Received: November 1, 2024. Revised: November 29, 2024. 754

International Journal of
Intelligent Engineering & Systems

http://www.inass.org/

Enhancing Surveillance Vision-Based Human Action Recognition Using Skeleton Joint Swing and Angle Feature and Modified AlexNet-LSTM

Riky Tri Yunardi[1,2] **Tri Arief Sardjono**[1,3] **Ronny Mardiyanto**[1]*

[1]*Electrical Engineering Department, Institut Teknologi Sepuluh Nopember, Surabaya, 60111, Indonesia*
[2]*Instrumentation and Control Engineering Technology, Faculty of Vocational, Universitas Airlangga, Surabaya, 60115, Indonesia*
[3]*Biomedical Engineering Department, Institut Teknologi Sepuluh Nopember, Surabaya, 60111, Indonesia*
* Corresponding author's Email: ronny@elect-eng.its.ac.id

Abstract: Human action recognition (HAR) identifies and classifies human activities by analyzing activity patterns, with feature extraction and classification accuracy being key challenges. In the field of surveillance vision-based, it requires the ability to accurately detect suspicious human activities to provide public safety. In the classifier model for action recognition, the steps start from feature extraction to classification. The classification step uses recurrent neural network architectures such as LSTM to handle sequential data. However, this approach struggles to process spatial information in video data, necessitating the need for a model to learn spatiotemporal patterns from feature data. To address these issues, this study proposes a novel method for classifying activities based on the pattern of 2D skeleton joint swing and angle features for each activity. Additionally, it introduces a novel modified AlexNet architecture with two LSTM layers, called AlexNet-2LSTM, to improve the accuracy of human activity classification. In the performance experiments, the proposed method was evaluated on the KTH and Weizmann datasets, both of which include videos of several people performing different actions. Moreover, to demonstrate the accuracy of the proposed classifier model, it was compared against other state-of-the-art (SOTA) deep learning classifiers, namely Optimized-LSTM, Triple Parallel LSTM, Hybrid CNN-LSTM, LCSWnet, and CC-LSTM-CNN, which the AlexNet-2LSTM achieved precision of 0.95, recall of 0.95, F1-score of 0.94, and accuracy of 0.96on the KTH dataset. Besides that, on the Weizmann dataset 0.95, 0.94, 0.94, and 0.93, respectively. These achievements highlight the proposed model contribution to improving feature extraction and classification accuracy in vision-based HAR systems.

Keywords: Vision-based human action recognition, Skeleton joint, Deep learning classifier, Long short-term memory.

1. Introduction

Human action recognition (HAR) is the process of identifying and classifying human activities or actions based on sensor data. It involves analyzing various types of data, such as visual sensors, motion sensors, and wearable devices, to discern patterns of activities [1]. HAR is widely employed in the field of computer vision for activity detection and analysis, including applications in surveillance and security systems [2], health rehabilitation monitoring [3], and video game controllers [4]. Surveillance systems use HAR to identify specific activities by detecting movement patterns in individuals. The motivation for implementing deep learning in surveillance vision-based systems to detect suspicious human activities is to enhance public safety. Typically, the HAR process begins with feature extraction from video input data, gathering crucial information that feeds into a classifier to identify the action [5, 6]. In the surveillance applications, human action recognition is beneficial for identifying threatening actions. The main challenge for HAR systems lies in the feature extraction methods and classifier models required to achieve high accuracy in identifying and classifying human actions.

Vision-based techniques focus on collecting image sequences using visual cameras to capture information about human motion. Previous studies [7,

8] have used multiple cameras to detect joint positions based on body markers. However, placing body markers on a subject's body makes this method unsuitable for surveillance systems. HAR systems often utilize image or video data captured to analyze body movements [9]. These studies frequently incorporate both RGB and depth information for more accurate body tracking based on skeleton models [10, 11]. They track the skeleton joints from video sequences and calculate vector features for each subject's movements. However, the use of depth cameras remains limited due to their cost. One study explored using a simple RGB camera to extract features from human skeleton motion data [12, 13]. Human motion refers to the movement of key body points used to represent the 2D image of body joints through pose estimation. Common pose estimation algorithms, such as OpenPose [14] and MediaPipe [15], are used to determine these joint positions in the human skeleton. These detected joints are then extracted to derive useful features for recognizing the actions of subjects in the image sequences, converting them into time-series data.

In action classification, HAR systems use deep learning to identify specific actions or activities based on joint position features. Recent work [11] has leveraged deep learning to classify human activities using deep convolutional neural networks with spatial features on image data, although the spatial component still limits its effectiveness. One of the deep learning algorithms for processing human motion time-series data is one-dimensional convolutional neural networks (1D-CNN) [16]. Using supervised learning methods, CNNs classify human activities such as walking, walking upstairs, walking downstairs, sitting, standing, and lying [17]. Long Short-Term Memory (LSTM) networks, widely used for time-series data, improve accuracy and reduce model complexity [18, 19]. The LSTM was implemented and trained with time-series data of joint positions to estimate walking patterns using markerless capture instead of whole images [20]. Integrating CNNs and LSTMs with temporal features can significantly enhance classification performance [21]. Although combining deep CNN and LSTM models requires higher computational costs [22, 23], this model has demonstrated satisfactory accuracy in classifying human activities. Technically, the choice of feature extraction methods and classifier models is crucial because it affects the accuracy of action recognition, and differentiating between various human activities remains a challenge.

In this study, a surveillance vision-based human action recognition system is proposed, utilizing skeleton joint swing and angle from 2D images, combined with modified deep convolutional neural networks and Long Short-Term Memory. The main contributions of this study are as follows:

- A feature extraction method for a human action recognition system is proposed, based on 2D skeleton joint positions, including the shoulder, wrist, hip, and knee. Pose estimation is used to detect these positions. Joint positions are extracted from a sequence of images to determine joint swing and angles, which serve as features for identifying specific activities performed by the subject.
- Based on the deep learning classifier, a novel deep convolutional neural network is proposed, incorporating a modified AlexNet model combined with Long Short-Term Memory to improve the accuracy of activity classification.
- The model is evaluated against other state-of-the-art deep learning classifiers to assess its accuracy. This classifier achieves the highest accuracy score, contributing to the enhancement of the human activity classification system.

This paper is organized as follows: In Section 2, related works on feature extraction and classifiers in activity recognition are reviewed. In Section 3, the proposed method model applied in this study are discussed. In Section 4, the proposed method is evaluated against other state-of-the-art deep learning classifiers to assess their accuracy. Finally, Section 5 provides conclusions and outline future work.

2. Related works

To identify specific actions or activities based on joint position features, Mundt *et al* [24] suggested using multiple calibrated 2D cameras to estimate human body poses and perform 3D tracking, enabling the calculation of human leg kinematics. Meanwhile, Yoo and Nixon [25] estimated body joint positions using 2D skeletal models with six lower body joints from each image, focusing on the hip, knee, and ankle joints as features. Jun *et al* [26] extracted hip, knee, and ankle angles to track leg movements derived from video sequences for detecting abnormal human movement. Incorporating joint position and angle feature extraction from skeleton data is essential for enhancing recognition performance. However, their studies focus on the lower body and leg joints.

