ISSN: 0067-2904

Improved RNN Model for Real-Time Human Activity Recognition

Azhee Wria Muhamad[1]*, Aree Ali Mohammed[2]

[1] *Department of Computer Science, College of Basic Education, University of Suleimani, Sulimania, Iraq*
[2] *Department of Computer Science, College of Science, University of Suleimani, Sulimania, Iraq*

Received: 7/12/2022 Accepted: 3/8/2023 Published: 30/9/2024

Abstract

This study focuses on the automatic detection of human actions in video streams. The requirement to detect what human activities happen in videos is recognition of human action due to significant differences in people's visual and motion appearance and actions, camera perspective shifts, moving background, occlusions, noise, and a massive amount of video data. The human activity recognition challenge involves identifying physical activities carried out by individuals or groups based on traces of movements, including gestures, actions, interactions, and group activities. The detection of concepts usually requires additional annotations for the training dataset. In this paper, useful methods for categorizing human action recognition are discussed. The current models are an accurate deep learning method that is based on models that have been changed to be more useful. The large disparities that result from the backdrop and the size of the objects have prevented the identification of activities in videos from being fully and effectively addressed. The main objective is to achieve better accuracy for the Long Short-Term Memory (LSTM) method, which was used to improve the Recurrent Neural Networks (RNN) model. In this paper, LSTM is used to come up with models for different action recognition tasks. The model was made better by making the LSTM have four layers and putting 128 units, 64 units, 32 units, and 16 units in each layer, respectively. In addition, the performance evaluation of deep learning-based approaches has been compared to other related works. Therefore, an improved approach to RNN is proposed to recognize human actions. To classify the videos, a multilayer RNN with a specific type of LSTM is used to extract features from video sequences. The UCF-101 and UCF Sports human action recognition datasets are utilized in this study for both training and assessment. Test findings demonstrate that the suggested strategy achieved increased accuracy. Finally, the enhanced RNN model's total model accuracy in the UCF-101 dataset is 93.78% and 95.70% for the UCF Sport dataset.

Keywords: Accuracy, Action detection, Deep learning, Human action recognition, Recurrent neural networks.

نموذج الشبكات العصبية المتكررة المحسن للتعرف على النشاط البشري في الوقت الحقيقي

ئەژى وريا محمد١*، ئارى على محمد٢

[1]قسم علوم الحاسب ، كلية التربية الأساسية ، جامعة السليمانية ، السليمانية، العراق
[2]قسم علوم الحاسب ، كلية العلوم ، جامعة السليمانية ، السليمانية، العراق

الخلاصه

تركز هذه الدراسة على الكشف التلقائي عن أفعال الإنسان في تدفقات الفديو. إن مطلب اكتشاف الأنشطة البشرية التي تحدث في مقاطع الفيديو هو التعرف على الإجراءات البشرية نظرًا للاختلافات الكبيرة

*Email: azhee.muhamad@univsul.edu.iq

في المظهر المرئي والحركي للأشخاص وأفعالهم ، وتحولات منظور الكاميرا ، والخلفية المتحركة ، والانسدادات ، والضوضاء ، وكمية هائلة من بيانات الفديو. تحدي التعرف على النشاط البشري يتضمن تحديد الأنشطة البدنية التي يقوم بها الأفراد أو المجموعات بناءً على آثار الحركة والتي تشمل الإيماءات والإجراءات والتفاعلات وأنشطة المجموعة. يتطلب اكتشاف المفاهيم عادةً تعليقات توضيحية إضافية لمجموعة بيانات التدريب. تصف هذه الورقة الاستراتيجيات العملية لتصنيف التعرف على العمل البشري. النماذج الحالية هي تقنية دقيقة للتعلم العميق تعتمد على نماذج مناسبة معدلة. منعت التباينات الكبيرة الناتجة عن الخلفية وحجم الأشياء تحديد الأنشطة في مقاطع الفديو من المعالجة الكاملة والفعالة. الهدف الرئيسي هو تحقيق دقة أفضل للنموذج المحسن للشبكات العصبية المتكررة (RNN) باستعمال طريقة الذاكرة طويلة المدى (LSTM). تقترح هذه الورقة نماذج لمختلف مهام التعرف على الإجراءات باستعمال LSTM. تم تحسين النموذج بزيادة LSTM إلى أربع طبقات وعدد الوحدات في كل طبقة إلى 128 و 64 و 32 و 16 لكل خلية على التوالي. بالإضافة إلى ذلك ، تمت مقارنة تقييم أداء المناهج القائمة على التعلم العميق بالأعمال الأخرى ذات الصلة. لذلك ، تم اقتراح نهج محسن لـ RNN للتعرف على الأعمال البشرية. لتصنيف مقاطع الفيديو ، استعمل RNN متعدد الطبقات بنوع معين من LSTM لاستخراج الميزات من تسلسلات الفيديو. تستعمل مجموعات بيانات التعرف على العمل البشري UCF−101 و UCF Sports في هذه الدراسة لكل من التدريب والتقييم. تظهر نتائج الاختبار أن الاستراتيجية المقترحة حققت زيادة في دقة الأداء. أخيرًا ، الدقة الإجمالية لنموذج RNN المحسّن في مجموعة بيانات UCF−101 هي 93 و 95 لمجموعة بيانات UCF Sport.

1. Introduction

Recognizing human activity is one of automation's most important responsibilities, particularly for intelligent applications. For instance, based on action detection, robots' systems in smart homes or intellectual industries might assist or work with people [1]. When combined with cyber-physical systems, action recognition has been helpful in various applications, including health care [2] and [3]. It has been utilized in cloud computing technologies for analyzing social behavior. However, with backdrop clutter and occlusions in the actual world, human action recognition is still a long way off [4], especially in complicated dynamic systems. Recognizing human activity in a video is the aim of human action recognition (HAR). Human activities have been split into two categories: human-human contact and human-object interaction. Due to differences in speed, light, partial occlusion of individuals, perspective, and the anthropometry of those taking part in the numerous exchanges, it is also difficult to recognize human activity. Some of the recognition problems are more directly related to the problem of locating and tracking people in videos. However, other people are more focused on locating the action. Human activity recognition involves several steps. First, the video data is captured using cameras or other imaging devices. Next, the video frames are pre-processed to remove noise and enhance the features of interest. Then, feature extraction techniques are used to extract relevant information from the video frames. Once the features are extracted, machine learning algorithms are used to classify the activities being performed in the video. These algorithms can be trained using labeled data sets that contain examples of different human activities, as shown in Figure 1.

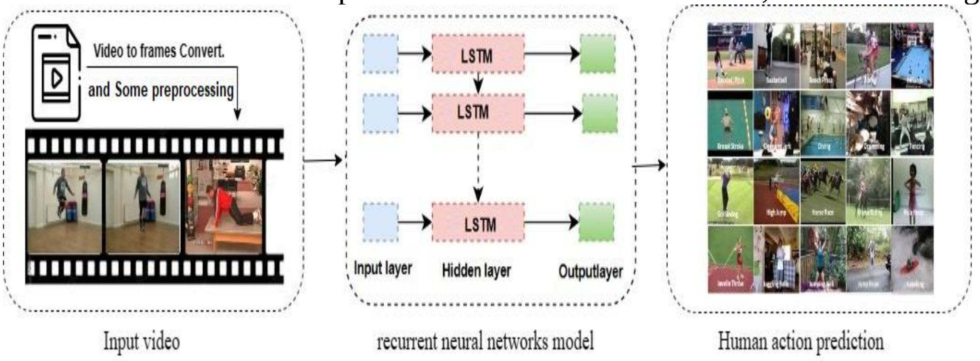


Figure 1: A description of the model for the prediction of human actions

Recognizing human actions can be hard, but it's important for processing video and classifying actions. It is being set up as part of human activity's ongoing intense care. has also been suggested for a number of other uses, such as health care and police work with older people, preventing sports injuries, figuring out where people are, and keeping an eye on people's homes. Although human action detection from streaming video sequences has improved, it is still not very strong due to factors including shifting angles and distances from the camera, context, and varying speeds. Finding good choices is the most difficult part of the whole process. It changes the speed of computer programs by making calculations take less time [5]. But local options made by hand from RGB video from second cameras can't handle the complex movements that the most common human action recognition systems can. In more than one way to find a moving object, background extraction is used.

The Human Activity Recognition (HAR) problem involves identifying physical activities carried out by individuals or groups based on traces of movement. Human activity can be recognized in a wide range of ways. Thus, action recognition has been categorized, and that typically refers to action primitives, actions, activities, interactions, and group activities depending on the complexity of the activity [6]. There are different approaches to solving each of these categories. When recognizing an interaction like "two people playing football," a high-level process would need information about what is occurring at the lower level, such as "extending a leg." But activities such as walking, lying, sitting, standing, and climbing stairs are classified as regular physical movements [7].

