DOI: 10.15625/vap.2023.0066

PERFORMANCE EVALUATION OF MEDIAPIPE AND OPENPOSE FOR SKELETON DATA EXTRACTION

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

[1]Faculty of Information Technology, University of Sciences, Hue University
[2]Faculty of Information Technology, HCMC University of Technology and Education
[3]University of Sciences, Hue University

pkanh.dhkh23@hueuni.edu.vn, dunghv@hcmute.edu.vn, lvtlan@husc.edu.vn

Corresponding: lvtlan@husc.edu.vn

ABSTRACT: *In the field of human action recognition, leveraging the proven advantages, the utilization of skeletal data extracted from various datasets as input for action recognition models has emerged as a prominent research direction in recent years. In this study, our focus lies in evaluating two approaches for extracting skeletal data based on OpenPose and Mediapipe frameworks across the KTH and UTD-MHAD datasets. The skeletal data extracted from both methods serves as input for the AcTv2 model, an enhanced version of the AcT model. Through training and evaluating on the AcTv2 model, we ascertain the effectiveness of both skeletal data extraction methods on specific datasets. The experimental results of this research contribute to a better understanding of the efficacy of these skeletal data extraction methods in providing informative data for the AcTv2 model to recognize human actions across different datasets.*

Keywords: *Deep learning, Human action regconition, Skeleton data.*

I. INTRODUCTION

In the realm of computer vision and artificial intelligence, the task of Human Action Recognition (HAR) has become a notable research focus in recent years. HAR is tasked with identifying and categorizing human actions from input data such as images or videos. Despite the emergence of numerous proposed methods to tackle the HAR problem in recent years, achieving high accuracy in action recognition still faces several challenges [1], among which the input data factor plays a crucial role in enhancing accuracy [2].

Among various prevalent forms of input data in HAR, skeleton data has been demonstrated as an effective option and was initially introduced in [3]. Skeleton data serves as an abstract representation of human shape and motion information, where each skeleton frame represents a specific position on the body and the relationships between them. Utilizing skeleton data brings forth numerous significant advantages and paves the way for a completely new direction in the field of human action recognition research.

Firstly, skeleton data focuses solely on the fundamental positions and relationships of the human body, eliminating irrelevant factors. Therefore, skeleton data reduces dependency on external elements like lighting, background, and surrounding environment. Thus, employing skeleton data enhances action recognition performance in scenarios with low lighting, complex backgrounds, or unfavorable external conditions.

Secondly, skeleton data reduces computational and storage costs. Compared to full image or video data, skeleton data is more compact in size. This efficiency facilitates data processing and analysis, while also decreasing the demand for computational resources and storage space.

The AcT (Action Transformer) model is a powerful architecture built upon the foundation of the Vision Transformer [4]. AcT was introduced in a distinct manner in the study [5], focusing on recognizing short actions based on skeleton data. The incorporation of the Transformer architecture [6] within the AcT model signifies a significant advancement in deep learning in recent years. The self-attention mechanism within the Transformer architecture, with multiple heads, has demonstrated effectiveness across various tasks such as image resolution enhancement [7, 8], image classification [4, 9, 10], and more. However, to optimize performance and generalization capability, in a recent research endeavor, we introduced the AcTv2 model, an improved version of the original AcT model, augmented with essential layers.

In this paper, we focus on comparing the performance of two skeletal data extraction methods - OpenPose and Mediapipe - on two widely used datasets: KTH and UTD-MFA. We conducted performance evaluations for both methods, followed by utilizing the extracted skeletal data as input for the AcTv2 model. Through a series of experiments on these datasets, our goal is to provide a deeper insight into the capabilities of the AcTv2 model and the effectiveness of the skeletal data extraction methods from OpenPose and Mediapipe.

II. RELATED WORKS

In 2016, Presti and La Cascia made significant contributions by shedding light on the 3D skeletal-based approach in the field of action recognition [11]. By synthesizing relevant studies, the author's team presented a

comprehensive perspective on preprocessing methods, descriptor sets tailored for skeletal data, along with performance evaluation through verification methods.

One of the challenges in human action recognition based on skeleton data through computer vision is ensuring that the number of input frames is sufficiently large for accurate action identification while also managing computational complexity. A study [12] proposed a solution by extracting representative frames for a group of closely related frames, thereby reducing the input data load and enhancing computational efficiency. Additionally, several other studies have focused on integrating skeleton data with various neural network models for action recognition [13, 14, 15].