In analyzing elderly actions, the system utilizes specific joint coordinates to monitor movements. The activity of crucial joints is referred to as motion analysis. The system models these movements through a series of joint positions along the x, y, and z axes. Oikonomou *et al* [27] explored this approach

Table 1. Previous studies problems related to our contribution

Studies	Problems	Our Contribution
Mundt *et al*. [24], Yoo and Nixon [25], Jun *et al* [26]	Identifying actions with human body pose and leg kinematics. Estimating body joint positions focusing on hip, knee, and ankle joints. Extracting the patterns of these joints movement to detect human motion.	Analyze the movement of body joint positions in a 2D framework, including not only the leg, hip knee, and ankle joints but also the shoulder and wrist.
Oikonomou *et al* [27],	Action analysis refers to the movement of key joints, and this approach uses SVM and CNN to recognize actions based on a subset of joint positions.	Add joint swing and angle features based on the movement position of joints to improve data validation for specific activities.
Tasnim *et al* [28], Challa *et al* [29], Shayestegan *et al* [30]	Using skeletal joint information along the x, y, and z axes is based on capturing temporal changes of video frames for action recognition. Action recognition utilizing temporal data changes through CNN-based and LSTM-based approaches.	Using temporal data of joint swing and angle features, propose a deep learning classifier for activity classification that draws inspiration from the AlexNet model and combines two LSTM layers.
Zheng *et al* [31], Cao *et al* [32], Alharthi and Ozanyan [33], Gao *et al* [34], Lis *et al* [35],	CNN, deep CNN, or LSTM can extract spatiotemporal features from skeleton angle data for human motion pattern analysis, but it often results in the loss of structural information and reduce accuracy. Combining CNN and LSTM can address this issue, but their architecture is complex.	Evaluate and demonstrate the proposed classifier architecture against other state-of-the-art deep learning classifier architectures to assess its classification performance using the proposed features.

by employing classifiers to assess the effectiveness of recognizing actions based on a subset of joints, comparing the performance of support vector machines (SVM) and convolutional neural networks (CNN) in improving recognition accuracy. The disadvantages of recognizing actions solely based on joint positions could include reduced accuracy and unreliability in data validation.

On the other hand, Tasnim *et al* [28] proposed an effective method for action recognition using 3D skeleton joint information. Their approach analyzes 3D skeleton data along the X, Y, and Z axes and captures temporal changes by subtracting consecutive frames. They designed a Deep CNN and evaluated prior models such as ResNet18 and MobileNetV2 for detection and classification. However, in CNN-based methods, the loss of temporal information is unavoidable when using sequential data. To overcome the challenges of sequence data and improve performance, various studies have proposed LSTM-based methods, such as Optimized-LSTM and Triple Parallel LSTM [29, 30]. The implementation of LSTM learning methods has achieved promising performance for action recognition. In their studies, they employ CNN-based and LSTM-based approaches for action classification using temporal data, respectively. So that, using an approach that combines CNN and LSTM to process temporal features can improve it.

Current methods mainly use CNN and LSTM to extract spatio-temporal features from skeleton data but often lose critical structural information. Zheng *et al* [31] highlighted that with CNN-LSTM models, both spatial and temporal by combining a CNN for spatial features and a multi-layer LSTM for temporal features. For instance, a combination of deep CNN and LSTM was evaluated for human movement analysis, such as ResNet and LSTM [32] to detect the movement phase, and Hybrid CNN and LSTM [33] for gait speed classification. In contrast, LSTM and deep CNN were used for the classification of movement patterns using angle information on the LCSWnet architecture [34] and to identify different users' movement data based on cross-correlation (CC) and the LSTM-CNN classification model [35]. While the studies present satisfactory results with high accuracy, there are also some limitations to consider, such as the complexity of the architecture used in the classifier model.

Based on the shortcomings of the studies mentioned above, this study proposes a method for human action recognition, which includes human detection and tracking, joint motion feature extraction, modification of deep CNN and LSTM models, and classification evaluation. The video sequences are processed using pose estimation with human body detection and tracking within a skeleton model, which includes the coordinates of each body joint. From the joint positions, features are extracted by calculating the swing and angle of the joints as temporal data. The feature data are then input into the human activity recognition. Finally, the proposed model is evaluated against other SOTA classifiers to compare their accuracy. Table 1 describes previous studies that motivated the contributions of this study for further work.

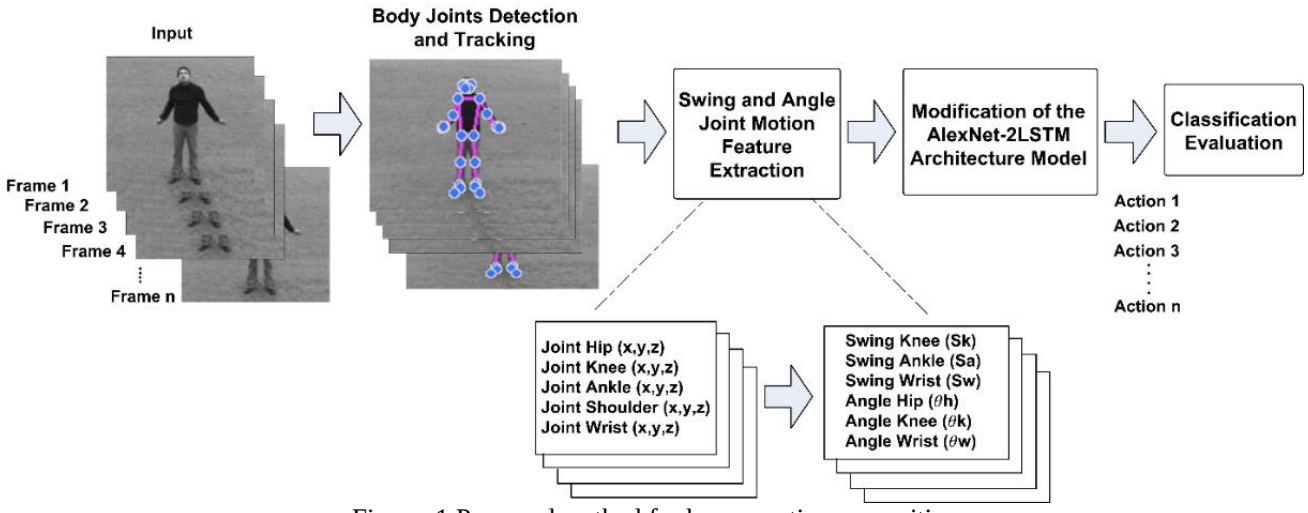


Figure. 1 Proposed method for human action recognition

3. Proposed method

The goal of this study is to improve classification accuracy in recognizing actions from video images by using joint swing and angle features along with a deep learning classifier to enhance a vision-based HAR system. To achieve this, the method is divided into four steps. The first step involves detecting humans and tracking the positions of body joints. The second step focuses on extracting joint positions to determine joint movement and angles. The next step is to build a classifier model based on a modified AlexNet model combined with LSTM. The final step involves evaluating the classification.

The proposed method for human action recognition is shown in Fig. 1. The input consists of a sequence of videos obtained from the KTH and Weizmann datasets featuring a subject performing a specific activity. The motion of the activity is represented using a human skeleton model through pose estimation. Pose estimation provides the coordinate data for each body joint to extract its features. For classification, the feature data is used as input for the proposed AlexNet-2LSTM classifier model. During the experiments, the accuracy of the proposed model is tested and compared with that of other classifiers.

3.1 Body joints detection and tracking

Human detection and tracking using pose estimation begins by capturing each frame from a video dataset [6]. The KTH dataset consists of recordings of grayscale videos at a resolution of 160×120 pixels, including six classes of activities performed by 25 subjects. Moreover, the Weizmann dataset is used for the detection of body joints, consisting of videos at a resolution of 180×44 pixels and including 10 classes performed by nine subjects. Fig. 2 shows the detection of body joints for various 2D actions from video images.

