International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING
ISSN:2147-6799 www.ijisae.org Original Research Paper

Identifying Complex Human Actions with a Hierarchical Feature Reduction and Deep Learning-Based Approach

Lakshmi Alekhya Jandhyam[*1], Ragupathy Rengaswamy[2] and Narayana Satyala[3]

Submitted: 06/02/2024 **Revised**: 14/03/2024 **Accepted**: 20/03/2024

Abstract: Among computer vision's most appealing and useful research areas is automated human activity recognition. In these systems, look and movement patterns in video clips are used to classify human behaviour. Nevertheless, the majority of previous research has either ignored or failed to employ time data to predict action identification in video sequences using standard techniques and classical neural networks. On the other hand, reliable and precise human action recognition requires a significant processing cost. To get over the challenges of the pre-processing stage, in this work, we choose a sample of frames at random from the input sequences. We only take the most noticeable elements from the representative frame rather than the entire set of attributes. We suggest a hierarchical approach in which bone modelling and a deep neural network are used first, followed by background reduction and HOG. For selecting features and historical data retention, a CNN along with LSTM recurrent network combo is taken into consideration; in the end, a SoftMax-KNN classification is employed to detect the human behaviours. The name of our model is represented by the abbreviation HFR-DL, which stands for a hierarchical Features Lowering & Deep Learning-based action detection approach. We utilize the UCF101 dataset, which is popular among action recognition researchers, for benchmarking in order to assess the suggested approach. There are 101 challenging tasks in the wild included in the dataset. When comparing the experimental results with eight cutting-edge methods, significant improvements in speed and accuracy are seen.

Keywords: *Identification of human actions, Deep neural networks, Oriented gradient histogram, HOGG, Skeletal model, Extracting features, Time-spatial data.*

1. Introduction

Despite the current activity in the field of human activity recognition, or HAR, there are still crucial considerations that must be made in order to comprehend how individuals engage with digital devices or with one another [11, 12, 63]. Recognising human activity involves a series of intricate sub-actions. Many academics from all around the world have recently looked into this utilizing various kinds of sensors. With constantly increasing needs in numerous industries, computer vision-based automatic recognition of human actions has proven more successful than in previous years. Healthcare systems, driver assistance programmes, autonomous cars, smart home activity monitoring, aforementioned security and environmental monitoring, intelligent meeting spaces, home automation, PDAs, and entertainment settings are a few examples of these that can alert the appropriate authorities to criminal or terrorist activity. Amidst the COVID-19 epidemic, they provide novel obstacles with the monitoring of social distancing measures [33].

Generally speaking, we may use a variety of sensors, including cameras and wearable sensors, to get the necessary data from a specific individual [1,39,58]. For real-time safety applications (such as intruder identification), cameras are excellent sensors. You may find tasks involving turning, backward or forward motion, and resting postures by concentrating on certain areas of the movie. For numerous researchers, the idea of motion and action recognition in clips from videos is an extremely intriguing and difficult study subject. Walking action identification using wearables and computer vision may be hindered by the sensor devices' visual limitations. Therefore, it's likely that insufficient data exist to account for the reasons behind objects' or individuals' movements [39, 42, 56].

On the other hand, video recording itself may be greatly impacted by light, visibility, size, and orientation, whereas advanced action identification using computer vision requires a very high computing cost [50]. Lowering the computing cost of a gesture identification system necessitates that it be able to detect the subjects' activities from little data because it is mostly reviewed online and must be examined in real time [16].

The person's physical attitude is represented by a sequence of directed rectangles in human posture evaluation; as a consequence, valuable frames and frame index knowledge may be employed. A histogram is produced by combining the locations and orientations of rectangles to provide a state representation for every frame. The backdrop is referred to

[*1,2] *Department of Computer Science and Engineering, Annamalai University, Annamalainagar, Tamil Nadu, INDIA*
[*1]*ORCID ID : 0000-0001-9880-7336*
[2]*ORCID ID : 0000-0001-7148-0345*
[3] *Department of Computer Science and Engineering, Seshadri Rao Gudlavalleru Engineering College, Andhra Pradesh, INDIA*
ORCID ID : 0000-0003-0227-6273
* *Corresponding Author Email: lakshmialekya22@gmail.com*

as the offset in background removal techniques (BGS). Action recognition systems based on films may perform better when techniques like the motion border the histogram (MBH), histograms of optical flow (HOF), and the previously described histogram, which of gradients with orientation (HOG) are used [30, 53]. The hands and arms of a person can be represented by a skeleton model, which can be used to classify human activities [17,30,40]. A variety of machine learning techniques, each with unique advantages, disadvantages, and capabilities, have been put forth to handle the aforementioned issues in action recognition.

By mixing a range of sections and constraints, convolutional neural networks, also known as (CNNs), a sort of deep-connected neural network, effectively classify the objects [51]. Some activity detection issues can be solved with the use of recurrent neural networks (RNN). In actuality, RNN recursive loops retain the information acquired from previous occurrences. The RNN's limited memory for previous steps is one of its drawbacks. Consequently, LSTM was created in order to safeguard data from multiple subsequent phases [8]. In theory, RNNs must be able to manage two common issues with future dependencies: disappearing gradients and exploding gradients; nonetheless, LSTM performs better in both situations [1,14,23,50].

We go into further detail and offer additional specifics about our concepts in the parts that follow. The following portions of this essay are arranged as follows: The most pertinent research on the subject is reviewed in the section that follows. The suggested approach and protocols are covered in depth in the section that follows. The experiment and assessment outcomes will be reviewed and compared with eight cutting-edge techniques in the last portion. The article is concluded in the last part, which also offers ideas for more research.

2. Literature Review

Over the past 10 years, several academics from diverse fields have shown interest in Human Action Recognition (HAR) due to its wide range of applications. The majority of currently used techniques rely on manually created characteristics, and deep neural networks are also able to identify people's movements in real-time movies because of advancements in GPU and extended memory technology. Human activity identification in a sequence of image frames is one field of automated vision research that focuses on precisely recognizing actions performed by humans from single-view photographs [44,50]. Using traditional manual methods, the video signal sequences were first processed to extract the low-level features linked to a particular action. Following that, a classifier such as K-means, decision trees, K-nearest neighbour (KNN), Support Vector machines (SVM), or Hidden Markov Models (HMMs) was used to identify these characteristics [50, 64]. Handcraft-based approaches are used to extract and display the characteristics; nevertheless, in order to locate and define features, descriptors, and processes for generating a dictionary, an expert is required. Several deep learning techniques have also used traditional hand-crafting techniques for the categorization of images, recognizing objects, HAR, or sound recognition, but they perform better than with conventional approaches [38].

The authors of [50] attempted to extract pertinent characteristics for action identification by analyzing every six frames of the input video clips using a pre-trained AlexNet Network. Using an in-depth LSTM that has two forward and reversed layers of data, this approach finds and extracts valuable details from a series of video frames.

In [38], features are extracted using an already trained deeper CNN, and actions are identified using a mix of SVM as well as KNN classifiers. A brief training dataset may be given to an action detection model that has already undergone considerable pre-training on a large annotation dataset. Consequently, deep neural network learning via transmission might be a helpful model-training strategy when the database volume is constrained.