These studies have primarily focused on utilizing skeleton data captured from sensor devices, proposing methods to enhance accuracy and reduce computational costs for action recognition. These solutions can prove effective when applied to systems where individuals actively carry sensor devices, such as healthcare monitoring systems or group exercise training applications. However, in specific scenarios where requiring individuals to wear sensor devices is infeasible, these solutions might not be applicable efficiently. In contrast, skeleton data is stable, less influenced by environmental factors, easily storable, and can be effectively employed for action classification without the need for individuals to wear any devices.

A novel idea that has emerged in recent years has introduced a fresh approach to human action recognition using skeleton data. Instead of relying on skeleton data collected from sensor devices, a recent study [16] has proposed a technique to transform video data into 3D body pose representations, which are then used to construct skeleton data. The results of this approach were compared with traditional skeleton data obtained from sensor devices.

In a study focused on the application of Human Action Recognition (HAR) in the medical field [17], the authors harnessed the capabilities of OpenPose to transform video data and consecutive frames into skeletal information. This information was then represented as three-dimensional vectors. Innovatively, the authors employed a Generative Adversarial Network (GAN) along with the depth-first search (DFS) algorithm to recognize actions and gestures of patients based on these vectors. The model developed in this study demonstrates high applicability in supporting patients during physical therapy exercises.

Another study [18] introduced a series of feature extraction techniques from skeletal data, including dynamic skeleton, skeleton superposition, and body articulations. These features were then utilized as input data to train CNN models such as MobileNet, VGG16, and ResNet 50.

Another study [19] addressed the inconvenience of using sensor devices for remote elderly care. The authors proposed utilizing optical streams to extract skeletal data from users. They constructed the REMS (Real-time Elderly Monitoring for senior Safety) model, achieving experimental results with high accuracy and real-time performance.

The study [20] proposed a solution for building a mobile patient observation system to assist in diagnosing and treating neurodegenerative disorders related to central nervous system degeneration. The system utilizes video data from low-cost cameras, extracting skeletal data during the patient's performance of the Sit-to-Stand (StS) test. Machine learning methods are employed to compare healthy individuals and those with neurodegenerative diseases, aiding in the timely detection of disease symptoms. The system was tested in two nursing homes and achieved promising results, with a system accuracy of up to 95.2%.

III. SKELETION EXTRACTION SOLUTIONS FOR ACTION RECOGNITION

A. Architecture of action recognition based on sekeletion data

Based on skeleton data, there have been various proposed solutions for human action recognition. Common approaches include using deep learning models like CNNs and RNNs, utilizing feature-based classification techniques such as SVM and Random Forest, employing probabilistic graphical models like CRFs to model interactions between skeletons, and developing probability-based action recognition models, etc. Among these solutions, those utilizing deep learning models have received significant attention and were proposed early on. These deep learning models use skeleton data as input for training and are capable of recognizing human actions after the training process. The AcTv2 model is constructed in line with this approach.

The AcTv2 model, which has demonstrated effectiveness in human action recognition from skeleton data, is an enhancement of the AcT model which was proposed in a recent study [21]. In this study, this model is utilized as a tool for reevaluating the skeleton datasets extracted by OpenPose and MediaPipe, as previously described.

Based on the AcT model, several modifications were introduced to enhance its performance as follows in Fig 1. We added a BatchNormalization1 layer to stabilize the input of the preceding Dense layer in the AcT model. This layer adjusts the output values, ensuring stability and consistency within the model. Subsequently, we applied a Dropout layer to randomly deactivate a portion of the output units during training. This helps mitigate overfitting issues and improves the model's generalization ability. The Dense layer establishes full connections between the layers, allowing for the learning of intricate features. Finally, the BatchNormaalization2 layer normalizes the output of the Dense layer

before making the ultimate predictions. This ensures stable and accurate prediction results. These enhancements collectively contribute to the improved performance of the AcTv2 model in capturing complex temporal and spatial patterns from skeleton data for action recognition.

