2024 IEEE 12th International Conference on Serious Games and Applications for Health (SeGAH) | 979-8-3503-8438-3/24/$31.00 ©2024 IEEE | DOI: 10.1109/SeGAH61285.2024.10639556

Exploring Human Action Recognition for Rehabilitation Game Application

Júlio Castro Lopes
CeDRI, Instituto Politécnico de Bragança, Portugal
SusTEC, Instituto Politécnico de Bragança, Portugal
Email: juliolopes@ipb.pt

Isaac Van-Deste
CeDRI, Instituto Politécnico de Bragança, Portugal
Email: isaac.marcelino@ipb.pt

Rui Pedro Lopes
CeDRI, Instituto Politécnico de Bragança, Portugal
SusTEC, Instituto Politécnico de Bragança, Portugal
Email: rlopes@ipb.pt

***Abstract*—Through computer vision algorithms motion can be computed, which can be a crucial element to be integrated with serious game environments. To evaluate the efficacy of motion detection algorithms, the information of these algorithms can be used to perform Human Action Recognition (HAR). There are several algorithms to perform HAR, although skeleton approaches can be seen as the best way to isolate human motion. To extract the human skeleton representation, the work described in this paper evaluates three distinct methods: OpenPose (2D), YOLO-Pose (2D) and BlazePose (3D). The information translated by the skeleton representations is normalized by lightweight normalization algorithms (for further real-time application). To classify the video sequence and further action identification, a Long Short Term Memory network (LSTM), was used. Using the N-UCLA dataset, the highest F1 score of 0.745 was achieved using OpenPose skeleton extraction (2D), followed by the computation of the angles in each joint, demonstrating that the OpenPose skeleton representation can be the most viable solution for computing human motion in serious games.**

***Index Terms*—Human Action Recognition, Skeleton extraction, OpenPose, YOLO-Pose, BlazePose, Moving Joints Descriptor**

I. Introduction

Human Action Recognition (HAR) is receiving considerable attention from the scientific community, owing to its importance in a variety of applications and scenarios [1]. The ability to automatically detect a person's actions is essential in several situations [2]. Besides from the corresponding action-detection system, data acquisition is also a critical step, because of the difficulties associated with the environment.

HAR can be addressed using standard RGB images, depth and RGB (RGB-D), optionally followed by skeleton extraction with 2D or 3D joints coordinates, or even other methods, such as accelerometers or others. The primary goal of RGB-based action recognition is to analyze human actions using standard visual data, such as images or video frames. This method can be computationally efficient and applicable in scenarios where depth data is unavailable or impractical to obtain. However, it may struggle to capture fine-grained motion details and recognize actions under changing viewpoints or occlusions [3]. The incorporation of depth information in addition to RGB data addresses some of these drawbacks. The spatial relationships and 3D structure of human motions can be explicitly captured by 3D action recognition techniques by utilizing depth sensors. As a result, actions can be modeled with greater accuracy, occlusions can be handled better, and performance in challenging situations is enhanced [4].

The application of HAR algorithms can be seen in distinct scenarios, for example in surveillance [5], healthcare [6], sports [7] and even gaming [8]. These algorithms can have a significant impact in gaming scenarios, as they can identify which of the supposed actions a player is performing and whether or not the player is performing them correctly. Aside from that, using skeleton tracking algorithms allows for the isolation of human motion by analyzing all of the skeleton points or by only considering specific skeleton points.

Having this in mind, this work evaluates several methods of skeleton tracking (both 2D and 3D), for application in a rehabilitation game, developed in association. Using a sequence network to understand the sequences generated by each video in performing different activities, allows the determination of the most viable method for detecting and analyzing human motion.

The work described in this paper presents three distinct methods to track human skeletons and also feature extraction methods to reach useful information for training a Long-Short Term Memory (LSTM) neural network. The primary goal of this work is to use feature extraction methods with low computational requirements and evaluate which skeleton extraction model can prove to be the most effective one. As such, it is not expected to achieve State-Of-The-Art (SOTA) performance, but a simple and quick system to detect human actions, as part of the game requirements include the possibility for executing in limited resources devices, such as smartphones or VR goggles. Skeleton-based methods were the preferable choice as they allow not only the classification of the action but also for a simple computation of Body Motion Rate (BMR) [9], being able to isolate and detect human motion. The most viable skeleton algorithm and feature extraction method were determined by the combination that produces the highest accuracy.

This paper is structured in 5 sections, starting with this introduction. Section II follows, evaluating recent scientific developments and SOTA in the field of HAR, detailing authors'

methods and results. Section III briefly describes each method used to solve the proposed problem and also the experimental setup. The paper continues with Section IV, with results and associated discussion, and ends with some conclusions In Section V.

II. Related Work

During recent years, HAR has been widely discussed and used to solve problems in distinct scenarios, by researchers and companies [10]. There are several ways to process and identify human actions [11], although, this work will be focused on the skeleton extraction approach, by isolating human's skeleton and respective keypoints.