An extended LSTM variant was presented in [25] elements known to be C2 LSTM is shown, allowing for the feeling of temporal interdependence, spatial features, and moving data. They created a special deep network architecture for HAR by using the movement and geographical characteristics of the video input. We used HMDB51 & UCF101 to verify the brand-new network.

[57] Offers a unique Reaffirming each other using Spatio-Temporal Convolution-based Tunnel (MRST) for Handheld Augmentation. The model minimizes the complexity of structure by emphasizing relevant spatial gaze and immediate mobility while streamlining the framework. It breaks down information in three dimensions into simultaneous temporal and spatial depictions, which are then jointly strengthened, using the interplay of geographical and temporal data.

Increased dataset size prevents overfitting, however providing a significant amount of tagged data is costly and challenging. In these cases, learning by transfer generates a lot of sense The method suggested in [38] seeks to use a successful, trained model to construct an innovative architecture.

In certain research, acknowledgment of hand gestures and human behaviors are studied using 3-D information sequences of the entire body and bones. Additionally, A learning-based method that combines CNN and LSTM is used to handle challenges related to three-dimensional temporal identification and posture recognition [50]. After putting out a structure for background separation (BGS) and

a feature retrieval tool, Singh et al. [44] ultimately used HMMs for action identification.

Using a range of extraction of features and fusion techniques on the UCF dataset, a method for movement detection is demonstrated in [30]. The study presents six different fusion models, all of which are inspired by the transitional, late, and early stages of fusion [34]. In the first two models, the system takes use of a preliminary fusion process. The third and fourth designs both make use of intermediate fusion techniques. The system is shown in the fourth model, which employs a kernel-based SVM classification system in a kernel-based fusion technique. The 6th and just fifth designs depict late fusion procedures.

[64] Completed the processing of just a single image inside a temporal neighbourhood utilizing a two-dimensional multilayer architecture to gather appearance attributes of the input frames. Nevertheless, merely aggregating scores is insufficient to depict historical connections across remote frames.

To improve the belief formed from a single picture, they integrate the feature drawings of distant photographs into a three-dimensional framework specifically for complicated long-term responsibilities. This structure estimates the time interval between the frames.

In their work, [32] introduced an effective and adaptive information transfer method known as R-NKTM, designed to identify human movements from various, unpredictable viewing angles. Using a fully linked neural network made up of neurons, the suggested R-NKTM finds a disorganized artificial path that connects different viewpoints. With this functionality, data created by humans may be easily moved from any unknown perspective to an alone, fictitious, high-level view. As a movement capture system, the R-NKTM trains on complex patterns of artificial three-dimensional human characters and then applies the learned behavior to actual human action scenes. Notably, the suggested approach involves training a single R-NKTM for all action outcomes and viewing angles, obviating the need for model refinement or retraining when transmitting information from any human behavior footage.

A probabilistic methodology for determining the dynamic information connected to a human stance is presented in [21]. The model estimates the test sample density in order to design a data-driven strategy. From the statistical conclusion drawn on the predicted density, they may derive values of attention, such the probability of the human's eventual movement and the amount of information about motion a pose carries.

Suggests ReHAR, a unique, reliable, and effective human activity identification system that may be used to forecast group activities as well as single-person activities. For every video frame, they first create an optical flow picture. Subsequently, the recreated model is fed with the real video segments together with the associated visual flow pictures to produce representations. Lastly, depending on the created representations, an LSTM network is employed to forecast the next tasks.

3. Methodology

Several subsections make up this section because the approach consists of several steps and sub-modules. The initial learning stage is followed by the transfer acquisition phase, and finally the general understanding of the model's architecture is given.

3.1. Model Architecure

FIGURE 1 depicts the architecture of the suggested technique known as Depth Learning with Organisational Feature Filtering (HFR-DL). The material that goes into the learning process and the output are the three main components of the proposed system. The KNN & SoftMax as its input layer are constructed as the separate modules for human behaviour classification after the CNN-LSTM model, which serves as the deep computational module, is first established. The learning technique module, also called the feature diminished module, is composed of three components: the human bone skeleton (BGS-HOG-SKE), the surroundings removal, and the histogram of transitions with orientation. The approach also makes use of AlexNet, for short & the UCF101 data collection. One is an extensive collection of complex video, and the second is a trained beforehand system to increase the efficiency of the device's motion detection. As an internet-based hub for transferable learning, we utilize AlexNet [50].

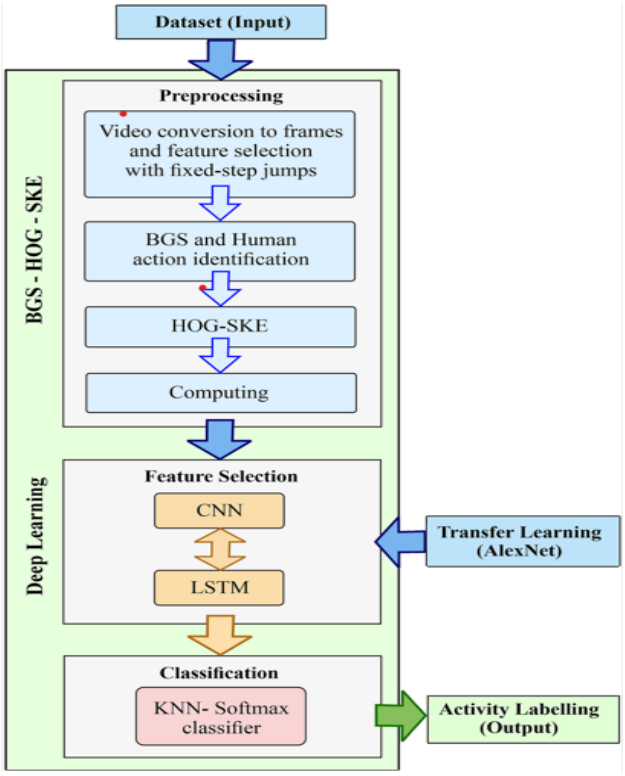
The three components that make up the action recognition component are classifications, feature selection, and preprocessing. Converting the video clips into a series of frames is the preprocessing stage. Just a few frames are used for the processes, though, which may improve both performance and cost. Optimized weight features are chosen using two deep CNNs and LSTM neural networks. In contrast to earlier approaches that did not rely on deep learning, the parameters are fine-tuned after being trained on a range of datasets.

When compared to other deep neural network models, the primary benefit of RNNs and deep LSTM in complicated action identification is their greater accuracy rate, which will be demonstrated later in "Experimental Results". The output as an action is labelled and classified using two SoftMax and KNN algorithms in the "KNN-SoftMax Classifier." The instructional phase of the developed movement detection system is followed by a system-wide test and efficiency analysis stage that determines the accuracy and inaccuracy of the system. Further information is given in "Experimental Results". Prior to delving deeper into the technical aspects, we go over the common symbols

used in the parts that follow. The notations and symbols used in this article are described in Table 1.

