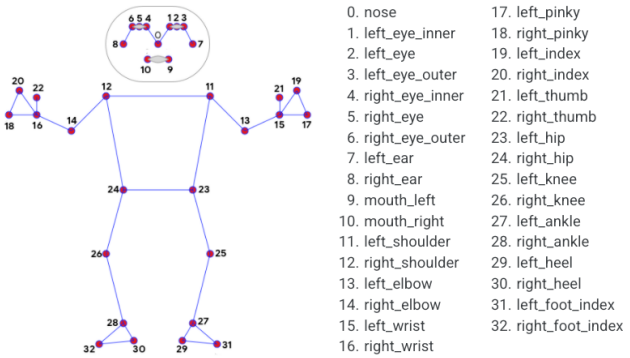
3.2.1. Trích xuất đặc trưng tư thế người

Ước tính tư thế người (human pose estimation) là quá trình nhận dạng và xác định vị trí của các điểm quan trọng trên cơ thể con người trong ảnh hoặc video. Mục tiêu chính là định vị và ước tính vị trí của các điểm như đầu, vai, khuỷu tay, khuỷu tay, đầu gối và chân,.. nhằm phân tích hành vi và hành động của con người.

Có hai kỹ thuật chính trong đó các mô hình ước lượng tư thế có thể phát hiện các tư thế của con người [12]:

- Ước tính tư thế 2D: Trong loại ước tính tư thế này, bạn chỉ cần ước tính vị trí của các khớp cơ thể trong không gian 2D so với dữ liệu đầu vào (tức là khung hình ảnh hoặc video). Vị trí được biểu thị bằng tọa độ X và Y cho từng điểm chính.
- Ước tính tư thế 3D: Trong loại ước tính tư thế này, bạn chuyển đổi hình ảnh 2D thành đối tượng 3D bằng cách ước tính một kích thước Z bổ sung cho dự đoán. Ước tính tư thế 3D cho phép chúng tôi dự đoán vị trí không gian chính xác của một người hoặc vật thể hiện.

Kỹ thuật ước tính tư thế 2D đơn giản, cần ít tài nguyên tính toán hơn kỹ thuật ước tính tư thế 3D. Tuy nhiên nó chỉ cung cấp thông tin trong không gian 2D và không thể xác định chính xác độ sâu và góc quay của cơ thể. Điều này có thể gây ra sự mất mát thông tin quan trọng trong quá trình phát hiện té ngã và làm giảm độ chính xác của hệ thống. Ngoài ra, nó còn nhạy cảm với góc nhìn, do đó có thể dẫn đến sai sót trong việc xác định tư thế và làm giảm độ tin cậy của hệ thống. Do đó, chúng tôi quyết định sử dụng kỹ thuật ước tính tư thế 3D trong dự án. Cụ thể trong dự án sử dụng thư viện MediaPipe Pose để trích xuất đặc trưng khung xương nhằm ước tính tư thế con người.



Hình 5: *Các vị trí khung xương trong MediaPipe Pose*

Quá trình nhận diện của Pose Classification dựa trên solution về BlazePose, nó sẽ bóc tách các điểm trên cơ thể trong không gian 3 chiều (x, y, z) từ một RGB video

Kiến trúc của BlazePose gồm 2 thành phần chính:

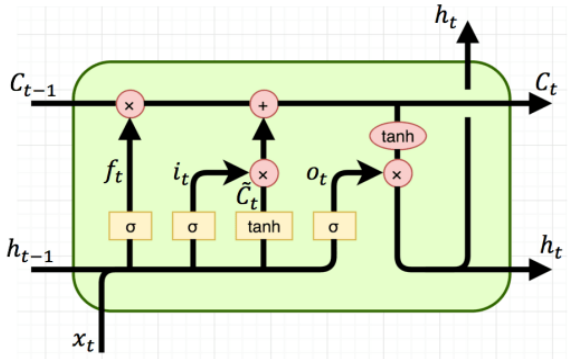
- Pose Detector: Để detect ra vùng chứa person trên bức ảnh
- Pose Tracker: Để trích xuất ra các keypoints trên vùng chứa vị trí của person đã được crop ra từ bức ảnh đồng thời dự đoán vị trí của person trong frame tiếp theo.