In addition to the Gaussian distribution, there are a number of other ways to relate human behavior, such as by using different motion tracking methods. Humans watch how it moves and make plans for all possible paths. For most people, the next steps are generally easy to understand. Things that move quickly make the problem harder. Still, you need a strong features vector in order to have a good classifier that can give the class name [8]. Features are the most important parts of any data set. In fact, feature extraction can show both the method and how hard it is to compute. In this study, one can talk about how important traits are for classifying and recognizing human behavior. Still, this situation has a strong link to the methods and algorithms used to pull out complex features from live video loops [9].

Vision-based human activity recognition is an important research area in the fields of computer vision and machine learning. The study of computer vision and automatic detection with classification is one of the most important fields today. The process of creating models that interpret and understand visual data is known as "computer vision," which is part of artificial intelligence (AI). Computer vision is important for training visual perception-based deep learning models for use in artificial intelligence (AI). Deep learning models are improved by utilizing high-quality training data, according to [10]. Computer vision gives machines an artificial vision that helps them analyze situations and make the most informed decisions.

Also, to better describe human actions, you should come up with a new method that uses motion tracking by watching a person and spatial feature extraction from a video series. It has a robust feature vector for classification. A modified RNN model for human action recognition is proposed, which involves recognizing and monitoring humans and comprehending human activities from a video sequence. Developing strategies for an automated visual surveillance system is the primary focus of research in this field. These are the contributions made by this research: (1) When selecting a dataset for deep learning, several factors need to be considered. These include dataset size, diversity, annotation quality, task relevance, and availability. A large dataset allows the model to learn more complex representations and generalize better to unseen data. (2) The purpose of the data augmentation approaches is to increase the accuracy of the machine learning algorithms by producing additional copies of the actual dataset. (3) The proposed LSTM model has a high capacity for learning sequences and changing features from frame to frame. (4) Some preprocessing techniques are applied to improve the performance of the RNN architecture model for feature extraction and others for normalization of the features before feeding them into the model. The remainder of the paper is organized as follows: In Section 2, a number of related works for human action recognition are presented. The proposed model architecture of the network model is discussed in Section 3, while the obtained results are addressed in Section 4. Finally, the study's conclusion is provided in Section 5.

2. Related works

Human action recognition has been well worked upon and studied in recent years. In the beginning, the methods used to find out interest points of action were mainly done manually by looking at various feature points for multiple activities. But in the last few years, a machine has made efforts to detect those feature points rather automatically. Before moving further, there are also some design issues with systems, such as the selection of different types of neural network models and data collection-related rules [11].

Researchers have been attempting to identify human motion from images and videos since the 1980s. One of the major approaches that academics have been researching for action recognition is similar to how the human visual system works. The human visual system is able to receive a number of observations about the motion and form of the human body in a short amount of time when it is working at a low level. Then, these data are sent to the intermediate human perception system so that the classification of these observations, such as walking, jogging, and running, may be further recognized [12]. In actuality, the identification of observed movement and human actions is made by the perception system, which is strong and very accurate. Researchers have put in a lot of work over the past few decades to get a computer-based recognition system to work as well as a human system. Still, the researchers are a long way from the level of the visual system in humans. However, there are several difficulties and problems associated with HAR, including environmental complexity, the non-rigid shapes of humans and objects, and another problem associated with human recognition [13]. The two main types of vision-based human activity identification techniques may be determined by a thorough analysis of the literature. First, the traditional method, which is based on handcrafted representations; second, a representational strategy based on learning.

2.1 Handcrafted representation-based technique

The classic technique for action recognition is based on constructed action representations. This approach has been popular among the HAR community and has produced amazing results on a number of famous public datasets. In this way, the important features from the sequence of videos are extracted, and then classification is executed by training a basic classifier such as a Support Vector Machine (SVM) [14]. Recent years have seen much work and study done on HAR. The methods used to find out interest points of action were mostly done manually in the beginning by looking at various features of many activities. In order to extract distinguishing features, methods based on handcrafted features rely on human creativity and prior knowledge. There are three major phases to these methods: 1) action segmentation corresponds to foreground detection; 2) expert feature selection and extraction; and 3) action classification is represented by the extracted features [15]. The most significant features are first extracted from the input video frames before the human action label is built.
The orientation and position of the limb in space are used to describe static activities, while dynamic activities are described as movements of these static activities. Three different types of spatial and temporal representations can be classified [16]. Dense trajectories, which

consist of histograms of orientation (HOG), histograms of optical flow (HOF), and motion boundary histograms (MBH), have recently been identified as a successful method. Improved Dense Trajectories (IDT) features [17] consider camera motion to improve action detection using a D.T. technique.

2.2 Deep learning for action recognition

In recent years, feature learning has become popular in a wide variety of computer vision applications, including pedestrian detection, image classification, vision-based anomaly detection, etc. [18]. Several learning-based approaches to action recognition involve converting pixels into action classes through end-to-end learning. The feature-learning-based HAR task is constructed using deep learning. The multilayer convolutional neural network with long short-term memory combined in a deep neural network could automatically extract action characteristics and categorize them based on some of the model parameters [19]. Recurrent neural networks, such as the LSTM, are particularly good at processing temporal sequences. A successful case study of frame classification is what inspired action recognition in video sequences [20].

Convolutional layers with long short-term memory combined in a deep neural network could mechanically extract action characteristics and categorize them using only a few model parameters. The LSTM is not to be confused with the recurrent neural network, which is better suited to processing temporal sequences. The success of image classification has inspired many advances in video action recognition [21]. The accomplishments in the image domain have rekindled interest in deep learning for video identification. When compared to manually created features, RNNs with long short-term memory [22] [23] perform much better on a variety of visual tasks. LSTM-enabled RNNs leverage a gating architecture consisting of a succession of memory blocks to control the input flow, thereby mitigating the gradient vanishing issue and improving the extraction of long-term features. LSTM neural networks are suggested to capture more vigorous and extended spatiotemporal representations. On the demanding benchmark databases, UCF Sport and UCF101, when an LSTM model is trained and tested at the same time, it tends to overfit, and the accuracy of identification is lower than with manual feature methods.

Existing approaches to recurrent neural networks have a few drawbacks. During training, RNNs are susceptible to the vanishing or explosive gradient problem, in which gradients become vanishingly small or exploringly large, resulting in unstable learning and difficulty capturing long-term dependencies. Standard RNNs have very little contextual memory, so it may be difficult for them to remember information from previous time steps when processing lengthy sequences. Human actions can look very different depending on things like camera angle, lights, occlusions, and how people are standing. This diversity makes it difficult for deep learning models to generalize effectively across various action instances and situations. This limitation hinders their ability to capture long-term dependencies effectively. Traditional RNNs need input sequences of a set length, which may be difficult when working with variable-length input data like words of various lengths or videos with differing frame counts [24].

RNNs are excellent at modeling temporal or sequential data because they can capture dependencies between various time stages. They are ideal for applications like video recognition, language modeling, and time series prediction because they have the ability to learn patterns and dynamics found in time series data. Researchers continue to address the shortcomings of current RNN techniques by proposing cutting-edge designs such as LSTM and GRU (Gated Recurrent Unit), which alleviate the vanishing/exploding gradient issue and enhance contextual memory [25]. Alternative models, including transformer-based architectures, have also gained popularity due to their ability to enable parallel computations and better capture long-term relationships. By addressing these issues, conventional RNNs are expected to exhibit improved performance across a range of applications.

Based on an extensive study of both vision and machine learning approaches, many gaps may be identified that other relevant publications do not consider. The central gap is that many researchers have used the UCF dataset, but they did not include video in the dataset. Several processing methods are utilized in the proposed approach to make up for the fact that they did not produce the file to store all information about movies in a CSV format for training and testing their models. For real-time human action recognition that includes action feature representation techniques, interaction recognition methods, and action detection methods, the authors proposed a modified RNN based on LSTM architecture in comparison to current RNN designs.

3. Methodology

An improved RNN based on a modified long-short-term memory network model for enhancing HAR accuracy is presented, which can be easily applied to automatically recognize human actions. Some data pre-processing algorithms have primarily been performed on both the UCF-Sport and UCF-101 datasets, including color-to-grayscale conversion, histogram equalization, filtering, and normalization. Then, data augmentation is used to increase the quantity of training data to prevent overfitting. The efficacy of the model is evaluated using scores for overall accuracy, precision, and recall, then using the data augmentation as shown in Figure 2.

