GIẢI PHÁP PHÂN LOẠI TƯƠNG TÁC GIỮA 2 NGƯỜI TRONG CHUỖI ẢNH RỜI RẠC (PHẦN I)

TS. Đỗ Văn Khánh, TS. Lê Xuân Đức, TS. Nguyễn Anh Tú
Phòng Thí nghiệm trọng điểm An toàn thông tin, Bộ Tư lệnh 86

Ngày nay, công nghệ trí tuệ nhân tạo (AI) có vai trò hết sức quan trọng trong mọi lĩnh vực của đời sống. Trong đó, lĩnh vực an toàn thông tin, giám sát an ninh thông minh có tiềm năng ứng dụng rất lớn. Bên cạnh các giải pháp như phát hiện mạng Botnet [1], phát hiện tấn công trinh sát mạng [2], việc ứng dụng AI trong giám sát an ninh, hỗ trợ điều tra tội phạm cũng đang được nghiên cứu, phát triển và ứng dụng rộng rãi. Trong bài báo này, nhóm tác giả đề xuất giải pháp sử dụng mô hình mạng nơ-ron tinh gọn phân loại tương tác giữa 2 người trong chuỗi ảnh rời rạc. Kết quả nghiên cứu có vai trò quan trọng làm cơ sở xây dựng và phát triển các mô hình phân loại hành động bất thường, phát hiện xâm nhập.

GIỚI THIỆU

Mục tiêu của bài báo là phân tích tương tác của con người trong các chuỗi ảnh rời rạc hoặc video. Dữ liệu có thể được trích ra từ các đoạn video hoặc các bộ sưu tập ảnh trên Internet. Có 2 cách tiếp cận chính trong bài toán nhận dạng hành động là nhận dạng hành động trực tiếp qua dữ liệu video hoặc dựa vào các điểm chính trên khung xương (skeleton-based methods) [3], trong đó tọa độ các điểm chính trên khung xương được xác định trước. Một số giải pháp phổ biến xác định tọa độ các điểm chính trên khung xương người (skeleton estimation) có thể kể đến như: OpenPose [4], DeepPose và DeeperCut.

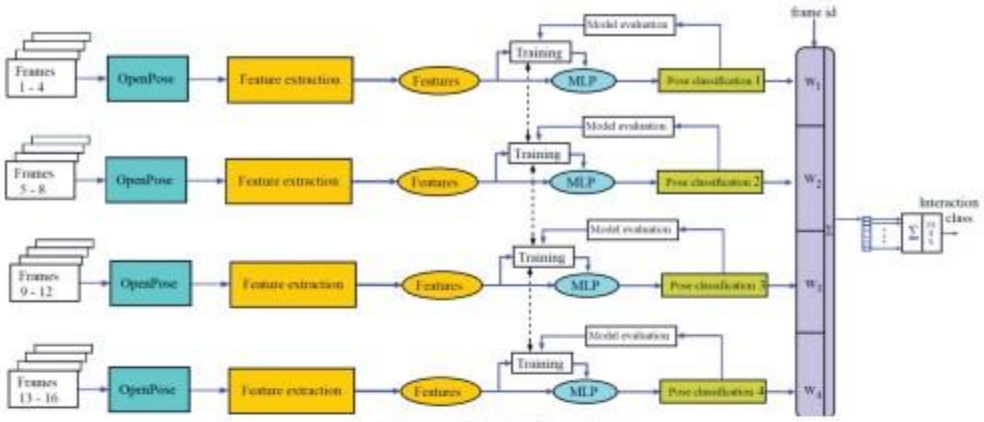
Trong nghiên cứu này, nhóm tác giả sẽ tập trung nhận dạng tương tác giữa 2 người trong chuỗi ảnh rời rạc, giả định rằng dữ liệu về các điểm chính trên khung xương được xác định trước. Nhóm tác giả đã xây dựng các bộ phân loại chuyên gia (expert classifier) cho các phân đoạn khác nhau của hành động. Mỗi hành động được chia ra thành 4 phân đoạn, gồm có: giai đoạn bắt đầu, giai đoạn trung gian thứ nhất, giai đoạn trung gian thứ hai và giai đoạn kết thúc. Kết quả cuối cùng sẽ được tính toán dựa trên kết quả phân loại hành động của 4 bộ phân loại chuyên gia tại mỗi phân đoạn.