MediaPipe Pose [23] trích xuất 33 điểm theo vị trí khung xương (hình 3) để ước tính tư thế con người. Thư viện tính toán trên không gian ba chiều với bốn đặc trưng cho mỗi vị trí khung xương bao gồm: x, y, z, visibility (khả năng nhìn thấy). Khả năng nhìn thấy là một giá trị trong khoảng 0-1 thể hiện một điểm có khả năng được nhìn thấy hay bị che khuất trong không gian.

3.3. Phân loại với mạng LSTM

3.3.1. Kiến trúc

Mạng LSTM được cải tiến từ mạng thần kinh hồi quy (RNN–Recurrent Neural Network) nhằm khắc phục những nhược điểm về phụ thuộc xa (Long–term Dependency) của mạng RNN truyền thống. LSTM được giới thiệu bởi [8] và càng ngày càng được cải tiến [9].



Hình 6: *Cấu trúc một nút mạng trong mạng LSTM*

Về mặt lý thuyết, RNN có khả năng xử lý các phụ thuộc theo thời gian (temporal dependencies) bằng việc sử dụng bộ nhớ ngắn hạn và dựa trên việc xác định (luyện) các tham số một cách hiệu quả [10]. Tuy nhiên, đáng tiếc trong thực tế RNN không thể giải quyết các phụ thuộc theo thời gian khi chuỗi số liệu có các phụ thuộc xa (long–term dependencies) do vấn đề độ dốc biến mất (vanishing gradient problem) [18, 19].

LSTM có cấu trúc dạng chuỗi các nút mạng như RNN, nhưng cấu trúc bên trong thì lại phức tạp hơn, bao gồm 4 tầng tương tác với nhau (Hình 3). Điểm đặc biệt của mạng LSTM nằm ở trạng thái ô C (cell state), nơi lưu trữ các trọng số dài hạn của mô hình. Các thông số trạng thái ô C, trạng thái ẩn h (hidden state), đầu vào tại thời điểm t là x_t được đưa vào nút mạng. Sau khi được xử lý qua các hàm kích hoạt sigmoid σ, tanh và các phép toán véc–tơ, kết quả đầu ra là trạng thái ô C và trạng thái ẩn h tại thời điểm t sẽ được sử dụng cho nút mạng t+1 tiếp theo.

Với bài toán phát hiện té ngã mang đặc trưng thông tin chuỗi thời gian để xác định hành động con người, do đó nhóm đã chọn mô hình LSTM để huấn luyện.

3.3.2. Huấn luyện

Quá trình huấn luyện mô hình LSTM (Long Short-Term Memory) là một bước quan trọng trong việc xây dựng mô hình nhận diện hành động. Trong báo cáo này, chúng ta sẽ xem xét các thông số và kết quả quan trọng trong quá trình huấn luyện mô hình.

Thông số huấn luyện:

a) Optimizer: Mô hình sử dụng thuật toán tối ưu hóa Adam để điều chỉnh các trọng số và cập nhật chúng trong quá trình huấn luyện.
b) Hàm mất mát: Hàm mất mát được sử dụng là binary_crossentropy. Đây là một hàm mất mát phổ biến được sử dụng trong các bài toán phân loại nhị phân.
c) Độ đo được sử dụng trong quá trình huấn luyện là accuracy. Độ đo này đánh giá độ chính xác của mô hình trên các lớp đầu ra.
d) Số epochs: Mô hình được huấn luyện trong 16 epochs. Mỗi epoch đại diện cho việc đưa toàn bộ dữ liệu huấn luyện qua mạng một lần.