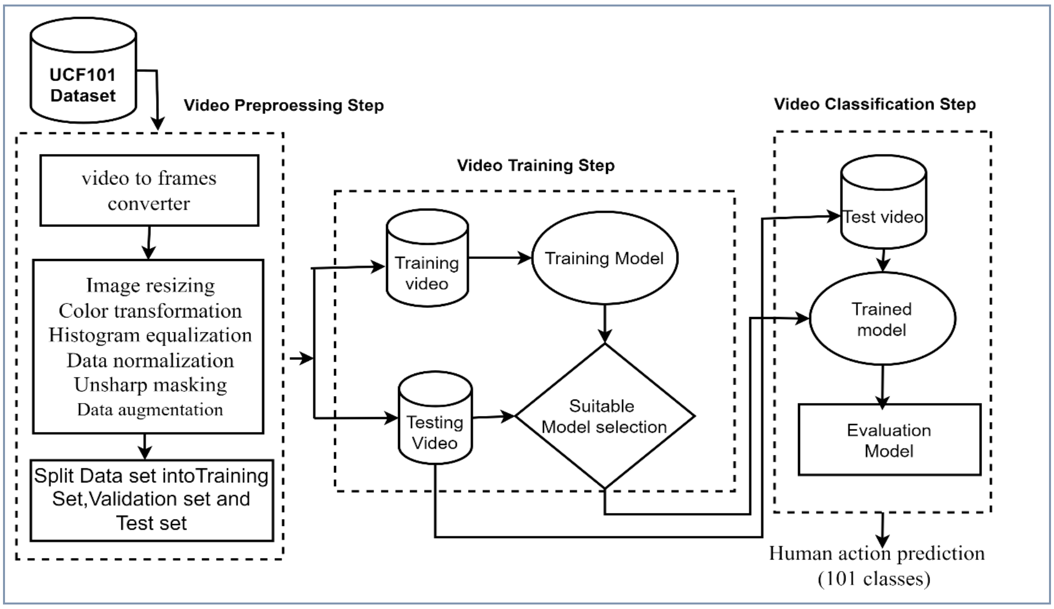


Figure 2: A general block diagram of the proposed model

3.1 Dataset preparation

The University of Central Florida's Department of Electrical Engineering and Computer Science created the UCF datasets, a collection of information that is gradually becoming more and more useful and that uses unscripted video from challenging to compile unscripted sources. 13320 YouTube video clips with a fixed frame rate of 25 frames per second and 320 × 240 pixels are included in the UCF101 database. 101 categories of human activity are included in the database [8].

Each action class is divided into 25 groups, with four to seven video clips in each group. According to the literature, UCF101's action recognition uses the train and test sections to avoid using the same video clips for training and testing. The data sets do not sufficiently simplify actual data since the majority of contemporary approaches are accurate to at least the 95th percentile or better [4]. The entire length of the video clips for each class in the UCF 101 dataset is shown in Figure 3. The datasets are divided into three sets: %75 train, %5 validation, and %20 test. Each set contains 10656, 660, and 2664 videos of human actions, respectively.

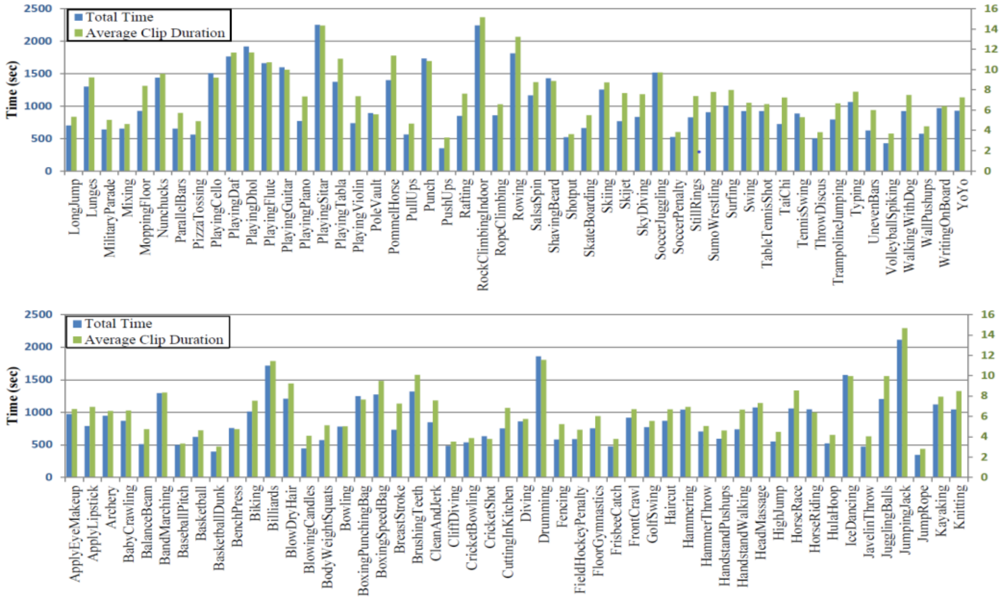


Figure 3: UCF101 has 101 actions; the total time of video clips for each human action class is blue. The average length of clips for each action is green [4].

UCF sports dataset This data was obtained from multiple sporting events broadcast on television channels such as ESPN, and this dataset included a total of 150 recordings. The sample labeled the proposed scheme's results “Multimedia Tools and Applications.” The number of action classes is ten, including diving, horseback riding, and walking, among others. Each video has a resolution of 720 by 480 pixels [26]. As shown in Figure 4, the duration of each video segment in the UCF sport dataset.

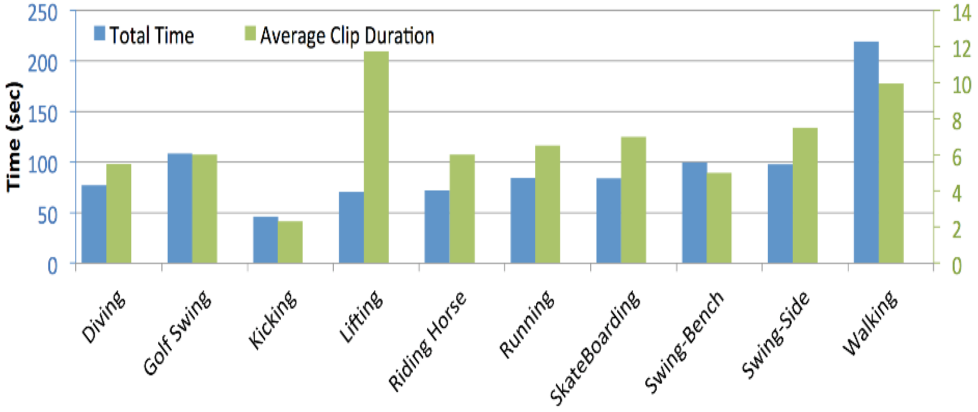


Figure 4: The total time of video clips for each human action class is blue. The average length of clips for each action is green in UCF sports [26].

3.2 *Data* pre-processing

In human action recognition, pre-processing the dataset is a crucial step in building an accurate and reliable model. The pre-processing pipeline typically involves several steps such as data cleaning, normalization, feature extraction, data augmentation, data splitting, label encoding, and data visualization. Previously, the model was trained and tested. The quality of the videos needs to be improved. This will help the model better classify and recognize human actions in videos. To reach this goal, different picture processing methods have been built into the suggested model. As a result, the following image pre-processing techniques are used:

3.2.1 Image resizing

Many deep learning model designs need identical-sized input images. As a result, the RNN model resizes the 320x240 pixel pictures from the UCF-101 and UCF sport datasets to 128x128 pixel images.

3.2.2 Color conversion

All video frames in both datasets are subjected to the color conversion from RGB to grayscale representation.

3.2.3 Histogram equalization

Histogram normalization is a way to make an image's histogram have a wider dynamic range, making the picture stand out more. The goal of histogram normalization is to get a histogram that looks the same everywhere.

3.2.4 Unsharp masking

The unsharp filter is a simple sharpening operator that enhances edges to solve blurry images. The improved model has not perfectly detected and recognized human actions in both the training and testing stages because there are several blurry frames in the datasets.

3.2.5 Data normalization

Data normalization is a process that standardizes data by transforming it into a uniform scale or range, thereby removing any differences in magnitude or units among different variables or features. This involves adjusting the data to fit a particular distribution, often ranging from 0 to 1, or having a mean of 0 and a standard deviation of 1. Normalizing data is advantageous for tasks like data analysis and machine learning because it minimizes the effects of dissimilar scales or units on model performance, improves feature extraction, and enhances the precision of classification or other analytical techniques.

These preprocessing techniques can be used individually or combined, depending on the specific requirements of the task and the characteristics of the image. The goal is to enhance the image quality, reduce noise, standardize features, and extract relevant information that can be used for subsequent analysis or machine learning tasks. Figure 5 displays some of the results obtained from preprocessing techniques.