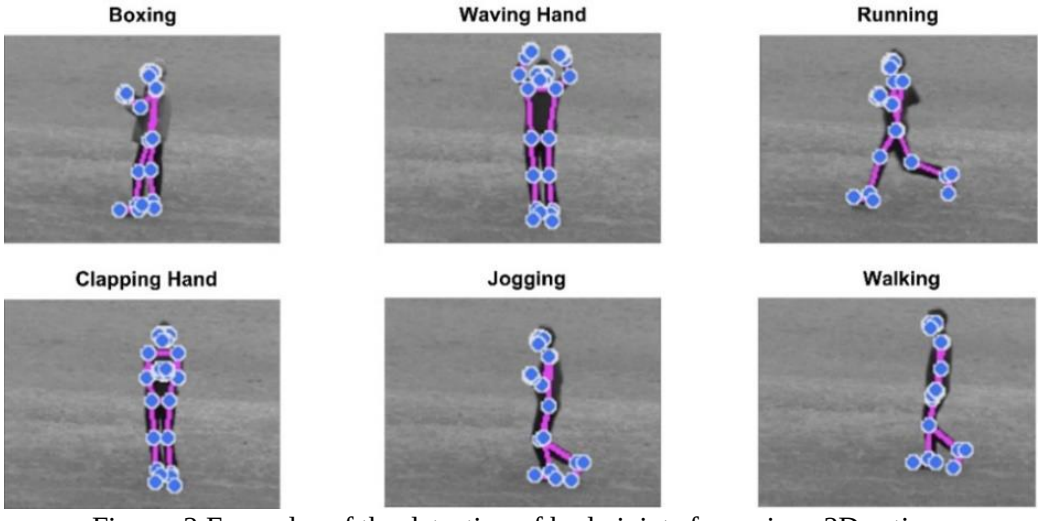
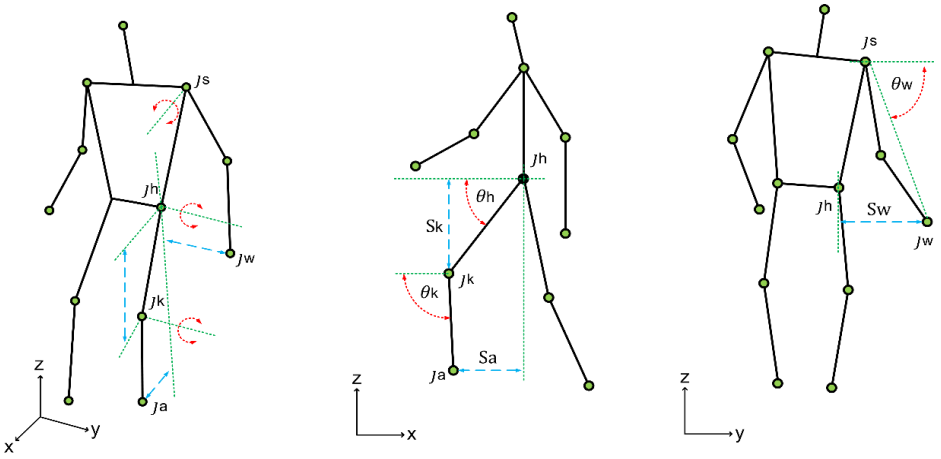


Figure. 2 Examples of the detection of body joints for various 2D actions



(a) (b) (c)

Figure. 3 The selected joint points for features on the skeleton model

3.2 Swing and angle joint motion feature extraction

When a person performs an action, body parts move and experience changes in position, especially at the joints. These changes in joint positions allow us to identify an action by comparing it to its position pattern during the activity. Temporal position data describes the activity's characteristics for the selected joints. Each joint point J consists of joint position coordinate data on the skeleton, where P is the total number of joint points, and $P = J_1, J_2, \dots, J_P$. At each point J_P, the coordinates consist of the x-axis, y-axis, and z-axis values (J_{Px}, J_{Py}, J_{Pz}). Fig. 3(a) shows the five joint points selected in this study: the hip (J_h), knee (J_k), ankle (J_a), shoulder (J_s), and wrist (J_w) joints. The selection of these joints is crucial for tracking the overall movement of the body's joints. By utilizing these reference points, the movement patterns of the human body during activities can be effectively observed. Joint features are extracted from each movement using the reference joint points, which are represented as vectors.

Joint swing refers to the distance between the joint reference point and the hip joint. To obtain the swing distances for the knee $(|S_k|)$ and ankle $(|S_a|)$, as shown in Fig. 3(b), refer to Eq. (1) and Eq. (2). The distance of the knee swing S_k measured from point J_h to J_k on the z-axis, while S_a is the distance of the ankle swing measured from J_k to J_a on the x-axis. Both S_k and S_a are absolute values.

$$|S_k| = \sqrt{\left(J_{h_z} - J_{k_z}\right)^2} \tag{1}$$

$$|S_a| = \sqrt{\left(J_{h_x} - J_{a_x}\right)^2} \tag{2}$$

Referring to Fig. 3(b), the hip joint angle (θ_h) between the vectors $\overrightarrow{J_h}$ and $\overrightarrow{J_k}$ can be calculated using Eq. (3).

$$\theta_h = \cos^{-1}\left(\frac{\overrightarrow{J_h} \cdot \overrightarrow{J_k}}{|\overrightarrow{J_h}| \cdot |\overrightarrow{J_k}|}\right) \tag{3}$$

Where,

$$\overrightarrow{J_h} \cdot \overrightarrow{J_k} = J_{h_x} J_{k_x} + J_{h_z} J_{k_z} \tag{4}$$

$$|\overrightarrow{J_h}| \cdot |\overrightarrow{J_k}| = \sqrt{{J_{h_x}}^2 + {J_{h_z}}^2} \cdot \sqrt{{J_{k_x}}^2 + {J_{k_z}}^2} \tag{5}$$

Then, the ankle swing (θ_k), which involves the vectors $\overrightarrow{J_k}$ and $\overrightarrow{J_a}$, is calculated using Eq. (6).

$$\theta_k = \cos^{-1}\left(\frac{\overrightarrow{J_k} \cdot \overrightarrow{J_a}}{|\overrightarrow{J_k}| \cdot |\overrightarrow{J_a}|}\right) \tag{6}$$

$$\overrightarrow{J_k} \cdot \overrightarrow{J_a} = J_{k_x} J_{a_x} + J_{k_z} J_{a_z} \tag{7}$$

$$|\overrightarrow{J_k}| \cdot |\overrightarrow{J_a}| = \sqrt{{J_{k_x}}^2 + {J_{k_z}}^2} \cdot \sqrt{{J_{a_x}}^2 + {J_{a_z}}^2} \tag{8}$$

Using the same calculations, the wrist swing $(|S_w|)$ on the y-axis, as shown in Fig. 3(c), can be computed using Eq. (9).

$$|S_w| = \sqrt{\left(J_{h_y} - J_{w_y}\right)^2} \tag{9}$$

The feature of ankle swing $(\theta_s)$is obtained using the vectors $\overrightarrow{J_s}$ and $\overrightarrow{J_w}$ as specified in Eq. (10).

$$\theta_s = \cos^{-1}\left(\frac{\overrightarrow{J_s}\cdot\overrightarrow{J_w}}{|\overrightarrow{J_s}|\cdot|\overrightarrow{J_w}|}\right) \quad (10)$$

$$\overrightarrow{J_s}\cdot\overrightarrow{J_w} = J_{s_x}J_{w_x} + J_{s_z}J_{w_z} \quad (11)$$

$$|\overrightarrow{J_s}|\cdot|\overrightarrow{J_w}| = \sqrt{{J_{s_x}}^2 + {J_{s_z}}^2}\cdot\sqrt{{J_{w_x}}^2 + {J_{w_z}}^2} \quad (12)$$

3.3 Modification of the AlexNet-2LSTM architecture model

The deep learning architecture is designed with neural network structures to extract spatio-temporal features. In the context of swing and angle motion joint features, detecting human activities in video recordings relies on extracting information from sequential data. CNN-LSTM models can significantly enhance classification performance, but they require substantial computational resources and training time. This means that the choice of deep learning architecture is based on specific needs. This study highlights the selection of the architectural model that provides the best accuracy, focusing on the optimal structure of the deep learning architecture to predict human actions, as demonstrated by its performance. The models are trained using state-of-the-art deep learning architectures with swing and angle features and their performance is compared. The architectural descriptions used in each SOTA model are shown in Table 2.

In this study modified the deep CNN architecture from AlexNet to train the action recognition model using one-dimensional time series data features. The proposed modified architecture is shown in Fig. 4. AlexNet consists of four 1D convolutional layers, followed by two max pooling layers and two fully connected layers with dense layers. The Softmax activation function in the final layer is used for classifying probabilities. In this stage, LSTM is utilized to enhance performance. The modified model includes two LSTM layers inserted between the convolutional layers and the fully connected layers to effectively handle data features during training.