A. HAR Fundamentals

Malik et al. [12] introduced an innovative extraneous frame scrapping technique that harnesses 2D skeleton features and a Fine-KNN classifier-based HAR system to alleviate dimensionality challenges. The proposed method, evaluated on the Multi-Camera Action Dataset (MCAD) and INRIA Xmas Motion Acquisition Sequences (IXMAS) dataset, achieves compelling results, surpassing existing techniques. Specifically, the OpenPose-FineKNN with Extraneous Frame Scrapping Technique attains an accuracy of 89.75% on MCAD and 90.97% on IXMAS, outperforming CNN-LSTM and other SOTA methods. The work not only contributes a novel approach to HAR, but also addresses concerns related to computational complexity and feature extraction. The paper concludes with a comprehensive discussion of results, emphasizing the potential and superiority of the proposed technique over existing methods.

Kamel et al. [13] introduced an innovative method for Human Activity Recognition (HAR), utilizing three RGB channels (Red, Green, and Blue) of deep Convolutional Neural Networks (CNNs) applied to posture data and depth maps. The authors utilized the Moving Joints Descriptor (MJD) to represent body posture sequences, offering essential information about joint movement directions and changes in joint poses based on angle size. For representing depth map sequences, they employed the Depth Motion Image (DMI) descriptor. Unlike previous works such as Kim et al. (2014) which used two views, the authors employed DMI solely from the front view, aided by the MJD descriptor, resulting in reduced computational complexity. For feature extraction and classification, three CNN channels were used, trained with DMI and MJD descriptors. The first channel was trained by DMI; the second channel functioned as a link between two sub-channels, one of which was trained by MJD and the other by DMI. The third channel was exclusively trained with MJD. To enhance the accuracy of action recognition, the authors implemented score fusion operations. Every action was given a score by each CNN channel, and the action that received the highest score was deemed to be accurate. Three public datasetsMSRAction3D, UTD-MHAD, and MADwere used to evaluate their approach. The results showed competitive accuracy rates of 94.50%, 88.14%, and 91.86%, respectively.

In order to determine if OpenPose [14] and BlazePose [15] could produce clinically reliable body keypoints for virtual motion assessment, Mroz et al. [16] compared the two models. The study used root-mean-square error measures and Pearson correlation to assess the effectiveness of keypoint detection. BlazePose demonstrated a faster runtime compared to OpenPose, approximately 6 times quicker. However, the comparison revealed that BlazePose exhibited more instances where keypoints deviated from anatomical joint centers, suggesting that it might not be the optimal solution for clinically meaningful evaluations. Despite this, BlazePose provides an additional advantage by incorporating the z coordinate, offering valuable information about the human pose. In conclusion, the authors determined that, for their specific application, OpenPose was the preferred model for pose estimation. Nevertheless, the inclusion of the z coordinate in BlazePose could potentially offer significant insights into human pose that may be valuable in other contexts.

Ravipati et al. [17] proposed a combined CNN-LSTM model named Long-term Recurrent Convolutional Networks (LRCN) for HAR classification. The CNN is used for spatial feature extraction, while the LSTM is employed for temporal sequence modeling. They achieve good results demonstrating the effectiveness of the CNN-LSTM model, achieving an accuracy of 85.25%. The proposed LRCN shows robustness and high accuracy compared to other DL strategies. Overall, the study contributes to the ongoing advancements in technology for human activity recognition and envisions further improvements in deep learning models for this purpose.

B. Motion Algorithms - Applications in Serious Games

Although it was not possible to find games integrating HAR algorithms, the scientific literature already approaches this possibility.

Raffe and Garcia [18] investigated the use of commercial skeletal tracking and Virtual Reality (VR) technologies to improve gameplay interfaces in fall prevention exercise games for older adults. The study used the StepKinnection game as a reference and investigated the use of Microsoft Kinect for skeletal tracking and HTC Vive for head tracking and VR visualization. The authors carried out a self-reflective study to assess various avatar positioning modes for accurate stepping motions, physical comfort, and user engagement. While the study reveals promising opportunities for developing engaging step training games, it also identifies limitations, particularly in terms of accessibility, safety concerns, and technical challenges associated with the combined technologies. The paper emphasized the need for additional research into alternative interaction modalities to improve accessibility and engagement in game-based fall prevention training programs for the elderly.

Ma et al. [19] assessed the accuracy and validity of Mystic Isle, a rehabilitation game, using the Microsoft Kinect V2 sensor, for assessing motion in stroke patients'. In this study, the authors compared the data from the Kinect V2 with the Vicon system, the industry-standard optical motion capture system, for a range of full-body movements. The participants

completed trials in sitting and standing positions, with the results showing high correlation coefficients and signal-to-noise ratios for arm joints, while hip joints showed less stability. Mystic Isle showed potential for clinical assessments and home-based rehabilitation, especially with the potential for remote monitoring and data collection, despite certain limitations, such as sample homogeneity and possible difficulties with stroke patients' movement patterns.

Lidstone et al. [20] presented a pilot study for automated and scalable computerized assessment of motor imitation in children with Autism Spectrum Disorder (ASD) using a single 2D camera. Motor imitation is the ability to observe and mimic the actions or movements of others. For children with ASD, it is a crucial skill for the development of social skills, forming social bonds, and observant learning. The goal of their research was to compare the OpenPose skeleton (2D method) to the well-established Kinect 3D and Human Observation Coding (HOC), evaluating the feasibility of using only a 2D camera. Significant correlations were found between OpenPose 2D, Kinect 3D, and HOC, indicating concurrent and construct validity, though the Kinect 3D CAMI method demonstrated superior discriminating ability. Furthermore, motor imitation scores from all methods were significantly associated with social-communication impairments in children with ASD. According to the study, OpenPose 2D provides a scalable and accessible method for assessing motor imitation in children with ASD, potentially facilitating therapeutic interventions and the development of serious games that target social and motor skills.