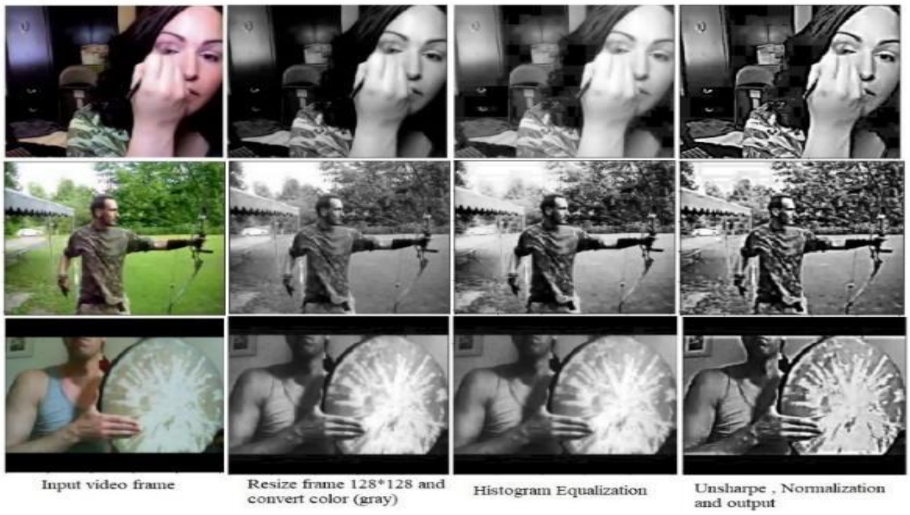

Input video frame
Resize frame 128*128 and convert color (gray)
Histogram Equalization
Unsharpe , Normalization and output

Figure 5: Outcomes of the Pre-processing Method

3.2.6 Data augmentation

Data augmentation methods create new variations of a real dataset to improve the precision of machine learning algorithms. The initial training images were randomly selected. The training images were (1) horizontally displaced and (2) rotated to enhance their dimensions. In addition, the parameters utilized for augmentation are arbitrarily chosen from 10.0, 10.0%, -10.0, -10.0%, 0.0, 0.2 radians CCW (counter clockwise) and CW (clockwise), (3) zoomed-in, (4) and angle 20 for shear. Figure 6 demonstrates several examples of data augmentation applied to the UCF101 dataset.

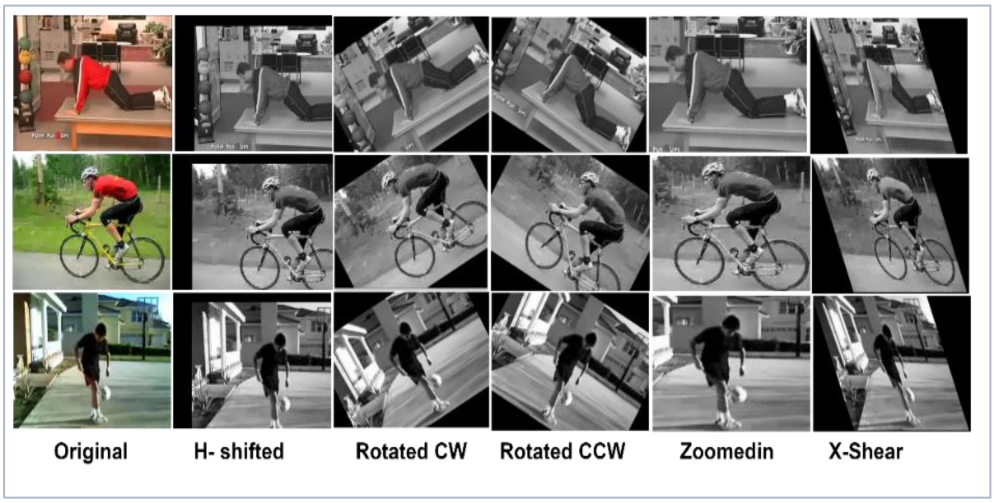


Figure 6: Augmented Human Action Recognition from the UCF101 Dataset

3.3 Improved RNN architecture

The pre-processed frames from previous steps are fed to the training phase using an improved RNN network model to enhance the video's human action recognition. The updated LSTM design, which has four hidden layers instead of one, is the basis on which the improved RNN model is based. Figure 7 shows a single cell in an LSTM memory block, where the dashed lines denote weight between the cell and the gates and the solid lines represent multiplication. The block's other connection weights are all set to 1. Each memory block in the LSTM architecture consists of a number of memory cells connected by internal connections, as well as input, output, and forget gates, which are multiplicative units [27].

The logistic sigmoid is frequently used as the gate activation function; the hyperbolic tangent or logistic sigmoid are the activation functions for cell input and output; and three gates provide continuous equivalents of the cell's write, read, and reset operations. The LSTM's particular memory architecture is useful for solving the disappearing and expanding gradient problem.

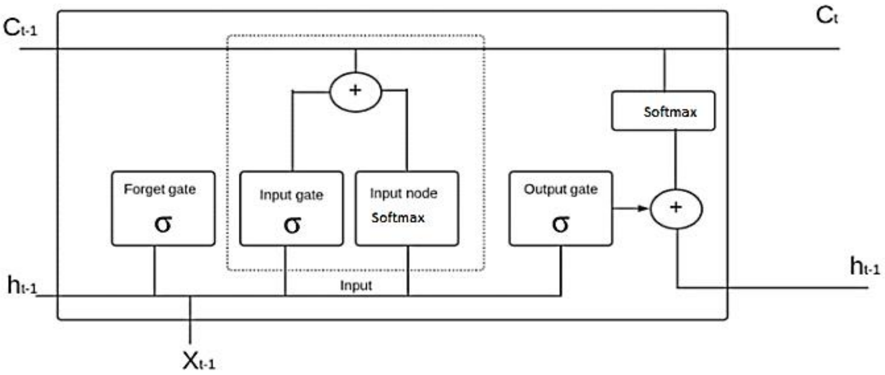


Figure 7: One cell in an LSTM memory block [25].

HAR performance is evaluated using an RNN-based deep learning architecture that demonstrates significant video classification. The improved RNN model works on multiple layers of LSTM rather than one layer of 32 and 64 bits, as illustrated in Table 1. The internal structure of the LSTM contains four thick layers. To calculate the number of learnable parameters correctly with the following equations [28]:

$$\text{Calculate layers} = (h+i)h + Bias = (h+i)*h+h \tag{1}$$

$$\text{Number of parameters} = 4[(h+i)*h+h] \tag{2}$$

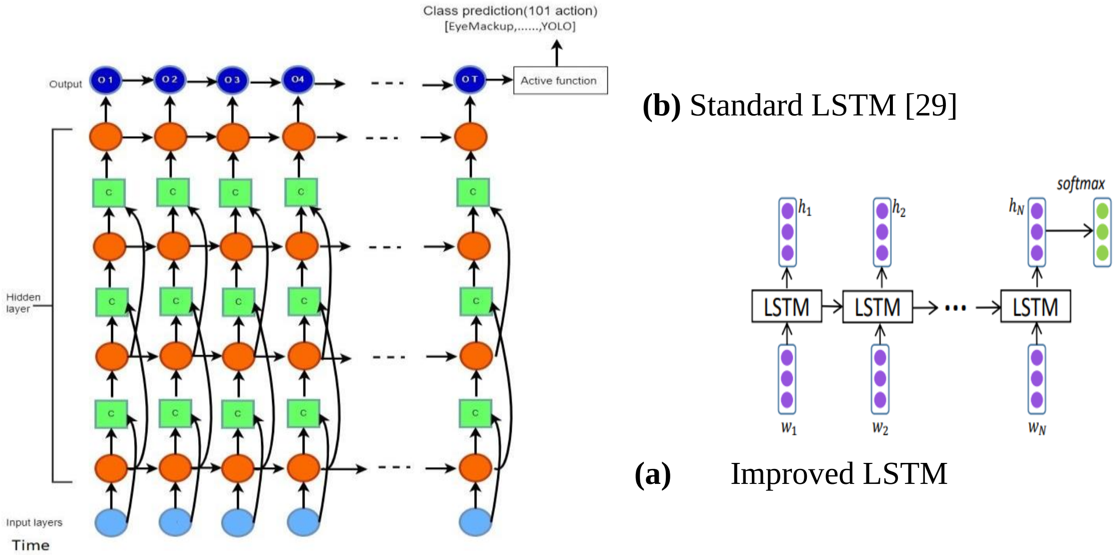
where the input layer is denoted by i and the hidden and output layers are represented by h.

