MULTI-VIEWS HUMAN DETECTION AND ACTIVITY RECOGNITION USING 3D SKELETON MODEL

Sandar Win
Faculty of Computing,
University of Information Technology, Yangon
Myanmar
sandarwin@ucsy.edu.mm

Thin Lai Lai Thein
Faculty of Information Science,
University of Computer Studies, Yangon
Myanmar
tllthein@ucsy.edu.mm

Abstract

Skeleton-based human detection and activity recognition are greatly affected by computer vision and are available in real-world applications such as security, the physical condition of human movement, robotics, smart home systems, and visual reality applications. Currently, several approaches have been established by using 3D skeleton data and it is directly taken from sensor devices. That is cost-effective and constraints on different lighting conditions, distance scales, and apparatus devices in real-life applications. To overcome these conditions, the proposed approach is considered on 2D to 3D skeleton model with deep learning approach to recognize more reliable information between skeleton joint sequences and human activity changes over time. Our method is used with OpenPose detector to extract effective 2D key points of body parts and reconstruct 3D skeleton model to recognize occlusion of body parts and more accurately on human activities. The experiment is tested with HMDB 51 dataset and our daily-life data and gets efficient results as 94% for human activity recognition.

***Keywords*: OpenPose detector; 2D to 3D Skeleton model; Joint scale distance; Deep learning.**

1. Introduction

Despite encouraging development in the past decade, human detection and activity recognition are active research and important application areas in computer vision [Pham *et al.* (2022)] and have many problems due to different body scales, occlusion of body parts, influence of environments, appearance of background and recording situations. Sudden changes in the object's motion are challenges in object detection and recognition. Early efforts of human activity recognition systems have used shape-based matching techniques, but later attempts have focused on spatial-temporal appearance. An action can be defined by spatial-temporal sequence of human body parts' movement.

Many of the researchers applied depth maps from depth sensors, the key idea is when a person performs any activity, sensors will capture depth data and then use it to extract skeleton points and process for activity recognition. There is specific correspondence between depth maps and calibration of edges are required and that have limited range and lighting changes in indoor and outdoor application areas. In most activities, joint appearance is very little deviation and brings noise in skeleton configurations. That can reduce the overall performance of different activities. Many recent works used convolutional neural networks to get human visual perception along the extraction of features and focused to detect locations of joints and that is applied to describe the analysis of activity. One main problem for the joint detection model is that the positions of joints are estimated from moving body parts and that leads to missing results on activity recognition.

Advance in the development of the deep learning approach [Sharma *et al.* (2022)], human detection is more reliable, and 3D skeleton view of motion recognition is more effective in the appearance of different human bodies and has great success in human activity. The 3D human skeleton comprises compact information about body joints connected to limbs that support viewpoint changes and can extract relevant information about body parts. Our system is developed with three approaches: the first one is considered on 2D skeleton point of body parts based on OpenPose detector, and second one is 2D to 3D configuration and the third one is human activity recognition on 3D skeleton joint sequences with deep learning approaches. That can focus better perception and recognition of human activity on realistic variation

This research effort is arranged as follows. Section 2, includes related work that supports the method of the development system. Section 3, illustrates and explains the proposed architecture of human detection and activity recognition with the skeletal model. Section 4, describes about deep learning process along with the performance and accuracy of the results. Finally, the conclusion and future work are made in Section 5.

2. Related Work

Human detection and activity recognition from video sequences have a considerable amount of work due to their real-world applications. Many techniques have been stated in the literature, yielding increasingly effective recognition results based on their respective approaches. Most of the early systems with handcrafted features with edges and corners and later developed with learning approaches. The system [Asghari-Esfeden *et al.* (2020)] applied dynamic motion captured from spatial-temporal information of skeleton joint movements. They used Mask R-CNN as a pose extractor to extract body joint key points and classified the activity with CNN. As the experimental result, the combination of dynamic motion with the spatial-temporal convolution method is best achieved mean accuracy as 87.2% on JHMDB, 84.2% on HMDB, and 98.4% on the UCF-101 dataset. [Luvizon *et al.* (2018)] proposed 2D joint and 3D skeleton pose by using a deep multitask convolutional neural network. The recognition process is based on pose estimation and aggregated the visual features. They trained multiple types of datasets to generate 3D predictions from 2D annotated data and proved an efficient way for action recognition based on skeleton information.