3.4. Đánh giá hiệu năng

3.4.1. Các độ đo đánh giá hiệu năng segmentation

Các độ đo segmentation phổ biến bao gồm độ chính xác điểm ảnh (pixel accuracy), mIoU (mean Intersection over Union), AP (Average Precision) hoặc mAP (mean Average Precision),... Chúng tôi sử dụng độ đo mAP để đánh giá hiệu năng segmentation bởi vì mAP kết hợp giữa precision, recall và các giá trị IoU (Intersection over Union), nó cung cấp một cái nhìn toàn diện về hiệu năng của mô hình phân đoạn. Điều này giúp mAP trở thành một phép đo được công nhận và sử dụng phổ biến để so sánh và đánh giá các phương pháp và mô hình trong bài toán segmentation [6, 7].

Công thức tính độ chính xác trung bình (mAP):

$$mAP = \sum_{q=1}^{Q} \frac{AP(q)}{Q} \qquad (1)$$

Trong đó Q là số lượng lớp đối tượng có trong tập dữ liệu, AP là độ chính xác trung bình của từng lớp được tính bằng công thức như sau:

$$AP = \frac{1}{11} \sum_{r\epsilon(0,0.1,\dots,1)} p_{interp}(r) \qquad (2)$$

Với $p_{interp}(r)$ được tính bằng công thức:

$$p_{interp}(r) = max_{\tilde{r}:\tilde{r}\geq r}\, p(\tilde{r}) \qquad (3)$$

Trong đó $p(\tilde{r})$ là độ chính xác (precision) được đo tại độ phủ (recall) $\tilde{r}$.

3.4.2. Các độ đo đánh giá hiệu năng phân loại

Sử dụng các độ đo bao gồm accuracy, precision, recall và confusion matrix để đánh giá hiệu năng phân loại của mô hình. Trong đó, precision giúp đánh giá được trong số dự đoán positive thực sự đúng, và kể cả trong trường hợp có imbalance class thì vẫn hiểu được hiệu năng của mô hình. Recall giúp hiểu được mức độ bỏ lỡ những điểm dữ liệu thực sự positive. Và confusion matrix sẽ thể hiện được mô hình đang gặp khó khăn khi phân định những lớp cụ thể nào.

4. Thực nghiệm và đánh giá kết quả

4.1. Dữ liệu

4.1.1. Bộ dữ liệu

Bộ dữ liệu bao gồm 430 video trong đó có 235 video mô tả các hoạt động bình thường như đi, đứng, ngồi, quỳ gối, nhảy, nhặt đồ và 195 video ngã được thu thập từ các nguồn: CAUCAFall Dataset [11], NTU RGB+D Dataset [12], Fall Dataset ImVia [14] và nhóm tiến hành tự quay. Môi trường của dataset có ánh sáng vừa đủ với tốc độ khung hình là 30fps.

Dữ liệu sau khi thu thập, chúng tôi đã chọn lọc những video có độ tương đồng góc nhìn camera quan sát.

Hình 7: *Vị trí đặt góc quay camera trong dataset*
(a) Góc quay ngang; (b) Góc quay từ trên xuống

Dữ liệu đã được xử lý để trong video đó chỉ chứa hành động ngã để tránh đánh nhãn fall cho những phân đoạn có các hành động khác như đi, ngồi,... Ngoài ra, để đồng bộ thông số các video thì dữ liệu đã được xử lý chuyển về fps bằng 30 và kích thước khung hình là 640 x 480.

Dữ liệu được nhóm phân chia thành các tập train, valid và test với tỉ lệ 70%, 10% và 20%.

4.1.2. Gán nhãn

Dữ liệu video được phân tách thành các quan sát với mỗi quan sát gồm có n frame, với n được thử nghiệm với các giá trị 30, 35, 40. Số lượng frame trong 1 quan sát phải phù hợp với lượng thời gian té ngã của một người để đảm bảo quá trình phát hiện té ngã không bị nhầm lẫn. Cụ thể, với tập dataset của chúng tôi thì video ngã có thời lượng từ 1-1.5s với fps bằng 30, do đó chúng tôi thử nghiệm số frame trong 1 quan sát lần lượt là 30, 35, 40 để so sánh kết quả.

