KHẢO SÁT MỘT SỐ MÔ HÌNH RÚT TRÍCH DỮ LIỆU KHUNG XƯƠNG TRONG BÀI TOÁN NHẬN DIỆN HÀNH ĐỘNG NGƯỜI

Phù Khắc Anh[1], Hoàng Văn Dũng[2*], Lê Văn Tường Lân[3]

[1]Khoa Công nghệ thông tin, trường Đại học Khoa học, Đại học Huế, Huế

[2]Khoa Công nghệ thông tin, trường Đại học Sư phạm kỹ thuật, Tp Hồ Chí Minh

[3]Trường Đại học Khoa học, Đại học Huế, Huế

[*]Email: dunghv@hcmute.edu.vn

Ngày nhận bài: 23/01/2024; ngày hoàn thành phản biện: 26/02/2024; ngày duyệt đăng: 5/3/2024

TÓM TẮT

Trong lĩnh vực nghiên cứu về thị giác máy tính, nhận diện hành động người từ dữ liệu hình ảnh, video là một thách thức được nhiều nhà nghiên cứu quan tâm trong thời gian gần đây. Trong đó, giải pháp sử dụng dữ liệu khung xương làm dữ liệu đầu vào để huấn luyện cho các mô hình học máy là một giải pháp được đánh giá có nhiều tiềm năng. Trong nghiên cứu này, chúng tôi khảo sát tổng quan về bài toán nhận dạng hành động người và giới thiệu giải pháp sử dụng dữ liệu khung xương cho bài toán này. Bên cạnh đó, chúng tôi cũng khảo sát chi tiết hơn các hướng tiếp cận để xây dựng mô hình rút trích dữ liệu khung xương cùng với những mô hình tiêu biểu trong từng hướng tiếp cận. Điều này mang lại cái nhìn toàn diện hơn về bài toán nhận diện hành động người và làm rõ được tầm quan trọng của việc chọn lựa giải pháp rút trích dữ liệu khung xương phù hợp trong từng trường hợp cụ thể.

Từ khóa: Dữ liệu khung xương, học sâu, nhận dạng hành động, thị giác máy tính.

1. GIỚI THIỆU

Trong cuộc sống hàng ngày, con người không chỉ tương tác với môi trường xung quanh mà còn kết nối với nhau thông qua việc thực hiện nhiều hành động đa dạng khác nhau. Nếu máy tính hoặc robot có khả năng nhìn nhận và hiểu biết như con người về những hành động này, đó sẽ là một cơ hội để phát triển một loạt những hướng ứng dụng tiềm năng đáng kể. Trong bài toán nhận diện hành động người, hành động của con người được phân chia thành 4 nhóm chính dựa trên bộ phận cơ thể tham gia vào hành động: Gesture, Action, Interactions và Group activities [1]. Quá trình nhận diện hành động người (Human Action Recognition - HAR) là quá trình sử dụng hình ảnh và video thu thập từ nhiều nguồn để xác định và phân loại các hành động mà con người đã thực

hiện. HAR có ứng dụng rộng rãi trong nhiều lĩnh vực như giao tiếp người-máy, giám sát an ninh, chăm sóc người cao tuổi từ xa, các hệ thống nhà thông minh, v.v… [2,3].

Hiện nay, các thuật toán tiên tiến được đề xuất để giải quyết bài toán nhận diện hành động người vẫn gặp phải những khó khăn như: sự đa dạng về ngữ nghĩa của hành động, ảnh hưởng của bối cảnh và góc quay, dữ liệu đầu vào không đầy đủ, sự lộn xộn của cảnh nền phía sau, hạn chế về từ vựng hành động, v.v… Dữ liệu về khung xương với tính chất đơn giản, gọn nhẹ, được biểu diễn dựa trên các khớp quan trọng của cơ thể con người, có thể giúp tránh được các vấn đề về ánh sáng cũng như sự thay đổi của cảnh nền phía sau. Hành động của con người có thể được nhận diện thông qua việc phân tích quỹ đạo di chuyển chung của các khớp nối theo thời gian, qua đó nhóm tác giả trong nghiên cứu [4] cũng đã xây dựng lý thuyết về dữ liệu khung xương với số điểm khớp xương được đề xuất ban đầu là 12 điểm.

Trong nghiên cứu này, chúng tôi thực hiện phân tích và đánh giá kỹ lưỡng về các phương pháp và mô hình liên quan đến việc rút trích dữ liệu khung xương cho bài toán nhận dạng hành động của con người. Nội dung bao gồm:

- Các phương pháp tiếp cận trong bài toán nhận dạng hành động của con người.
- Giới thiệu các mô hình rút trích dữ liệu khung xương được áp dụng cho bài toán nhận dạng hành động con người, bằng cách sử dụng hai phương pháp chính là Top-down và Bottom-up. Đồng thời, chúng tôi cũng trình bày về các thang đo và bộ dữ liệu tiêu biểu phổ biến được dùng trong các thực nghiệm liên quan.