The dual-source network developed by [Iqbal *et al.* (2018)], they collected data from independently two integrated training sources of 3D motion data and 2D annotated position and combined the estimated 2D pose with the retrieval of the nearest 3D pose and reconstructed 3D human pose. The system achieved qualitative results on the Human 3.6M dataset and an efficient method for pose estimation error. Authors [Yu *et al.* (2020)] proposed multimodal feature fusion with skeleton and RGB, they utilize a graph convolutional subnetwork for skeleton representation. For the RGB modal, use the spatial-temporal part from the region of interest and apply the attention feature from the skeleton model. In this system, the fixed attention mechanism achieved the performance of the system. In [Khan et al. (2024)], authors applied a feature fusion mechanism with Deep Neural Network (DNN) and multi-view by using pre-trained VGG-19. They combine gradient information which has a high probability consisting of relative entropy, mutual information, and a strong correlation coefficient, and classify the activity using the Naïve Bayes classifier. As a test result, the system achieved high accuracy as 93.7% on HMDB51, 98% on UCF sports, 99.4% on YouTube, 95.2% on IXMAS, and 97% on KTH datasets.

The spatial-temporal image formation (STIF) technique on 3D skeleton-joint sequences in [Tasnim et al. (2021)] used three fusion mechanisms to determine feature distribution. The spatial-temporal representation is obtained along with the temporal changes of joint mapping and line mapping in two consecutive frames. There are many attempts have been made to develop methods in past research fields. However, human action is harder to define under different conditions, and ambiguous input situations remain a considerable factor. The recognition results can be affected by human occlusion with one another, self-occlusion, ambiguity, and so on. In our system, the system is identified to get better performance on occlusion of body part detection and activity recognition.

3. Proposed System and Methodology

This section designates the process and method of the proposed motion detection and activity recognition technique using with video sequences. Our contributions are described as follows:

- Extract 2D key points from the body parts in video sequences
- Reconstruct 3D skeletal joints and define activity recognition for moving objects with deep learning framework
- Searching the missing joint areas and solving similar poses for different activities

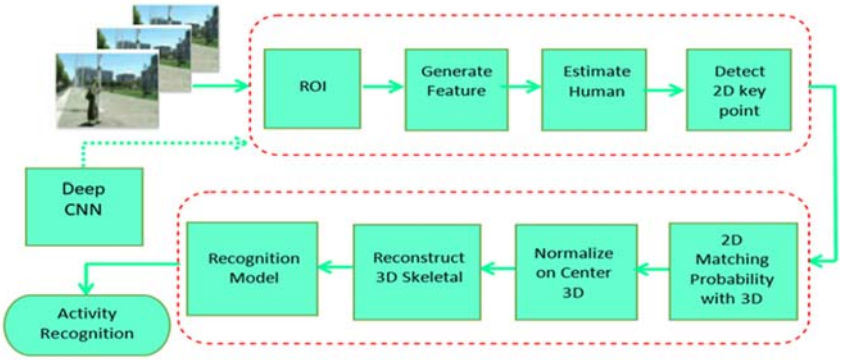

Fig. 1. Block diagram of system model

- Consider on necessary joint with Joint Collection Distance and classify more accurate results on joint sequence activity.

To illuminate the involvement of the proposed method, a block diagram of process flow is described in Fig. 1.

3.1. Extract 2D key points from body parts

In general, a set of body key point information such as shoulder, elbow, wrist, waist, hip, knee, ankle, etc. represents the organization of the human skeleton. Human poses can be affected by occasional missing data which occurs in isolated frames, or persistent data which happens from the entire sequences. Firstly, the proposed system detects regions of interest (ROI) to segment region by region and by finding the ratio between the intersection of the target mask and prediction pixels of human body parts. And extracts geometric features consisting of joint indices and distances and coordinate features of 2D joint locations. The timeline of 2D skeleton point and the segmentation result appear the digitized human motion, and that can track from which paths and trajectories are moving.

3.1.1. Region of Interest Extraction

The region where the segment can be found is contained within the ROI mask. The Intersection-over-Union (IoU) is used to get the performance of object segmentation and to extract ROI [Abuowaida *et al*. (2021)]. If the detection of the target is less than the IoU minimum, that is defined as a reject assessment. The bounding box is computed maximum value between the last detection of an active track and candidate detections in the next frame. The intersection of the target mask and prediction pixels based on the ROI mask is shown in Fig.2.