The first step uses a Conv1D layer with 32 filters activated by the ReLU activation function. Next, the output passes through a max pooling layer with a pooling window size of 1, which helps to retain all the dimensional information of the data after convolution and improves the processing accuracy on smaller data. The second layer uses a Conv1D layer with 64 filters, followed by max pooling. The third and fourth layers, each consisting of a Conv1D layer with 128 filters, aim to extract more features than the previous layers.

LSTM is chosen for its reliability in handling sequential data because it has a learning parameter called the hidden state h_t, which is updated repeatedly based on both current and previous information. The input of LSTM is the current sequence input data x_t, which includes the previous hidden layer state h_{t-1} at time t. Where the W is the weight, and the b is the bias. An LSTM updating the hidden states using the forget gate f_t, input gate i_t, output gate o_t, and the memory cell state value C_t refer to Eq. (13) – Eq. (17). Updating the hidden h_t as the result of the output gate and the memory cell state [30].

Table 2. Architectural descriptions used in each SOTA model

Model	Reference	Architectural Layers
Optimized-LSTM	[29]	Conv1D, MaxPooling1D, Reshape, LSTM, Dropout, LSTM, Dropout, Dense, Dense, Dense(Softmax)
Triple Parallel LSTM	[30]	Conv1D, lstm_1 = LSTM, lstm_2 = LSTM, lstm_3 = LSTM, concatenate([lstm_1, lstm_2, lstm_3]), Dropout, Dense, Dense, Dense(Softmax)
Hybrid CNN-LSTM	[33]	Conv1D, MaxPooling1D, Reshape, conv1_1 = Conv1D, conv1_2 = Conv1D, concatenate([conv1_1, conv1_2]), lstm_1 = LSTM, concatenate([merged_conv1, lstm_1]), LSTM, Dropout, Dense, Dense(Softmax)
LCSWnet (LSTM-CNN)	[34]	Conv1D, LSTM, Dropout, Conv1D, Conv1D, Conv1D, MaxPooling1D, Reshape, Dense, Dense(Softmax)
CC-LSTM-CNN	[35]	Conv1D, LSTM, Dropout, MaxPooling1D, Conv1D, MaxPooling1D, Conv1D, MaxPooling1D, Conv1D, MaxPooling1D(1), Conv1D, MaxPooling1D, Conv1D, Dense, Dense(Softmax)

$$f_t = \sigma\left(W_f \cdot [h_{t-1}, x_t] + b_f\right) \quad (13)$$

$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i) \quad (14)$$

$$o_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o) \quad (15)$$

$$C_t = f_t \circ C_{t-1} + i_t \circ \tanh(W_c \cdot [h_{t-1}, x_t] + b_c) \quad (16)$$

$$h_t = o_t \circ \tanh(C_t) \quad (17)$$

The output of the LSTM is fed into two dense layers, with 128 and 64 neurons respectively, for change the output dimension of the LSTM matrix to linear and enlarge the dimension of data output before classification. Finally, the output data from the dense layer is multiplied by a matrix in Softmax to get probability values. These values are used to classify activities into their respective classes as output in vector using Eq. (18). The W_{fc} is the weight, and the b_{fc} is the bias.

$$\hat{y} = Softmax\left(W_{fc} \cdot h_t + b_{fc}\right) \quad (18)$$

3.4 Classification evaluation

The performance of the classification model was evaluated using metrics such as precision, recall, F1-score, and accuracy [36]. The confusion matrix indicates the number of correct and incorrect of human action predictions. Based on the confusion matrix, where true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN) are obtained, the evaluation metrics are calculated as shown in Eq. (19) – Eq. (22).

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (19)$$

$$Precision = \frac{TP}{TP+FP} \quad (20)$$

$$Recall = \frac{TP}{TP+FN} \quad (21)$$

$$F1-Score = \frac{2 \times Precision \times Recall}{Precision+Recall} \quad (22)$$

4. Experiments and results

In this experiment, to generate features for each action class in the form of swing joint patterns and angles, the coordinates of each body joint are extracted from the tracked joint positions representing the subject performing the action. The proposed method is implemented on the KTH and Weizmann datasets, which are video datasets of subjects performing specific actions. To evaluate the performance of the AlexNet-2LSTM model for action recognition, it is analyzed based on precision, recall, F1 score, and accuracy values for each dataset. A confusion matrix is used to compare the performance of the proposed model with the baseline model and the state-of-the-art (SOTA) model.