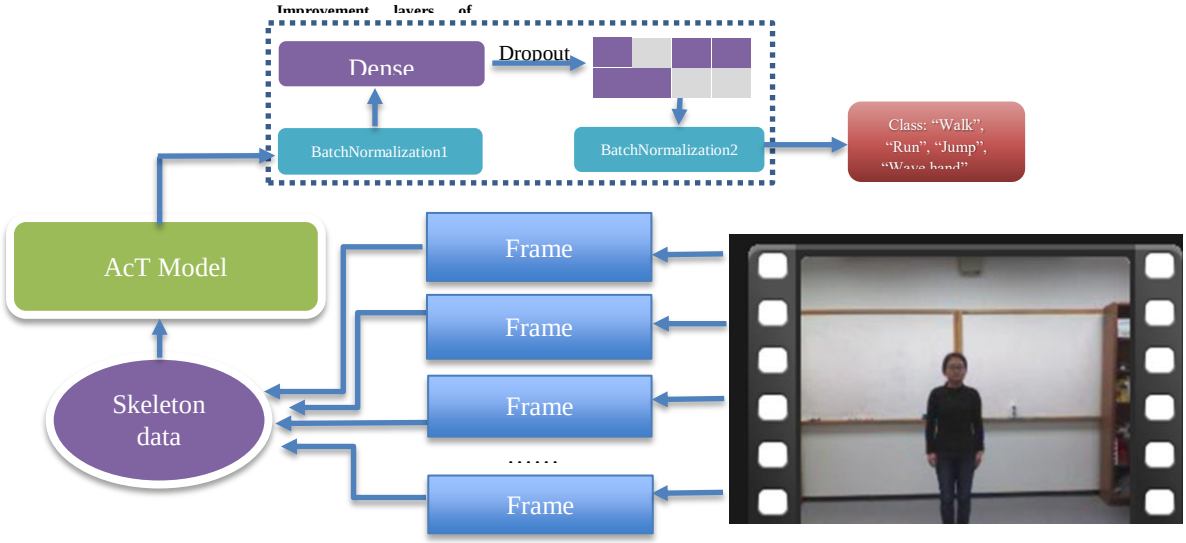

Fig 1. Architecture of AcTv2 model, an enchanced of AcT model

B. OpenPose and Mediapipe for skeletion extraction

OpenPose

As discussed in Section I, utilizing skeletal data brings significant advantages to human action recognition, including high accuracy and efficient real-time processing. This has spurred the development of numerous studies and models aimed at extracting skeletal data from various datasets. Among these, OpenPose [22] stands out as a model of great interest and has become a foundational resource for many researchers to enhance. OpenPose was developed based on the PAFs (part affinity fields) model [23] by its own authors. Both of these models have been tested and refined using the COCO dataset [24]. Figure 2 illustrates an example of extracting skeleton data, Part Confidence Maps and Part Affinity Fields (PAF) from an input image with OpenPose.

(a) (b) (c) (d)

Fig 2. Extract (b) skeleton data, (c) Part Confidence Maps and (d) Part Affinity Fields from an (a) image

The input data of the OpenPose model consists of images with dimensions w × h. OpenPose utilizes two essential concepts in the process of extracting skeleton data, namely Part Confidence Maps - representing a set S that depicts the joints of the human body, and Part Affinity Fields - a set containing vectors used to represent the movement directions of body parts based on their connected joints. To extract skeleton data, the OpenPose model employs 25 skeleton joints as follows in Table 1 and Fig 3:

Table 1. List of 25 skeleton joints used by Openpose model

Skeleton joints	Human part	Skeleton joints	Human part	Skeleton joints	Human part	Skeleton joints	Human part	Skeleton joints	Human part
0	Nose	5	Left Shoulder	10	Right Heel	15	Right Eye	20	Left Small toe
1	Neck	6	Left Elbow	11	Right Ankle	16	Left Eye	21	Left Ankle
2	Right Shoulder	7	Left Wrist	12	Left Hip	17	Right Ear	22	Right Big toe

3	Right Elbow	8	Mid Hip	13	Left Knee	18	Left Ear	23	Right Small toe
4	Right Wrist	9	Right Knee	14	Left Heel	19	Left Big toe	24	Right Ankle

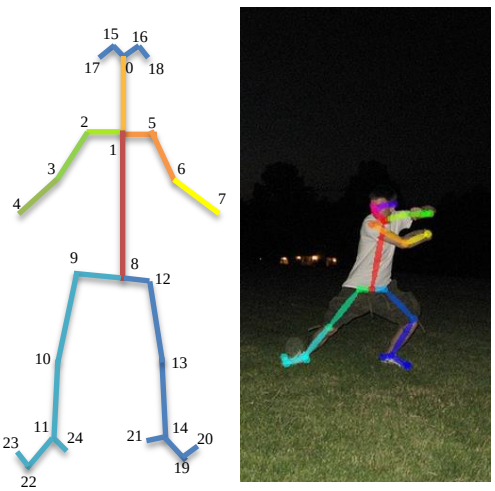

Fig 3. 25 Skeleton joints of the OpenPose model and an example of extracted Skeleton data from input image

Furthermore, with the latest improvements, OpenPose also supports extracting skeleton data from faces and hands. Another notable advantage of OpenPose is its ability to extract skeleton data for multiple individuals within the same frame. However, the OpenPose model still has several limitations. In certain specific cases, the extracted skeleton data might not be accurate or even empty.