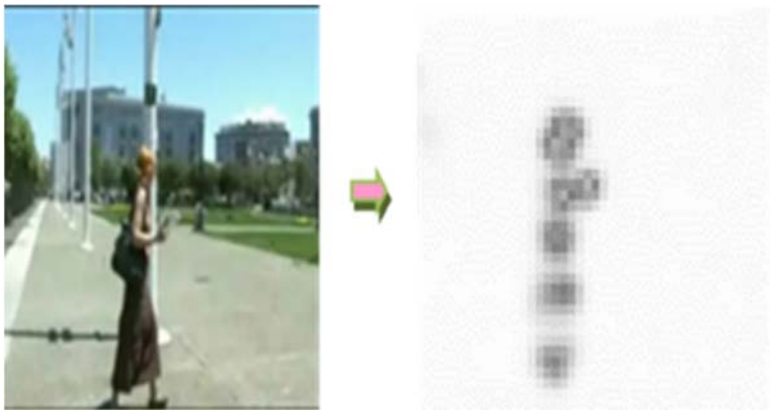

Fig. 2. Example of input video and prediction pixels with ROI mask

3.1.2. Feature generation

To extract effective results for 2D key points from the input sequences, the image is analyzed by a pre-trained convolutional neural network to produce a set of feature maps (F). In the first stage, the function variables are used to extract feature maps that have the highest activated value of the joint locations between corresponding limbs and the associated joints of human body parts. The feature maps are processed on two branches convolutional neural network and continue to produce 2D key points. At the next stage, the prediction of the previous stage along with image features are concatenated and produce more refined predictions. The confidence values indicate the configurations of the skeletal points. The process continues to define the confidence maps and part affinity fields that are concatenated and produce 2D key points of each region in the detected image and generate feature pass through the ROI pooling. Finally, the corresponding mask with fixed-sized feature representation is predicted by the joint location score and position of the bounding box statement.

3.1.3. Human estimation

To estimate the human structure, the human pose skeleton is a more effective type of detecting human shapes and is more robust to occlusions. Since relative positions of human joints may be very diverged concerning their neighboring. At the stage of feature processing, the estimation of the confidence level of each pixel has distinct values of joint score and spatial dependency. To refine the confidence level of joints, by exploiting the spatial configuration of the human body that is defined by the joint detection model. The positions of joints consist of the score maps and joint locations concerning their neighboring, and that is determined with features learning by using neural networks.

In human detection, missing of body parts can be either self-occlusions or other objects in the field of view during the video capture. The system is considered with a non-maximum suppression method to pick each output

prediction bounding box with maximal probability [Hosang *et al*. (2017)]and it can detect partial occlusions and improve the interactions of previously unseen joints feature and prediction pixel acquired along the video sequences.

3.1.4. OpenPose detector

OpenPose is a real-time multi-person 2D pose detector. It can provide human poses represented by 2D coordinates made from a root-centered graph and allows for the recognition of skeletons of multiple persons in the same scene. OpenPose detects key points that belong to the same skeleton of the human body. That obtains two associated joints belonging to each limb of the person. It might be able to provide viewpoints changes and detect all valid joints that have a high confidence score [Fang *et al*. (2022)]. OpenPose network extracts features from an image and feeds them into two parallel branches of convolutional layers. The first branch predicts a key point of confidence maps, each of which represents a specific part of the human skeleton. The second branch predicts a set of Part Affinity Fields (PAFs) that are found in all subsequent stages and which determine the degree of association between different pairs of body parts.

3.1.5. Analyze the spatial-temporal information

Human action recognition can be improved by the analysis of spatial-temporal skeleton joint sequences and define as $J_s(t)$. Each frame consists of body joint sequences as nodes and connected joints over time as edges in a spatial-temporal graph, and generating higher-level feature maps through the operation of convolution [Yang et al. (2018)]. To obtain effective joint sequences from the configuration of body parts, the main idea is to determine the minimum distance across the frame between the target pose and OpenPose detector by "Eq. (1)".

$$\mathrm{D}_{\mathrm{V}_l,\mathrm{J_s}}(\mathrm{t}) = \min_{\mathrm{v}\in\mathrm{V}_l} (\mathrm{p_i}(\mathrm{t}), \mathrm{v}) \tag{1}$$

The meaningful sequences that can be defined with "Eq. (2)".

$$\mathrm{Seq}\ (\mathrm{V}_l) = \left\{ \mathrm{D}_{\mathrm{V}_l,\mathrm{J}_1}(\mathrm{t}), \mathrm{D}_{\mathrm{V}_l,\mathrm{J}_2}(\mathrm{t}), \dots, \mathrm{D}_{\mathrm{V}_l,\mathrm{J}_n}(\mathrm{t}) \right\} \forall l \in \mathcal{L} \tag{2}$$

where the target pose is $\mathrm{p_i(t)}$and V_l be prototypes for action $\mathcal{L}$.