Table 1. The explanation of each symbol

Symbols	Descriptions
J	Jump length in input frames
N_F	Number of representative frames
f_k	k^{th} representative frame
$v_i(u)$	Value of pixel u in block i
$NG_i(u)$	Spatial neighbourhood of pixel u in block i
$sk = \{f_1, f_2, ..., f_{N_F}\}$	The sequence of skeleton with N_F frames
$p^l = [p_1, ..., p_l]$	Output feature vector of the deep network and input of classifications
V	Video activity
TF(.)	Conversion function



2. Module wise architecture of the HAR system

3.2. Learning Phase

Features picking, categorizing, and preparation are the three stages that make up the learning phase (refer to Fig. 1). The highly sensitive stage of preprocessing affects the performance of the model and may increase the accuracy of the HAR output. Further actions and information will be provided in the next subsections.

3.3. Preprocessing

The preprocessing step involves turning the input films into a series of frames, as seen in Fig. 2. The sample frames will then be chosen from the provided frame sequences. In this investigation, we used the BGS approach to eliminate the background from typical frames (bottom row, Fig. 2). Subsequently, the representative frames undergo the deep and skeleton approach, wherein motion depictions are expressly generated from the source frames via depth motion mapping. Below, we provide a step-by-step explanation of each of the four pre-processing stages:

(1) Making a film into frames and choosing frames At the beginning, the provided videos are converted into a set of frames [2], where each frame is portrayed by a matrix, as demonstrated by Eq. 1,

$$f_k = \begin{bmatrix} f_{11} & f_{12} & \cdot & \cdot & f_{1m} \\ f_{21} & f_{22} & \cdot & \cdot & f_{2m} \\ \cdot & \cdot & f_{ij} & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ f_{n1} & f_{n2} & \cdot & \cdot & f_{nm}, \end{bmatrix} \quad (1)$$

Where f_k, with n rows and m columns, is the kth representative frame. The feature values (intensities of each pixel) for the matching frame k are denoted by f_{ij}. The process of translating footage to frames involves a substantial computational expenditure, which leaves us with a huge number of still images and pixels that lower the overall effectiveness of the entire system. To address the problem, we provide an easy-to-implement but efficient method for getting rid of the unnecessary photos. Fixed-step jumps J can be used to perform this in order to remove comparable consecutive frames [52]. Our testing indicates that choosing one frame out of every six will greatly increase system speed without appreciably lowering quality.

Later on in "Experimental Results," we go into further depth about this. Therefore, just NF frames [17,20,30,40] were chosen for feature extraction rather than all features from all frames. As a result, our CNN network operates more effectively for the upcoming stages.

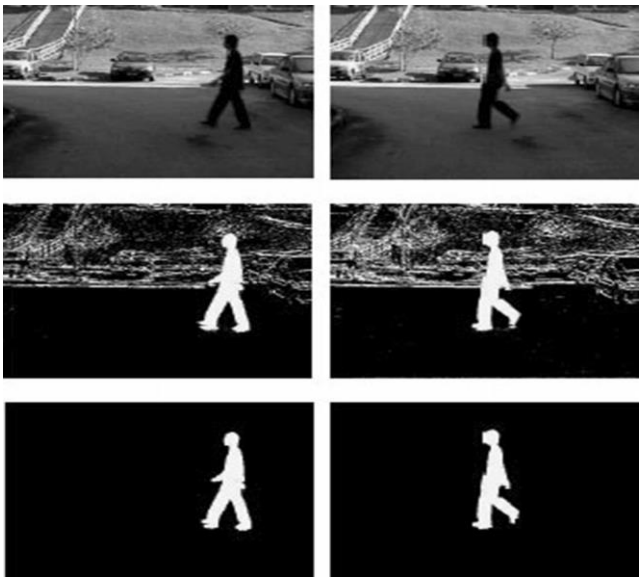


Fig. 2. A sample representation of BGS technique to pre-process the extracted frames from a video

(2) BGS and identification of human actions: BGS, statistical approaches, temporal differencing, and optical flow comprise most of the moving object recognition methods. To identify foreground items, we employ methods similar to background modelling [61]. Backing subtraction-based algorithms have been used to identify moving components in a video series, preserving a single the context within which model is constructed from previous frames. [6].

The BGS scheme is a widely used technique for separating moving portions of a picture into background and foreground, and it may be applied both indoors and outdoors [2,7,48]. Once the pixels have been extracted from the scene's static background, the areas may be categorized into groups, like human groups. The classification approach is based on comparing the silhouettes of the recognized things with pre-labeled patterns in the object silhouette database. To create the template database, examples of object outlines from video samples are collected and suitably classified. After that, a contour-based tracking method is used to get the object areas' outlines using the forefront pixel-map [2, 7]. The BGS stages are explained in [55], where the temporal probability of the pixels that surround the representative frame of the video material are assumed to be equal. The parameters $V_i(u)$ & $NG_i(u)$ compute the values and positional surrounds of a pixel at u within the ith chunk of the picture. As a result, Eq. 2 makes it clear that v, which is randomly selected in $NG_i(u)$ (the indicative frame), has precisely the same value as the portion of the sample that surrounds every Pixel u with bi,j(u):

$$b_{i,j}(u) = v(u|u \in NG_i(u)) \quad j = 1, 2, ..., l. \tag{2}$$

The backdrop model of every pixel in the ith block may then be used to initialize Ai, The backdrop model associated with pixel u is:

$$A_i = [\{b_{i,1}(u)\}\{b_{i,2}(u)\}, ..., \{b_{i,l}(u)\}], \quad u \in \text{Block } i. \tag{3}$$

Using this method, the foregrounds of certain frames from short video series or from integrated gadgets with limited memory and computing capacity may be retrieved. Furthermore, small yet effective data sizes are chosen since excessively high data sizes might destroy statistical association among the pixels at various places. You may get more details on following the foreground extraction procedures in [55]. This will also lessen the difference in the strength among each of the pixels in the present image and its matching value in the surrounding reference image.

Fig. 2 displays an example of a walking BGS step sequence. Human form is crucial for identifying human activity, which may be used to extract blobs from BGS, as the centre and bottom rows of Figure 2 demonstrate. The shape of humans in a scene may be represented using a variety of techniques that are based on global characteristics, boundaries, and skeletal traits [48]. A number of noises may go away after applying the BGS, although additional noises may still emerge in different areas [15,46]. We employ erosion and dilation morphological operators, utilizing 3 × 3 structural components, to eliminate these artefacts. The diagnostic data required to characterise the human silhouette is established during the feature extraction stage. Broadly speaking, we can state that BGS preserves the most significant portions of the embedded information while extracting valuable characteristics from an item that improve our model's performance by reducing the quantity of the original raw data.

(3) The Histogram of Skeletal and Gradients with Orientation, or HOG-SKE: We provide an approach that employs four distinct techniques to assess the place descriptor's effectiveness: frame voting, wide histogram, SVM classification, and spontaneous periodic deviation. Then, intricate screws or geometric shapes like spheres, cones, and extended cylinders are used to extract the human body. A common method for -+

+extracting features from educational and assessment videos is called HOG. It produces a fixed-sized vector called a histogram of words. A word histogram illustrates the frequency of each recognisable word that appears in a segment of the movie [9].

As Fig. 3 [30] strongly shows, we can extract HOG properties from the silhouette we formed from the BGS phase. The method divides the area inside each frame across a n × n regular grid of the histograms, which is used to compute regions of concern. Each grid cell creates a regular histogram, which shows the size and direction of the the outside region in each unique cell [3].