Để có thể cung cấp chi tiết kết quả nghiên cứu, nhóm tác giả sẽ chia nội dung trình bày giải pháp thành hai phần. Trong phần I này, nhóm tác giả sẽ giới thiệu khái quát vấn đề nghiên cứu; các giải pháp truyền thống để giải quyết bài toán; kỹ thuật phát hiện điểm chính trên khung xương, thuật toán gộp điểm và trích xuất đặc trưng cho mô hình.

GIẢI PHÁP

Kiến trúc

Dữ liệu vào cho bộ phân loại hành động là một chuỗi các khung ảnh (frame) rời rạc từ video. Giả



Hình 1. Kiến trúc của mô hình

định rằng thời điểm bắt đầu và kết thúc hành động của một đoạn video được xác định trước. Đoạn video sẽ được chia thành M đoạn ngắn liên tiếp (chẳng hạn $M = 16$). Chọn 1 frame từ mỗi đoạn ngắn này để phân loại hành động. Giả sử rằng $M=N*m$, trong đó: N là số phân đoạn của hành động, m là số frame được chọn trong mỗi phân đoạn này.

Trong nghiên cứu này, nhóm tác giả chọn $N=4$, bao gồm các phân đoạn: bắt đầu, giai đoạn trung gian thứ 1, giai đoạn trung gian thứ 2 và giai đoạn kết thúc. Các bộ phân loại yếu được xây dựng để phân loại hành động cho mỗi phân đoạn này. Kiến trúc tổng thể của giải pháp được thể hiện trong Hình 1, bao gồm một số bước xử lý sau:

Bước 1. Ước lượng tọa độ các điểm chính trên khung xương: Sử dụng thư viện OpenPose [5] để phát hiện tọa độ các khớp xương chính trên cơ thể người.

Bước 2. Trích xuất vectơ đặc trưng: Thuật toán *"keypoint enhancement algorithm"* được đưa ra để trích xuất ra 2 tập điểm đáng tin cậy tương ứng với tọa độ các khớp xương chính của 2 người trong các frame ảnh; tiếp theo, các vectơ đặc trưng tương ứng được trích xuất từ 2 tập điểm đầu ra của thuật toán.

Bước 3. Huấn luyện và đánh giá các bộ phân loại chuyên gia: Huấn luyện các mạng nơ-ron tinh gọn để phân loại hành động trong mỗi phân đoạn bằng bộ phân loại yếu; tối ưu siêu tham số các mạng nơ-ron sử dụng thư viện Keras-tuner; đánh giá các mô hình.

Bước 4. Bộ phân loại kết hợp: Xây dựng mạng nơ-ron kết hợp từ các bộ phân loại yếu để phân loại tương tác giữa 2 người. Mỗi dữ liệu vào, mô hình sẽ đưa ra phân bố xác suất trên tập hợp các hành động đầu ra; hành động có xác suất lớn nhất được lựa chọn làm kết quả phân loại cuối cùng.

Bước 5. Kiểm thử mô hình: Các mô hình xây dựng được kiểm thử trên 2 bộ dữ liệu, bao gồm bộ dữ liệu tự xây dựng humiact5 dataset và bộ dữ liệu NTU RGB+D.

Ước lượng tọa độ các điểm chính trên khung xương

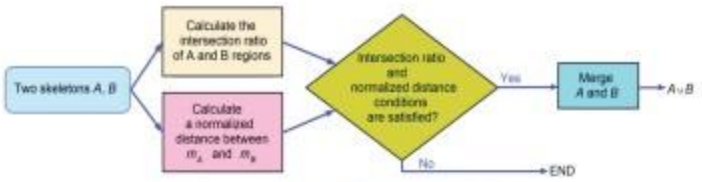
Trong nghiên cứu này, nhóm tác giả sử dụng mô hình *body_25_model* của thư viện OpenPose để trích xuất tọa độ 25 điểm chính trên mỗi bộ khung xương (có chỉ số tham chiếu từ 0 đến 24), tương đương với 1 mảng 25 phần tử. Mỗi phần tử của mảng gồm thông tin tọa độ và điểm tin cậy.