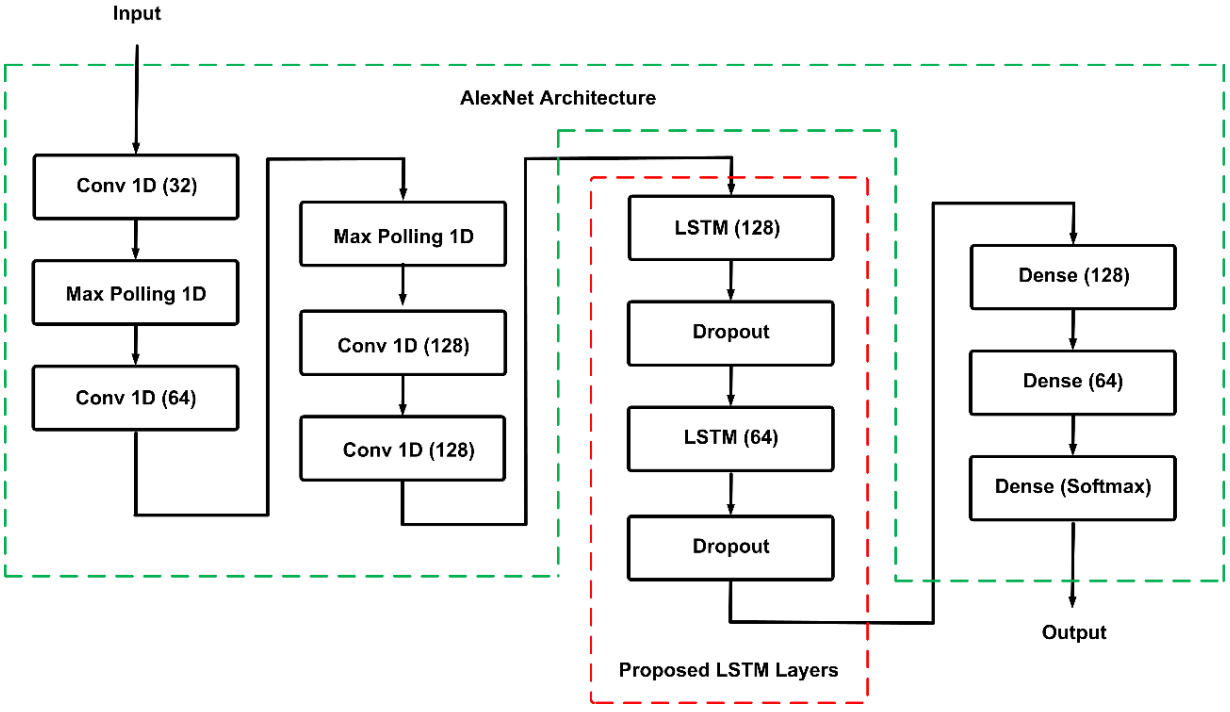


Figure. 4 Proposed of the AlexNet-2LSTM architecture model

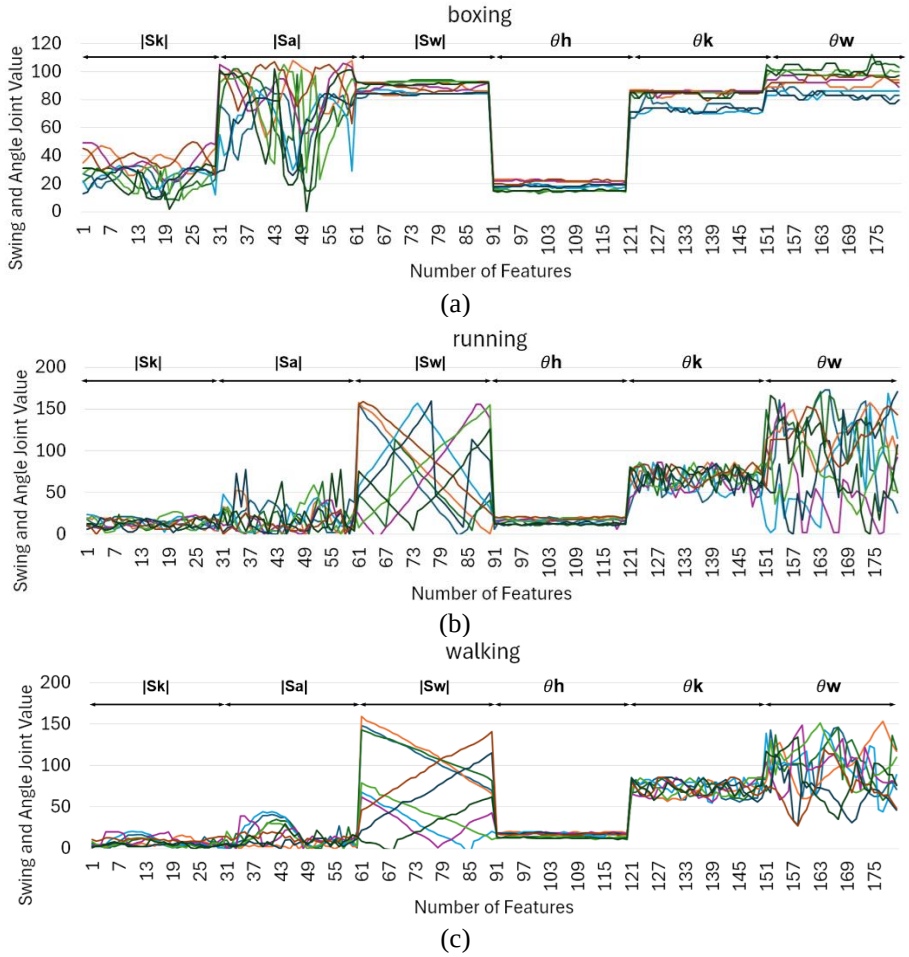


Figure. 5 Samples the pattern of swing and angle joint feature for activities: (a) boxing, (b) running, and (c) walking

4.1 Pattern of the swing and angle joint feature

The feature extraction method for a human action recognition system is proposed to use swing and angle joints as features to identify specific activities performed by the subject. The detected joints focus on the movement of the 2D skeleton joint positions, specifically the shoulder, wrist, hip, and knee, using pose estimation. As long as the subject's entire body is captured by the camera, the joint positions will be detected and then extracted to obtain the swing and joint angle features from the time series data. From the dataset, the feature data patterns are used to identify human activities such as boxing, waving hands, running, clapping hands, jogging, and walking.

Fig. 5 shows samples the pattern of swing joints and joint angles derived from the activity features, which are used as input data for the classifier model. The feature pattern includes the swing and joint angles obtained from the time series data, consisting of normalized values for $|S_k|$, $|S_a|$, $|S_w|$, θ_h, θ_k, and θ_w . Significant differences exist in the feature distribution across various actions. However, the distribution of feature values for running and walking actions demonstrates some similarities. Nevertheless, most of the amplitudes differ, indicating the effectiveness of the proposed feature extraction method.

4.2 Performance of the swing and angle joint feature and AlexNet-2LSTM model with KTH dataset and Weizmann sataset

In the performance experiments of the modified AlexNet-2LSTM model for action recognition, the model was trained on various video datasets specifically designed for surveillance systems. These datasets feature a constant recording angle and a specific set of actions. The training utilized the KTH and Weizmann datasets [37]. The results were analyzed based on the precision, recall, F1-Score and accuracy values for each dataset, as shown in the classification reports in Table 3 and Table 4.

The initial experiments, we applied k-fold 2 to maximize both the training and testing data, minimize computational time, and ensure the system's generality. Table 3 presents the evaluation metrics for the KTH dataset, including precision, recall, F1-

Score and accuracy. The results indicate that the average values for all action classes in the dataset are 0.95 for precision, 0.95 for recall, 0.94 for F1-Score and 0.96 for accuracy. However, the system shows lower performance in distinguishing between running, jogging, and walking, as these activities are quite similar.

Next, the performance of the modified model was tested using the Weizmann dataset, with results shown in Table 4. The findings indicate that, despite having nine classes, the system performs well. The average values for the three metrics are 0.97, 0.96, 0.96, and 0.93, respectively. However, the Weizmann dataset presents challenges in classifying the actions of jumping in place, running, and walking.

In addition, the visualized curves of the training history provide information about the loss and accuracy for both training and validation. To evaluate the performance of the model, the dataset was input into the model using the AlexNet-2LSTM architecture until the optimal weights were achieved. The first test was conducted on the KTH dataset with up to 100 epochs of iterations. Fig. 6 shows the accuracy and loss curves for the action recognition model trained on the KTH dataset. Accuracy increased from 0.81 to 0.96 between epochs 64 to 100, while the loss value decreased from 0.39 to 0.14 starting at epoch 65.

Fig. 7 shows the accuracy and loss curves for the Weizmann dataset, spanning up to 150 epochs of iterations. The accuracy value increases from 0.83 to 0.97 on the training data, starting from epoch 120. Starting at epoch 145, the loss curve shows a value lower than 0.18. The amount of data and the similarity of the classes used in the training process can influence the accuracy and loss values, thereby affecting the overall accuracy.

4.3 Performance of the AlexNet-2LSTM with baseline deep CNN and LSTM models using KTH sataset

To demonstrate the performance of the proposed model, this test compares it with baseline deep CNN and LSTM models using the same swing and angle joint feature data. The models are trained using five different baseline architectures: CNN, CNN-LSTM, LeNet-LSTM, AlexNet-LSTM, VGG-LSTM, and AlexNet-2LSTM. The resulting confusion matrix, which represents different types of actions for the KTH dataset is shown in Fig. 8.

Table 3. Classification reports of AlexNet-2LSTM model using KTH dataset

Action	Precision	Recall	F1-Score	Accuracy
boxing	1.00	0.67	0.80	
waving hand	1.00	1.00	1.00	
running	0.69	1.00	0.82	
clapping hand	1.00	1.00	1.00	**0.96**
jogging	1.00	1.00	0.89	
walking	1.00	0.78	0.88	
Average	**0.95**	**0.95**	**0.94**	

Table 4. Classification reports of AlexNet-2LSTM model using Weizmann dataset

Action	Precision	Recall	F1-Score	Accuracy
bending	1.00	1.00	1.00	
jumping jack	1.00	1.00	1.00	
jumping	1.00	1.00	1.00	
jumping in place	0.77	1.00	0.87	
running	1.00	0.67	0.80	**0.93**
galloping sideways	1.00	1.00	1.00	
skipping	1.00	1.00	1.00	
walking	1.00	0.87	0.95	
two hand waving	1.00	1.00	1.00	
Average	**0.95**	**0.94**	**0.94**	