3.2. 2D Matching Probability on 3D Space

For a pair of input 2D poses (x_i, x_j), the proposed system defines the probability distribution $p(m\backslash x_i, x_j)$ in "Eq. (3)" and their corresponding poses as matching of $(\mathrm{y_i}, \mathrm{y_j})$ are visually similar. i.e. $p(m\backslash x_i, x_j) \sim (\mathrm{y_i}, \mathrm{y_j})$. By mapping 2D pose to 3D probabilistic embedding is based on $x \to p(z\backslash x)$.

$$p(\mathrm{m}\backslash x_i, x_j) = \int p\,(m\backslash z_i, z_j) p(z_i\backslash x_i) p(z_j\backslash x_j) dz_i\, dz_j \tag{3}$$

Each distribution is defined with K samples by "Eq. (4)".

$$p(\mathrm{m}\backslash x_i, x_j) \approx \frac{1}{k^2} \sum_{k_1=1}^{K} \sum_{k_2=1}^{K} p(z_i^{(k_1)}, z_j^{(k_2)}) \tag{4}$$

Normalizing 3D pose from 2D skeleton involves the 2D points based on triangulating their projections onto the image by scaling the entire pose and it conforms to a standardized unit of measurement. One common approach is to scale the center 3D pose, so that the distance between specific joints (e.g., shoulders or hips) matches a predefined length. This is ensured by defining the mean position of all joints across frames. Specify the dispersal of the latent variables x_i relied on the model parameter $w$from 2D to 3D. This is the expectation "E-step" $q(z_i)$ and the estimates nearest 3D pose is defined as "Eq. (5)".

$$q(z_i) = p(z_i\backslash x_i, w) = N(z_i\backslash \mu_i, \sum i) \tag{5}$$

Maximize the expected likelihood concerning updating the model parameter w by maximization "M-step" with mean μ_i and variance σ^2 is constructed by using "Eq. (6)".

$$w = [\sum_{i=1}^{n} x_i \mu_i^T][\sigma^2 I + \sum_{i=1}^{n} (\mu_i \mu_i^T + \sum i)]^{-1} \tag{6}$$

Iterate between the E-step and M-step until convergence to describe point estimation and posterior distribution on each z_i. By using normalization on image gradient, can predict new locations of features and be more robust for tracking.

3.3. 3D Point Reconstruction

3D point reconstruction focuses on the development of 2D joint information that can be more effective in recognizing scenes and objects. That consists of projecting the 3D position of key point transformation from the 2D image. The probability of 2D heat maps from perceived 2D joints and trained by processing of convolutional neural networks and lifting the semantic pose with human order in 3D space as shown in Fig. 3.

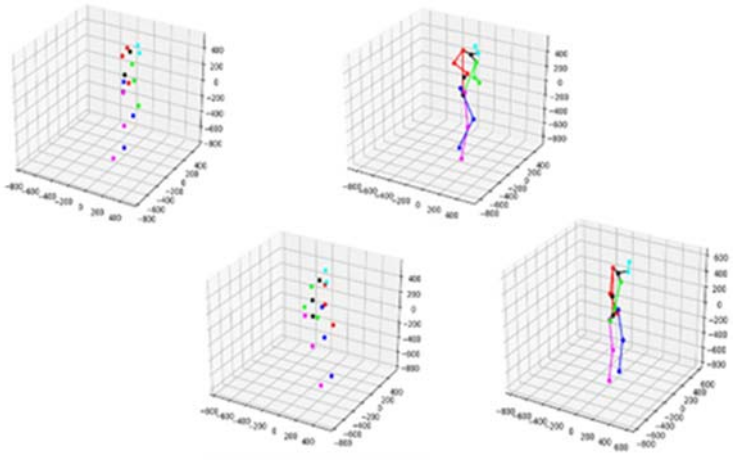

Fig. 3. Joint sequences extraction with human order and reconstruction on 3D space

During this process, the relative positions of the limbs are used to control the motion model and the visible points in 3D space are the projections of the real points on the image. The 3D point in space is determined as "Eq. (7)".