Trích xuất vectơ đặc trưng

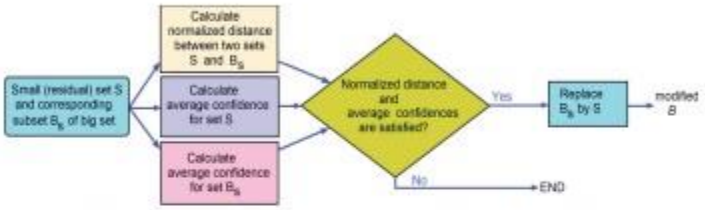
Từ các tập điểm chính trên khung xương trích xuất bởi OpenPose, thuật toán gộp điểm được sử dụng để trích xuất ra 2 tập điểm đại diện cho mỗi frame dựa trên thước đo về độ lớn, cụ thể: 2 tập điểm có sự thay đổi về tọa độ lớn nhất được chọn để trích xuất đặc trưng.

Skeleton enhancement: Có những trường hợp OpenPose nhận dạng sai và chia tách các điểm chính của 1 người thành các tập điểm khác nhau do người bị che khuất hoặc ảnh có độ phân dải thấp. Vì vậy, nhóm tác giả đã phát triển một thuật toán cho phép tìm kiếm, thay thế và gộp

các tập điểm có khả năng cao là thuộc về một người. Đầu tiên, cố gắng gộp các tập điểm khi có thể để tạo ra các tập điểm đáng tin cậy hơn (Hình 2).



Hình 2. Các bước tính toán gộp 2 tập điểm



Hình 3. Các bước thay thế các điểm độ tin cậy thấp trên khung xương

Các công thức tính toán cho mỗi 2 tập điểm bao gồm các chỉ số được chuẩn hóa về số lượng điểm giao nhau và khoảng cách giữa 2 tập điểm. Các chỉ số tính toán được so sánh với các ngưỡng tham số thực nghiệm tương ứng để quyết định xem 2 tập đang xem xét có thể gộp lại với nhau hay không. Nếu các chỉ số này thỏa mãn điều kiện, khi gộp 2 tập điểm, các điểm có xác suất cao hơn tại các điểm giao sẽ được chọn.

Trình bày một cách rõ ràng hơn, mỗi tập điểm ứng với một bộ khung xương từ mô hình *body_25_model* là một mảng gồm cố định 25 phần tử. Tuy nhiên, phần tử tại chỉ số *i* của mảng có các giá trị tọa độ và điểm tin cậy đều bằng 0 sẽ được coi là phần tử khuyết trong tập điểm đó. Một tập điểm có thể bị khuyết một hoặc nhiều phần tử. Ví dụ, có 2 tập điểm *A* và *B* như sau: *A={0,1,2,5,8,23}* ,*B={0,5,6,7,8,9,10,24}*. Các chỉ số khuyết của tập *A* và tập *B* (các chỉ số có các giá trị tọa độ và điểm tin cậy đều bằng 0) sẽ không xuất hiện trong 2 tập này. Với 2 tập đã cho, các điểm giao là các điểm có chỉ số 0, 5, 8.

Giả sử, hai tập điểm *A* và *B* được xem xét để gộp lại với nhau. Không mất tính tổng quát, giả sử tập *A* là tập bé, tập *B* là tập lớn (tập bé là tập có số lượng điểm nhỏ hơn tập lớn). Hai tập *A*, *B* có điểm trung bình lần lượt là m_A và m_B. Độ lệch chuẩn của các tập điểm *A*, *B* theo trục O_x và O_y lần lượt là $[std_{A,x}, std_{A,y}]$ và $[std_{B,x}, std_{B,y}]$. Điều kiện để gộp như sau:

$$\frac{|A \cap B|}{|B|} \leq \theta_1 \quad (1)$$

$$\frac{|m_{A,x} - m_{B,x}|}{std_{A,x} + std_{B,x}} + \frac{|m_{A,y} - m_{B,y}|}{std_{A,y} + std_{B,y}} \leq \theta_2 \quad (2)$$