Sau khi phân tách video thành các quan sát, các quan sát được gán nhãn thành 2 lớp là 0 (no fall) và 1 (fall).

4.2. Kết quả phương pháp Top-down

4.2.1. Kết quả segmentation với YOLOv8

Với mô hình yolov8n-seg mặc định [13] cho thấy kết quả hoạt động tốt ở các hành động đứng, ngồi,... Tuy nhiên, với 1 chuỗi hành động trong thế giới thực thì có rất nhiều trường hợp nó không tìm được vùng chứa người trong khung hình, điều này dẫn đến sự mất mát thông tin, cũng như có thể làm quá trình phát hiện té ngã tăng khả

năng nhầm lẫn. Do đó, nhóm đã huấn luyện lại mô hình yolov8n-seg dựa trên dữ liệu nhóm tự tạo bao gồm 1273 ảnh được chia thành 3 phần: tập train (874 ảnh), tập valid (183 ảnh) và test (116 ảnh).

Hình 8: *Kết quả segmentation ở tập valid*

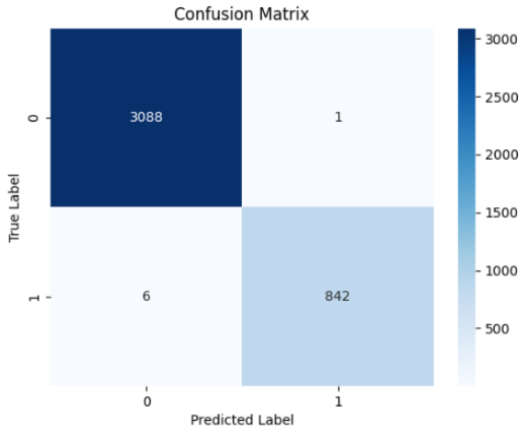
Độ chính xác trung bình mAP thu được là 0.995 với ngưỡng IoU là 0.5.

4.2.2. Kết quả phân loại với mô mình LSTM

Bảng 1: *Độ chính xác mô hình LSTM với phương pháp Top-down*

Số frame trong 1 quan sát	Accuracy test	F1 score
30	99.58%	0.9916
35	99.79%	0.9956
40	99.82%	0.9959

Kết quả tốt nhất với số frame trong 1 quan sát là 40 frame, ma trận nhầm lẫn thu được:



Hình 9: *Ma trận nhầm lẫn phương pháp Top-down*

Kết quả nhận diện trên video:

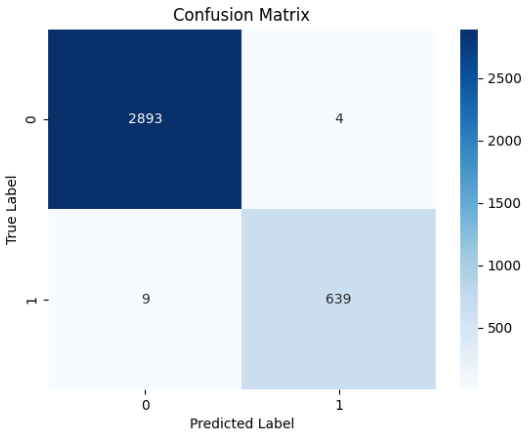
Hình 10: *Kết quả nhận diện với phương pháp Top-down*

4.3. Kết quả phương pháp Bottom-up

Bảng 2: *Độ chính xác mô hình LSTM với phương pháp Bottom-up*

Số frame trong 1 quan sát	Accuracy Test	F1 score
30	99.26%	0.9818
35	99.63%	0.9899
40	99.59%	0.9875

Tương tự như phương pháp Top-down, chúng tôi cũng tiến hành thử nghiệm với số frame trong 1 quan sát lần lượt là 30, 35, 40. Kết quả chính xác nhất thu được với số frame trong 1 quan sát là 35 frame, ma trận nhầm lẫn thu được:



Hình 11: *Ma trận nhầm lẫn phương pháp Bottom-up*

Kết quả nhận diện trên video:

Hình 12: *Kết quả nhận diện với phương pháp Bottom-up*

4.4. Kết quả so sánh hai phương pháp

Số frame trong 1 quan sát là 30 frames

Kết quả được thử nghiệm ở bộ vi xử lý: Intel(R) Core(TM) i7-10750H CPU @ 2.60GHz 2.59 GHz

Bảng 3: *So sánh kết quả hai phương pháp*

Tiêu chí	Top-down	Bottom-up
Thời gian trung bình trích xuất đặc trưng 1 frame	0.12s	0.02s
Thời gian trung bình dự đoán 1 quan sát	4.21s	1.11s
Độ chính xác trên tập test	99.58%	99.26%
F1 score	0.9916	0.9818

Từ bảng kết quả trên, ta nhận thấy rằng với hướng tiếp cận Bottom-up thì ta thu được kết quả tốt hơn về mặt hiệu suất, tuy độ chính xác có thấp hơn nhưng nhìn chung với các phiên bản huấn luyện mà nhóm thu được thì độ chính xác giữa phương pháp Top-down và Bottom-up tương đối gần bằng nhau.

5. Bàn luận

5.1. Bối cảnh, giả thuyết, mục tiêu nghiên cứu

Nghiên cứu này tập trung vào việc phát triển hệ thống phát hiện té ngã trong phòng sử dụng camera. Mục tiêu

của nghiên cứu là đánh giá hiệu quả của hai phương pháp tiếp cận phổ biến, đó là phương pháp Top-down và phương pháp Bottom-up, khi được kết hợp với mô hình học máy LSTM. Phát hiện và dự đoán hành động té ngã là phát hiện chính của nghiên cứu.

5.2. So sánh kết quả với các nghiên cứu trước

Các nghiên cứu khác thường chỉ thực hiện trên một phương pháp là Top-down hoặc Bottom-up, nghiên cứu này thực hiện trên cả hai phương pháp để so sánh ưu và nhược điểm của hai phương pháp này. Ngoài ra, bài báo này giúp độc giả có cái nhìn tổng quan các phương pháp trong việc phát hiện té ngã dựa trên camera.

Phương pháp Top-down đã được sử dụng nhiều trong các nghiên cứu và được chứng minh là có hiệu quả trong việc phát hiện và dự đoán té ngã. Cụ thể trong bài báo [20], tác giả đã sử dụng biểu đồ giám sát MEWMA và bộ phân loại SVM cho bài toán phát hiện té ngã và cho độ chính xác lên đến 96.66%. Chúng tôi dựa trên cách trích xuất vector đặc trưng đề xuất trong bài báo [20], với việc sử dụng mô hình YOLOv8 để khử nền và mô hình LSTM để phân loại đã cho kết quả chính xác cao nhất lên đến 99.82%.

Phương pháp Bottom-up mới mẻ hơn và chưa được nghiên cứu sâu trong lĩnh vực này. Nghiên cứu của các tác giả trong tài liệu [9], [24], [25] về ước tính tư thế 2D có độ nhầm lẫn cao khi thay đổi về góc quay đặt camera, với độ chính xác cao nhất là 94.6%. Trong nghiên cứu này, chúng tôi thử nghiệm ước tính tư thế 3D đem lại kết quả tốt hơn so với các góc camera khác nhau với độ chính xác cao nhất lên đến 99.63%.

5.3. Bàn luận về điểm mạnh, điểm yếu của việc nghiên cứu

Phương pháp Top-down có một số điểm mạnh đáng kể. Phương pháp Top-down ổn định hơn và phương pháp này có khả năng kết hợp với phân vùng hình ảnh, cho phép xác định các phần thân thể và tạo ra đặc trưng tư thế cơ thể để hỗ trợ trong việc nhận diện hành động té ngã.