Table 1: Proposed RNN model summary

Layer No.	Layer (type)	Output Shape	Param
1st Layer	lstm_1 (LSTM)	(None, 128, 128)	131584
2nd Layer	dropout_1 (Dropout)	(None, 128, 128)	0
3rd Layer	lstm_2 (LSTM)	(None, 128, 64)	49408
4th Layer	dropout_2 (Dropout)	(None, 128, 64)	0
5th Layer	lstm_3 (LSTM)	(None, 128, 32)	12416
6th Layer	dropout_3 (Dropout)	(None, 128, 32)	0
7th Layer	lstm_4 (LSTM)	(None, 16)	3136
8th Layer	dropout_4 (Dropout)	(None, 16)	0
9th Layer	dense (Dense)	(None, 3)	51
Total params: 196,595 Trainable params: 196,595 Non-trainable params: 0			

The traditional LSTM architecture comprises one layer, which is a kind of RNN that potentially overcomes this constraint because of its particular memory cells, and LSTM outperforms recurrent neural networks in feature extraction of sequence data. The input data first passes through two layers of LSTMs to make it easier to pull out time information from sequence data. Different gates receive the inputs. A sequential model with layers stacked in a straight line is used. The first layer is an LSTM layer with 128 memory units that returns sequences.

This approach makes sure that the subsequent LSTM layer gets sequential sequences of 64, 32, and 16 memory units for each LSTM layer rather than data that is dispersed randomly. To prevent overfitting the model, a dropout layer is included after each LSTM layer. The model was enhanced, as shown in Figures 7 (A) and 8. This is followed by a completely linked layer that uses a SoftMax activation mechanism as the last layer. Figure 8 (B) shows both classic and updated LSTM designs based on recurrent neural networks. When {w1, w2, ..., wN} represents the word vector in a sentence whose length is N, {h1, h2, ..., hN} is the hidden [29].



(b) Standard LSTM [29]

(a) Improved LSTM

Figure 8: Two RNN architectures for action recognition (a) Improved LSTM architecture with four hidden layers (b) Standard LSTM architecture with one hidden layer

Since the traditional RNN system is not very good at predicting, the LSTM architecture is used instead. LSTMs are great for remembering data for long and brief periods. The LSTM blocks have either 3 or 4 gates, which employ the logistic function to ascertain values from 0 to 1. This value decides how much information enters or leaves the memory. The model has an input gate to control the flow of data, a forget gate to control how long memory is kept, and an output gate to control how much data is used to make the block's output [30]. The modified RNN model uses a higher level of LSTM architecture, with four layers compared to just one in the traditional model, as shown in Figure 9.

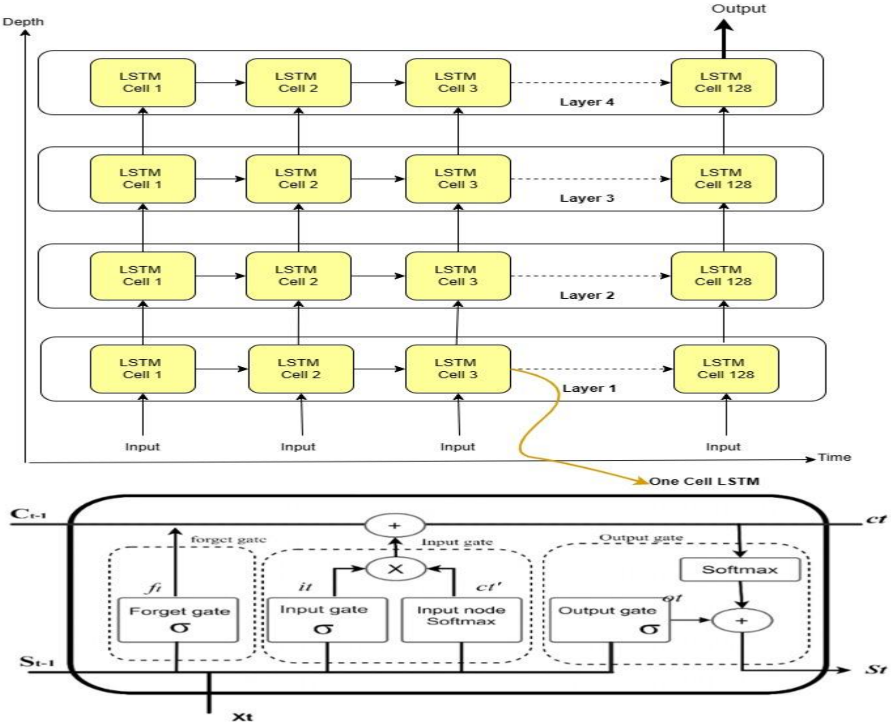


Figure 9: Block representation of the LSTM model

The LSTM network model has been shown to work better than the standard model. It is made up of the following parts:

1-The LSTM model has more layers than it did before. Long-short-term memory algorithms are good at recognizing temporal information in sequential data, but they can be affected by the gradient vanishing problem. By adding more LSTM layers to the model, the neural networks can be trusted, making it more accurately considered a deep learning model. This type of depth has been associated with successful predictions when it comes to human action in videos. The LSTM model is better than basic LSTMs because it has memory cells that help pull features out of sequential data. The model in the study runs the raw data through four layers of LSTM in order to figure out how the sequence data relates to time.

2- Adding more units to every layer of this model can be compared to describing the hidden state or output. Memory cells exist in this model's concealed state. The number of units shows how many neurons are connected to the layer that has the secret state vector and input.

3- By changing the parameters or settings of neural networks, various optimization strategies may be utilized to reduce mistakes. This optimization process involves narrowing the gap between predicted and actual outcomes. To evaluate the consequences of using Adam optimizers, an LSTM model was utilized.

4. Test results

Two publicly available datasets (UCF1-sport and UCF101) are used to evaluate the performance of a modified RNN model based on enhanced LSTM for human activity recognition. The RNN model has been modified by changing the number of layers, filters, and filter size for each model layer. Also, some filters are implemented to remove the noise and enhance the quality of the video's frames for better object detection. Experiments are conducted on Windows 10 (64 bits) with an Intel Core i7-3540m and 16 GB of memory. Finally, compare the proposed model to the state-of-the-art research works that have been recently published.

4.1 Evaluation on recurrent units in improved RNNs

The UCF 101 and UCF Sports datasets are used to train and evaluate the proposed RNN model. The data was distributed as follows: 80% were utilized for training, 20% for testing, and 10% for validation tests using the training data. The validation data and test data are used to monitor the model's performance while it is being trained, and the training data and validation data are used to assess the model's performance. The enhanced model also trains, tests, and validates using the hyperparameters shown in Table 2.

Table 2: The proposed model's hyperparameters

Epoch No.	**10-50**
Batch Size	64
Loss Function	categorical_crossentropy
Optimizer	Adam
Activation Functions	ReLU and SoftMax
Dropout	50%

Different types of recurrent units can considerably increase the complexity and performance of RNNs. LSTM recurrent units have been tested on the UCF101 dataset. Test results show that LSTM units achieve better performance in the UCF101 dataset. According to Table 4, the optimal accuracy is reached with an epoch of 50 and a batch size of 64. Consequently, when the epochs and batch sizes have increased, the model's loss has decreased. Besides, Table 4 also compares the results of the improved model's accuracy with and without the un-sharp mask algorithm. Although the result of the improved model for both datasets has provided different accuracy without data augmentation because data augmentations increase the accuracy of data in training and testing for datasets, as shown in Table 3, the outcomes of accuracy without data augmentation.

Table 3: The model's accuracy without the data augmentation

Model	**Datasets**	**Accuracy**
LSTM	UCF 101	90.6
	UCF Sport	93.2

Table 4: The model's accuracy and loss with and without the unsharp mask technique

No. of Epochs	Batch Size	Test Loss UCF101	Test loss without Unsharp Masking UCF101	Test Accuracy UCF101	Test Accuracy without Unsharp Masking UCF101
10	16	0.501430630683898	0.729885399341583	0.81176471710205	0.741176486015319
	32	0.419181823730468	0.530194699764251	0.87411765336990	0.823529422283172
	64	0.192385643720626	0.382819414138794	0.91294117927551	0.870588243007659
20	16	0.397985076904296	0.464682459831237	0.91764706373214	0.870588243007659
	32	0.497392421960830	0.386070048809051	0.92941176891326	0.884117653369903
	64	0.260645824670791	0.219330394268035	0.92764706373214	0.897647063732147
30	16	0.528318941593170	0.530194699764251	0.91764706373214	0.889000011920929
	32	0.433211743831634	0.538281941413879	0.92000001192092	0.891764717102050
	64	0.384885680675506	0.464682459831237	0.92895294818878	0.895823537826538
40	16	0.458709397315979	0.414638072252273	0.91882353782653	0.902411765336990
	32	0.452400714159011	0.386854112148284	0.92058824300765	0.906411768913269
	64	0.427272200584411	0.365486562252044	0.93031176891326	0.908176474094390
50	16	0.574198842048645	0.345061641931533	0.92941176891326	0.905882358551025
	32	0.4475834667682647	0.341889500617980	0.92764706373214	0.911176474094390
	64	0.400051653385162	0.292526954412460	0.93484706373214	0.911176474094390

Figure 8 displays the accuracy and loss curve for the proposed RNN model when it has been trained using the training data with the same model hyperparameters in the same environment. Among the different outcomes of tuning the set of hyperparameters explained in the improved model, the output shown in Figure 10 has provided the best results in terms of accuracy.