2. NHẬN DẠNG HÀNH ĐỘNG NGƯỜI VỚI DỮ LIỆU KHUNG XƯƠNG

Trong bài toán nhận diện hành động, nhóm tác giả phân loại 3 hướng tiếp cận khác nhau, bao gồm: sử dụng các thuật toán thị giác máy tính; sử dụng dữ liệu khung xương được ghi nhận từ các thiết bị cảm biến; sử dụng dữ liệu khung xương được trích xuất từ hình ảnh, video, v.v… Hướng tiếp cận thứ nhất trong bài toán nhận diện hành động người tập trung vào việc sử dụng các thuật toán thị giác máy tính, chuyển bài toán nhận diện hành động người về dạng bài toán phân lớp từ một ảnh tĩnh. Cách tiếp cận này có nhược điểm rõ rệt khi xử lý dữ liệu video, nơi sự biến đổi của hành động con người được đánh giá mật thiết qua thời gian.

Một số nghiên cứu theo hướng tiếp cận này đã được đề xuất nhằm giải quyết vấn đề trên như: nghiên cứu của Poppe năm 2010 [5], các phương pháp để mô hình hóa các lớp hành động và xử lý chuyển động phức tạp của Turaga [6], phương pháp Khối lượng - Không gian - Thời gian của Zhu [7], các phương pháp sử dụng Deep Learning của Herath [8], phân loại bằng cách kết hợp mô tả BoW và mô hình phân loại riêng biệt [9], gia tăng tốc độ xử lý trong nhận diện đối tượng với mô hình cấu trúc phân tầng nhị

phân [10], cải thiện khả năng dự đoán hành động thông qua việc kết hợp hai lớp phân loại của tác giả Van – Dung Hoang [11], trích xuất đặc trưng có thể mở rộng (SPE) với cấu trúc dữ liệu để lưu trữ thông tin về luồng quang học và dải màu của đối tượng [12], bộ mô tả MOMP [13], mô hình ghi nhãn giả chéo CMPL [14], v.v…. Ngoài ra, phương pháp Vision Transformers [15] đang thu hút sự chú ý trong giải quyết vấn đề thị giác máy tính. Một nghiên cứu năm 2016 [16] giới thiệu mô hình Future Transformers có thể nhận dạng và dự đoán hành động từ dữ liệu video có thời gian dài.

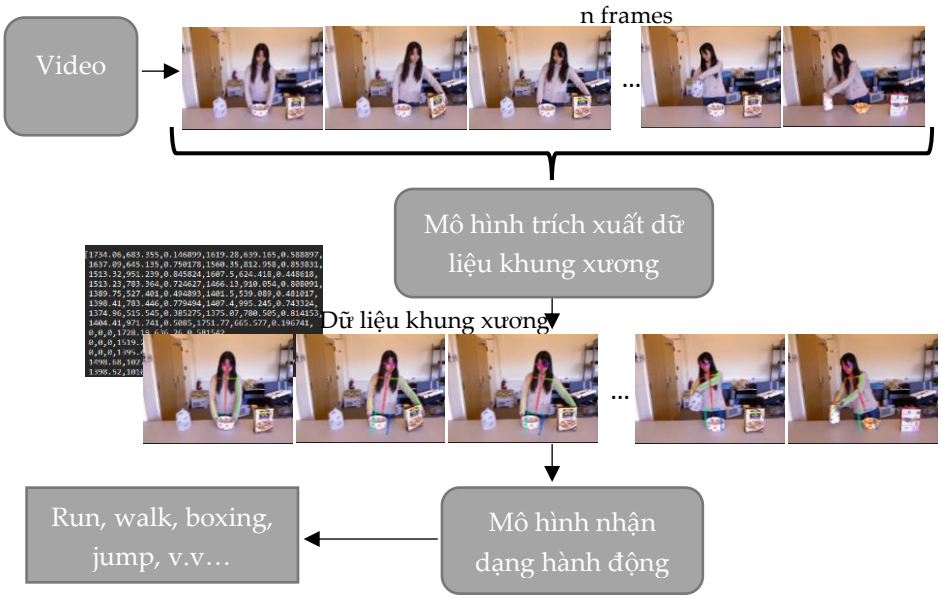
Trong hướng tiếp cận thứ hai, những nghiên cứu về nhận dạng hành động người với dữ liệu khung xương được ghi nhận trực tiếp từ các thiết bị cảm biến thường được thực hiện trên những bộ dữ liệu có đầy đủ video, hình ảnh cùng với dữ liệu khung xương - ground truth thực tế tương ứng. ***Bảng 1*** mô tả tổng quan về một số bộ dữ liệu phổ biến cho bài toán nhận diện hành động người

Bảng 1. Thông tin tổng quan một số bộ dữ liệu về nhận diện hành động người.