$$X_i = [x_i \; y_i \; z_i]^T \tag{7}$$

$$X = [X_1^T \ldots X_s^T] \approx \theta_1\beta_1 + \cdots + \theta_{3k}\beta_{3k} = \theta\beta \tag{8}$$

where $x_i, y_i, \; z_i \; as$ component vectors are along the X-axis, Y-axis, Z-axis, and T as the transpose operation. X represents the 3D point in space and can be approximated by the linear combination of coefficients and features vectors $\theta\beta$in "Eq. (8)". The reconstruction of the human skeleton joint from 2D to 3D space along the order sequences of human motion is shown in Fig. 4.

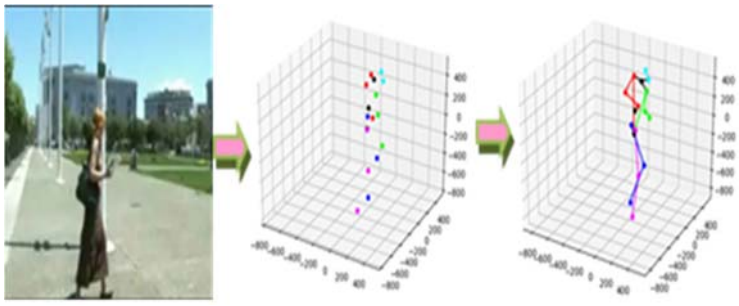

Fig. 4. Input 2D image and skeleton joint on 3D space

3.4. Activity Recognition

For skeleton-based action recognition, geometric features and point-of-coordinate features such as joint indices and location distances are important to get accurate results for the deep learning process. The geometric features are invariant from a location viewpoint but not stable from one data to another. Joint indies (i.e., the IDs of defining key points which consist of shoulders, elbow, wrist, etc.) can be dynamically changed in various activities. Hence, recognition errors arise for requiring the correlation of joints to be predefined by the ordering of their indices.

Using Joint Collection Distances (JCD) with Euclidean Distance matrix along the frame sequence k is defined by $s_k = \{J_1^k, J_2^k, \ldots, J_N^k\}$ and that can solve for missing joint errors and can improve the recognition result. The system generates a transformation process from 2D to 3D pose with joint sequences to make the skeleton more realistic and to represent human activity. Since each joint is associated with one or more other joints in the same skeleton neighborhood and that appears to form the order of distribution [Ci *et al.* (2019)]. The encoded structure network and model representation ability of weight matrix, which are dependent features and activity recognition is observed by "Eq. (9)".

$$\mathcal{L}_{activity} = \sum_I \sum_{t=1}^{T} \sum_{k=1}^{K_1 \times K_2} (N_t^I(k) - \alpha_t^I(k))^2 \quad (9)$$

where $T$is total time steps, k is spatial location, N_t^I is the observation of high visual data for all joints, α_t^I and is joint attention score.

4. Deep Learning Process

Deep learning techniques are valuable in activity recognition due to automatically learning hierarchical representations from raw data and accurate results. The architecture of the deep learning process consists of the input layer, hidden layers, fully connected layer, and output layer [Shrestha and Mahmood (2019)]. In the input layer, the normalization process is used to reduce the impact of scale differences and to assist the optimization process. Each hidden layer comprises three main tasks, such as convolution, max pooling, and neuron activation with Rectified Linear Unit (ReLU) to speed up the convergence and to prevent the vanishing of the gradient.

In the proposed system, activations are to be zero-mean and unit standard deviation, which enables faster convergence, accelerates training, and reduces the generalization error. The system performs the randomly initialized weight vectors and the network is trained forward and backward multiple times until it meets minimum loss. All weight values are updated based on loss value using the backpropagation algorithm with the gradient descent method. The system predicts the desired format along the network and 3D skeleton information of the human pose is interpreted with the distribution of joint features by fully connected layers of deep neural network. At the output layer, the probability of each possible action is classified by using the soft-max function and recognizes the human activity.