III. Methodology

This section will go over the steps used to perform HAR, including feature extraction, image normalization (based on keypoint normalization, angles and MJD), and the classification algorithm.

A. Feature Extraction

AI systems generally use frame-by-frame analysis to process video sequences, applying algorithms to each frame of the video or to a sequence of them. The capacity to recognize motion patterns and accurately detect humans are critical components of the HAR algorithms' performance. Over time, numerous algorithms have been put forth to deal with these issues [21]. While skeleton-based approaches remain widely used in HAR, alternative methods exist for tackling the classification problem [22], [23]. However, this work exclusively relies on skeleton extraction algorithms, using OpenPose, YOLO-Pose and BlazePose for that purpose.

OpenPose is currently one of the most popular skeleton extraction algorithms. After processing each frame, the main keypoints, that make the body image skeletons, is obtained. Each skeleton is represented by 18 points in a 2D coordinate system, where each coordinate is the pixel position of a joint (keypoint) in the frame [24].

YOLO-Pose is also a 2D algorithm based on the YOLO object detection framework [25]. The skeleton detected is represented by 17 keypoints, each one having a pair of 2D coordinates and a confidence value. For this work, the version 7 of YOLO-Pose was used.

BlazePose is a real-time inference 3D skeleton technique designed for mobile devices [15]. The model presents high-fidelity body posture tracking, by using RGB video frames, to infer 33 3D landmarks and a background segmentation mask for the entire body.

Figure 1 shows the skeleton representation of the three algorithms.

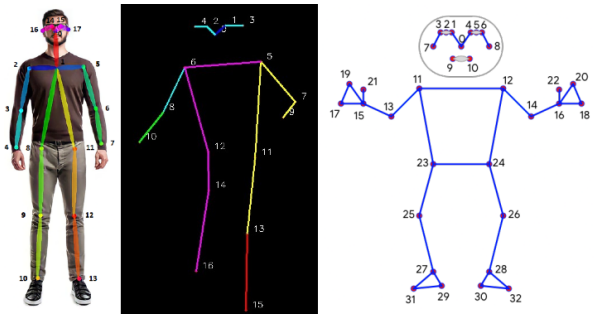


Fig. 1: OpenPose, YOLO-Pose and BlazePose skeleton representations

B. Skeleton representation

After skeleton extraction, three different methods were investigated for representing the detected joints (Figure 2).

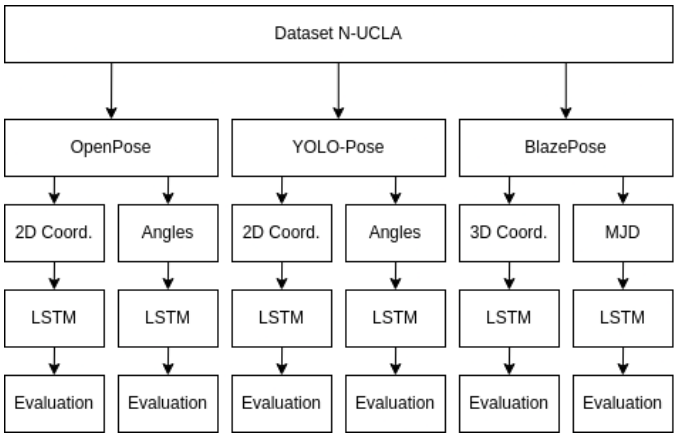


Fig. 2: Evaluation methodology

The first technique (Coordinates) normalizes the 2D coordinates extracted by OpenPose and YOLO-Pose. The second (Angles) computes the angles of each joint's segment when determining the angles between joints. The third method makes use of the MJD descriptor [13]. All of the algorithms that were discussed, employed the coordinates of every skeletal joint of its respective original skeleton. It is important to acknowledge that not every feature extraction technique discussed in this section was utilized for every skeleton algorithm. Since the coordinates are already normalized in BlazePose, the keypoint normalization was only performed in the OpenPose

and YOLO-Pose algorithms. Since the 3D formulation would be significantly more complicated, the calculation of the angles between the joints was restricted to the 2D techniques. BlazePose normalized coordinates are the sole sets of data to which the MJD has been applied because this descriptor depends on 3D data for classification.

OpenPose algorithm outputs 18 joints and YOLO-Pose outputs 17 joints. Image coordinates are used to represent both joints. Finding the lowest and greatest values of x and y in each frame may assist in normalizing them. Every real coordinate can be normalized with these four points. A standing static individual is used to show this reference in Figure 3.

Fig. 3: Representation of the rectangle referential for 2D skeleton normalization.

Equation 1 is used to compute the new coordinates. For each joint:

$$(new_x, new_y) = (\frac{x - min_x}{max_x - min_x}, \frac{y - min_y}{max_y - min_y}) \quad (1)$$

One possible method of deriving significant information from a human's movement is through the angles that the skeleton creates. To this purpose, the primary joints in the skeleton are represented by 12 defined angles for OpenPose and 8 formed angles for YOLO-Pose. Each angle is the result of a set of three vertices. All the keypoints are presented in Figure 1.