TT	Tên bộ dữ liệu	Số lượng hình ảnh/video	Số lớp hành động	Định dạng dữ liệu
1	MPII (17)	24984	410	S
2	UTD - MAH (18)	861	27	RGB, D, S
3	MSR – Action3D (19)	557	20	D, S
4	UT – Kinect (20)	200	10	RGB, D, S
5	Berkeley MHAD (21)	660	11	RGB, D, S, Acc
6	MSRAction Pair (22)	360	12	RGB, D, S
7	NTU + RGB + D (23)	56880	60	RGB, D, S
8	CMDFALL (24)	1963	20	RGB, D, S, Acc
9	NTU + RGB + D 120 (25)	114480	120	RGB, D, S
10	UAV-Human (26)	67428	155	RGB, D, S

** S: Skeleton data, RGB: Red green blue, D: Depth, Acc: Acceleration*

Trong một số tình huống đặc thù, việc yêu cầu đối tượng cần được nhận diện phân loại hành động phải đeo các thiết bị cảm biến trên người để ghi nhận dữ liệu khung xương là bất khả thi, ví dụ trong các bài toán về camera giám sát an ninh, truy vết tội phạm, tái nhận diện người với đặc trưng là tư thế người, v.v… thì cách tiếp cận thứ 3 lại được nhiều nhà nghiên cứu quan tâm, cụ thể được trình bày trong ***Hình 1.***



Hình 1. Nhận diện hành động người với dữ liệu khung xương trích xuất từ hình ảnh/video.

Dữ liệu khung xương trích xuất được từ hình ảnh hoặc video sẽ được sử dụng làm dữ liệu đầu vào để huấn luyện bởi các mô hình khác nhau, kết quả đầu ra sẽ là tên gọi hành động của các đối tượng trong các video hoặc hình ảnh ghi nhận được. Các mô hình, giải pháp được xây dựng theo hướng tiếp cận này sẽ thường được thực hiện trên các tập dữ liệu ở Bảng 1 nhằm để so sánh độ chính xác giữa dữ liệu khung xương rút trích được từ giải pháp của tác giả và dữ liệu khung xương thực tế ghi nhận từ thiết bị cảm biến. Nếu như ở cách tiếp cận thứ 2, hiệu quả của giải pháp được đánh giá đơn giản thông qua độ chính xác của mô hình nhận dạng hành động thì với cách tiếp cận thứ 3, hiệu quả của giải pháp đề xuất ra được tính theo công thức sau:

$$D = d_1 * d_2 * d_3 \quad (1)$$

Trong đó D là độ chính xác của giải pháp, d_1 là tỷ lệ trích xuất được dữ liệu khung xương từ dữ liệu hình ảnh, video; d_2 là độ chính xác của dữ liệu khung xương trích xuất được khi so với dữ liệu khung xương thực tế - ground truth; d_3 là độ chính xác khi sử dụng dữ liệu khung xương rút trích được để huấn luyện trong một mô hình nhận dạng hành động cụ thể. Để đánh giá độ chính xác d_2, nhiều thang đo đã được các nhà nghiên cứu đề xuất như PCK, OKS, PDJ, mAP, v.v…

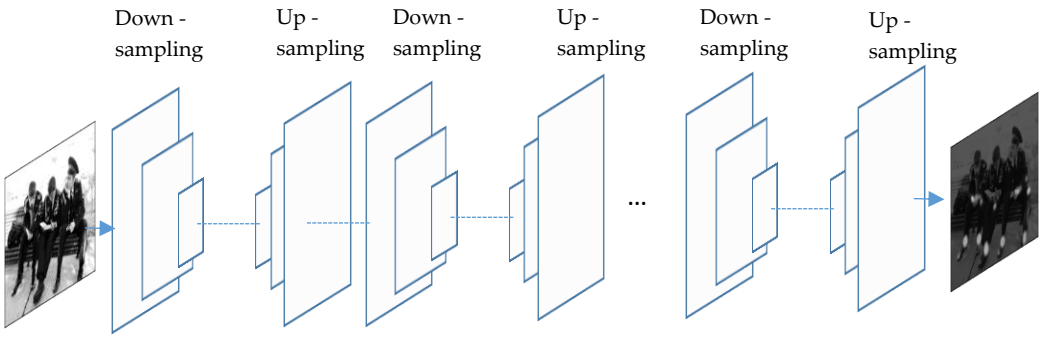
3. CÁC MÔ HÌNH TIÊU BIỂU TRÍCH XUẤT DỮ LIỆU KHUNG XƯƠNG