4.1. Training on Network

Advance in the progress of learning methods, human detection is more accurate and the 3D skeleton view of motion recognition is more effective in the appearance of information. In the training process, Visual Geometry Group (VGG) networks [Vaghela *et al.* (2023)] are used to perform easy scalability and better feature processing on additional layers. The network consists of 3x3 convolution with stride 1 and followed by 2x2 max-pooling with stride 2, and ReLU activation function controls better gradient flow in convolution which provides the dense prediction for all joints. Each layer is composed of a pre-trained set of weights and adjust hyperparameters to improve the model's accuracy and performance metrics are measured based on the validation set. The system generates a heat map to recognize the likelihood per pixel for joint sequence localization and precisely captures the hidden spatial dependency with the achievement of the desired result in the deep neural network.

4.2. Experimental Result and Discussion

The visualized and experimental results can describe the effectiveness of the system. We tested experiments with the HMDB51 dataset [Kuehne *et al.* (2011)] which consists of the full complexity around of 7000 realistic video clips from movies, Google videos, YouTube, and public databases. For the training and testing phase, the proposed system is considered for model accuracy with the four state-of-the-art optimization algorithms [Soydaner (2020)] such as Adamax, Nada, Adam, and RMSprop are used and shown in Fig. 5. To improve a good performance, the

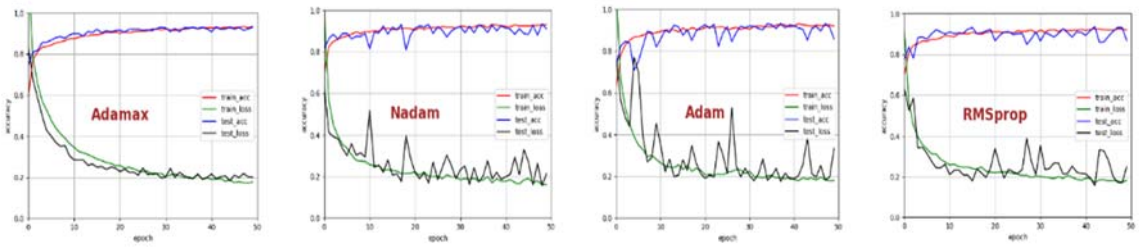

Fig. 5. Training and testing with accuracy and loss on four optimization algorithms

system accuracy with fine-tuning hyper parameter by taking different optimization algorithms are used for the learning process. We split the test size dataset as 80% for training and 20% for testing, repeating 50 epochs with a batch size of 32 and a learning rate is 0.001. As a result, Adamax achieves the best results up to 94.5% on training accuracy and 93.2% on testing accuracy among all. AdaMax is a modification to the Adam version of gradient descent, and that adjusts to accelerate optimization and leads to a more effective and efficient accuracy in the learning process.

To know the efficient method of classification and to define overall performance, the proposed system tested the experiments including each of the 20 participants around the environment with different activities on multi-view. The video clip is recorded frame width and height as 1920x1080p @ 30 fps with an average of 150 frames per video and contains the forward, backward, and side views for different actions. Each activity was repeated 10 times from different views in each video stream. The experiment result is described in Fig. 6.

The system is implemented on Intel Core i7 with NVIDIA GeForce GTX 1080 using Python 3.6 with Tensorflow framework. To identify the performance of our proposed method, focus on different views of human activities under real-life environments, and we measure in terms of Precision, Recall, F1-Sore, and Accuracy percentages are described in Table 1, and performance results of forward, backward, lateral and frontal are illustrated in Fig. 7. The overall accuracy of multi-views is 94% and our system outperforms satisfactory result for human detection and activity recognition using 3D skeleton model.