Mediapipe

Mediapipe is an open-source library developed by Google, designed to facilitate the analysis of image and video data by extracting information about the positions and features of skeletons and keypoints within images. This library provides powerful tools for recognizing and tracking crucial elements in images. It has been extensively used by various researchers to create diverse applications in the field of computer vision related to image and video processing, such as gesture recognition and real-time video processing. Initially introduced in 2019 through a research publication [25], Mediapipe has undergone multiple advancements. It has now become one of the popular libraries utilized in numerous studies focused on human action recognition. To extract skeleton data, the MediaPipe model employs 33 landmarks as follows in Table 2 and Fig 4.

Table 2. List of 33 landmarks in the MediaPipe library

Landmarks - Human part	Landmarks - Human part	Landmarks - Human part	Landmarks - Human part
0 - nose	9 - mouth right	17 - right pinky knuckle #1	25 - right knee
1 - right eye inner	10 - mouth left	18 - left pinky knuckle #1	26 - left knee
2 - right eye	11 - right shoulder	19 - right index knuckle #1	27 - right ankle
3 - right eye outer	12 - left shoulder	20 - left index knuckle #1	28 - left ankle
4 - left eye inner	13 - right elbow	21 - right thumb knuckle #2	29 - right heel
5 - left eye	14 - left elbow	22 - left thumb knuckle #2	30 - left heel
6 - left eye outer	15 - right wrist	23 - right hip	31 - right foot index
7 - right ear	16 - left wrist	24 - left hip	32 - left foot index
8 - left ear			

In contrast to OpenPose, which works with input data containing multiple individuals, MediaPipe focuses solely on recognizing input data from a single human subject, which can be in the form of images or videos. Additionally, instead of using the term "Skeleton Joints" like OpenPose, MediaPipe refers to the human body's key points as "Landmarks," with a total of 33 Landmarks across the human body. These Landmarks can be utilized to extract data for both facial and foot features.

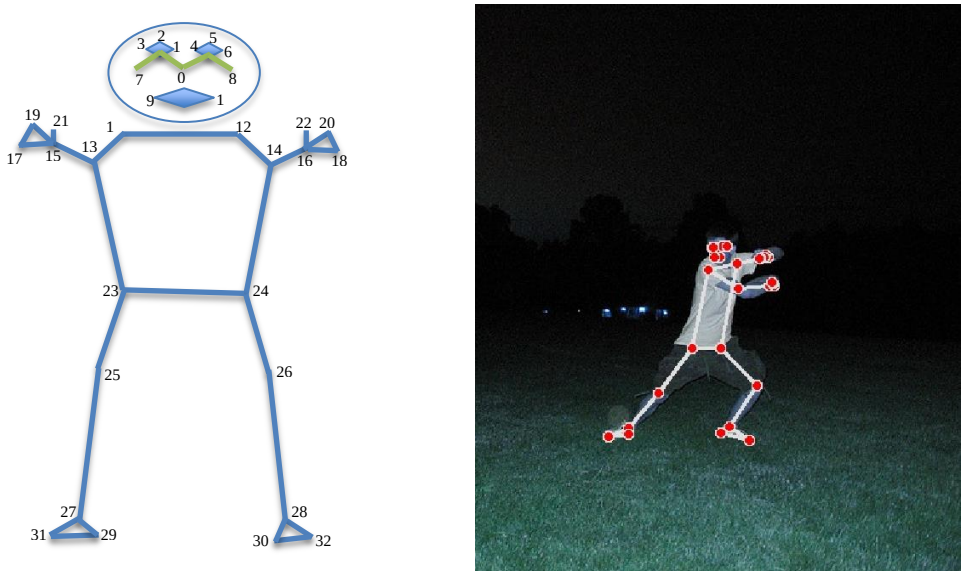

Fig 4. Extracted landmarks results using MediaPipe

IV. EXPERIMENTS OF SKELETON DATA EXTRACTION

This section presents the experimental results of two main aspects: the extraction of skeleton data from the KTH and UTD-MAH datasets using both OpenPose and MediaPipe methods, and the utilization of the extracted skeleton data for training the AcTv2 model. The conclusions are drawn based on the training outcomes of the AcTv2 model. The experiments conducted in this section were performed on a computer equipped with an Intel i5-13600K CPU and an Nvidia 3090 GPU.