Formulas four and five demonstrate how, because each cell is attached, HOG determines the value of the variation of each individual cell or sub-image, I, with honour to X and Y:

$$I_X = I \times DX, \tag{4}$$

$$\text{where } DX = [+1 \quad 0 \quad -1],$$

$$I_Y = I \times DY, \tag{5}$$

$$\text{where } DY = \begin{bmatrix} +1 \\ 0 \\ -1 \end{bmatrix},$$

I_X and I_Y represent their byproducts of the image with respect to X and Y, respectively. The vertically and horizontally oriented Sobel purification equipment or D_X and D_Y, are placed together on the picture to get these derivatives. Every video has hundreds of frames, therefore using the HOG will result in an enlarged vector and increased computational expenses. To alleviate these issues, a crossover and 6-step frame leaps are used. Next, each cell's

magnitude and angle are determined using Equations 6 and 7, respectively. Cell histograms will eventually be normalized.

This article uses a basic skeletal view for action recognition in addition to the HOG approach. Commercial deep integrate cameras employ algorithms for skeletal estimation in real-time. The human body's joints may be removed quickly and easily thanks to this technology [3]. Certain studies exclusively employ the hands or other skeletal regions when analysing data. However, the whole body is used in this study to increase overall accuracy. Figure 4-left illustrates a skeleton approach for the three movements of sitting down, standing up, and lifting the hand; Figure 4-right focuses upon additional hand activity detection.

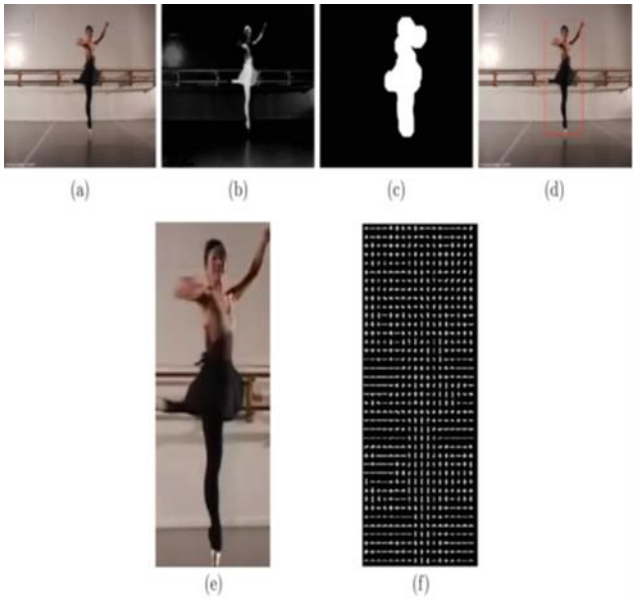


Fig. 3. HOG stages for identifying an instance of "dancing" movement [30]

$$|G| = \sqrt{I_X^2 + I_Y^2} \tag{6}$$

$$\phi = \arctan\left(\frac{I_X}{I_Y}\right). \tag{7}$$

Deep data and skeleton data have an advantage over typical RGB data in that they are less susceptible to variations in illumination [17]. To increase the precision of our action recognition model, we leverage skeleton and inertia data at the feature extraction and decision-making stages.

The skeleton's N_F frame sequences, denoted as $s_k = \{f_1, f_2, ...f_{NF}\}$, are shown. Same notations as in [31] are used here.

It is thought that the coordinate skeleton sequence (X_i, Y_i, Z_i) can be used to convey time and geographical information. For every f_i skeleton, $i \in [1, NF]$, in the interval [0, 255], the normalisation process is carried out using the TF(.) conversion function in accordance with Eq. 8:

$$\begin{aligned}
(X_i', Y_i', Z_i') &= TF(X_i, Y_i, Z_i) \\
X_i' &= 255 \times \frac{X_i - \min\{C\}}{\max\{C\} - \min\{C\}}, \\
Y_i' &= 255 \times \frac{Y_i - \min\{C\}}{\max\{C\} - \min\{C\}}, \\
Z_i' &= 255 \times \frac{Z_i - \min\{C\}}{\max\{C\} - \min\{C\}},
\end{aligned} \tag{8}$$

Where the minimum and maximum of each coordinate value are denoted by min{C} and max{C}, respectively. In the newly constructed interface the universe, computed to integral image depiction, three points (Xi^1, Yi^1, Zi^1) are considered as well as the three parts (R, G, & B) of a color pixel:

$$(X_i' = R,\ Y_i' = G,\ Z_i' = B), \tag{9}$$

The new coordinate for the picture display is (Xi^1, Yi^1, Zi^1). In Fig. 5, the stages are displayed. The skeleton sequence's raw data is transformed into 3-D tensors using the aforementioned processes and conversions before being fed as inputs into the learning model. Figure 5's F_N indicates the quantity of frames for every skeletal series, whereas K, or the amount of bones in a frame, is a function of the depth sensor and data recording parameters.

(4) ROI calculation involves the designation of the field which is relevant for the extraction of characteristics [17, 30, 40, 44]. A combination of contour-based features for distance signals, flow-based characteristics for motion [50, 53], and consistent rotation local binary configurations may be used to represent the action throughout the method of obtaining features. Thus, at this point, areas that are appropriate for feature extraction are identified. The input videos may contain specific multi-view activities, depending on the dataset's makeup, which improves the classification's accuracy. For the extraction of entropy-based silhouettes, a comparable technique is provided in [18, 32].

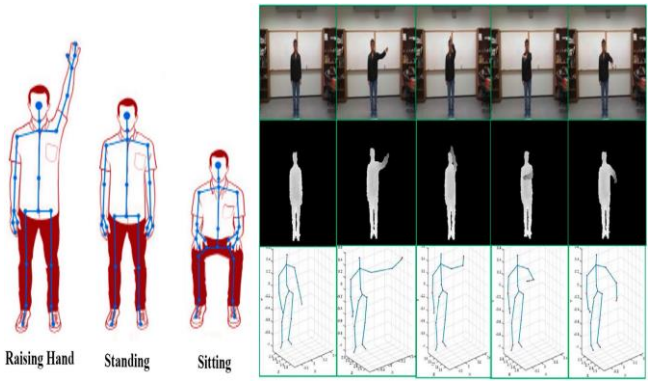


Fig. 4. The processes of choosing the right frame area and extracting the movement of the skeletal system [17, 31]

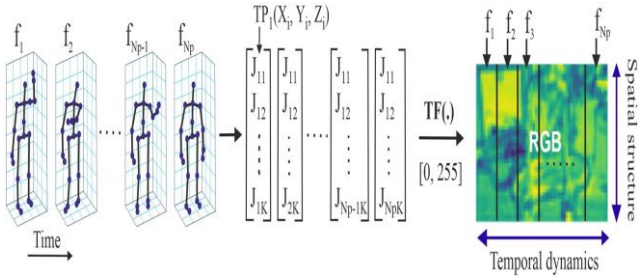


Fig. 5. The steps involved in training the model by transforming kinetic patterns into spatiotemporal data.