For the OpenPose angle computation, the following set of joints was used: 1 - (0, 1, 2); 2 - (1, 2, 3); 3 - (2, 3, 4); 4 - (0, 1, 5); 5 - (1, 5, 6); 6 - (5, 6, 7); 7 - (0, 1, 8); 8 - (1, 8, 9); 9 - (8, 9, 10); 10 - (0, 1, 11); 11 - (1, 11, 12); 12 - (11, 12, 13).

For YOLO-Pose, the following set of joints was used: 1 - (6, 8, 10); 2 - (6, 12, 14); 3 - (12, 14, 16); 4 - (5, 6, 8); 5 - (6, 5, 7); 6 - (5, 7, 9); 7 - (5, 11, 13); 8 - (11, 13, 15).

As a next step, the three points were used to calculate the angles formed by each vertices sequence. In order to achieve this, two new corresponding vertex definitions were generated:

Transformed Vertex 1:

$$newvertex_1x = vertex_2[x] - vertex_1[x] \quad (2)$$

$$newvertex_1y = vertex_2[y] - vertex_1[y] \quad (3)$$

Transformed Vertex 2:

$$newvertex_2x = vertex_3[x] - vertex_2[x] \quad (4)$$

$$newvertex_2y = vertex_3[y] - vertex_2[y] \quad (5)$$

Each skeleton segment is represented by a vector, and NumPy's function $arccos$ is used to calculate the angle between vectors (Figure 4).

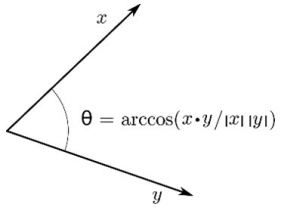


Fig. 4: Angle calculation.

The methodology employed by the MJD bears some changes to the work of Kamel et al. [13]. Out of the 20 joints, they have only utilized 13, and the BlazePose algorithm's original 33 joints were used in this work. Moreover, they employed a central critical point, the hip joint, that BlazePose is unable to identify. This was addressed by utilizing the middle point between the closest joints to generate a hip point estimator (Figure 5). The point centered on the hip is one of the most stable regions of the human body, so it makes sense to choose it as the origin of the spherical coordinates. Based on this, each subsequent joint will have two angles and the distance from the reference point.

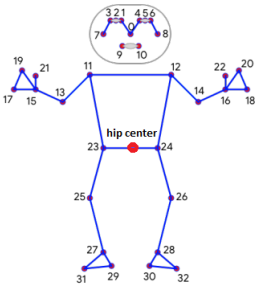


Fig. 5: Hip Center Point - BlazePose Skeleton.

The z coordinate can be traduced in a boost on the detected human skeleton as it also brings depth into account. The real coordinates of every joint are automatically normalized by BlazePose's method, therefore the coordinates were converted back to the real coordinates. After this restoration procedure, Equation 6 was used to convert the data into spherical coordinates.

$$(r, \theta, \phi) = (\sqrt{x^2 + y^2 + z^2}, \arccos(z/r), \arctan 2(y, x)) \quad (6)$$

C. Sequence Classification

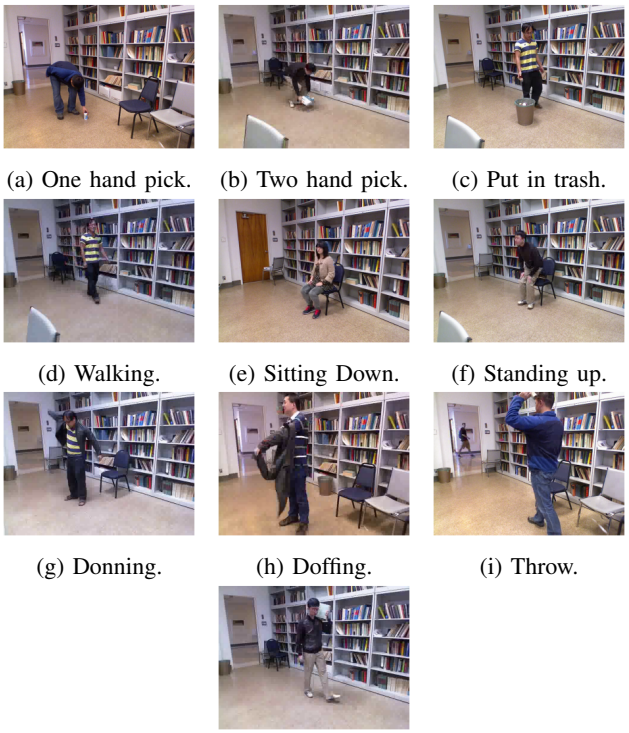
Since the DNN classification approach necessitates a prompt and precise response upon training, it is also a crucial step in

the HAR process. LSTM was selected based on the literature review and its sequence-handling capabilities.

LSTM, a variant of Recurrent Neural Network (RNN) utilized in Deep Learning, possesses memory capabilities to retain pertinent information and discard obsolete data. Unlike traditional RNNs, LSTMs feature a gate that receives the input that is treated by a hidden state using the functions sigmoid and tanh, leading to the production of an output gate. This result is computed repeatedly, creating a loop, and is added as a new instant (input) each time the classifier achieves the expected result. The forget gate determines the information deemed irrelevant and eligible for removal, the input gate identifies information suitable for inclusion in the cell, and the output gate specifies the data to be output during each loop iteration. Each gate can transmit either a portion or the entirety of the information, allowing flexibility. For instance, the forget gate can discard all or part of the information, as illustrated by [26].