Extracting skeleton data from the KTH and UTD-MAH datasets

Within the scope of this study, we collectively chose to extract skeleton data with a consistent number of frames, specifically 30 frames, for each video on both datasets. To evaluate the performance of the feature extraction model for human action recognition, we propose three evaluation criteria: the successful extraction rate of skeleton data from the input data, the accuracy of the extracted skeleton data, and the accuracy of the human action recognition model using the extracted skeleton data. Within the scope of this study, our focus is on investigating criteria 1 and criteria 3. For the KTH dataset: This dataset is characterized by low-quality input video data. The study investigates the feasibility of extracting skeleton data using OpenPose and Mediapipe under limited input data quality conditions (criterion 1). Additionally, it evaluates the effectiveness of the extracted skeleton data when applied to the AcTv2 model for action recognition (criterion 3). For the UTD-MAH dataset: This dataset contains real skeleton data recorded using sensor devices. The research also examines criteria 1 and 3, but in the case of criterion 3, there is a difference in that, apart from using skeleton data extracted from OpenPose and Mediapipe, real skeleton data from the dataset is also utilized in the training of the AcTv2 model. This allows for a comparison of effectiveness and differences in all three scenarios.

The KTH dataset is a widely used collection in various studies related to artificial intelligence and computer vision. This dataset comprises videos capturing human actions, totaling 617 videos across 12 action classes, with an initial size of 1.11GB. In the context of this research, for the extraction method using OpenPose, we extracted data from 25 skeleton joints that describe the human pose in the videos, excluding facial and foot landmarks. Subsequently, we stored the skeleton data as numpy arrays, which serve as input data for the AcTv2 model. Similarly, for the extraction method using Mediapipe, we utilized landmarks numbered from 11 to 32, excluding facial landmarks. The following parameters are used consistently across both Table 2 and Table 3 to evaluate the performance of the feature extraction methods and the AcTv2 model across different datasets and scenarios. Extraction Method parameter refers to the method used to extract skeleton data from the input data. Extraction Ratio from Videos represents the proportion of video data from which skeleton data has been successfully extracted. This measurement is based on criterion 1. Execution Time indicates the amount of time taken for a particular process or operation, typically measured in seconds. Size of Skeleton Data denotes the storage size required for the extracted skeleton data, usually measured in megabytes (MB). The specific experimental results are as follows in Table 3:

Table 3. Evaluation of Skeleton data extraction results on the KTH dataset

Extraction Method	Openpose	Mediapipe
Extraction Ratio from Videos	100%	98.3%
Execution Time (s)	3000	2280
Size of Skeleton Data (MB)	10.5	9.38

Based on the extraction results obtained from the KTH dataset, we observed that Mediapipe has the advantage of faster skeleton data extraction times and smaller storage sizes for the skeleton data. However, when considering the successful extraction ratio, Mediapipe has a lower rate compared to OpenPose. Some videos in the surveyed dataset were unable to extract skeleton data using Mediapipe (10 videos with the percentage is 1.17%). Specifically, this includes 1 video from the "boxing" action class, 1 video from the "jogging" action class, and 8 videos from the "running" action class.

Meanwhile, the UTD - MAH (UTD Multimodal Human Action Dataset) is a valuable resource in the field of human action recognition, developed by the University of Texas at Dallas. This dataset contains videos capturing various human actions, including 861 videos across 27 action classes, with an initial size of 1.12GB, similar to the KTH dataset. However, the number of samples for each action class in UTD - MAH is fewer, while the number of action classes is greater. Additionally, the video quality of the UTD - MAH dataset is superior to that of the KTH dataset. This factor also raises the question of whether Mediapipe will perform better in extracting skeleton data from a dataset with higher video quality. The specific experimental results are as follows in Table 4:

Table 4. Evaluation of Skeleton data extraction results on the UTD - MAH dataset

Extraction Method	Openpose	Mediapipe
Extraction Ratio from Videos	100%	100%
Execution Time (s)	2700	3360
Size of Skeleton Data (MB)	14.7	13.3