3.1 Các phương pháp rút trích dữ liệu khung xương

Trong lĩnh vực ước lượng tư thế người và rút trích dữ liệu khung xương, có hai phương pháp tiếp cận chính được sử dụng: Top-down và Bottom-up. Phương pháp Top-down phát hiện toàn bộ cơ thể người của từng đối tượng, sau đó phân chia thành các phần nhỏ hơn như đầu, vai, cổ, cổ tay và từ đó nhận diện khung xương của từng đối tượng. Ngược lại, phương pháp Bottom-up tập trung vào việc nhận dạng từng điểm chính (keypoint) trên cơ thể của tất cả các đối tượng mà không quan tâm đến cấu trúc cơ thể tương ứng. Các điểm chính này sau đó được kết hợp lại với nhau để tạo ra dữ liệu khung xương tổng thể. Cả hai phương pháp này đều đóng góp vào sự phát triển của lĩnh vực ước lượng tư thế và đều có ưu điểm và hạn chế riêng. Sự lựa chọn giữa Top-down và Bottom-up thường phụ thuộc vào yêu cầu cụ thể của ứng dụng và đặc tính của dữ liệu đang được xử lý. Trong khuôn khổ nghiên cứu này, với mỗi hướng tiếp cận, nhóm tác giả giới thiệu ba mô hình tiêu biểu đại diện cho hướng tiếp cận đó.

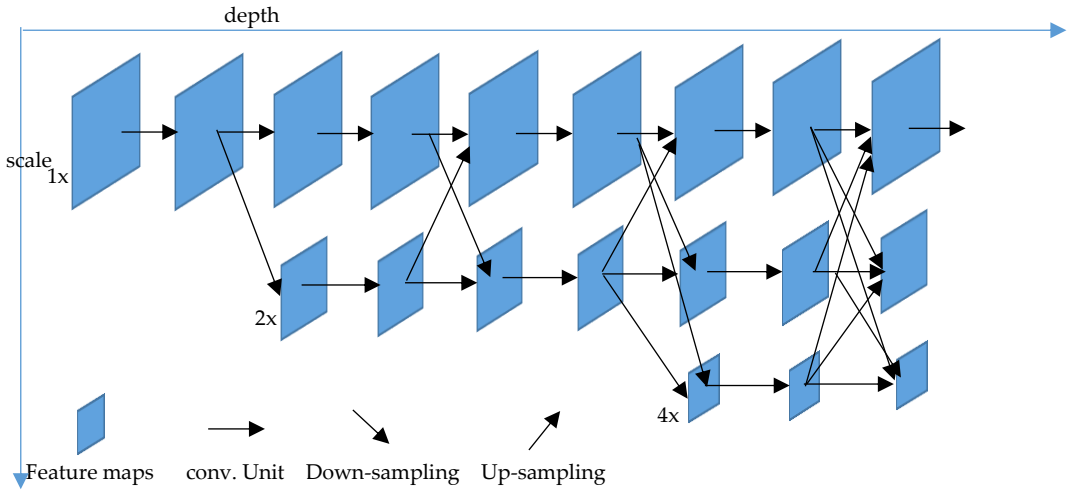
3.2 Phương pháp Top-down: mô hình Hourglass, mô hình HRNet và mô hình VITPose

Mô hình Hourglass là một kiến trúc mạng neural network được thiết kế đặc biệt cho ước lượng tư thế người [27]. Kiến trúc này được gọi là "Hourglass" do nó bao gồm nhiều mô-đun, mỗi mô-đun có cấu trúc tương tự như đồng hồ cát với các lớp upsampling và downsampling. Mỗi mô-đun trong Hourglass thực hiện nhiệm vụ chính là đầu vào upsampling và downsampling. Lớp upsampling giúp tăng kích thước đầu vào, trong khi lớp downsampling giúp giảm kích thước đầu vào. Để nâng cao quá trình học, mỗi mô-đun so sánh dự đoán của heatmap với vị trí thực tế.



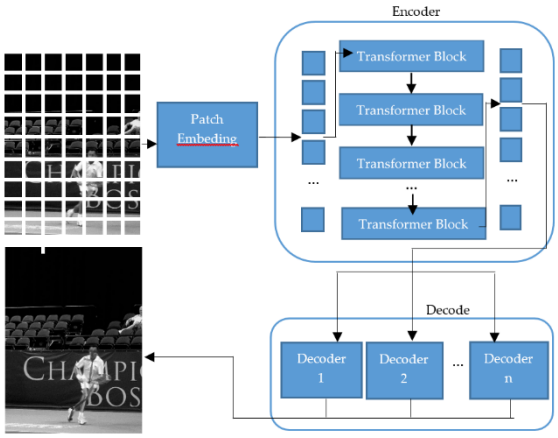
Hình 2. Mô hình Hourglass

Mô hình HRNet (High-Resolution Network) là một kiến trúc mạng neural network được thiết kế để giải quyết vấn đề giảm độ phân giải trong quá trình xử lý thông tin [28]. Khác với Hourglass, HRNet không giảm độ phân giải từ cao xuống thấp và sau đó tăng trở lại, mà thay vào đó, mô hình này cố gắng duy trì độ phân giải cao suốt quá trình huấn luyện, đồng thời mở rộng thêm thông tin của dữ liệu đầu vào với các độ phân giải thấp hơn. **Hình 3** trình bày cụ thể kiến trúc của mô hình HRNet.