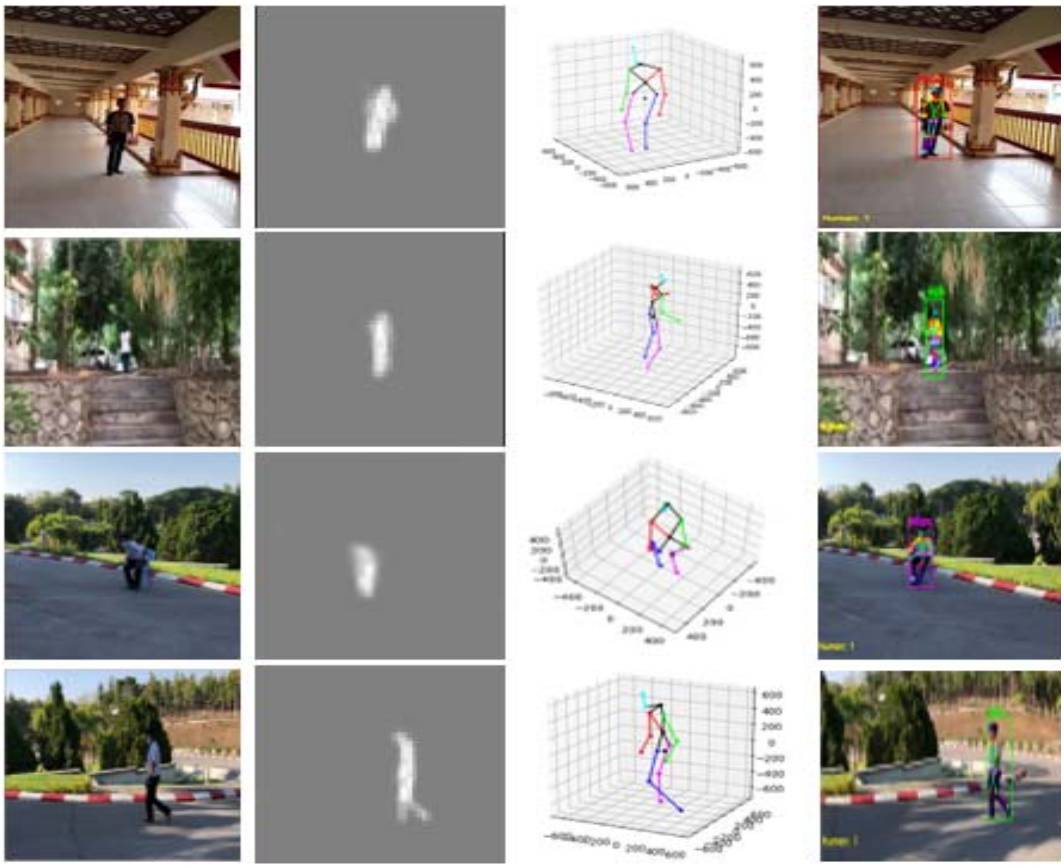

Fig. 6. Experiment of human detection and activity recognition in a real-life environment

Multi Views	Precision	Recall	F1-Score	Accuracy
forward	**94.118**	**88.889**	**91.429**	**89.474**
backward	**95.000**	**100.000**	**97.436**	**95.000**
lateral	**100.000**	**95.000**	**97.436**	**95.000**
frontal	**94.444**	**82.353**	**94.436**	**95.000**

Table 1. Performance metric for different actions with multi views

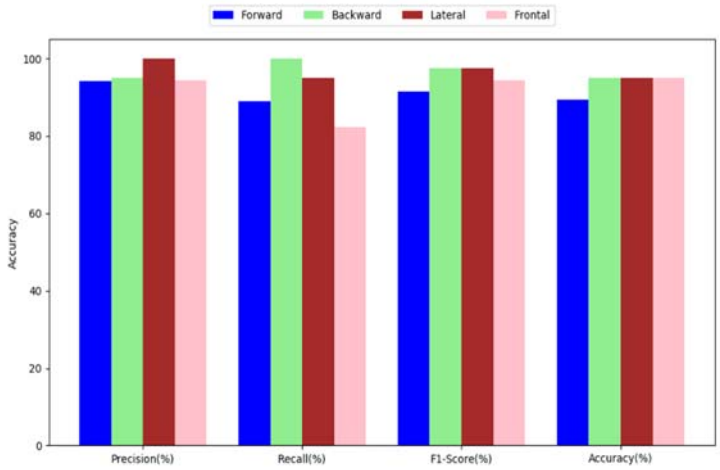

Fig. 7. Performance results on multi-views

5. Conclusion

The system was achieved and improved to produce a high confidence score of 2D key points by OpenPose detector and reconstructs 3D skeleton model to recognize occlusion of body parts and more accurate results on human activities changes over time. The proposed system recognizes the activities by using the deep learning method and the experimental results can prove the good results for daily life environment. The experiment concludes that the proposed method outperforms human detection and activity recognition in a real-life environment. As a result, skeleton model with deep learning supports high accuracy recognition of human movement and extracts valuable information about skeletal joints and activity.

Future direction will continue to define the state of overlapping areas with pre-trained models and configure more joint sequences for human activities. The research work will study multiple human activities in 3D space through the development of a recognition system for real-life environments.

Acknowledgments

This work was supported by Image Processing Lab, University of Computer Studies, Yangon, Myanmar. I would like gratitude to people who have participated in performances in videos with real-life activities through the development of our system adequately.

Conflicts of Interest

The authors have no conflicts of interest to declare.