Based on the observed results from extracting skeleton data on the UTD - MAH dataset, we noticed that Mediapipe no longer encounters cases where it fails to extract skeleton data when the quality of the input video is improved. Furthermore, another noteworthy point is that the advantage of faster skeleton data extraction time with Mediapipe has diminished. In our assessment, this could be due to OpenPose facing more time-related challenges when attempting to extract skeleton data from low-quality videos. In other words, when selecting a skeleton data extraction solution for low-quality input videos to ensure data completeness, OpenPose is considered more favorable than Mediapipe. When the input video data quality is satisfactory, considering the criterion of time, we rate OpenPose's skeleton data extraction method higher. However, in terms of compactness of the extracted skeleton data, we value the Mediapipe approach more. It's important to note that our evaluation is based solely on the context of recognizing actions of a single individual within the video.

Training Skeleton Data with AcTv2 Modele

Next, we proceed to train the AcTv2 model using the extracted skeleton data as mentioned above. This step aims to re-evaluate the effectiveness of the two skeleton data extraction methods. The specific parameters used for training the AcTv2 model are as follows in Table 5:

Table 5. Hyperparameters used for the training experiment on AcTv2 Model

Trainning	Training epochs	7000 (KTH), 12000 (UTD)
	Batch size	512
	Optimizer	AdamW
	Warmup epochs	40%
Leguralization	Weight decays	1e-4
	Label smoothing	0.1
	Dropout-AcTv2	0.5
	Randomflip	50%
	Randomnoise	0.03

Based on the training results of the AcTv2 model with skeleton data, we have drawn two observations as follows in Table 5 and Table 6. Firstly, for the KTH dataset, when utilizing OpenPose-extracted skeleton data for training the AcTv2 model, we achieved significantly superior performance compared to using skeleton data extracted with Mediapipe, despite the relatively low accuracy rate. This reaffirms our evaluation from section 2 that, for low-quality video data, Mediapipe is not a suitable choice for extracting skeleton data, even though it has a faster processing time.

Table 6. Training results of AcTv2 Model with Skeleton data from KTH dataset Extracted by OpenPose and MediaPipe

	Openpose		**Mediapipe**	
	Accuracy mean	Highest accuracy	Accuracy mean	Highest accuracy
Split data1	82.1 %	88.7 %	70.0 %	77.0 %
Split data2	81.3 %	88.7 %	71.0 %	73.8 %
Split data3	81.3 %	87.1 %	70.0%	75.4%

Table 7. The training results of the AcTv2 model with skeleton data extracted from the UTD-MAH dataset using OpenPose, Mediapipe, and real skeleton data from the UTD-MAH dataset

	Openpose		**Mediapipe**		**Real skeleton data**	
	Accuracy mean	Highest accuracy	Accuracy mean	Highest accuracy	Accuracy mean	Highest accuracy
Split data1	95.3 %	97.7 %	95.2 %	97.7%	98.8 %	99.0 %
Split data2	97.0 %	98.9 %	96.3 %	98.9 %	97.8 %	99.2 %
Split data3	96.1 %	97.7 %	95.7 %	96.6 %	96.9 %	98.1 %

Furthermore, concerning the UTD - MAH dataset, we observed that both techniques for extracting skeleton data deliver highly promising outcomes, as demonstrated by the achieved accuracy when employed in training the AcTv2 model. However, taking into consideration the extraction time for skeleton data from the UTD - MAH dataset, as illustrated in Table 3, and the accuracy in split data3, OpenPose emerges as the preferable choice in this context. Notably, the skeleton data extracted using OpenPose exhibits training results remarkably close to those obtained with the real-world skeleton data from the UTD-MAH dataset. Given the inherently compact size of the skeleton data, the difference in data size is less significant compared to execution time and accuracy when employed for action recognition.

V. CONCLUSION

In this study, we conducted an evaluation of the effectiveness of two methods for extracting skeletal frame data from human action videos using OpenPose and Mediapipe on two datasets: KTH and UTD - MAH. The research findings contribute valuable insights into the applicability and efficiency of these two skeletal frame data extraction tools for human action recognition.

We derived several important observations from the experimental results. For the KTH dataset, we verified that utilizing OpenPose for skeletal frame data extraction yields superior performance compared to Mediapipe. Although Mediapipe generally processes faster, accurate skeletal frame data extraction remains a drawback for it in this context. For the UTD - MAH dataset, both methods produced favorable results with the AcTv2 model. Notably, Mediapipe improved its ability to extract accurate skeletal frame data and no longer encountered issues with data extraction when video quality improved. However, when prioritizing real-time execution and accuracy optimization, OpenPose remains the preferred choice.