Hình 3. Mô hình HRNet

Với sự xuất hiện của kiến trúc Vision Transformers và sự phổ biến ngày càng tăng của kiến trúc này trong lĩnh vực thị giác máy tính, không lâu sau đó các nhà nghiên cứu đề xuất mô hình "Transformer cho Estimation Pose" hay còn được gọi là VITPose [29]. Cấu trúc của mô hình này bao gồm một tập hợp các TransformerBlocks, (mỗi khối là sự kết hợp của Layer Normalization, Multi Headed Self Attention, và Feed Forward Network), và một mô-đun giải mã. Sau khi trích xuất đặc trưng trong bộ mã hóa, một bộ giải mã khá đơn giản được sử dụng. Bộ giải mã này bao gồm 2 lớp: Deconvolution Layer, tiếp theo là Batch Normalization và Relu, và một lớp dự đoán tuyến tính. ***Hình 4*** mô tả chi tiết về cấu trúc của mô hình VITPose.

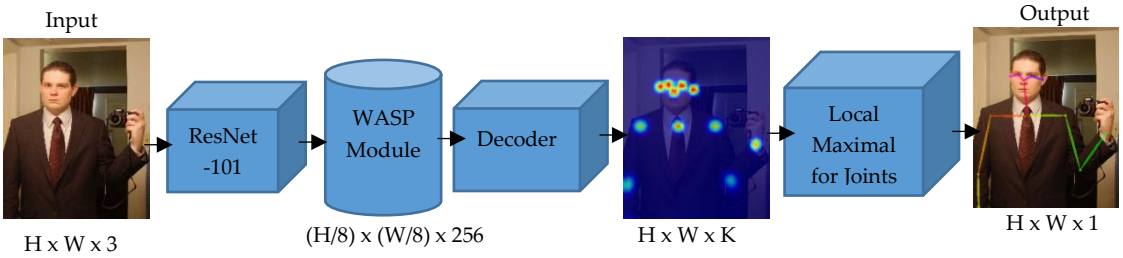


Hình 4. Mô hình VITPose

3.3 Phương pháp Bottom-up: Mô hình UniPose, mô hình Openpose và mô hình Omni-Pose

UniPose là một mô hình đặc biệt trong lĩnh vực bài toán trích xuất dữ liệu khung xương. Mô hình này đề xuất sử dụng một kiến trúc mới gọi là "Waterfall Atrous Spatial

Pooling" [30]. UniPose không dựa nhiều vào việc xử lý số liệu thống kê hoặc các tư thế khớp được xác định trước đó để biết vị trí của các phần khớp mới. Thay vào đó, UniPose kết hợp cả việc phân đoạn ngữ cảnh và xác định vị trí khớp trong một lượt duy nhất. Kiến trúc UniPose được mô tả cụ thể ở ***Hình 5.***



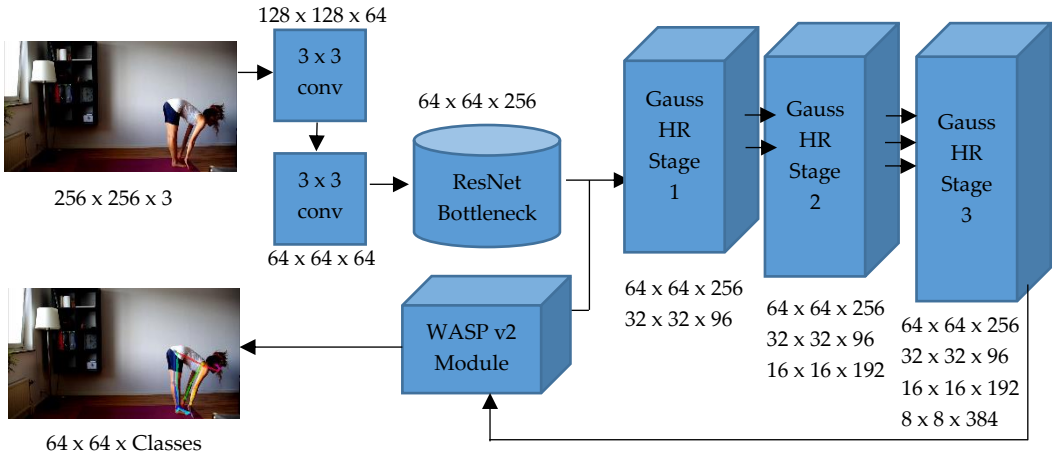
Hình 5. Mô hình UniPose

Mô hình OpenPose đến là một trong những mô hình theo phương pháp Bottom-up phổ biến nhất [31]. Mô hình này hoạt động dựa trên hai nhánh chính của các lớp tích chập. Nhánh đầu tiên tạo ra 18 bản đồ tin cậy (PCMs – Part Confidence Maps), mỗi bản đồ đại diện cho một phần cụ thể của khung xương đối tượng. Nhánh thứ hai tạo ra 38 bản đồ liên kết giữa các bộ phận cơ thể (PAFs – Part Affinity Fields), biểu thị mức độ liên kết giữa các phần khác nhau của cơ thể. ***Hình 6*** thể hiện các kết quả trích xuất dữ liệu khung xương từ mô hình Openpose.