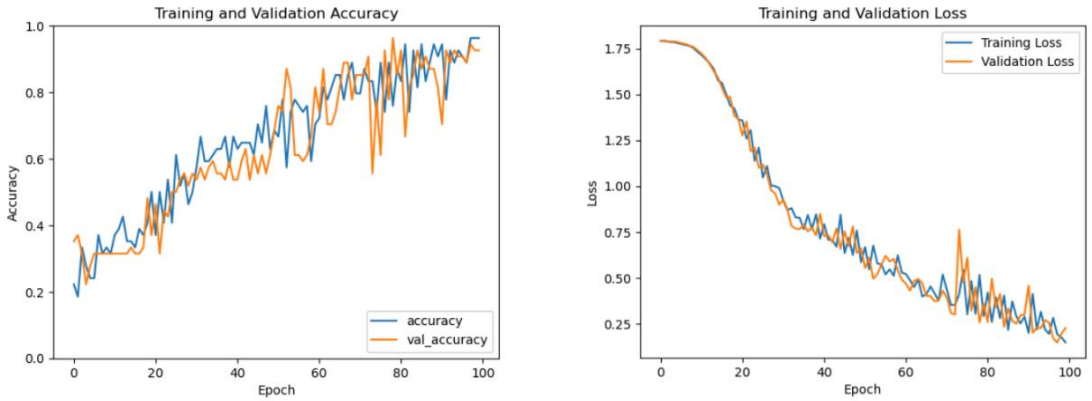


Figure. 6 The training accuracy and loss curve of AlexNet-2LSTM using KTH dataset

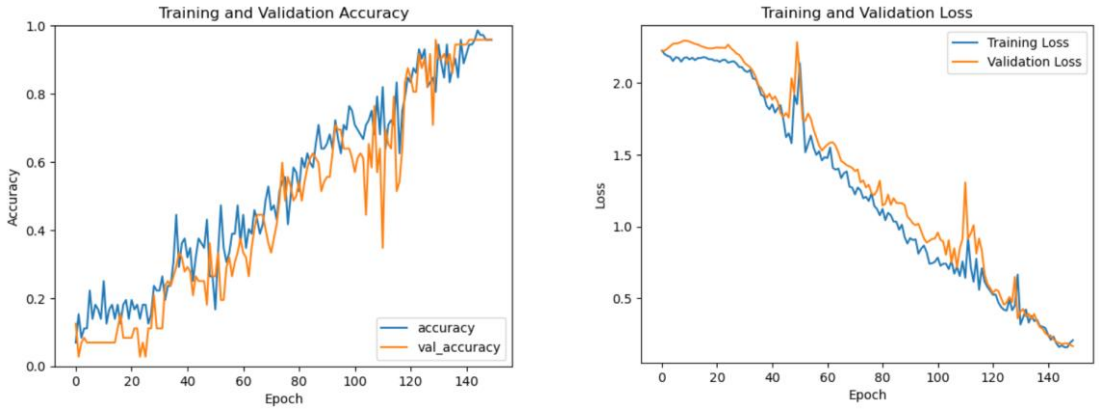


Figure. 7 The training accuracy and loss curve of AlexNet-2LSTM using Weizmann dataset

The results presented in the confusion matrix for the CNN-LSTM and LeNet-LSTM models show that confusion is higher for the "jogging" and "running" actions. This is because the classifier treats the "jogging" and "running" actions as similar, but the accuracy remains above 0.75. In the CNN-LSTM, LeNet-LSTM, AlexNet-LSTM, and VGG-LSTM models, the accuracy for the "running" action is low and is frequently misclassified as the "boxing" action, with the highest error rate reaching 0.44. This misclassification occurs because during the "boxing" action, the foot angle remains relatively still, whereas in the "running" action, the foot angle changes due to the subject's speed. As a result, the foot angle data in both actions does not exhibit significant movement. Under certain conditions, compared to other baseline models, the "jogging" action is often misclassified as the "clapping hand" action, since the hand movements in both actions are similar. Overall, the results indicate that the accuracy of the AlexNet-2LSTM model is superior to that of the baseline models, achieving a score of 0.96, which demonstrates the reliability of the proposed model.

4.4 Comparison of the AlexNet-2LSTM with SOTA deep learning models

To validate and evaluate the performance of the proposed AlexNet-2LSTM model, experiments were conducted by comparing the precision, recall, F1-score, and accuracy values with those of other state-of-the-art (SOTA) deep learning models, namely Optimized-LSTM, Triple Parallel LSTM, Hybrid CNN-LSTM, LCSWnet, and CC-LSTM-CNN. Tables 5 and 6 present a performance comparison of the SOTA models on the KTH and Weizmann datasets. The results of the tests performed on each model were compared using the same feature set of swing and angle data.

Using VGG-LSTM and LCSWnet for the classification of the most relevant activity features achieves accuracies of 0.95 and 0.94, respectively. Meanwhile, the accuracy of AlexNet-LSTM, Optimized-LSTM, Triple Parallel LSTM, and Hybrid CNN-LSTM exceeds 0.91. Evaluation of the proposed model on the KTH dataset shows that

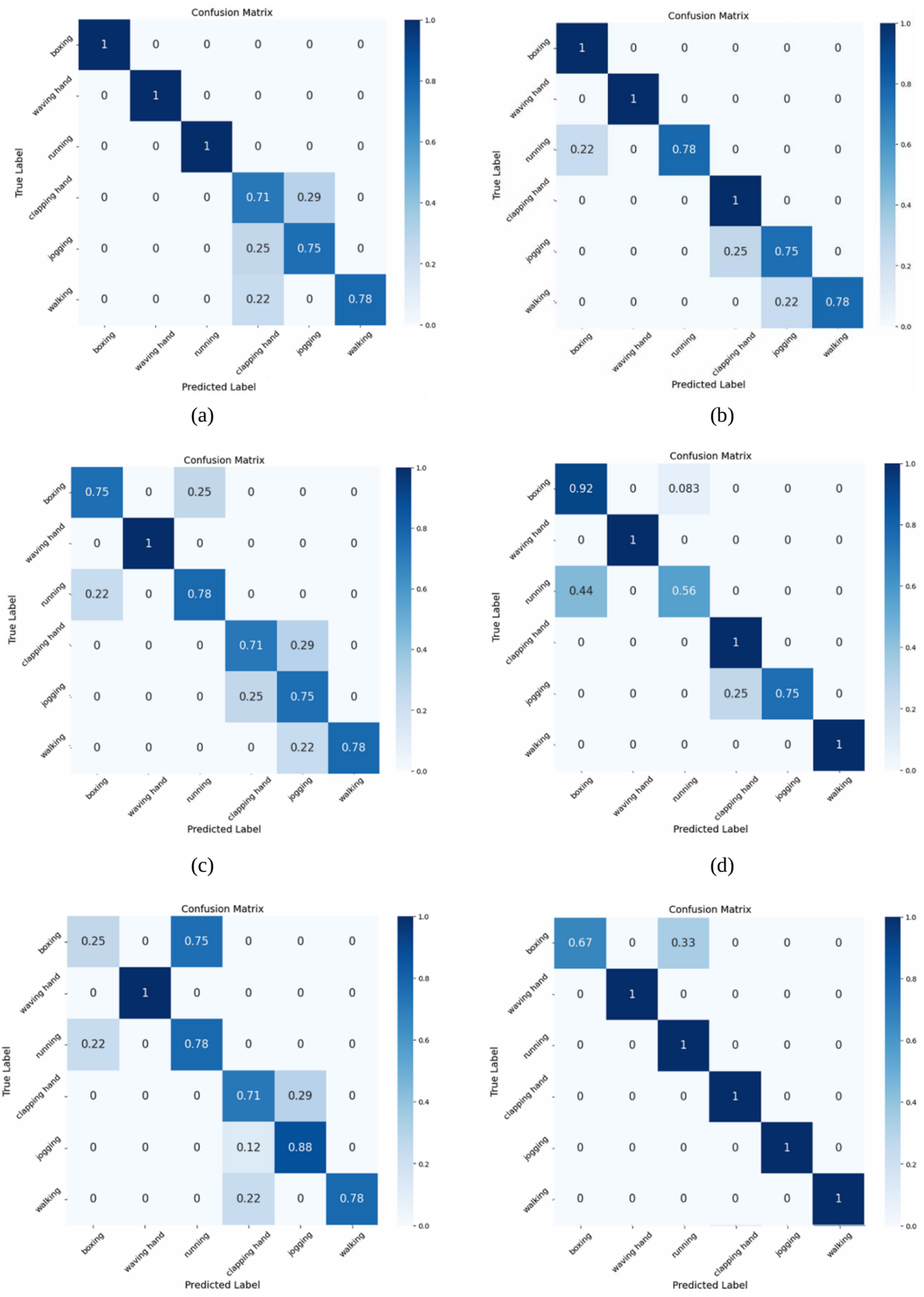


Figure. 8 Confusion matrix for architecture model: (a) CNN, (b) CNN-LSTM, (c) LeNet-LSTM, (d) AlexNet-LSTM, (e) VGG-LSTM, and (f) AlexNet-2LSTM using the KTH dataset