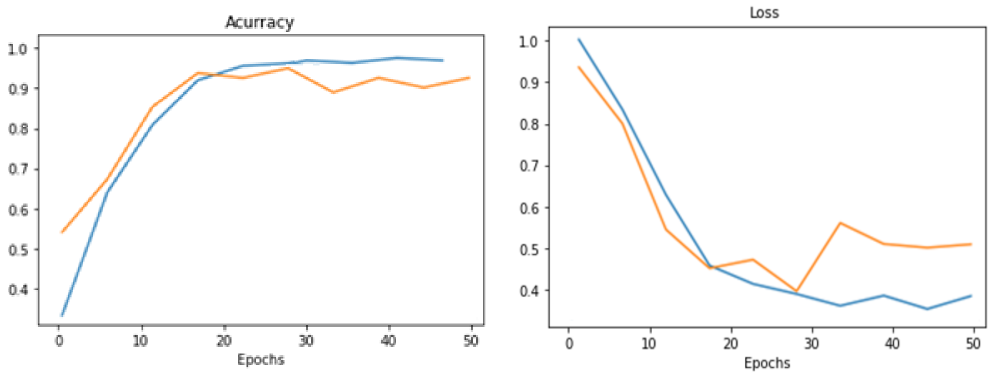


Figure 10: Recurrent neural network accuracy and loss with an unsharp filter on the UCF-101 dataset

4.2 Performance evaluation metrics

A number of performance evaluation metrics may be used to evaluate the performance of the LSTM network classifiers used in HAR. For performance evaluation, four assessment measures are selected: accuracy, precision, recall, and AUC [31] [32].

True Positive (TP): The prediction is positive, and it is actually positive.
True Negative (TN): The prediction is Negative, and it is actually negative.
False Positive (FP): The prediction is Positive, but in fact it is negative.
False Negative (FN): The prediction is Negative, but in fact it is positive.
The accuracy rate shows how many of the original samples may have been correctly predicted.

$$\text{Accuracy} = \frac{(\text{TP}+\text{TN})}{(\text{TP}+\text{TN}+\text{FP}+\text{FN})} \tag{3}$$

where accuracy is proportional to the predicted results. It indicates the proportion of positive samples among those anticipated to be positive. Then, there are two positive prediction possibilities. The first is to predict a positive class as positive (TP). The alternative is to predict the negative class as positive (FP).

$$P\text{recision} = \frac{\text{TP}}{(\text{TP}+\text{FP})} \tag{4}$$

The recall rate is based on the original positive sample. This is the number of positive samples that were accurately predicted. There are two possibilities for results. One is to predict that the initial positive class will remain positive (TP). The alternative is to predict the positive class as a negative class (FN).

$$\text{True Positive Rate/ Recall/ Sensitivity} = \frac{\text{TP}}{(\text{TP}+\text{FN})} \tag{5}$$

Area under the curve of receiver operating characteristics (AUC-ROC) curve is used in this study as a measurement to evaluate the effectiveness of the categorization process. The probability distribution is represented graphically by the ROC curve, and the separability of the model is measured by the AUC, which measures the indicator of separability. The model's classification of human activity is more accurate with a higher AUC value. As a result, the best value achieved for the AUC for the proposed model is 0.93 for the UCF-101 dataset, as shown in Figure 11. Typically, the TPR (true positive rate), recall, or sensitivity are displayed on the y-axis of the ROC curve, while the FPR (false positive rate), or specificity, is shown on the x-axis [31] [32].

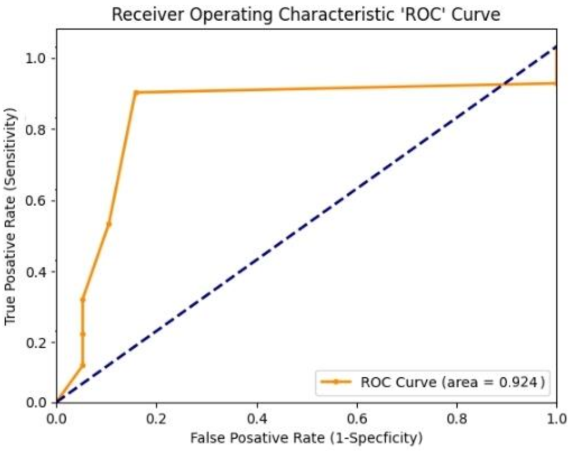


Figure 11: The AUC-ROC curve for the proposed LSTM model

$$Specificity = \frac{TN}{TN+FP} \quad (6)$$

$$FPR = 1 - Specificity \quad (7)$$

4.3 Impact of hidden layers number on the accuracy of improved RNN

Each deep neural network technique has more than one hidden layer that can store various types of data. When the number of layers increases, the overfitting problem decreases. Hence, the overall accuracy of the model reaches its optimal value. Table 5 shows the results of action recognition accuracy with various numbers of hidden layers applied to the UCF-101 dataset.

Table 5: Evaluation of the different numbers of hidden layers (UCF-101 dataset)

Reference	Models	No of hidden layers	Accuracy %
[33]	Convolutional Recurrent Neural Network (ConvLSTM)	One Layer	91.30
[34]	Regularizing long short-term memory (LSTM)	Two Layers	86.90
[33]	Convolutional Recurrent Neural Network (ConvLSTM)	Two Layers	92.40
[33]	Convolutional Recurrent Neural Network (ConvLSTM)	Three Layers	91.10
[35]	Spatio–Temporal Differential LSTM (ST-D LSTM)	Four Layers	74.48
[36]	Deep Bi-directional LSTM with CNN	Five Layers	91.20
[37]	Convolutional Neural Networks	Six Layers	85.1
Proposed LSTM Model	Improved LSTM	Four Layers	93.78

4.4 State-of-the-art comparison

In general, there have been several previous implementations of video classification, particularly for videos of human actions. These works were chosen based on the use of the UCF101 dataset, which is larger and more public, and the RNN architecture, which is more precise. The proposed model has been compared with other state-of-the-art strategies. Some models are tested on UCF 101, whereas others are tested on UCF Sports. 6 shows how the proposed method compares to other deep learning methods in terms of performance. The proposed method got an average accuracy of 93% with four LSTM layers, which is better than other deep learning methods.

Table 6: Performance Accuracy Comparison

Reference	Methodology	Dataset	Accuracy %	Precision %	Recall %
[38]	Hand-crafted	UCF101	83.80	NA	NA
[39]	Support Vector Machine (SVM)	UCF101	86.71	87.28	86.71
[40]	Multi-level Representation for Action Recognition (MoFAP)	UCF101	88.30	NA	NA
[41]	Trajectory Rejection	UCF101	85.70	NA	NA
[42]	Convolution Neural Network	UCF101	90.855	90.848	95.918
[43]	Convolutional Neural Networks	UCF Sports	88.20	NA	NA
[44]	Two-stream Convolutional Neural Networks	UCF101	92.50	NA	NA
[45]	Principal component analysis network (PCANet)	UCF Sports	92.67%	NA	NA
[46]	Convolutional Neural Networks	UCF Sports	87.20	NA	NA
[47]	Convolution Neural Network, Recurrent Neural Networks	UCF Sports	89.10	NA	NA
[48]	Convolutional Neural Networks	UCF101	65.40	NA	NA

[49]	ConvLSTM (Convolutional Long Short-Term Memory) and 3DCNN (3D Convolutional Neural Network)	UCF101	85.2	78.5%	81.2%
[50]	Recurrent Neural Networks	UCF101	89.20	NA	NA
[51]	Sequential convolutional neural network	UCF101	92.00	NA	NA
[52]	Front Door Security (FDS)	UCF101	71.55	70.05	69.02
[53]	Pseudo Recurrent Residual Neural Networks	UCF101	87.60	NA	NA
[54]	Temporal Convolutional Networks (TCNs)	UCF101	84.50	79.21	48.79
[55]	Convolutional Neural Networks + Recurrent Neural Networks	UCF101	92.00	NA	NA
[56]	Convolution Neural Network, Recurrent Neural Networks	UCF 101	89.10	NA	NA
Proposed LSTM Model	Recurrent Neural Network	UCF101	93.78	92.01	92.4
		UCF Sports	95.70	94.40	95.03