a) b) c) d)

Hình 6. Kết quả trích xuất dữ liệu khung xương từ mô hình Openpose với a) Ảnh đầu vào b) Part Confidence Maps, c) Part Affinity Fields, d) Skeleton data

Omnipose hiện nay được xem là một trong những mô hình hiệu suất cao nhất trong phương pháp bottom-up, với cấu trúc đơn giản và khả năng dễ dàng mở rộng. Từ dữ liệu đầu vào, mô hình sử dụng hai lớp tích chập 3x3, tiếp theo là một khối ResnetBottleneck. Sau đó, mô hình tích hợp ba khối HRNet, mỗi khối kèm theo cải tiến modul hóa heatmap Gaussian, như được đề xuất trong bài báo [32]. Một điểm đặc biệt nổi bật khác của mô hình Omnipose là mô-đun Waterfall Atrous Spatial Pyramid (WASPv2) đã được cải tiến từ môđun WASP của mô hình UniPose. Cấu trúc của mô hình Omnipose được thể hiện qua ***Hình 7***.



Hình 7. Mô hình OmniPose

4. TỔNG KẾT

Trong nghiên cứu này, chúng tôi đã trình bày những khái niệm cơ bản về dữ liệu khung xương, cách vận dụng vào các mô hình nhận diện hành động, cũng như đã giới thiệu một loạt các mô hình trích xuất dữ liệu khung xương đại diện cho cả hai phương pháp tiếp cận chính: Top-down và Bottom-up. Các mô hình mà chúng tôi đã trình bày đều có ưu điểm và hạn chế riêng, và sự lựa chọn giữa chúng có thể phụ thuộc vào mục tiêu và yêu cầu cụ thể của ứng dụng. Nếu mục tiêu là ước lượng tư thế người trong các tình huống tự nhiên và đòi hỏi độ chính xác cao, phương pháp Top-down có thể là sự lựa chọn tốt. Ngược lại, nếu ứng dụng đặt yêu cầu cao về tính linh hoạt và khả năng xử lý sự che khuất, phương pháp Bottom-up có thể là sự lựa chọn phù hợp.

Bên cạnh đó, chúng tôi cũng đề xuất những hướng phát triển cụ thể như sau: tạo ra một mô hình hợp nhất giữa hai phương pháp Top-down và Bottom-up nhằm đem lại khả năng ước lượng tư thế người chính xác cao trong nhiều tình huống phức tạp; áp dụng công nghệ Edge Computing để phát triển mô hình nhẹ có thể chạy trên các thiết bị đầu cuối và nâng cao khả năng ứng dụng trong thời gian thực; khai thác và tích hợp dữ liệu từ nhiều nguồn và định dạng khác nhau để mô hình có thể hiểu và xử lý tốt hơn các tư thế hành động đa dạng; nâng cao tính bảo mật và riêng tư khi xử lý thông tin nhạy cảm liên quan đến hình ảnh của con người, v.v...

TÀI LIỆU THAM KHẢO

[1]. Aggarwal JK, Ryoo MS (2021), Human activity analysis: a survey. *ACM Comput Surv (CSUR)*, vol. 43, No. 3, pp. 1-43, Association for Computing Machinery.

[2]. M. H. Siddiqi et al. (2021). A Unified Approach for Patient Activity Recognition in Healthcare Using Depth Camera, in *IEEE Access*, vol. 9, pp. 92300-92317, doi: 10.1109/ACCESS.2021.3092403.

[3]. Kim, K., Jalal, A. & Mahmood, M (2019). Vision-Based Human Activity Recognition System Using Depth Silhouettes: A Smart Home System for Monitoring the Residents. *J. Electr. Eng. Technol*, vol. 14, pp. 2567–2573.

[4]. Johasson G. (1973). Visual perception of biological motion and a model for its analysics, *Perception and psychophysics*, vol. 14, No. 2, pp. 201-211.

[5]. Poppe R (2010), A survey on vision-based human action recognition. *Image and Vision Computing*, vol. 2, No. 6, pp. 976–990.

[6]. Turaga P, Chellappa R, Subrahmanian VS, Udrea O (2008). Machine recognition of human activities: a survey, *IEEE Trans Circuits Syst Video Technol*, vol. 18, no. 11, pp. 1473 – 1488.

[7]. Zhu F, Shao L, Xie J, Fang Y (2016). From handcrafted to learned representations for human action recognition: a survey, *Image and Vision Computing*, vol. 55, pp. 42–52, ScienceDirect.

[8]. Herath S, Harandi M, Porikli F (2017). Going deeper into action recognition: a survey, *Image and Vision Computing*, vol. 60, pp. 4–21.

[9]. Van-Huy Pham, My-Ha Le, and Van-Dung Hoang (2017). Boosting Discriminative Models for Activity Detection Using Local Feature Descriptors, *ACIID 2017*, LNAI vol. 10191, pp. 609 – 618, Springer, 2017, Doi: 10.1007/978-3-319-54472-4_57.