Table 5. Performance comparison of SOTA model using KTH dataset

Classifiers Model	Author	Precision	Recall	F1-Score	Accuracy
CNN	baseline	0.91	0.88	0.89	0.91
CNN-LSTM	baseline	0.81	0.78	0.79	0.82
LeNet-LSTM	baseline	0.83	0.83	0.83	0.86
AlexNet-LSTM	baseline	0.93	0.89	0.90	0.91
VGG-LSTM	baseline	0.96	0.95	0.94	0.95
Optimized-LSTM	[29]	0.94	0.88	0.87	0.92
Triple Parallel LSTM	[30]	0.91	0.88	0.89	0.91
Hybrid CNN-LSTM	[33]	0.91	0.88	0.89	0.91
LCSWnet (LSTM-CNN)	[34]	0.95	0.91	0.92	0.94
CC-LSTM-CNN	[35]	0.86	0.79	0.78	0.83
AlexNet-2LSTM	our proposed model	0.95	0.95	0.94	0.93

Table 6. Performance comparison of SOTA model using Weizmann dataset

Classifiers Model	Author	Precision	Recall	F1-Score	Accuracy
CNN	baseline	0.95	0.88	0.88	0.91
CNN-LSTM	baseline	0.92	0.89	0.88	0.91
LeNet-LSTM	baseline	0.91	0.88	0.87	0.91
AlexNet-LSTM	baseline	0.90	0.85	0.85	0.88
VGG-LSTM	baseline	0.94	0.88	0.88	0.91
Optimized-LSTM	[29]	0.93	0.91	0.89	0.93
Triple Parallel LSTM	[30]	0.90	0.86	0.83	0.87
Hybrid CNN-LSTM	[33]	0.81	0.88	0.84	0.91
LCSWnet (LSTM-CNN)	[34]	0.87	0.86	0.85	0.87
CC-LSTM-CNN	[35]	0.93	0.91	0.89	0.93
AlexNet-2LSTM	our proposed model	0.95	0.94	0.94	0.93

AlexNet-2LSTM is able to accurately recognize human actions, achieving a precision of 0.95, recall of 0.95, F1-score of 0.94, and accuracy of 0.96.

In addition, it can be observed that the performance evaluation results for the Weizmann dataset show that the accuracy of CNN, CNN-LSTM, LeNet-LSTM, VGG-LSTM, and Hybrid CNN-LSTM exceeds 0.91. Accuracy of the AlexNet-2LSTM model is very close to that of Optimized-LSTM and CC-LSTM-CNN, reaching 0.93. In this case, AlexNet-2LSTM outperforms these models in precision, recall, and F1-score, with values of 0.95, 0.94, and 0.94, respectively.

The simpler model may be more suitable for the classification task using the proposed features, resulting in better outcomes, even though it is less complex than the other SOTA models. Comparison experiments show that the proposed model, by utilizing the swing and angle joint features, outperforms existing deep learning models and improves vision-based human action recognition.

5. Conclusions

A surveillance vision-based human action recognition (HAR) system is widely used in computer vision applications, including surveillance and security systems. However, it requires the ability to accurately detect suspicious human activities to ensure public safety. This study proposes a novel feature extraction method based on joint swing patterns and angle features, along with a new classifier model, AlexNet-2LSTM, to improve the accuracy of human activity classification. A pose estimation algorithm is employed to create a 2D skeleton, which contains the coordinates of the body's joint positions. These joint positions are extracted from the image sequence to analyze joint swings and angles as features of human activity. The extracted features are then fed into a modified AlexNet model, which includes four 1D convolutional layers, two max-pooling layers, two fully connected dense layers,

and a Softmax layer. Additionally, two LSTM layers are incorporated to enhance performance during training. To evaluate the proposed method, its performance is analyzed based on precision, recall, F1 score, and accuracy using the KTH and Weizmann datasets. To further demonstrate the effectiveness of the proposed classifier model, it is compared against other state-of-the-art (SOTA) models, including Optimized-LSTM, Triple Parallel LSTM, Hybrid CNN-LSTM, LCSWnet, and CC-LSTM-CNN. The results show that the AlexNet-2LSTM model achieves a precision of 0.95, recall of 0.95, F1-score of 0.94, and accuracy of 0.96 on the KTH dataset, and a precision of 0.95, recall of 0.94, F1-score of 0.94, and accuracy of 0.93 on the Weizmann dataset. These results highlight the contribution of the proposed model in improving feature extraction and classification accuracy for vision-based HAR systems. Further studies could explore significant performance improvements with different datasets and more varied activities. Additionally, the incorporation of attention mechanisms could enhance the performance of classifier models.

Conflicts of interest

The authors declare no conflict of interest.

Author contributions

Conceptualization, R.T. Yunardi, T.A. Sardjono and R. Mardiyanto; methodology, R.T. Yunardi and R. Mardiyanto; validation, R.T. Yunardi and R. Mardiyanto; software and resources, R.T. Yunardi; investigation, R.T. Yunardi and R. Mardiyanto; writing-original draft preparation, R.T. Yunardi; writing-review and editing, R.T. Yunardi, and R. Mardiyanto; visualization, R.T. Yunardi; supervision T.A. Sardjono and R. Mardiyanto.

Acknowledgement

This work was supported in part by the Ministry of Education, Culture, Research, and Technology, Republic of Indonesia, and the Balai Pembiayaan Pendidikan Tinggi (BPPT) through the Lembaga Pengelola Dana Pendidikan (LPDP) Scholarship under Grant No. 00991/J5.2.3/BPI.06/9/2022.

References

[1] M. A. R. Ahad, M. Ahmed, A. D. Antar, Y. Makihara, and Y. Yagi, "Action recognition using kinematics posture feature on 3D skeleton joint locations", *Pattern Recognition Letters*, Vol. 145, pp. 216-224, 2021.

[2] A. Srivastava, T. Badal, A. Garg, A. Vidyarthi, and R. Singh, R, "Recognizing human violent action using drone surveillance within real-time proximity", *Journal of Real-Time Image Processing*, Vol. 18, pp. 1851-1863, 2021.

[3] K. Kim, A. Jalal, and M. Mahmood, "Vision-based human activity recognition system using depth silhouettes: A smart home system for monitoring the residents", *Journal of Electrical Engineering & Technology*, Vol. 14, No. 6, pp. 2567-2573, 2019.

[4] S. K. Adapa, P. Panapana, J. S. Boddu, R. B. Gathram, and M. L. N. Atyam, "Multi-player gaming application based on human body gesture control", In: *Proc. of International Conference on Intelligent Systems Design and Applications*, 2023, pp. 1-30.

[5] N. Jaouedi, N. Boujnah, and M. S. Bouhlel, "A new hybrid deep learning model for human action recognition", *Journal of King Saud University-Computer and Information Sciences*, Vol. 32, No. 4, pp. 447-453, 2020.

[6] A. C. Cob-Parro, C. Losada-Gutiérrez, M. Marrón-Romera, A. Gardel-Vicente, and I. Bravo-Muñoz, "A new framework for deep learning video based human action recognition on the edge", *Expert Systems with Applications*, Vol. 238, pp. 122220, 2024.

[7] B. Kwolek, A. Michalczuk, T. Krzeszowski, A. Switonski, H. Josinski, and K. Wojciechowski, "Calibrated and synchronized multi-view video and motion capture dataset for evaluation of gait recognition", *Multimedia Tools and Applications*, Vol. 78, No. 22, pp. 32437-32465, 2019.