5. Conclusions

Human action recognition has been a major area of study in computer vision, with the goal of making algorithms and models that can automatically analyze and understand human actions based on visual data. This study describes practical strategies for classifying human action recognition. The present model is an accurate deep learning technique based on a modified LSTM model. The hidden layers of LSTM deep learning have been modified into four layers. Additionally, several preprocessing algorithms have been applied to the UCF101 and UCF Sports datasets to reduce the overfitting problem using different transformations such as translation, scaling, and rotation. According to the results, the Adam optimizer with a SoftMax activation function is an effective way to optimize an LSTM model. Data augmentation, which employs transformations such as translation, scaling, and rotation, has also been used to mitigate the overfitting problem. In this research, an enhanced RNN model is provided in which the LSTM deep learning technique's hidden layers have been expanded from two to four. Tuning the model's parameters allows us to examine the impact of pre-processing and model design. The results of experiments show that the improved model is more accurate at recognizing actions than previous deep learning methods. Table 6 compares the model to state-of-the-art studies that have used conventional, machine, and deep learning approaches and shows that the suggested methodology obtains a superior accuracy rate with modified LSTM utilizing four hidden layers. The results of the experiments indicate that the proposed approach improves accuracy. The total model accuracy of the enhanced RNN model is 93.78% for the UCF-101 dataset and 95.70% for the UCF Sport datasets. The model can be effectively fused with another deep-learning algorithm. Also, the proposed model is enhanced to classify human actions throughout the video more reliably. In terms of classification and recognition, the improved model must be used with VGG-16, Inception-V4, and others in future research.

Conflicts of Interest

Each author confirms that there is no financial or personal bias in their work.

References

[1] R. Poppe, “A survey on vision-based human action recognition,” *Image and Vision Computing*, vol. 28, no. 6, pp. 976–990, 2010. doi:10.1016/j.imavis.2009.11.014

[2] Y. Guo, X. Hu, B. Hu, J. Cheng, M. Zhou, and R. Y. K. Kwok, “Mobile Cyber Physical Systems: Current Challenges and Future Networking Applications,” *IEEE Access*, vol. 6, no. c, pp. 12360–12368, 2017, doi: 10.1109/ACCESS.2017.2782881.

[3] M. Khamees, I. Mishkhal, and H. H. Saleh, “Enhancing the accuracy of Health Care Internet of medical things in real time using CNNets,” *Iraqi Journal of Science*, vol. 62, no. 11, pp. 4158–4170, 2021. doi:10.24996/ijs.2021.62.11.34

[4] S. Kansal, S. Purwar, and R. K. Tripathi, “Image contrast enhancement using unsharp masking and histogram equalization,” *Multimed. Tools Appl.*, vol. 77, no. 20, pp. 26919–26938, 2018, doi: 10.1007/s11042-018-5894-8.

[5] K. Soomro, A. R. Zamir, and M. Shah, “UCF101: A Dataset of 101 Human Actions Classes From Videos in The Wild,” arXiv:1212.0402, 2012. [Online] Available: https://arxiv.org/pdf/1212.0402.pdf

[6] Y. Nan et al., “Deep learning for activity recognition in older people using a pocket-worn smartphone,” *Sensors*, vol. 20, no. 24, p. 7195, 2020. doi:10.3390/s20247195

[7] O. Elharrouss, N. Almaadeed, and S. Al-Maadeed, “A review of Video Surveillance Systems,” *Journal of Visual Communication and Image Representation*, vol. 77, p. 103116, 2021. doi:10.1016/j.jvcir.2021.103116

[8] E. P. Ijjina and C. K. Mohan, “Hybrid deep neural network model for human action recognition,” *Appl. Soft Comput. J.*, vol. 46, pp. 936-952, Spt. 2016, doi: 10.1016/j.asoc.2015.08.025.

[9] A. Wria and A. A. Mohammed, “Review on recent Computer Vision Methods for Human Action Recognition,” *Advances in Distributed Computing and Artificial Intelligence Journal*, vol. 10, no. 4, pp. 361–379, 2021.

[10] I. Jegham, A. Ben Khalifa, I. Alouani, and M. A. Mahjoub, “Vision-based Human Action Recognition: An overview and real world challenges,” *Forensic Science International: Digital Investigation*, vol. 32, p. 200901, 2020. doi:10.1016/j.fsidi.2019.200901

[11] A. Muhamad and A. Mohammed, “A comparative study using improved LSTM /GRU for human action recognition,” *Journal of Theoretical and Applied Information Technology*, vol. 101, no. 5, pp. 1863–1879, 2022. doi:10.21203/rs.3.rs-2380406/v1

[12] M. A. Quaid and A. Jalal, “Wearable sensors based human behavioral pattern recognition using statistical features and reweighted genetic algorithm,” *Multimedia Tools and Applications*, vol. 79, no. 9–10, pp. 6061–6083, 2019. doi:10.1007/s11042-019-08463-7

[13] M. Al-Faris, J. Chiverton, D. Ndzi, and A. I. Ahmed, “A review on computer vision-based methods for human action recognition,” *Journal of Imaging*, vol. 6, no. 6, p. 46, 2020. doi:10.3390/jimaging6060046

[14] M. M. Hossain Shuvo, N. Ahmed, K. Nouduri, and K. Palaniappan, “A hybrid approach for human activity recognition with support vector machine and 1d Convolutional Neural Network,” *2020 IEEE Applied Imagery Pattern Recognition Workshop (AIPR)*, 2020. doi:10.1109/aipr50011.2020.9425332

[15] F. Afza et al., “A framework of human action recognition using length control features fusion and weighted entropy-variances based feature selection,” *Image and Vision Computing*, vol. 106, p. 104090, 2021. doi:10.1016/j.imavis.2020.104090

[16] H. Mohammadzade, S. Hosseini, M. R. Rezaei-Dastjerdehei, and M. Tabejamaat, “Dynamic time warping-based features with class-specific joint importance maps for action recognition using Kinect depth sensor,” *IEEE Sensors Journal*, vol. 21, no. 7, pp. 9300–9313, 2021. doi:10.1109/jsen.2021.3051497

[17] R. Huang, C. Chen, R. Cheng, Y. Zhang, and J. Zhu, “Human action recognition based on three-stream network with frame sequence features,” *2022 7th International Conference on Image, Vision and Computing (ICIVC)*, 2022. doi:10.1109/icivc55077.2022.9887162

[18] S. K. Abaas and L. E. George, “The performance differences between using recurrent neural networks and feedforward neural network in sentiment analysis problem,” *Iraqi Journal of Science*, vol. 61, no. 6, pp. 1512–1524, Jun. 2020, doi:10.24996/ijs.2020.61.6.31

[19] S. Ghimire et al., “Streamflow prediction using an integrated methodology based on convolutional neural network and long short-term memory networks,” *Scientific Reports*, vol. 11, Article number: 17497, Spt. 2021. doi:10.1038/s41598-021-96751-4

[20] W. Chu, H. Xue, C. Yao, and D. Cai, “Sparse coding guided spatiotemporal feature learning for abnormal event detection in large videos,” *IEEE Transactions on Multimedia*, vol. 21, no. 1, pp. 246–255, 2019. doi:10.1109/tmm.2018.2846411

[21] N. Crasto, P. Weinzaepfel, K. Alahari, and C. Schmid, “Mars: Motion-augmented RGB stream for action recognition,” *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. doi:10.1109/cvpr.2019.00807

[22] J. Wensel, H. Ullah, and A. Munir, “VIT-ret: Vision and recurrent transformer neural networks for human activity recognition in videos,” *IEEE Access*, vol. 11, pp. 72227-72249, 2023, doi: 10.1109/ACCESS.2023.3293813

[23] E. A. Al-Zubaidi and M. M. Mijwil, “Medical Image Classification for coronavirus disease (COVID-19) using Convolutional Neural Networks,” *Iraqi Journal of Science*, vol. 62, no. 8, pp. 2740–2747, Aug. 2021. doi:10.24996/ijs.2021.62.8.27

[24] A. Rehmer and A. Kroll, “On the vanishing and exploding gradient problem in gated recurrent units,” *IFAC-PapersOnLine*, vol. 53, no. 2, pp. 1243–1248, 2020. doi:10.1016/.ifacol.2020.12.1342

[25] P. B. Weerakody, K. W. Wong, G. Wang, and W. Ela, “A review of irregular time series data handling with gated recurrent neural networks,” *Neurocomputing*, vol. 441, pp. 161–178, 2021.