Trong đó, θ_1, θ_2 là các ngưỡng về số lượng điểm giao và về khoảng cách để có thể gộp 2 tập. Gọi tập các điểm giao trong tập nhỏ *A* là *S*, tập các điểm giao trong tập lớn *B* là B_s. Tập *S* sẽ được xem xét để thay thế tập B_s. Để quyết định điều này, các chỉ số sau cần được tính toán: Khoảng cách Euclidean chuẩn hóa giữa 2 tập điểm *S* và B_s, điểm tin cậy trung bình của tất cả các điểm trong tập *S* và tập B_s (Hình 3). Các giá trị phương sai tiêu chuẩn của tập nhỏ là $[Std_{(S,x)}, Std_{(S,y)}]$. Giá trị xác xuất độ tin cậy của điểm *j* thuộc tập điểm được ký hiệu bởi *P(j)*. Điều kiện thay thế các điểm độ tin cậy thấp như sau:

$$\frac{1}{Std_{S,x} + Std_{S,y}} \sum_{i=1}^{N} \sqrt{\left((x_{S_i} - x_{B_{S_i}})^2 + (y_{S_i} - y_{B_{S_i}})^2\right)} \leq \theta_3 \quad (3)$$

$$\frac{1}{N} \sum_{i=1}^{N} P(S_i) \geq \theta_4 \quad (4)$$

$$\frac{1}{N} \sum_{i=1}^{N} P(B_{S_i}) \leq \theta_5 \quad (5)$$

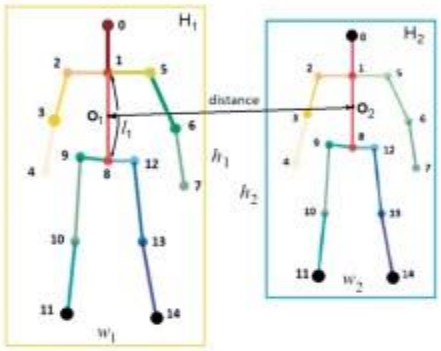
Trong đó, θ_3 là ngưỡng khoảng cách Euclidean tiêu chuẩn, θ_4 là ngưỡng độ tin cậy cận dưới cho tập *S*, và θ_5 là ngưỡng độ tin cậy cận trên cho tập B_s.

Sau khi thực hiện thuật toán gộp, các tập điểm còn lại sẽ được sắp xếp từ lớn đến bé bởi kích thước hình chữ nhật bao (bounding box) các tập điểm này. Với *(w,h)* biểu diễn chiều rộng và chiều cao của bounding box, điểm *score=w·h*. Hai tập

điểm có điểm số cao nhất sẽ được giữ lại và sử dụng trong bước trích xuất đặc trưng.

1) Trích xuất đặc trưng: Cho H_1,H_2 lần lượt là tập điểm trên bộ khung xương của người thứ nhất và người thứ hai (Hình 4). O_1O_2 là điểm chính giữa xương cột sống, α_1,α_2 là góc nghiêng của xương cột sống lần lượt của người thứ nhất và người thứ hai, l_1 là độ dài xương cột sống của người thứ nhất. Đặc trưng khoảng cách được tính toán là khoảng cách giữa hai gốc tọa độ O_1O_2 và được chuẩn hóa bằng cách chia cho l_1:

$$d = \frac{distance}{l_1} \quad (6)$$



Hình 4. Đặc trưng khoảng cách giữa 2 bộ khung xương

Việc chuẩn hóa tọa độ của mỗi tập điểm được thực hiện độc lập. Đặt $p_i = (p_{(i,x)},p_{(i,y)})$ là tọa độ của 1 điểm trên bộ khung xương H_i (i=1,2). Công thức chuẩn hóa của điểm này được tính toán theo các công thức sau:

$$p'_i = (p'_{i,x},p'_{i,y}) = (p_{i,x} - o_{i,x}; p_{i,y} - o_{i,y}), \quad i = 1,2 \quad (7)$$

$$\begin{pmatrix} p''_{i,x} \\ p''_{i,y} \end{pmatrix} = \begin{bmatrix} \cos(\alpha_i) & -\sin(\alpha_i) \\ \sin(\alpha_i) & \cos(\alpha_i) \end{bmatrix} \begin{pmatrix} p'_{i,x} \\ p'_{i,y} \end{pmatrix}, \quad i = 1,2 \quad (8)$$

$$(p'''_{i,x}, p'''_{i,y}) = \left(\frac{p''_{i,x}}{w_i}, \frac{p''_{i,y}}{h_i}\right), \quad i = 1,2 \quad (9)$$