Tuy nhiên, phương pháp Top-down cũng có một số điểm yếu. Thứ nhất, phương pháp đòi hỏi quá trình phát hiện và theo dõi hộp giới hạn của con người một cách chính xác. Điều này có thể gặp khó khăn trong các tình huống mà con người không được phân tách rõ ràng hoặc bị che bởi các vật thể khác. Thứ hai, phương pháp Top-down có thể đòi hỏi công nghệ phân vùng hình ảnh phức tạp để đảm bảo tính chính xác trong việc xác định các phần thân thể, trong nghiên cứu này chúng tôi sử dụng YOLOv8. Điều này có thể làm tăng độ phức tạp và tốn kém về thời gian tính toán.

Đối với phương pháp Bottom-up cũng có những điểm mạnh và điểm yếu khác biệt. Đầu tiên, phương pháp này không yêu cầu quá trình phát hiện hộp giới hạn của con người. Thay vào đó, nó tập trung vào việc định vị các điểm chính không có danh tính và xây dựng các mô hình người từ các liên kết giữa chúng. Điều này giúp giảm bớt sự phụ thuộc vào quá trình phát hiện hộp giới hạn và có thể áp dụng trong các tình huống mà con người không được phân tách rõ ràng. Thứ hai, phương pháp Bottom-up có thể nhận biết các hành động phức tạp và nhóm chúng thành các phiên bản người dùng khác nhau, đồng thời hỗ trợ trong việc phát hiện hành động té ngã.

Phương pháp Bottom-up cũng có một số điểm yếu. Đầu tiên, việc xây dựng các mô hình người từ các điểm chính không có danh tính có thể gặp khó khăn trong việc xác định các liên kết chính xác và đảm bảo tính chính xác trong việc nhận diện hành động. Thứ hai, phương pháp này có thể đòi hỏi phân đoạn và phân loại các phiên bản người dùng để xác định xem có ngã hay không, và điều này có thể tốn kém về thời gian tính toán.

6. Kết luận

Trên cơ sở các nghiên cứu và phân tích trong lĩnh vực phát hiện té ngã dựa trên video, bài báo nghiên cứu này đã tìm hiểu và đánh giá hiệu quả của hai phương pháp tiếp cận phổ biến, đó là phương pháp Top-down và phương pháp Bottom-up, kết hợp với mô hình học máy LSTM.

Việc sử dụng mô hình LSTM cho phép mô hình học được các quan hệ dữ liệu thời gian và hỗ trợ trong việc dự đoán hành động té ngã. Đồng thời, việc sử dụng các đặc trưng tư thế cơ thể đã cung cấp thông tin quan trọng về vị trí và chuyển động của con người trong video. Nghiên cứu này đã đạt được kết quả đáng chú ý trong việc phát hiện té ngã trên video với độ chính xác trên 99%.

Bài báo nghiên cứu này đã đóng góp vào việc phát triển các phương pháp và kỹ thuật phát hiện té ngã dựa trên video. Trong nghiên cứu này vẫn còn một số vấn đề chưa được xử lý, bao gồm sự hiện diện của vật cản, sự thay đổi độ sáng, và tình huống có nhiều người trong video. Các nghiên cứu trong tương lai có thể tìm hiểu mở rộng phạm vi của bài toán trong trường hợp có nhiều hơn một người trong video, cũng như xử lý các vấn đề có nhiều yếu tố nhiễu từ môi trường.

Tài liệu tham khảo

[1] "Phòng chống té ngã ở người cao tuổi", https://moh.gov.vn/web/phong-chong-tai-nan-thuong-tich/tin-noi-bat/-/asset_publisher/iinMRn208ZoI/content/phong-chong-te-nga-o-nguoi-cao-tuoi, truy cập ngày

[2] O. Levy and A. amp, " Falls are leading cause of death for seniors 65 and older, according to CDC" CDC, 03-Feb-2020.

[3] Medrano, C., Igual, R., García-Magariño, I., Plaza, I., & Azuara, G. (2017). Combining novelty detectors to improve accelerometer-based fall detection. Medical & Biological Engineering & Computing, 55(10), 1849–1858.