[26] Y. Tian and L. Pan, “Predicting short-term traffic flow by long short-term memory recurrent neural network,” *2015 IEEE International Conference on Smart City/SocialCom/SustainCom (SmartCity)*, 2015. doi:10.1109/smartcity.2015.63

[27] S. Tiwari et al., “A smart decision support system to diagnose arrhythmia using ensembled ConvNet and ConvNet-LSTM model,” *Expert Systems with Applications*, vol. 213, part A, p. 118933, 2023. doi:10.1016/j.eswa.2022.118933

[28] R. Karim, “Counting no. of parameters in deep learning models by hand,” Medium, https://towardsdatascience.com/counting-no-of-parameters-in-deep-learning-models-by-hand-8f1716241889 (accessed Jul. 20, 2023).

[29] Y. Wang, M. Huang, L. Zhao, and X. Zhu, “Attention-based LSTM for aspect-level sentiment classification,” *EMNLP 2016 - Conf. Empir. Methods Nat. Lang. Process. Proc.*, pp. 606–615, 2016, doi: 10.18653/v1/d16-1058.

[30] F. Aksan, Y. Li, V. Suresh, and P. Janik, “CNN-LSTM vs. LSTM-CNN to predict power flow direction: A case study of the high-voltage subnet of Northeast Germany,” *Sensors*, vol. 23, no. 2, p. 901, 2023. doi:10.3390/s23020901

[31] S. Mekruksavanich and A. Jitpattanakul, “LSTM networks using smartphone data for sensor-based human activity recognition in Smart Homes,” *Sensors*, vol. 21, no. 5, p. 1636, 2021. doi:10.3390/s21051636

[32] R. S. Abdul Ameer and M. Al-Taei, “Human action recognition based on bag-of-words,” *Iraqi Journal of Science*, vol. 61, no. 5, pp. 1202–1214, May 2020. doi:10.24996/ijs.2020.61.5.27

[33] W. Ye, J. Cheng, F. Yang, and Y. Xu, “Two-Stream Convolutional Network for Improving Activity Recognition Using Convolutional Long Short-Term Memory Networks,” *IEEE Access*, vol. 7, pp. 67772–67780, 2019, doi: 10.1109/ACCESS.2019.2918808.

[34] B. Mahasseni and S. Todorovic, "Regularizing Long Short Term Memory with 3D Human-Skeleton Sequences for Action Recognition," *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, NV, USA, 2016, pp. 3054-3062, doi: 10.1109/CVPR.2016.333.

[35] D. Avola, M. Cascio, L. Cinque, G. L. Foresti, C. Massaroni, and E. Rodola, “2-D Skeleton-Based Action Recognition via Two-Branch Stacked LSTM-RNNs,” *IEEE Trans. Multimed.*, vol. 22, no. 10, pp. 2481–2496, 2020, doi: 10.1109/TMM.2019.2960588.

[36] A. Basnet and A. K. Timalsina, “Improving Nepali News Recommendation Using Classification Based on LSTM Recurrent Neural Networks,” *Proc. 2018 IEEE 3rd Int. Conf. Comput. Commun. Secur. ICCCS 2018*, pp. 138–142, 2018, doi: 10.1109/CCCS.2018.8586815.

[37] Cheng Wang, Haojin Yang, and C. Meinel, “Exploring multimodal video representation for action recognition,” *2016 International Joint Conference on Neural Networks (IJCNN)*, Vancouver, BC, 2016, pp. 1924-1931, doi: 10.1109/IJCNN.2016.7727435.

[38] H. Wang, A. Kläser, C. Schmid, and C. L. Liu, “Action recognition by dense trajectories,” *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, Colorado Springs, CO, USA, 2011, pp. 3169-3176, doi: 10.1109/CVPR.2011.5995407.

[39] R. A. Minhas, A. Javed, A. Irtaza, M. T. Mahmood, and Y. B. Joo, “Shot classification of field sports videos using AlexNet Convolutional Neural Network,” *Applied Sciences*, vol. 9, no. 3, p. 483, 2019. doi:10.3390/app9030483

[40] L. Wang, Y. Qiao, and X. Tang, “MoFAP: A Multi-level Representation for Action Recognition,” *Int. J. Comput. Vis.*, vol. 119, no. 3, pp. 254–271, 2016, doi: 10.1007/s11263-015-0859-0.

[41] J. J. Seo, H. Il Kim, W. De Neve, and Y. M. Ro, “Effective and efficient human action recognition using dynamic frame skipping and trajectory rejection,” *Image Vis. Comput.*, vol. 58, pp. 76–85, 2017, doi: 10.1016/j.imavis.2016.06.002.

[42] R. A. Minhas, A. Javed, A. Irtaza, M. T. Mahmood, and Y. B. Joo, “Shot classification of field sports videos using AlexNet Convolutional Neural Network,” *Applied Sciences*, vol. 9, no. 3, p. 483, 2019. doi:10.3390/app9030483

[43] S. Yu, Y. Cheng, L. Xie, Z. Luo, M. Huang, and S. Li, “A novel recurrent hybrid network for feature fusion in action recognition,” *J. Vis. Commun. Image Represent.*, vol. 49, pp. 192–203, 2017, doi: 10.1016/j.jvcir.2017.09.007.

[44] Y. Zhao, K. L. Man, J. Smith, K. Siddique, and S. U. Guan, “Improved two-stream model for human action recognition,” *Eurasip J. Image Video Process.*, vol. 2020, no. 24, 2020, doi: 10.1186/s13640-020-00501-x.

[45] Abdelbaky, A. and Aly, S. ,“Human action recognition using three orthogonal planes with unsupervised deep convolutional neural network*,” Multimedia Tools and Applications*, vol. 80, no. 13, pp. 20019–20043, 2021, doi.org/10.1007/s11042-021-10636-2.

[46] F. An, “Human Action Recognition Algorithm Based on Adaptive Initialization of Deep Learning Model Parameters and Support Vector Machine,” *IEEE Access*, vol. 6, pp. 59405–59421, 2018, doi: 10.1109/ACCESS.2018.2874022.

[47] S. Nazir, M. H. Yousaf, J.-C. Nebel, and S. A. Velastin, “A bag of expression framework for improved human action recognition,” *Pattern Recognition Letters*, vol. 103, pp. 39–45, 2018. doi:10.1016/j.patrec.2017.12.024

[48] C. Feichtenhofer, A. Pinz, and R. P. Wildes, “Spatiotemporal multiplier networks for Video action recognition,” *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR),* Honolulu, HI, USA, 2017, pp. 7445-7454. doi:10.1109/cvpr.2017.787

[49] R. Vrskova, R. Hudec, P. Kamencay, and P. Sykora, “Human activity classification using the 3DCNN architecture,” *Applied Sciences*, vol. 12, no. 2, p. 931, Jan. 2022, doi: 10.3390/app12020931. [Online]. Available: http://dx.doi.org/10.3390/app12020931

[50] Leng, B. et al., “A 3D model recognition mechanism based on Deep Boltzmann machines,” *Neurocomputing*, 151, part 2, pp. 593–602. Available at: https://doi.org/10.1016/j.neucom.2014.06.084.

[51] H. Yang, C. Yuan, J. Xing, and W. Hu, “SCNN: Sequential convolutional neural network for human action recognition in videos,” 2017 IEEE International Conference on Image Processing (ICIP), Beijing, China, 2017, pp. 355-359, doi: 10.1109/ICIP.2017.8296302

[52] L. P. Paula et al., “A novel front door security (FDS) algorithm using GoogleNet-BiLSTM hybridization,” *IEEE Access*, vol. 11, pp. 19122–19134, 2023. doi:10.1109/access.2023.3248509

[53] L. I. N. Liu, “Learning Long-Term Temporal Features With Deep Neural Networks for Human Action Recognition,” *IEEE Access*, vol. 8, pp. 1840–1850, 2020, doi: 10.1109/ACCESS. 2019 .2962284.

[54] C. Lea, M. D. Flynn, R. Vidal, A. Reiter, and G. D. Hager, “Temporal convolutional networks for action segmentation and detection,” *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR),* Honolulu, HI, USA, 2017 pp. 1003-1012. doi: 10.1109/CVPR.2017.113

[55] J. J. Seo, H. Il Kim, W. De Neve, and Y. M. Ro, “Effective and efficient human action recognition using dynamic frame skipping and trajectory rejection,” *Image Vis. Comput.*, vol. 58, pp. 76–85, 2017, doi: 10.1016/j.imavis.2016.06.002.

[56] L. Wang, Y. Xu, J. Cheng, H. Xia, J. Yin and J. Wu, "Human Action Recognition by Learning Spatio-Temporal Features With Deep Neural Networks," *IEEE Access*, vol. 6, pp. 17913-17922, 2018, doi: 10.1109/ACCESS.2018.2817253.