References

[1] Abuowaida, S. F. A.; Chan, H. Y.; Alshdaifat, N. F. F.; Abualigah, L. (2021): A novel instance segmentation algorithm based on improved deep learning algorithm for multi-object images, 7(1), pp. 10-5455.
[2] Asghari-Esfeden, S.; Sznaier, M.; Camps, O. (2020): Dynamic motion representation for human action recognition, pp. 557-566.
[3] Ci, H.; Wang, C.; Ma, X.; Wang, Y. (2019): Optimizing network structure for 3d human pose estimation. pp. 2262-2271.
[4] Fang, H. S.; Li, J.; Tang, H.; Xu, C.; Zhu, H.; Xiu, Y.; Lu, C. (2022): Alphapose: Whole-body regional multi-person pose estimation and tracking in real-time. IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 45, pp. 7157-7173.
[5] Hosang, J.; Benenson, R.; Schiele, B. (2017): Learning non-maximum suppression. In: Proc. of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4507-4515.
[6] Iqbal, U.; Doering, A.; Yasin, H.; Krüger, B.; Weber, A.; Gall, J. (2018): A Dual-Source Approach for 3D Human Pose Estimation from a Single Image. International Journal of Computer Vision and Image Understanding, vol. 172, 37-49.
[7] Khan, M. A.; Javed, K.; Khan, S. A.; Saba, T.; Habib, U.; Khan, J. A.; Abbasi, A. A. (2024): Human action recognition using fusion of multiview and deep features: an application to video surveillance. Multimedia tools and applications, 83(5), pp. 14885-14911.
[8] Kuehne, H.; Jhuang, H.; Garrote, E.; Poggio, T.; Serre, T. (2011): HMDB: a large video database for human motion recognition. Proceeding of the IEEE Conference on Computer Vision, Barcelona, Spain, pp. 2556-2563.
[9] Luvizon, D. C.; Picard, D.; Tabia, H; (2018): 2d/3d pose estimation and action recognition using multitask deep learning. pp. 5137-5146.
[10] Pham, H. H.; Khoudour, L.; Crouzil, A.; Zegers, P.; Velastin, S. A. (2022): Video-based human action recognition using deep learning: a review. arXiv preprint arXiv: pp. 2208.03775.
[11] Sharma, V.; Gupta, M.; Pandey, A. K.; Mishra, D.; Kumar, A. (2022): A review of deep learning-based human activity recognition on benchmark video datasets. Applied Artificial Intelligence, 36(1), 2093705.

[12] Shrestha, A.; Mahmood, A. (2019): Review of deep learning algorithms and architectures. International Journal of IEEE access, vol. 7, pp. 53040-53065.
[13] Soydaner, D. (2020): A comparison of optimization algorithms for deep learning. International Journal of Pattern Recognition and Artificial Intelligence, 34(13), pp. 2052013.
[14] Tasnim, N.; Islam, M. K.; Baek, J. H. (2021): Deep learning based human activity recognition using spatio-temporal image formation of skeleton joints. Applied Sciences, 11(6), 2675.
[15] Vaghela, R.; Labana, D.; Modi, K. (2023): Efficient I3D-VGG19-based architecture for human activity recognition", The Scientific Temper, 14(4), pp.1185-1191.
[16] Yan, S.; Xiong, Y.; Lin, D. (2018): Spatial temporal graph convolutional networks for skeleton-based action recognition. In: Proc. of the Association for the Advancement of Artificial Intelligence, 32(1), pp. 7444-7452.
[17] Yu, B. X.; Liu, Y.; Chan, K. C. (2020): Skeleton focused human activity recognition in rgb video. arXiv preprint arXiv:2004.13979.

Authors Profile

Sandar Win received the M.I.Sc. degree from the University of Computer Studies, Mandalay, Myanmar in 2003. She is currently working as an Associate Professor at the University of Information Technology, Yangon, Myanmar. Her research interests are Image Processing, Computing, Machine Learning, and Deep Learning. She can be contacted at email: sandarwin@ucsy.edu.mm.

Dr. Thin Lai Lai Thein got Ph.D (IT) from the University of Computer Studies, Yangon, Myanmar. Currently, she is working as a professor of Data Analytics Lab at the University of Computer Studies, Yangon, Myanmar. She has been supervising Master's thesis and Ph.D. research on Image Processing and Geographic Information Systems. She can be contacted at email: tllthein@ucsy.edu.mm.