D. Dataset

The N-UCLA dataset [27] is composed of ten action categories: picking up with one hand, picking up with two hands, dropping trash, walking around, sitting down, standing up, donning, doffing, throwing, and carrying. Each action was performed by ten actors. This dataset has the particularity that most videos are short, containing just the number of frames from the beginning of the action to the end (Figure 6).



(a) One hand pick. (b) Two hand pick. (c) Put in trash.

(d) Walking. (e) Sitting Down. (f) Standing up.

(g) Donning. (h) Doffing. (i) Throw.

(j) Carry.

Fig. 6: N-UCLA Dataset - Set of Actions.

E. HAR Setup

This paper provides numerous distinct approaches to classify the activity a human is performing, following the methodology outlined in Figure 2:

1) The extraction of the OpenPose skeleton is followed by the 2D coordinate normalization.
2) After extracting the skeleton from OpenPose, the angles formed by the chosen sets of joints were calculated.
3) YOLO-Pose skeleton extraction, then the 2D coordinates are normalized.
4) YOLO-Pose skeleton extraction, and the angles formed by the chosen joint sets were then calculated.
5) The BlazePose skeleton extraction process is succeeded by the 3D coordinate normalization.
6) The BlazePose skeleton extraction process is then followed by the MJD descriptor computation.

All these frame sequences were used as individually as input to each respective LSTMs.

An LSTM was used to categorize a given sequence, identifying it as a single entity among the ten actions in the N-UCLA dataset. Various hidden layer sizes, namely 48, 100, and 300, were tested; nonetheless, these values turned out to be the optimal answer. The F1 metric was employed to assess the efficacy of HAR approaches. F1 metric can be seen as the best acceptable metric for unbalanced datasets, which functions as the balanced average of Recall and Precision. Accuracy may be helpful in situations when the class distribution is similar, but the F1-score is a better measure in cases where the classes are not balanced.

Additionally, for training, the Adam optimizer, a batch size of 16, a learning rate of 0.001, and 300 epochs were employed for each HAR variation. The dataset was divided, using a split of 80% to train the model and 20% to validate the efficacy of the proposal. Although this configuration produced the best result, other configurations were also evaluated. A 32-core AMD Ryzen Threadripper 3970X processor paired with an NVIDIA GeForce RTX 3090 graphics card and 64GB of RAM was utilized for this endeavor in terms of software.

IV. Discussion and Results

After combining all the scenarios described in Figure 2, the results of the F1 score were reported (Table I).

TABLE I: Results on N-UCLA dataset (F1 score)

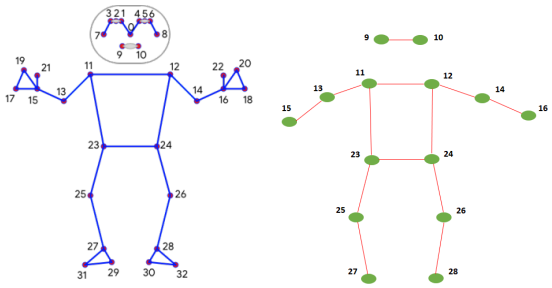
	OpenPose		YOLO-Pose		BlazePose	
LSTM	2D Coord.	Angles	2D Coord.	Angles	3D Coord.	MJD
48	0.733	0.745	0.718	0.735	0.584	0.476
100	0.722	0.711	0.715	0.724	0.536	0.511
300	0.711	0.703	0.678	0.661	0.557	0.444

The best result was obtained with an LSTM utilizing 48 as the hidden size and pre-processing with angle computation, resulting in an F1 score of 0.745.

It is evident from BlazePose's classification performance that this skeleton extraction technique was insufficient to

outperform the 2D skeleton techniques. This led to more experiences to try to understand the low score provided by BlazePose approaches.

Since there are many keypoints in BlazePose's original skeleton, a lower number was tried, aimed at keeping only the data required to categorize each action. From the 33, a total of 14 were selected (Figure 7).



(a) BlazePose skeleton full representation [15] (b) BlazePose skeleton simplified representation

Fig. 7: Full and partial keypoint representation

Utilizing only 14 segments of BlazePose's identified skeleton, did not result in an improvement in the F1-Score; the best F1 score was 0.574 (Table II).

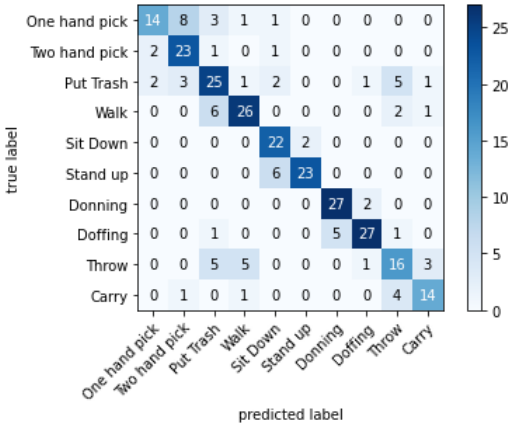
TABLE II: F1 score reported using N-UCLA dataset using BlazePose simplified representation