[4] Šeketa, G., Vugrin, J., & Lacković, I. (2017). Optimal threshold selection for acceleration-based fall detection. In International Conference on Biomedical and Health Informatics (pp. 151–155). Springer.

[5] Chaudhry, H. N., Javed, Y., Kulsoom, F., Mehmood, Z., Khan, Z. I., Shoaib, U., & Janjua, S. H. (2021). Sentiment analysis of before and after elections: Twitter data of US election 2020. Electronics, 10(17), 2082.

[6] Hsi-Chun Kao; Jui-Chung Hung; Chih-Pang Huang, GA-SVM applied to the fall detection system, 13-17 May 2017

[7] Nguyễn Việt Anh, "Phát hiện ngã sử dụng đặc trưng chuyển động và hình dạng cơ thể dựa trên camera đơn", 2016.

[8] Jesús Gutiérrez, Víctor Rodríguez and Sergio Martin, "Comprehensive Review of Vision-Based Fall Detection Systems", 2021.

[9] Qi Dang, Jianqin Yin, Bin Wang, and Wenqing Zheng, "2D Human Pose Estimation: A Survey", 2019.

[10] Evaluation metrics for object detection and segmentation: mAP, "https://kharshit.github.io/blog/2019/09/20/evaluation-metrics-for-object-detection-and-segmentation", truy cập ngày

[11] M. Everingham, L. Van Gool, C. K. I. Williams, and J. Winn, "The PASCAL Visual Object Classes (VOC) Challenge," pp. 303-338, 2010.

[12] S. Hochreiter and J. Schmidhuber, "Long short-term memory," Neural Computation, vol. 9, pp. 1735–1780, 1997.

[13] Kaisheng Yao, Trevor Cohn, Katerina Vylomova, Kevin Duh, Chris Dyer, "Depth-Gated LSTM", 2015.

[14] Jan Koutnik, Klaus Greff, Faustino Gomez, Jurgen Schmidhuber, "A Clockwork RNN", 2014.

[15] Alex Sherstinsky, "Fundamentals of Recurrent Neural Network (RNN) and Long Short-Term Memory (LSTM) Network", 2020.

[16] A Comprehensive Guide on Human Pose Estimation, "https://www.analyticsvidhya.com/blog/2022/01/a-comprehensive-guide-on-human-pose-estimation".

[17] Segment - Ultralytics YOLOv8 Docs, https://docs.ultralytics.com/tasks/segment/#models, truy cập ngày

[18] Y. Bengio, P. Y. Simard, and P. Frasconi, "Learning long-term dependencies with gradient descent is difficult," IEEE Trans. Neural Networks, vol. 5, pp. 157–166, 1994.

[19] J. F. Kolen and S. C. Kremer, Gradient Flow in Recurrent Nets: The Difficulty of Learning LongTerm Dependencies. WileyIEEE Press, 2001, pp. 237–243. [Online]. Available: http://ieeexplore.ieee.org/xpl/articleDetails.jsp?arnumber=5264952

[20] Harrou, Fouzi; Zerrouki, Nabil; Sun, Ying; Houacine, Amrane, "Vision-based fall detection system for improving safety of elderly people", 2017.

[21] Chatchai Wangwiwattana, "Fall Detection with a Single Commodity RGB Camera Based-on 2D Pose Estimation", 2019.

[22] S. Minaee, Y. Boykov, F. Porikli, A. Plaza, N. Kehtarnavaz, and D. Terzopoulos, "Image Segmentation Using Deep Learning: A Survey," IEEE Trans Pattern Anal Mach Intell, vol. 44, no. 7, pp. 3523–3542, Jan. 2020, doi: 10.48550/arxiv.2001.05566.

[23] "MediaPipe Solutions", https://developers.google.com/mediapipe

[24] Chuan-Bi Lin, "A Framework for Fall Detection Based on OpenPose Skeleton and LSTM/GRU Models"

[25] Cao, Z.; Simon, T.; Wei, S.; Sheikh, Y. Realtime multi-person 2D Pose Estimation Using Part Affinity Fields. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017