3.4. Feature Selection

Since every movement in a video can be illustrated by an ordered set of frames, we may do recognize an activity by looking at the structure of several images in a series. In order to identify activities that are similar to how people perceive them in real life, we provide a number of approaches and strategies. Predicting future actions from past sequences of events is one of human skills. Deep neural networks, which are modelled after the brain networks of humans, are therefore very suitable to create a system with such properties. These networks comprise, among others, LSTM, RNN, and CNN. At the feature and decision levels, CNN streams are frequently merged with RGB frames and skeletal sequences in research projects. Voting strategies are also used for classification at the decision-making level. The availability of multidimensional visual data, as was previously indicated, helps us to integrate all visual signals, including skeleton and depth information, as in [17]. Numerous research concentrate on CNN architecture's enhanced skeleton presentation.

CNN segments that are trained to identify different walkers exhibit substantial activation values that depend on the human area rather than the background. An individual commuter may be placed and lined up within the boundaries of the image by using this attention method [61]. “One of the primary challenges in adopting CNN-based techniques for skeletal-based detection of action is how to quickly display a time-based skeleton order and integrate them into the CNN for component studying and categorization. In order to address this difficulty, we capture the dynamics of skeletal sequences in both space and time within 2-D picture structures”. To find the original skeletal sequences, CNN is trained to recognise the image's characteristics and classify it [28]. Convolutional, pooling, and fully linked layers are the typical components of a CNN. The convex layer's masks are very useful for determining the image borders [5, 37, 52, 58, 60]. The completely linked stages of the Max-type are used to convert cubic multiple-dimensional information into a series of vectors with dimensions comparable to one, whereas the layers that pool information are frequently used to decrease dimension [27].

This neural network is unlikely to be able to detect complicated sequences of videos with sophisticated activities, such as feeding or jumping over hurdles, since it learns to enhance the filter assembly weight on the basis of a collection of NF incoming frames. This difficulty can be solved by RNNs [24, 26, 50], which avoid the exploding and vanishing gradient problem by storing just the preceding step. By retaining a small memory for an extended period of time, the LSTM network—which addresses the aforementioned problems—can be considered a kind of RNN. For choosing attributes and accurate action detection, we combine CNN as well as LSTM in our research because to their enhanced efficiency in apparent and continuous data. AlexNet is also included in the attribute selection procedure in order to uncover undetectable trends in photographic information. Parallel duplex LSTMs are used for the feature selection process, which is carried out in parallel to expedite processing. [13, 19, 50, 64], and [29] all take a similar tack. That is, there are two primary purposes for which we employ LSTM:

1.Keeping the crucial data from succeeding frames for an extended period of time would increase system efficiency since every frame in a movie matters. This is a suitable use for the "LSTM" approach.

2.Leveraging long short-term memory (LSTM) & artificial neural networks, text data analysis, picture processing, speech recognition, processed digital signals, and intelligent neural systems are all having achieved exceptional success in utilizing consecutive multimedia information [28, 50, 62].

The suggested CNN and dual LSTM network deep learning model's architecture is shown in Figure 6. Long-term dependencies may be learned by LSTM, and its unique structure, which consists of inputs, outputs, and forget gates to govern long-term sequence recognition, has been demonstrated in studies [27, 31, 49, 50, 54]. During the training, the Sigmoid unit opens and closes the gates. Equations 10, 11, 12, 13, 14, 15, and 16 are used to calculate each LSTM unit.

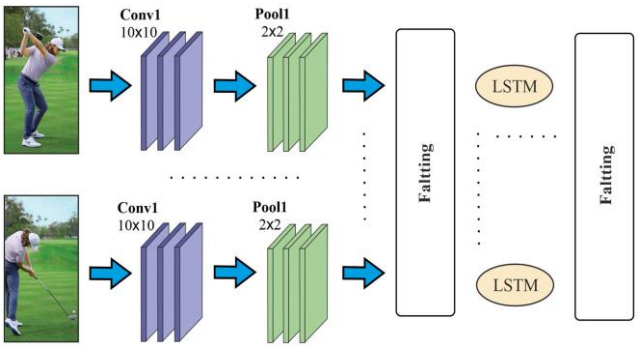


Fig. 6. The suggested Deep Learning model's simplified architecture, It uses an amalgam of The CNN network and simultaneous LSTM to choose the deep characteristics of a particular frame set (e.g. Golf swing motion detection employing spatio-temporal metadata)

$$i_t = \sigma((x_t + s_{t-1})W^i + b_i), \tag{10}$$

$$f_t = \sigma((x_t + s_{t-1})W^f + b_f), \tag{11}$$

$$o_t = \sigma((x_t + s_{t-1})W^o + b_o), \tag{12}$$

$$g = \tanh((x_t + s_{t-1})W^g + b_g), \tag{13}$$

$$c_t = c_{t-1} \odot f_t + g \odot i_t, \tag{14}$$

$$s_t = \tanh(c_t) \odot o_t, \tag{15}$$

$$Final\text{state} = \text{Softmax}(Vs_t), \tag{16}$$

If f_t at time t represents the gate known as the forget gate, which keeps the information of the last frame and, if needed, removes the data from the stored data cell, then x_t represents the input. With the currently displayed frame intake and the condition of the previous s_{t-1} frame, the ability to activate tanh of a return value unit, g, is computed. Information regarding the subsequent step is included in the outcome of the gate OT. The output of the current mode's RNN is represented by s_t. Tanh and C_t memory cells are turned on in order to determine the hidden mode from a single RNN step. The weights of the input, output, forget, and returning units from the LSTM cell are represented by the letters W^i, W^o, W^f, and W^g, respectively. The data being provided, outcome, neglect and transferring unit gates have four biases: b_i, b_o, b_f, & b_g.

Utilizing a SoftMax filter on the very last state of the RNN system, we arrived at a final decision since motion recognition not requires the middle ground result of the LSTM. To understand long-term relationships in video footage, we stack many LSTM cells together since, when training massive volumes of data, such as video information, only one LSTM cell is unable to recognise complex sequence structures.

3.5. KNN-Softmax Classifier

Usually, deep neural systems based on the Softmax functioning are employed for categorization. The Softmax classification technique in a network of deep neurons is effectively positioned after the last layer. The pattern of features $p_l = [p_1,...,p_l]$ that is generated by the layers based on convolution and pool corresponds to the Softmax's feed [5, 60]. After propagation through forward motion, a stochastic descent of gradients (SGD) optimization procedure is used on numerous training examples and cycles to update weights and minimise errors. Return propagation distributes the weights by calculating the variation of the convolution strengths. When there are numerous classes, the Softmax works badly. Usually, there are two main causes for this: first, an extensive amount of parameters inhibits the final layer from accelerating the forward-backward speed; second, GPU synchronization becomes difficult [52, 60].