	BlazePose	
LSTM	3D Coord.	MJD
48	0.521	0.456
100	0.562	0.464
300	0.501	0.574

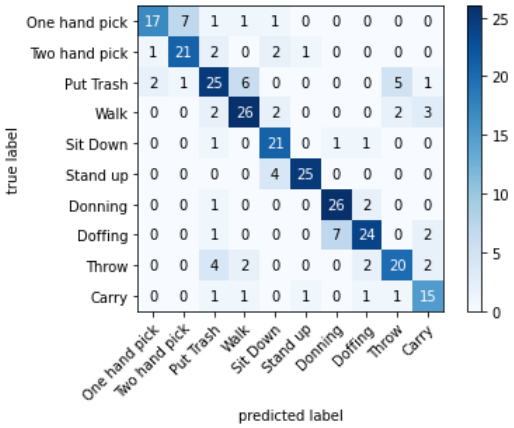
The 2D algorithms proved to be the more robust models for these classification methods. It is possible to observe on Table I that OpenPose and YOLO-Pose have very similar F1 scores.

In the other hand, these models outperformed BlazePose by a considerable margin. Not a single BlazePose technique surpasses an F1 of 0.60, which is a relatively low number in contrast with 2D approaches. Even though it was anticipated that more information about the generated skeleton (z coordinate in BlazePose) would boost accuracy, OpenPose and YOLO-Pose demonstrated significantly more dependable and consistent, obtaining F1 values greater than 0.70. The confusion matrices for the classification of the optimal techniques utilizing the OpenPose skeleton were also displayed (Figure 8) in order to provide a deeper understanding of the best model's recognition capacity.

Through a comparison of the confusion matrix between the Angles methods (Figure 8b) and the 2D Coordinates techniques (Figure 8a), it is feasible to deduce that various classification abilities were obtained using the same skeleton extraction approach. It is important to highlight that, because these classes share many characteristics, both configurations



(a) 2D coordinates



(b) Angles

Fig. 8: Confusion Matrix - OpenPose most effective methods.

find it difficult to predict when someone will "don" or "doff." This also holds true for when someone will "sit down" or "stand up," "pick with one hand" or "pick with two hands," and when someone will "walk" or "put trash" because both activities require walking.

Table III shows the accuracy attained by multiple methods using the N-UCLA dataset, allowing us to validate the results of our approach with SOTA. Since the majority of cutting-edge methods employ accuracy rather than F1-score, our top solution's accuracy is also shown in this table.

TABLE III: Proposed methodolody vs. SOTA using the N-UCLA dataset

Methods	Accuracy	F1-Score
HBRNN-L [28]	0.805	
Glimpse Clouds [29]	0.876	
GCNHCRF [30]	0.915	
VE-GCN [31]	0.918	
AGC-LSTM [32]	0.933	
CTR-GCN [33]	0.965	
LST [34]	0.972	
Our method	0.746	0.745

By the analysis of Table III, it is possible to conclude that

our method can still be improved to achieve SOTA results. It should be mentioned that, although GCNs are among the most accurate approaches, they have certain drawbacks that force the employment of LSTM networks. The reliability to coordinates, compatibility with different inputs, and scalability to several users are a few of these drawbacks.

V. Conclusions

This work described a HAR methodology and the validation of the methods in the N-UCLA dataset. A skeleton-based method was employed, with two algorithms (OpenPose and YOLO-Pose) for extracting 2D data and another algorithm (BlazePose) for extracting 3D data from the identified human skeleton. Several methods were applied to the extracted coordinates, such as the MJD descriptor, the angle between joint computation, and coordinate normalization.

Although our results fell short of SOTA goals, they did enable us to quickly and easily devise real-time methods for comprehending the obstacles and opportunities associated with HAR mitigation. On the N-UCLA dataset, OpenPose achieved a respectable F1-Score of 0.745, demonstrating their superior robustness over BlazePose (3D) and YOLO-Pose. Given the increased depth information, it was anticipated that the 3D algorithm would perform best; however, the 2D approach turned out to be more reliable.

Acknowledgements

The authors are grateful to the Foundation for Science and Technology (FCT, Portugal) for financial support through national funds FCT/MCTES (PIDDAC) to CeDRI, UIDB/05757/2020 (DOI: 10.54499/UIDB/05757/2020) and UIDP/05757/2020 (DOI: 10.54499/UIDP/05757/2020) and SusTEC, LA/P/0007/2020 (DOI: 10.54499/LA/P/0007/2020).

References

[1] D.-Q. Vu, T. Thu, N. Lê, and J.-C. Wang, "Deep Learning for Human Action Recognition: A Comprehensive Review," *APSIPA Transactions on Signal and Information Processing*, vol. 12, Jan. 2022. DOI: 10.1561/116.00000068.

[2] I. Jegham, A. Ben Khalifa, I. Alouani, and M. A. Mahjoub, "Vision-based human action recognition: An overview and real world challenges," *Forensic Science International: Digital Investigation*, vol. 32, p. 200 901, Mar. 2020, ISSN: 2666-2817. DOI: 10.1016/j.fsidi.2019.200901.

[3] M.-R. Tseng, A. Gupta, C.-K. Tang, and Y.-W. Tai, *HAA4D: Few-Shot Human Atomic Action Recognition via 3D Spatio-Temporal Skeletal Alignment*, arXiv:2202.07308 [cs], Feb. 2022.