The study focuses solely on single-individual action representation in video data. However, in reality, action recognition often occurs within videos featuring multiple individuals performing various actions simultaneously. Open Pose encounters numerous inaccuracies when extracting skeletal frame data from multiple individuals in the same video, as discussed in [23]. On the other hand, Mediapipe was initially designed for single-individual skeletal frame data extraction. When combined with suitable object detection methods to identify individuals within a video, a promising solution for accurate skeletal frame data extraction from videos with multiple individuals can be achieved.

In the future, we hope that this study will help shape and provide valuable insights for selecting the appropriate skeletal frame data extraction method for various scenarios in the field of human action recognition.

REFERENCES

[1] T.H. Thi, et al., "Human action recognition and localization in video using structured learning of local space-time features," *2010 7th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, pp. 204 - 211, IEEE, 2010, DOI 10.1109/AVSS.2010.76.

[2] Z. Sun, Q. Ke, H. Rahmani, M. Bennamoun, G. Wang and J. Liu, "Human Action Recognition From Various Data Modalities: A Review," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1-20, IEEE, 2022, Doi: 10.1109/TPAMI.2022.3183112.

[3] G. Johasson, "Visual perception of biological motion and a model for its analysics," *Perception and psychophysics*, vol. 14, no. 2, pp. 201-211, Springer, 1973.

[4] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," in *International Conference on Learning Representations*, 2021.

[5] Vittorio Mazzia, Simone Angarano, Francesco Salvetti, Federico Angelini, Marcello Chiaberge, "Action Transformer: A Self-Attention Model for Short-Time Pose-Based Human Action Recognition," *Computer Vision and Pattern Recognition*, vol. 124, April 2022, 108487, DOI: 10.1016/ j.patcog.2021.108487.

[6] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, and I. Polosukhin, "Attention is all you need," in I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, vol. 30. Curran Associates, Inc., 2017.

[7] F. Yang, H. Yang, J. Fu, H. Lu, and B. Guo, "Learning texture transformer network for image super-resolution," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5791-5800, June 2020.

[8] H. Zhu, H. Liu, C. Zhu, Z. Deng, and X. Sun, "Learning spatialtemporal deformable networks for unconstrained face alignment and tracking in videos," *Pattern Recognition*, 107:107354, 2020.

[9] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, and H. Jegou, "Training data-efficient image transformers & distillation through attention," in M. Meila and T. Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning, vol. 139 of Proceedings of Machine Learning Research*, pp. 10347-10357. PMLR, 18-24 Jul 2021.

[10] S. D'Ascoli, H. Touvron, M. L. Leavitt, A. S. Morcos, G. Biroli, and L. Sagun, "Convit: Improving vision transformers with soft convolutional inductive biases," in M. Meila and T. Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning, vol. 139 of Proceedings of Machine Learning Research*, pp. 2286-2296. PMLR, 18-24 Jul 2021.

[11] L.L. Presti et al., "3D skeleton-based human action classifcation: A survey," *Pattern Recognition*, vol. 53, pp. 130 -147, ScienceDirect, 2016.

[12] Guannan Chen and Shimin Wei, "Fusion sampling networks for skeleton-based human action recognition," *Journal of Electronic Imaging*, vol. 31, SPIE, 2022, DOI: 10.1117/1.JEI.31.5.053015.

[13] W. K. M. Mithsara, "Comparative Analysis of AI-powered Approaches for Skeleton-based Child and Adult Action Recognition in Multi-person Environment," *2022 International Conference on Computer Science and Software Engineering (CSASE)*, pp. 24-29, IEEE, 2022. DOI: 10.1109/CSASE51777.2022.9759717.

[14] S. Jang, H. Lee, S. Cho, S. Woo and S. Lee, "Ghost Graph Convolutional Network for Skeleton-based Action Recognition," *2021 IEEE International Conference on Consumer Electronics-Asia (ICCE-Asia)*, pp. 1-4, IEEE, 2021, DOI: 10.1109/ICCE-Asia53811.2021.9641919.

[15] B. Shi, L. Wang, Z. Yu, S. Xiang, T. Liu and Y. Fu, Zero-Shot Learning for Skeleton-based Classroom Action Recognition, 2021 International Symposium on Computer Science and Intelligent Controls (ISCSIC), pp. 82-86, IEEE, 2021, doi: 10.1109/ISCSIC54682.2021.00026.