2) Vectơ đặc trưng: Kết quả nhận dạng điểm chính của OpenPose trên bộ dữ liệu *humiact5* và của thiết bị chuyên dụng Kinect v2 trên bộ dữ liệu *NTU RGB+D* cung cấp 25 điểm chính trên bộ khung xương. Bằng việc phân tích một tập con dữ liệu khung xương, nhóm tác giả nhận thấy rằng dữ liệu các điểm được đánh số từ 15 tới 24, tương ứng với các phần nhỏ trên cơ thể chẳng hạn như ngón tay, thường xuyên bị thiếu dữ liệu. Vì vậy, nhóm tác giả chỉ sử dụng những điểm trên khung xương có chỉ số từ 0 đến 14. Kết quả, vectơ đặc trưng thu được từ dữ liệu khung xương của một frame có 61 chiều; trong đó, 15 điểm x 2 *giá trị tọa độ* x 2 *tập điểm* và 1 đặc trưng khoảng cách. Giả sử rằng chúng ta đã chọn được *m* frame để phân tích, kết quả sẽ có một ma trận vectơ đặc trưng kích thước *m x 61*.

Trên cơ sở vectơ đặc trưng được trích xuất, các bộ phân loại chuyên gia, bộ phân loại kết hợp sẽ được huấn luyện và đánh giá. Các nội dung này sẽ được trình bày trong phần II.

KẾT LUẬN

Nhóm tác giả đã trình bày khái quát các phương pháp chính nhận dạng hành động trong video và chuỗi ảnh rời rạc. Kiến trúc giải pháp đề xuất trong nghiên cứu được đưa ra, cùng với đó việc ước lượng các điểm chính trên khung xương con người sử dụng thư viện OpenPose, thuật toán gộp các tập điểm cũng như kỹ thuật trích xuất đặc trưng được trình bày chi tiết. Trên cơ sở các nội dung nghiên cứu trong Phần I, nội dung tiếp theo sẽ được trình bày tại phần II về phương pháp cũng như các bước tiến hành huấn luyện và đánh giá các bộ phân loại chuyên gia và đưa ra các hướng nghiên cứu tiếp để có thể ứng dụng sâu rộng hơn nữa AI vào cuộc sống.❖

TÀI LIỆU THAM KHẢO

[1]. *[Online], "Deep Learning Techniques to Detect Botnet," https://isj.vn/index.php/journal_STIS/article/view/846, (accessed on 13.06.2023).*

[2]. *[Online], "Ứng dụng mô hình học sâu trong phát hiện tấn công trinh sát mạng," https://isj.vn/index.php/journal_STIS/article/view/922, (accessed on 13.06.2023).*

[3]. *E. Cippitelli, E. Gambi, S. Spinsante, and F. Florez-Revuelta, "Evaluation of a skeleton-based method for human activity recognition on a large-scale RGB-D dataset," in 2nd IET International Conference on Technologies for Active and Assisted Living (TechAAL 2016), London, UK, 24-25 October 2016, pp.1-6, doi: 10.1049/ic.2016.0063.*

[4]. *Z. Cao, G. Hidalgo, T. Simon, S.-E. Wei, and Y. Sheikh, "OpenPose: Realtime Multi-Person 2D Pose Estimation Using Part Affinity Field," IEEE Trans. Pattern Anal. Mach. Intell., vol.43, no.1, pp.172-186, Jan. 2021, doi: 10.1109/TPAMI.2019.2929257.*

[5]. *[Online], "OpenPose", CMU-Perceptual-Computing-Lab, 2021 https:// github.com/CMU-Perceptual-Computing-Lab/openpose/, (accessed on 15.07.2022).*