[4] W. Lin and J. Yu, "Beyond 2D: Fusion of Monocular 3D Pose, Motion and Appearance for Human Action Recognition," in *2019 22th International Conference on Information Fusion (FUSION)*, Jul. 2019, pp. 1–8. DOI: 10.23919/FUSION43075.2019.9011279.

[5] M. A. Khan, K. Javed, S. A. Khan, *et al.*, "Human action recognition using fusion of multiview and deep features: An application to video surveillance," en, *Multimedia Tools and Applications*, Mar. 2020, ISSN: 1573-7721. DOI: 10.1007/s11042-020-08806-9.

[6] X. Zhou, W. Liang, K. I.-K. Wang, H. Wang, L. T. Yang, and Q. Jin, "Deep-Learning-Enhanced Human Activity Recognition for Internet of Healthcare Things," *IEEE Internet of Things Journal*, vol. 7, no. 7, pp. 6429–6438, Jul. 2020, Conference Name: IEEE Internet of Things Journal, ISSN: 2327-4662. DOI: 10.1109/JIOT.2020.2985082.

[7] M. Ullah, M. Mudassar Yamin, A. Mohammed, S. Daud Khan, H. Ullah, and F. Alaya Cheikh, "ATTENTION-BASED LSTM NETWORK FOR ACTION RECOGNITION IN SPORTS," in *Electronic Imaging*, ISSN: 2470-1173 Issue: 6 Journal Abbreviation: ei, vol. 33, Jan. 2021, pp. 302–1–302–6. DOI: 10.2352/ISSN.2470-1173.2021.6.IRIACV-302.

[8] V. Bloom, D. Makris, and V. Argyriou, "G3D: A gaming action dataset and real time action recognition evaluation framework," in *2012 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, ISSN: 2160-7516, Jun. 2012, pp. 7–12. DOI: 10.1109/CVPRW.2012.6239175.

[9] J. C. Lopes, J. Vieira, I. Van-Deste, and R. P. Lopes, "An Architecture for Capturing and Synchronizing Heart Rate and Body Motion for Stress Inference," in *2023 IEEE 11th International Conference on Serious Games and Applications for Health (SeGAH)*, ISSN: 2573-3060, Aug. 2023, pp. 1–7. DOI: 10.1109/SeGAH57547.2023.10253815.

[10] K. Gedamu, Y. Ji, L. Gao, Y. Yang, and H. T. Shen, "Relation-mining self-attention network for skeleton-based human action recognition," en, *Pattern Recognition*, vol. 139, p. 109 455, Jul. 2023, ISSN: 0031-3203. DOI: 10.1016/j.patcog.2023.109455.

[11] R. Kumar and S. Kumar, "A survey on intelligent human action recognition techniques," en, *Multimedia Tools and Applications*, Nov. 2023, ISSN: 1573-7721. DOI: 10.1007/s11042-023-17529-6.

[12] N. u. R. Malik, U. U. Sheikh, S. A. R. Abu-Bakar, and A. Channa, "Multi-View Human Action Recognition Using Skeleton Based-FineKNN with Extraneous Frame Scrapping Technique," en, *Sensors*, vol. 23, no. 5, p. 2745, Jan. 2023, Number: 5 Publisher: Multidisciplinary Digital Publishing Institute, ISSN: 1424-8220. DOI: 10.3390/s23052745.

[13] A. Kamel, B. Sheng, P. Yang, P. Li, R. Shen, and D. D. Feng, "Deep Convolutional Neural Networks for Human Action Recognition Using Depth Maps and Postures," *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, vol. 49, no. 9, pp. 1806–1819, Sep. 2019, Conference Name: IEEE Transactions on Systems, Man, and Cybernetics: Systems, ISSN: 2168-2232. DOI: 10.1109/TSMC.2018.2850149.

[14] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh, "Realtime Multi-person 2D Pose Estimation Using Part Affinity Fields," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, ISSN: 1063-6919, Jul. 2017, pp. 1302–1310. DOI: 10.1109/CVPR.2017.143.

[15] V. Bazarevsky, I. Grishchenko, K. Raveendran, T. Zhu, F. Zhang, and M. Grundmann, *BlazePose: On-device Real-time Body Pose tracking*, arXiv:2006.10204 [cs], Jun. 2020.

[16] S. Mroz, N. Baddour, C. McGuirk, *et al.*, "Comparing the Quality of Human Pose Estimation with BlazePose or OpenPose," in *2021 4th International Conference on Bio-Engineering for Smart Technologies (BioSMART)*, Dec. 2021, pp. 1–4. DOI: 10.1109/BioSMART54244.2021.9677850.

[17] A. Ravipati, R. K. Kondamuri, M. P. A, and A. M. J, "Vision Based Detection and Analysis of Human Activities," en, in *2023 7th International Conference on Trends in Electronics and Informatics (ICOEI)*, Tirunelveli, India: IEEE, Apr. 2023, pp. 1542–1547, ISBN: 9798350397284. DOI: 10.1109/ICOEI56765.2023.10125829.

[18] W. L. Raffe and J. A. Garcia, "Combining skeletal tracking and virtual reality for game-based fall prevention training for the elderly," in *2018 IEEE 6th International Conference on Serious Games and Applications for Health (SeGAH)*, ISSN: 2573-3060, May 2018, pp. 1–7. DOI: 10.1109/SeGAH.2018.8401371.