[16] Imran, J., Raman, B, "Evaluating fusion of RGB-D and inertial sensors for multimodal human action recognition," *Journal of Ambient Intell Human Comput*, vol. 11, pp. 189-208, Springer, 2020.

[17] J. Cha, M. Saqlain, D. Kim, S. Lee, S. Lee and S. Baek, "Learning 3D Skeletal Representation From Transformer for Action Recognition," *IEEE Access*, vol. 10, pp. 67541-67550, 2022, DOI: 10.1109/ACCESS.2022.3185058.

[18] Y. Segal and O. Hadar, "Constructing a skeleton database and enriching it using a Generative Adversarial Network (GAN) simulator to assess human movement," *2022 IEEE 38th International Conference on Data Engineering (ICDE)*, pp. 3226-3229, IEEE, 2022, DOI: 10.1109/ICDE53745.2022.00304.

[19] Snoun, A., Jlidi, N., Bouchrika, T. et al., "Towards a deep human activity recognition approach based on video to image transformation with skeleton data," *Multimed Tools Appl*, vol. 80, pp. 29675-29698, Springer, 2021.

[20] G. Cicirelli, T. D'Orazio, "A Low-Cost Video-Based System for Neurodegenerative Disease Detection by Mobility Test Analysis," *Appl. Sci.* 2023, 13, 278. https:// doi.org/10.3390/app13010278

[21] Khac-Anh Phu, Van-Dung Hoang, Van-Tuong-Lan Le, "Action transformer: Model Improvement And Effective Investigation With MPOSE2021 and MSR Action 3D Datasets," *Icon3E* 2023 in 28-29 August 2023, Accepted in 18 July 2023.

[22] Zhe Cao, Gines Hidalgo, Tomas Simon, Shih-En Wei, and Yaser Sheikh, "OpenPose: Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields," *Computer Vision and Pattern Recognition, arXiv:1812.08008*, 2019.

[23] Zhe Cao, Tomas Simon, Shih-En Wei, Yaser Sheikh, "Realtime Multi-Person2D Pose Estimation using Part Affinity Fields," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 7291-7299.

[24] T.Y. Lin, et al., "Coco 2016 keypoint challenge," (2016).

[25] Camillo Lugaresi, Jiuqiang Tang, Hadon Nash, Chris McClanahan, Esha Uboweja, Michael Hays, Fan Zhang, Chuo-Ling Chang, Ming Guang Yong, Juhyun Lee, Wan-Teh Chang, Wei Hua, Manfred Georg and Matthias Grundmann, "Google Reseach, MediaPipe: A Framework for Building Perception Pipelines," *arXiv:1906.08172*, 2019.

ĐÁNH GIÁ HIỆU SUẤT CỦA MEDIAPIPE VÀ OPENPOSE CHO VIỆC TRÍCH XUẤT DỮ LIỆU KHUNG XƯƠNG

Khắc Anh Phu, Văn Dũng Hoang, Văn Tường Lân Lê

TÓM TẮT: *Trong lĩnh vực nhận diện hành động người, dựa vào những ưu điểm đã được chứng minh, việc sử dụng dữ liệu khung xương trích xuất từ các tập dữ liệu khác nhau làm dữ liệu đầu vào cho các mô hình nhận diện hành động đã trở thành một hướng nghiên cứu nổi bật trong những năm gần đây. Trong nghiên cứu này, chúng tôi tập trung vào việc đánh giá hai phương pháp trích xuất dữ liệu khung xương là OpenPose và Mediapipe trên hai tập dữ liệu KTH và UTD-MHAD. Dữ liệu khung xương trích xuất từ cả hai phương pháp này được sử dụng làm dữ liệu đầu vào cho mô hình AcTv2, một phiên bản nâng cấp của mô hình AcT. Qua quá trình huấn luyện và đánh giá trên mô hình AcTv2, chúng tôi đánh giá hiệu suất của cả hai phương pháp trích xuất dữ liệu khung xương trên hai tập dữ liệu cụ thể. Kết quả thực nghiệm của nghiên cứu này đóng góp vào việc hiểu rõ hơn về tính hiệu quả của các phương pháp trích xuất dữ liệu khung xương này trong việc cung cấp dữ liệu hữu ích cho mô hình AcTv2 nhằm nhận diện hành động người trên các tập dữ liệu khác nhau.*

Từ khóa: *Deep learning, Human action regconition, Skeleton data.*