[10]. Van-Dung Hoang and Kang-Hyun Jo (2016). Accelerative Object Classification Using Cascade Structure for Vision Based Security Monitoring Systems, *ACIIDS 2016*, Part I, LNAI vol. 9621, pp. 790–800, Doi: 10.1007/978-3-662-49381-6_76.

[11]. Van-Dung Hoang (2017), Multiple classifier-based spatiotemporal features for living activity prediction, *Journal of Information and Telecommunication*, vol. 1, pp. 100-112, Taylor & Francis Online, 2017, Doi: 10.1080/24751839.2017.1295668.

[12]. Van-Huy Pham, Kang-Hyun Job and Van-Dung Hoang (2019). Scalable local features and hybrid classifiers for improving action recognition, *Journal of Intelligent & Fuzzy Systems*, vol. 36, no. 4, pp. 3357–3372, IOS Press, Doi: 10.3233/JIFS-181085.

[13]. Phan H.H, Vu N.S., Nguyen V.L., and Qouy M (2018). Action recognition based on motion of oriented magnitude patterns and feature selection. *IET Computer Vision*, vol. 12, pp. 735 – 743, IET, Doi: 10.1049/iet-cvi.2017.0282.

[14]. Xu, Y., Wei, F., Sun, X., Yang, C., Shen, Y., Dai, B., ... & Lin, S. (2022). Cross-model pseudo-labeling for semi-supervised action recognition. *In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 2959-2968).

[15]. Anwaar Ulhaq, Naveed Akhtar, Ganna Pogrebna, Ajmal Mian (2022). Vision Transformers for Action Recognition: A Survey, *Computer Vision and Pattern Recognition*, arXiv 2022. arXiv preprint arXiv:2209.05700.

[16]. Gong, D., Lee, J., Kim, M., Ha, S. J., & Cho, M. (2022). Future transformer for long-term action anticipation. *In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 3052-3061).

[17]. Mykhaylo Andriluka, Leonid Pishchulin and Peter Gehler and Schiele, Bernt (2014). 2D Human Pose Estimation: New Benchmark and State of the Art Analysis, *IEEE Conference on Computer Vision and Pattern Recognition (CVPR).*

[18]. C. Chen, R. Jafari and N. Kehtarnavaz (2015). UTD-MHAD: A multimodal dataset for human action recognition utilizing a depth camera and a wearable inertial sensor, *2015 IEEE International Conference on Image Processing (ICIP)*, pp. 168-172, doi: 10.1109/ICIP.2015.7350781.

[19]. I W., Zhang Z., and Liu Z. (2010). Action recognition based on a bag of 3D points. *In 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition-Workshops*, pp. 9-14. IEEE.

[20]. Xia L., Chen C.C., and Aggarwal J.K. (2012). View invariant human action recognition using histograms of 3D joints. *In 2012 IEEE Computer Society Conference on Computer Vision and Pattern Recognition-Workshops*, pp. 20-27.

[21]. Ofli F., Chaudhry R., Kurillo G. Vidal R., and Bajcsy R. (2013). Berkeley mhad: A comprehensive multimodal human action database. *In 2013 IEEE Workshop on Applications of Computer Vision (WACV)*, pp. 53-60.

[22]. Oreifej O. and Liu Z. (2013). *HON4D:* Histogram of oriented 4D normal for activity regcognition from depth sequences. *In Proceedings of IEEE conference on computer vision and patern recognition*, pp. 716 – 723.

[23] Shahroudy A., Liu J., Ng T.T, and Wang G. (2016). NTU RGB + D: A large scale dataset for 3D human activit analysis. *In Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1010 – 1019.

[24] Tran T.H, Le T.L, Pham D.T, Hoang V.N, Khong V.M, Tran Q.T, Nguyen T.S, and Pham C. (2018). A multi-modal multi-view dataset for human fall analysis and preliminary investigation on modality. *In 2018 24th International Conference on Pattern Recognition (ICPR)*, pp. 1947 – 1952.

[25]. Liu J., Shahroudy A., Perez M., Wang G., Duan L.Y., and Kot A.C. (2019). NTU RGB + D 120: A large-scale benchmark for 3D human activity understanding. *IEEE transactins on pattern analysis and machine intelligence*, 42(10): pp. 2684-2701.

[26]. Li T., Liu J., Zhang W., Ni Y., Wang W., and Li Z. (2021). UAV-Human: A large benchmark for human behavior understanding with unmanned aerial vehicles. *In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*

[27]. Newell, A., Yang, K., Deng, J. (2016). Stacked Hourglass Networks for Human Pose Estimation. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds) *Computer Vision – ECCV 2016*. Lecture Notes in Computer Science (), vol 9912. https://doi.org/10.1007/978-3-319-46484-8_29.

[28]. ke, Sun & Xiao, Bin & Liu, Dong & Wang, Jingdong (2019). *Deep High-Resolution Representation Learning for Human Pose Estimation*. arXiv:1902.09212.