[19] M. Ma, R. Proffitt, and M. Skubic, "Validation of a Kinect V2 based rehabilitation game," en, *PLOS ONE*, vol. 13, no. 8, e0202338, Aug. 2018, Publisher: Public Library of Science, ISSN: 1932-6203. DOI: 10.1371/journal.pone.0202338.

[20] D. E. Lidstone, R. Rochowiak, C. Pacheco, B. Tunçgenç, R. Vidal, and S. H. Mostofsky, "Automated and scalable Computerized Assessment of Motor Imitation (CAMI) in children with Autism Spectrum Disorder using a single 2D camera: A pilot study," *Research in Autism Spectrum Disorders*, vol. 87, p. 101 840, Sep. 2021, ISSN: 1750-9467. DOI: 10.1016/j.rasd.2021.101840.

[21] A. Sarkar, A. Banerjee, P. K. Singh, and R. Sarkar, "3D Human Action Recognition: Through the eyes of researchers," en, *Expert Systems with Applications*, vol. 193, p. 116 424, May 2022, ISSN: 0957-4174. DOI: 10.1016/j.eswa.2021.116424.

[22] Y. Zhang, Q. Guo, Z. Du, and A. Wu, "Human Action Recognition for Dynamic Scenes of Emergency Rescue Based on Spatial-Temporal Fusion Network," en, *Electronics*, vol. 12, no. 3, p. 538, Jan. 2023, Number: 3 Publisher: Multidisciplinary Digital Publishing Institute, ISSN: 2079-9292. DOI: 10.3390/electronics12030538.

[23] S. A. Mahmoudi, O. Amel, S. Stassin, M. Liagre, M. Benkedadra, and M. Mancas, "A Review and Comparative Study of Explainable Deep Learning Models Applied on Action Recognition in Real Time," en, *Electronics*, vol. 12, no. 9, p. 2027, Jan. 2023, Number: 9 Publisher: Multidisciplinary Digital Publishing Institute, ISSN: 2079-9292. DOI: 10.3390/electronics12092027.

[24] Z. Cao, G. Hidalgo, T. Simon, S.-E. Wei, and Y. Sheikh, "OpenPose: Realtime Multi-Person 2D Pose Estimation Using Part Affinity Fields," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 43, no. 1, pp. 172–186, Jan. 2021, Conference Name: IEEE Transactions on Pattern Analysis and Machine Intelligence, ISSN: 1939-3539. DOI: 10.1109/TPAMI.2019.2929257.

[25] D. Maji, S. Nagori, M. Mathew, and D. Poddar, *YOLO-Pose: Enhancing YOLO for Multi Person Pose Estimation Using Object Keypoint Similarity Loss*, arXiv:2204.06806 [cs], Apr. 2022. [Online]. Available: http://arxiv.org/abs/2204.06806 (visited on 09/27/2023).

[26] F. A. Gers, N. N. Schraudolph, and J. Schmidhuber, "Learning Precise Timing with LSTM Recurrent Networks," en, p. 29, 2002.

[27] J. Wang, X. Nie, Y. Xia, Y. Wu, and S.-C. Zhu, "Cross-view action modeling, learning and recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2014.

[28] Y. Du, Y. Fu, and L. Wang, "Representation Learning of Temporal Dynamics for Skeleton-Based Action Recognition," *IEEE Transactions on Image Processing*, vol. 25, no. 7, pp. 3010–3022, Jul. 2016, Conference Name: IEEE Transactions on Image Processing, ISSN: 1941-0042. DOI: 10.1109/TIP.2016.2552404.

[29] F. Baradel, C. Wolf, J. Mille, and G. W. Taylor, *Glimpse Clouds: Human Activity Recognition from Unstructured Feature Points*, arXiv:1802.07898 [cs], Aug. 2018.

[30] K. Liu, L. Gao, N. Mefraz Khan, L. Qi, and L. Guan, "Graph Convolutional Networks-Hidden Conditional Random Field Model for Skeleton-Based Action Recognition," in *2019 IEEE International Symposium on Multimedia (ISM)*, Dec. 2019, pp. 25–256. DOI: 10.1109/ISM46123.2019.00013.

[31] K. Liu, L. Gao, N. M. Khan, L. Qi, and L. Guan, "Integrating vertex and edge features with Graph Convolutional Networks for skeleton-based action recognition," en, *Neurocomputing*, vol. 466, pp. 190–201, Nov. 2021, ISSN: 0925-2312. DOI: 10.1016/j.neucom.2021.09.034.

[32] C. Si, W. Chen, W. Wang, L. Wang, and T. Tan, *An Attention Enhanced Graph Convolutional LSTM Network for Skeleton-Based Action Recognition*, arXiv:1902.09130 [cs], Mar. 2019. DOI: 10.48550/arXiv.1902.09130.

[33] Y. Chen, Z. Zhang, C. Yuan, B. Li, Y. Deng, and W. Hu, *Channel-wise Topology Refinement Graph Convolution for Skeleton-Based Action Recognition*, arXiv:2107.12213 [cs], Aug. 2021. DOI: 10.48550/arXiv.2107.12213.

[34] W. Xiang, C. Li, Y. Zhou, B. Wang, and L. Zhang, *Language Supervised Training for Skeleton-based Action Recognition*, arXiv:2208.05318 [cs], Aug. 2022. DOI: 10.48550/arXiv.2208.05318.