In this paper, we employ KNN in situations where Softmax performs poorly due to a large number of classes. KNN should only be used if Sofmax categorization fails (which in this case is, when there is a possibility that an event is strongly connected to multiple classes). KNN uses the Euclidean & Hamming distances to determine the similarity of two feature vectors are [41]. As mentioned before, $Pl = [p_1,..., p_l]$ is a sorting feed that retains

$$p_l = \{(x_i, y_j), i = 1, 2, ..., n_l\},$$

The amount of features retrieved is portrayed by xi, and the similar label is indicated by yj for every attribute set of xi. The KNN algorithm uses Euclidean distance with doubled negative path weights and k = 10. The distance calculated using the Euclidean concept is given in Equation 17.

$$d(x_i, x_{i+1}) = \min_i (d(x_i, x_{i+1})). \tag{17}$$

If u is an entirely novel instance having a label of y_j, then Eq. 18 can be used to find the distance calculation using $d(u, x_i)$. We will be able to determine v + 1 and u's closest neighbour as a result.

$$d(u, x_i) = \frac{d(u, x_i)}{d(u, x_{v+1})}. \tag{18}$$

Using the kernel function and weighting in accordance with Eq. 19, we normalise $d(u, x_i)$.

$$w(i) = k(d(u, x_i)). \tag{19}$$

The following is the formulation of the weighted K-nearest neighbour (W-KNN) final membership function:

$$\hat{y} = max_j \left(\sum_{i=1} w(i) I(y_i = j) \right). \tag{20}$$

4. Experimental Results

After applying the precision criteria to evaluate the proposed technique on the UCF101 databases, an usual evaluation dataset, this part presents the experimental results. The training, testing, and validation portions of the dataset are split into 60%, 20%, and 20% divisions, respectively. Examples of datasets are displayed in Fig. 7. TensorFlow, a deep learning framework, and Python 3 are used to implement the suggested model.

Fig. 7. A selection of shots and tasks obtained from the UCF101 footage collection

Using the accuracy criteria, we evaluate the suggested strategy against eight cutting-edge approaches.

4.1. Performance Evaluation

Table 2 displays the outcomes of our assessments on the classification accuracy of the proposed HRR-DL method. In comparison to eight distinct state-of-the-art approaches, our proposed technique demonstrates an enhancement ranging from 0.8% to 4.47%. A key advancement in the suggested HFRDL method lies in the effective utilization of the BGS, HOG, and Skeleton techniques during the preprocessing phase. This strategic application enhances outcomes from the initial stages, facilitating the extraction of highly informative features integrated into a customized DNN platform. This approach proves pivotal in achieving precise action recognition in real-world scenarios. The amalgamation of “convolution, pooling, fully linked, and LSTM units contributes to improved feature learning, feature selection, and classification”. Consequently, this integration significantly reduces the likelihood of errors during the classification stage and enhances the accuracy of identifying complex activities.

4.2. Optimum Frame Jumping

In the second phase of our experimentation, we focused on determining the most effective “jump length for the proposed HFR-DL approach. Each video was treated as a singular input, and frame characteristics were derived by selecting one frame out of every x frames”. The assessment, outlined in Table 3, delves into the impact of frame jumps at intervals of 4, 6, and 8 on the overall system performance. Opting for a frame jump of J = 6 yielded a substantial improvement in the system's speed and computational efficiency, registering an approximate 50% enhancement compared to J = 4. Remarkably, this boost in efficiency was achieved with only a marginal decline in accuracy, around 1.5%. Consequently, “our analysis led us to designate a frame jump of 6 as the optimal choice for our specific application. This preference not only outperformed the analogous state-of-the-art approach (DB-LSTM) [50] but also struck a favorable balance between speed and accuracy in our comprehensive trade-off analysis”.

4.3. Confusion Matrix

Utilizing a reference classification system, a confusion matrix offers both quantized and visual insights into various classifiers [13, 50]. Detailed results of the HFR-DL method on the UCF Sports dataset are depicted in Figure 8. In this representation, each row corresponds to the projected class, and each column showcases examples of the ground truth classes. The comprehensive examination of the confusion matrix reveals that HFR-DL consistently outperforms “the ReHAR method [50]. Even for actions such as Golf, Run, and Walk, which exhibit relatively weaker results at 82.6%, 83.42%, and 72.30%, respectively, compared to 83.33%, 75.00%, and 57.14% for the ReHAR method, HFR-DL demonstrates superior performance”. Figure 8 further illustrates instances of misclassification, with "walking," "running," and "golfing" erroneously labeled as "kicking," "skateboarding," and "walking," respectively.

Such misclassifications are anticipated, given that certain behaviors possess characteristics that can lead to errors. Factors like the presence of other objects and individuals in the background of the image contribute to potential misclassifications. For instance, in the movie "walk-front/006RF1-13902-70016.avi," a person walking on a golf course while holding a golf pole is depicted. The golf pole's movement in the distance may create the illusion of someone swinging it in front of them, contributing to the misclassification within the suggested HFR-DL approach [13].

Table 2. Performance assessment of the suggested approach in comparison to eight other approaches

Methods	Year	Accuracy %
Lohit et al. [21]	2018	57.90
Patel et al. [30]	2018	89.43
R-NKTM [32]	2018	90.00
DB-LSTM [50]	2018	91.22
SVM-KNN [38]	2017	91.47
MRST-T [57]	2019	92.20
ConvLSTM(Softmax)	2020	92.60
C^2LSTM [25]	2020	92.80
ECO [64]	2019	93.10
HFR-DL (Proposed method)	2020	**93.90**

Table 3. Assessment of the suggested approach with various leaps

Methods	Frame jump	Average time (S)	Average Acc. %
DB-LSTM [50]	4.0	1.72	92.2%
	6.0	1.12	91.5%
	8.0	0.9	85.34%
HFR-DL (Proposed method)	4.0	2.10	95.62%
	6.0	1.6	93.9%
	8.0	1.10	89.6%

5. Conclusion

As per Table 2's summary performance report, the HFR-DL approach that was suggested resulted in a better identification of human actions by utilizing spatiotemporal information that was concealed inside sequential patterns and characteristics.

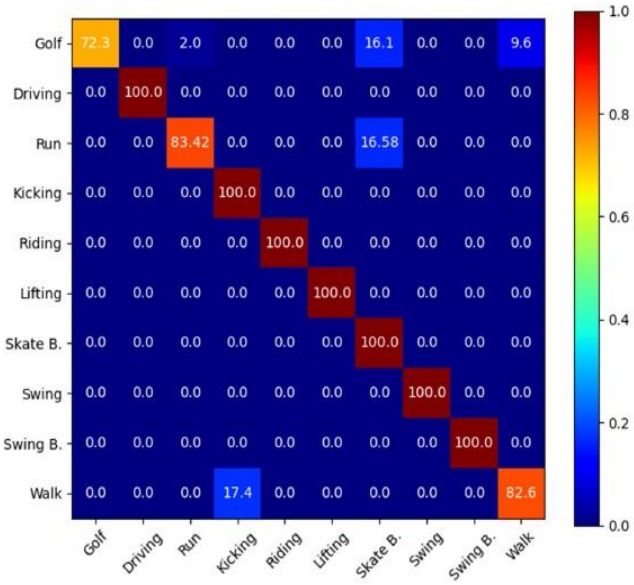


Fig. 8. Confusing matrix illustrating the HFR-DL technique's results for activities involving sports

The right frames for the model's preprocessing stage were identified and described using the BGS, HOG, and Skeletal combinations. The feature selection process then used an effective blend of LSTM and deep CNN. The network was trained during the training phase by randomly initializing the weights and repeating the training stage until the network had the fewest mistakes possible [10, 22, 31, 43, 59]. For both the training and test stages, the suggested system can lessen the consequences of the degradation phenomena. It should be mentioned that the size of the datasets has a significant impact on degradation phenomena. This explains why overly complex networks have greater error rates than medium-sized networks.