[29]. Yufei Xu, Jing Zhang, Qiming Zhang, Dacheng Tao. *ViTPose: Simple Vision Transformer Baselines for Human Pose Estimation*, https://doi.org/10.48550/arXiv.2204.12484.

[30]. B. Artacho and A. Savakis (2020)/ UniPose: Unified Human Pose Estimation in Single Images and Videos, *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7033-7042, doi: 10.1109/CVPR42600.2020.00706.

[31]. Z. Cao, G. Hidalgo, T. Simon, S. -E. Wei and Y. Sheikh (2021). OpenPose: Realtime Multi-Person 2D Pose Estimation Using Part Affinity Fields, *in IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 43, no. 1, pp. 172-186, 1 Jan. 2021, doi: 10.1109/TPAMI.2019.2929257.

[32]. F. Zhang, X. Zhu, H. Dai, M. Ye and C. Zhu (2020). Distribution-Aware Coordinate Representation for Human Pose Estimation, *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7091-7100, doi: 10.1109/CVPR42600.2020.00712.

SURVEYING SOME SKELETON DATA EXTRACTION MODELS IN HUMAN ACTION RECOGNITION PROBLEMS

Khac-Anh Phu[1], Van-Dung Hoang [2*], Van-Tuong-Lan Le[3]

[1] Faculty of Information Technology, University of Sciences, Hue University, Hue

[2] Faculty of Information Technology, HCMC University of Technology and Education

[3] University of Sciences, Hue University, Hue

[*]Email: dunghv@hcmute.edu.vn

ABSTRACT

In the realm of computer vision research, the detection of human actions from image and video data has emerged as a prominent challenge, capturing the interest of numerous researchers in recent time. One particularly promising solution involves the utilization of skeleton data as the primary input for machine learning models. In this research, we conduct a comprehensive survey on the issue of human action recognition and introduce approaches leveraging skeleton data to address this challenge. Furthermore, we conduct an in-depth examination of various approaches to construct models for skeleton data extraction, highlighting exemplary models representative of each distinct approach. This detailed analysis not only contributes to a more nuanced understanding of the human action recognition problem but also emphasizes the critical importance of carefully selecting an appropriate skeleton data extraction solution tailored to specific use cases. Our study also provides valuable insights for researchers and practitioners navigating the complexities of human action recognition, shedding light on the diverse landscape of skeleton data utilization in this domain.

Keywords: Skeletal data, deep learning, action recognition, computer vision.

Phù Khắc Anh sinh ngày 01/03/1990 tại Tân An, Long An. Ông tốt nghiệp cử nhân chuyên ngành Công nghệ thông tin tại trường Đại học Khoa Học Tự Nhiên, ĐH Quốc Gia Thành phố Hồ Chí Minh năm 2012, tốt nghiệp thạc sĩ chuyên ngành Khoa học máy tính năm 2016 tại trường Đại học Công Nghệ Thông Tin, ĐH Quốc Gia Thành phố Hồ Chí Minh năm 2016. Ông công tác tại trường Cao đẳng Kỹ thuật Cao Thắng Thành phố Hồ Chí Minh. Năm 2023 ông trở thành nghiên cứu sinh ngành Khoa học máy tính trường Đại học Khoa học, Đại học Huế.

Lĩnh vực nghiên cứu: Trí tuệ nhân tạo, thị giác máy tính

Hoàng Văn Dũng nhận bằng tiến sĩ ngành Kỹ thuật điện tử và Hệ thống thông tin tại Trường Đại học Ulsan, Hàn Quốc, năm 2015. Ông công tác tại Khoa Công nghệ thông tin, Trường ĐH Sư phạm Kỹ thuật TP. HCM. Ông là thành viên tích cực của các hiệp hội khoa học kỹ thuật như IEEE, IEEE Computer, ICROS. Ông nghiên cứu và xuất bản các công trình liên quan đến Trí tuệ nhân tạo, thị giác máy tính, xử lý ảnh y tế, xe tự hành, các hệ thống hiểu biết thông minh và giám sát thông minh.

Lĩnh vực nghiên cứu: Trí tuệ nhân tạo, Thị giác máy tính.

Lê Văn Tường Lân sinh ngày 10/11/1974 tại Thừa Thiên Huế. Năm 1996, ông tốt nghiệp Đại học ngành Toán - Tin tại Trường Đại học Khoa học, Đại học Huế. Ông nhận bằng thạc sỹ Công nghệ thông tin tại Trường Đại học Bách Khoa Hà Nội năm 2002 và nhận học vị Tiến sĩ ngành Khoa học máy tính tại Trường Đại học Khoa học, Đại học Huế năm 2018. Hiện công tác tại Ban Đào tạo và Công tác sinh viên, Đại học Huế.

Lĩnh vực nghiên cứu: Cơ sở dữ liệu, Công nghệ phần mềm, Khai phá dữ liệu.