[8] R. T. Yunardi and Winarno, "Marker-based motion capture for measuring joint kinematics in leg swing simulator", In: *Proc. of 2017 5th International Conference on Instrumentation, Control, and Automation (ICA)*, pp. 13-17, 2017.

[9] S. Rastogi and J. Singh, "Human fall detection and activity monitoring: a comparative analysis of vision-based methods for classification and detection techniques", *Soft Computing*, Vol. 26, No. 8, pp. 3679-3701, 2022.

[10] N. Lannan, L. Zhou, and G. Fan, "A multiview depth-based motion capture benchmark dataset for human motion denoising and enhancement research", In: *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 427-436, 2022.

[11] J. Zhang, A. Yang, C. Miao, X. Li, R. Zhang, and D. N. Thanh, "3D graph convolutional feature selection and dense pre-estimation for

skeleton action recognition", *IEEE Access*, Vol. 12, pp. 11733-11742, 2024.

[12] R. T. Yunardi, T. A. Sardjono, and R. Mardiyanto, "Motion capture system based on RGB camera for human walking recognition using marker-based and markerless for kinematics of gait", In: *Proc. of 2023 IEEE 13th Symposium on Computer Applications & Industrial Electronics (ISCAIE)*, pp. 262-267, 2023.

[13] J. Lee and B. Ahn, "Real-time human action recognition with a low-cost RGB camera and mobile robot platform", *Sensors*, Vol. 20, No. 10, pp. 2886, 2020.

[14] W. Chen, Z. Jiang, H. Guo, and X. Ni, "Fall detection based on key points of human-skeleton using OpenPose", *Symmetry*, Vol. 12, No. 5, pp. 744, 2020.

[15] X. L. Lau, T. Connie, M. K. O. Goh, and S. H. Lau, "Fall detection and motion analysis using visual approaches", *International Journal of Technology*, Vol. 13, No. 6, pp. 1173-1182, 2022.

[16] M. F. Trujillo-Guerrero, S. Román-Niemes, M. Jaén-Vargas, A. Cadiz, R. Fonseca, and J. J. Serrano-Olmedo, "Accuracy comparison of CNN, LSTM, and Transformer for activity recognition using IMU and visual markers", *IEEE Access*, Vol. 11, pp. 106650-106669, 2023.

[17] A. Hayat, F. Morgado-Dias, B. P. Bhuyan, and R. Tomar, "Human activity recognition for elderly people using machine and deep learning approaches", *Information*, Vol. 13, No. 6, pp. 275, 2022.

[18] R. Cui, A. Zhu, S. Zhang, and G. Hua, "Multi-source learning for skeleton-based action recognition using deep LSTM networks", In: *Proc. of 2018 24th International Conference on Pattern Recognition (ICPR)*, pp. 547-552, 2018.

[19] J. Liu, G. Wang, L. Y. Duan, K. Abdiyeva, and A. C. Kot, "Skeleton-based human action recognition with global context-aware attention LSTM networks", *IEEE Transactions on Image Processing*, Vol. 27, No. 4, pp. 1586-1599, 2017.

[20] R. Sugai, S. Maeda, R. Shibuya, Y. Sekiguchi, S. I. Izumi, M. Hayashibe, and D. Owaki, "LSTM network-based estimation of ground reaction forces during walking in stroke patients using markerless motion capture system", *IEEE Transactions on Medical Robotics and Bionics*, Vol. 5, No. 4, pp. 1016-1024, 2023.

[21] R. T. Yunardi, T. A. Sardjono, and R. Mardiyanto, "Skeleton-based gait recognition using modified deep convolutional neural networks and long short-term memory for person recognition", *IEEE Access*, Vol. 12, pp. 121131-121143, 2024.

[22] K. Xia, J. Huang, and H. Wang, "LSTM-CNN architecture for human activity recognition", *IEEE Access*, Vol. 8, pp. 56855-56866, 2020.

[23] N. A. Choudhury and B. Soni, "An efficient and lightweight deep learning model for human activity recognition on raw sensor data in uncontrolled environment", *IEEE Sensors Journal*, Vol. 23, No. 20, pp. 25579-25586, 2023.

[24] M. Mundt, Z. Born, M. Goldacre, and J. Alderson, "Estimating ground reaction forces from two-dimensional pose data: a biomechanics-based comparison of AlphaPose, BlazePose, and OpenPose", *Sensors*, Vol. 23, No. 1, pp. 78, 2022.

[25] J. H. Yoo and M. S. Nixon, "Automated markerless analysis of human gait motion for recognition and classification", *ETRI Journal*, Vol. 33, No. 2, pp. 259-266, 2011.

[26] K. Jun, D.-W. Lee, K. Lee, S. Lee, and M. S. Kim, "Feature extraction using an RNN autoencoder for skeleton-based abnormal gait recognition", *IEEE Access*, Vol. 8, pp. 19196–19207, 2020.

[27] K. M. Oikonomou, I. Kansizoglou, P. Manaveli, A. Grekidis, D. Menychtas, N. Aggelousis, and A. Gasteratos, "Joint-aware action recognition for ambient assisted living", In: *Proc. of 2022 IEEE International Conference on Imaging Systems and Techniques (IST)*, pp. 1-6, 2022.

[28] N. Tasnim, M. M. Islam, and J. H. Baek, "Deep learning-based action recognition using 3D skeleton joints information", *Inventions*, Vol. 5, No. 3, pp. 49, 2020.

[29] S. K. Challa, A. Kumar, V. B. Semwal, and N. Dua, "An optimized-LSTM and RGB-D sensor-based human gait trajectory generator for bipedal robot walking", *IEEE Sensors Journal*, Vol. 22, No. 24, pp. 24352-24363, 2022.

[30] M. Shayestegan, T. Zalabsky, and J. Mareš, "Triple parallel LSTM networks for classifying the gait disorders using Kinect camera and robot platform during the clinical examination", In: *Proc. of 2023 3rd International Conference on Electrical, Computer, Communications and Mechatronics Engineering (ICECCME)*, pp. 1-6, 2023.

[31] W. Zheng, L. Li, Z. Zhang, Y. Huang, and L. Wang, "Relational network for skeleton-based action recognition", In: *Proc. of 2019 IEEE International Conference on Multimedia and Expo (ICME)*, pp. 826-831, 2019.

[32] S. Cao, M. Ko, C. Li, D. Brown, X. Wang, F. Hu, and Y. Gan, "Single-belt versus split-belt:

intelligent treadmill control via microphase GAIT capture for poststroke rehabilitation", *IEEE Transactions on Human-Machine Systems*, Vol. 53, No. 6, pp. 1006–1016, 2023.

[33] A. S. Alharthi and K. B. Ozanyan, "Fusion from multimodal gait spatiotemporal data for human gait speed classifications", In: *Proc. of 2021 IEEE Sensors*, pp. 1-4, 2021.

[34] J. Gao, P. Gu, Q. Ren, J. Zhang, and X. Song, "Abnormal gait recognition algorithm based on LSTM-CNN fusion network", *IEEE Access*, Vol. 30, No. 7, pp. 163180-163190, 2019.

[35] R. Li, C. Song, D. Wang, F. Meng, Y. Wang, and Q. Tang, "A novel approach for gait recognition based on CC-LSTM-CNN method", In: *Proc. of 2021 13th International Conference on Intelligent Human-Machine Systems and Cybernetics (IHMSC)*, pp. 25-28, 2021.

[36] H. H. Tan, R. Shahid, M. Mishra, and S. L. Lim, "Malaysian vanity license plate recognition using convolutional neural network", *International Journal of Technology*, Vol. 13, No. 6, pp. 1271-1281, 2022.

[37] A. Nadeem, A. Jalal, and K. Kim, "Human actions tracking and recognition based on body parts detection via artificial neural network", In: *Proc. of 3rd International Conference on Advancements in Computational Sciences (ICACS)*, pp. 1-6, 2020.