By taking use of the orientation connection between the joints and the Euclidean distance, we further expanded the skeleton encoding approach. Table 3 shows that while J = 4 somewhat increases accuracy levels in both procedures, Jump 6's time complexity is noticeably lower than Jump 4's. As a result, the Jump 6 is regarded as the best compromise between accuracy and temporal complexity. Ultimately, the action identification and categorization process employed a hybrid Softmax-KNN method. The experiment was conducted using the widely available UCF dataset, which consists of 101 distinct human behaviours. After evaluation and analysis of the confusion matrix and accuracy metric, the overall findings demonstrated that the suggested approach performs better in human action recognition when compared to eight other cutting-edge studies in the same sector (Table 2). Table 3 demonstrates that the DB-LSTM approach exhibits a slightly higher level of accuracy than the proposed HFR-DL method when considering speed and computational expenses. Typically, processing a one-second HD video clip on a mid-range Core i7 PC takes 1.6 seconds when utilizing the recommended approach with a jump of 6 [50]. This can be readily adjusted to achieve a real-time action recognition solution tailored to specific speed and accuracy requirements. Achieving this entails adjusting the jump step, increasing CPU speed, reducing input video resolution, or implementing a combination of these modifications. The flexibility of the training dataset enables the established approach to be deployed across a wide spectrum of practical applications. These applications include but are not limited to home monitoring for the elderly and infants, accident detection, criminal identification and recognition in surveillance systems, detection of anomalous human behavior, human-computer interaction, and sports analysis. We propose an extension of this study to refine the current architecture, exploring avenues to predict a subject's future behavior based on spatiotemporal information, the ongoing activity, and semantic scene segmentation and comprehension.

Conflicts of interest

The authors declare no conflicts of interest.

References

[1] Angel B, Miguel L, Antonio J, Gonzalez-Abril L. Mobile activity recognition and fall detection system for elderly people using ameva algorithm. Pervasive Mob Comput. 2017;34:3–13.

[2] Anuradha K, Anand V, Raajan NR. Identification of human actor in various scenarios by applying background modeling. Multimed Tools Appl. 2019;79:3879–91.

[3] Bajaj P, Pandey M, Tripathi V, Sanserwal V. Efficient motion encoding technique for activity analysis at ATM premises. In progress in advanced computing and intelligent engineering. Berlin: Springer; 2019. p. 393–402.

[4] Chaquet JM, Carmona EJ, Fernández-Caballero A. A survey of video datasets for human action and activity

recognition. ComputVis Image Underst. 2013;117:633–59.

[5] Chavarriaga R, Sagha H, Calatroni A, Digumarti S, Tröster G, Millán J, Roggen D. The opportunity challenge: a benchmark database for on-body sensor-based activity recognition. Pattern Recognit Lett. 2013;34:2033–42.

[6] Chen BH, Shi LF, Ke X. A robust moving object detection in multi-scenario big data for video surveillance. IEEE Trans Circuits Syst Video Technol. 2018;29(4):982–95.

[7] Dedeoğlu Y, Töreyin BU, Güdükbay U, Çetin AE. Silhouettebased method for object classification and human action recognition in video. In European conference on computer vision. Berlin: Springer; 2006. p. 64–77.

[8] Donahue J, Hendricks LA, Guadarrama S, Rohrbach M, Venugopalan S, Saenko K, Darrell T. Long-term recurrent convolutional networks for visual recognition and description. In: Proceedings of the IEEE conference on computer vision and pattern recognition. 2015; pp. 2625–34.

[9] Ehatisham-Ul-Haq M, Javed A, Azam MA, Malik HM, Irtaza A, Lee IH, Mahmood MT. Robust human activity recognition using multimodal feature-level fusion. IEEE Access. 2019;7:60736–51.

[10] Fernando B, Anderson P, Hutter M, Gould S. Discriminative hierarchical rank pooling for activity recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016, pp. 1924–32.

[11] Gammulle H, Denman S, Sridharan S, Fookes C. Two stream LSTM: a deep fusion framework for human action recognition. In: Applications of Computer Vision (WACV), 2017 IEEE Winter Conference. 2017.

[12] Hegde N, Bries M, Swibas T, Melanson E, Sazonov E. Automatic recognition of activities of daily living utilizing insole based and wrist worn wearable sensors. IEEE J Biomed Health Inform. 2017;22:979–88.

[13] Huan RH, Xie CJ, Guo F, Chi KK, Mao KJ, Li YL, Pan Y. Human action recognition based on HOIRM feature fusion and AP clustering BOW. Plos One. 2019;14:e019910.

[14] Jain A, Zamir AR, Savarese S, Saxena A. Structural-RNN: Deep learning on spatio-temporal graphs. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016, pp. 5308–17.

[15] Ke S, Thuc H, Lee Y, Hwang J, Yoo J, Choi K. A review on videobased human activity recognition. Computers. 2013;2:88–131.

[16] Keyvanpour M, Serpush F. ESLMT: a new clustering method for biomedical document retrieval. Biomed Eng. 2019;64(6):729–41.

[17] Khaire P, Kumar P, Imran J. Combining CNN streams of RGB-D and skeletal data for human activity recognition. Pattern Recognit Lett. 2018;115:107–16.

[18] Kumari P, Mathew L, Syal P. Increasing trend of wearables and multimodal interface for human activity monitoring: a review. Biosens Bioelectron. 2017;90:298–307.

[19] Li X, Chuah MC. Rehar: Robust and efficient human activity recognition. In: 2018 IEEE Winter Conference on Applications of Computer Vision (WACV), 2018; pp. 362–71.

[20] Liu AA, Xu N, Nie WZ, Su YT, Zhang YD. Multi-domain and multi-task learning for human action recognition. IEEE Trans Image Process. 2019;28(2):853–67.

[21] Lohit S, Bansal A, Shroff N, Pillai J, Turaga P, Chellappa R. Predicting dynamical evolution of human activities from a single image. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. 2018; pp.383–92.

[22] Ma CY, Chen MH, Kira Z, AlRegib G. TS-LSTM and temporal inception: exploiting spatiotemporal dynamics for activity recognition. Signal Process. 2019;71:76–877.

[23] Ma S, Sigal L, Sclaroff S. Learning activity progression in LSTMs for activity detection and early detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016, pp. 1942–50.

[24] Mahasseni B, Todorovic S. Regularizing long short term memory with 3D human-skeleton sequences for action recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. 2016, pp. 3054–62.

[25] Majd M, Safabakhsh R. Correlational convolutional LSTM for human action recognition. Neurocomputing. 2020;396:224–9.

[26] Molchanov P, Yang X, Gupta S, Kim K, Tyree S, Kautz J. Online detection and classification of dynamic hand gestures with recurrent 3D convolutional neural network. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016, pp. 4207–15.

[27] Montes A, Salvador A, Pascual S, Giro-i Nieto X. Temporal activity detection in untrimmed videos with recurrent neural networks. 2016;1–5.

[28] Núñez JC, Cabido R, Pantrigo JJ, Montemayor AS, Vélez JF. Convolutional neural networks and long short-term memory for skeleton-based human activity and hand gesture recognition. Pattern Recognit. 2018;76:80–94.

[29] Park E, Han X, Berg T, Berg AC. Combining multiple sources of knowledge in deep CNNs for action recognition. In: Applications of Computer Vision

(WACV), 2016 IEEE Winter Conference. 2016, pp. 1–8.

[30] Patel C, Garg S, Zaveri T, Banerjee A, Patel R. Human action recognition using fusion of features for unconstrained video sequences. Comput Electr Eng. 2018;70:284–301.

[31] Pham HH, Khoudour L, Crouzil A, Zegers P, Velastin SA. Exploiting deep residual networks for human action recognition from skeletal data. Comput Vis Image Underst. 2018;170:51–66.

[32] Rahmani H, Mian A, Shah M. Learning a deep model for human action recognition from novel viewpoints. IEEE Trans Pattern Anal Mach Intell. 2018;40:667–81.

[33] Rezaei M, Azarmi M. Deep-SOCIAL: social distancing monitoring and infection risk assessment in COVID-19 pandemic. Appl Sci. 2020;10:1–29.

[34] Rezaei M, Fasih A. A hybrid method in driver and multisensory data fusion, using a fuzzy logic supervisor for vehicle intelligence. In: Sensor Technologies and Applications, IEEE International Conference. 2007, pp. 393–98.

[35] Rezaei M, Klette R. Look at the driver, look at the road: No distraction! no accident! In: 2014 IEEE Conference on Computer Vision and Pattern Recognition. 2014, pp. 129–36.

[36] Rezaei M, Shahidi M. Zero-shot learning and its applications from autonomous vehicles to COVID-19 diagnosis: a review. Intell-Based Med. 2020;3–4:1–27. https ://doi.org/10.1016/j.ibmed.2020.10000 5.

[37] Ronao CA, Cho SB. Deep convolutional neural networks for human activity recognition with smartphone sensors. International Conference on Neural Information Processing. Cham: Springer; 2015. p. 46–53.

[38] Sargano AB, Wang X, Angelov P, Habib Z. Human action recognition using transfer learning with deep representations. In: 2017 International joint conference on neural networks (IJCNN). IEEE. 2017, pp. 463–69.

[39] Schneider B, Banerjee T. Activity recognition using imagery for smart home monitoring. Advances in soft computing and machine learning in image processing. Berlin: Springer; 2018. p. 355–71.

[40] Shahroudy A, Ng T, Gong Y, Wang G. Deep multimodal feature analysis for action recognition in rgb+ d videos. IEEE Trans Pattern Anal Mach Intell. 2018;40:1045–58.

[41] Sharif M, Khan MA, Zahid F, Shah JH, Akram T. Human action recognition: a framework of statistical weighted segmentation and rank correlation-based selection. Pattern Anal Appl. 2019;23:281–94.

[42] Sharma S, Kiros R, Salakhutdinov R. Action recognition using visual attention. Int Conference ICLR, 2016; pp. 1–11.

[43] Singh B, Marks TK, Jones M, Tuzel O, Shao M. A multi-stream bi-directional recurrent neural network for fine-grained action detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016;pp. 1961–70

[44] Singh R, Kushwaha AKS, Srivastava R. Multi-view recognition system for human activity based on multiple features for video surveillance system. Multimed Tools Appl. 2019;78:17168–9.

[45] Soomro K, Zamir AR, Shah M. UCF101: A dataset of 101 human actions classes from videos in the wild. 2012;1–7. arXiv:1212.0402.

[46] Tao L, Volonakis T, Tan B, Jing Y, Chetty K, Smith M. Home activity monitoring using low resolution infrared sensor. 2018;1–8. https ://arxiv .org abs/1811.05416 .

[47] Tu Z, Xie W, Qin Q, Poppe R, Veltkamp R, Li B, Yuan J. Learning representations based on human-related regions for action recognition. Pattern Recognit. 2018;79:32–433.

[48] Turaga P, Chellappa R, Subrahmanian VS, Udrea O. Machine recognition of human activities: a survey. IEEE Trans Circuits Syst Video. 2008;18:1473.

[49] Ullah A, Muhammad K, Ser J, Baik SW, Albuquerque V. Activity recognition using temporal optical flow convolutional features and multi-layer LSTM. IEEE Trans Ind Electron. 2018;66:9692–702.

[50] Ullah JA, Muhammad K, Sajjad M, Baik SW. Action recognition in video sequences using deep Bi-directional LSTM with CNN features. IEEE Access. 2018;6:1155–66.

[51] Varior RR, Haloi M, Wang G. Gated siamese convolutional neural network architecture for human re-identification. In European Conference on Computer Vision. Cham: Springer; 2016. p. 791–808.

[52] Wang X, Gao L, Song J, Shen H. Beyond frame-level CNN: saliency-aware 3-d cnn with LSTM for video action recognition. IEEE Signal Process Lett. 2017;24:510–4.

[53] Wang X, Gao L, Song J, Zhen X, Sebe N, Shen H. Deep appearance and motion learning for egocentric activity recognition. Neurocomputing. 2018;275:438–47.

[54] Wang X, Gao L, Wang P, Sun X, Liu X. Two-stream 3-D convNet fusion for action recognition in videos with arbitrary size and length. IEEE Trans Multimed. 2018;20:634–44.

[55] Wang Y, Lu Q, Wang D, Liu W. Compressive background modeling for foreground extraction. J Electr Comput Eng. 2015;2015:1–9.

[56] Wang Y, Wang S, Tang J, O'Hare N, Chang Y, Li B. Hierarchical attention network for action recognition in videos. 2016; arXiv :1607.06416 . pp. 1–9.

[57] Wu H, Liu J, Zha ZJ, Chen Z, Sun X. Mutually reinforced spatio- temporal convolutional tube for human action recognition. In: IJCAI. 2019;pp. 968–74.

[58] Wu Y, Li J, Kong Y, Fu Y. Deep convolutional neural network with independent Softmax for large scale face recognition. In: Proceedings of the 2016 ACM on Multimedia Conference. 2016;pp. 1063–67.

[59] Ye J, Qi G, Zhuang N, Hu H, Hua KA. Learning compact features for human activity recognition via probabilistic first-takeall. IEEE Trans Pattern Anal Mach Intell. 2018;42:126–39.

[60] Zeng R, Wu J, Shao Z, Senhadji L, Shu H. Quaternion softmax classifier. Electron Lett. 2014;50:1929–31.

[61] Zhou Q, Zhong B, Zhang Y, Li J, Fu Y. Deep alignment network based multi-person tracking with occlusion and motion reasoning. IEEE Trans Multimed. 2018;21(5):1183–94.

[62] Zhu G, Zhang L, Shen P, Song J. Multimodal gesture recognition using 3-D convolution and convolutional LSTM. IEEE Access. 2017;5:4517–24.

[63] Ziaeefard M, Bergevin R. Semantic human activity recognition: A literature review. Pattern Recognit. 2015;48:2329–45.

[64] Zolfaghari M, Singh K, Brox T. Eco: Efficient convolutional network for online video understanding. In: In Proceedings of the European Conference on Computer Vision (ECCV). 2018; pp.695–712.