ORIGINAL RESEARCH

Complex Human Action Recognition Using a Hierarchical Feature Reduction and Deep Learning-Based Method

Fatemeh Serpush[1] · Mahdi Rezaei[2]

Abstract

Automated human action recognition is one of the most attractive and practical research fields in computer vision. In such systems, the human action labelling is based on the appearance and patterns of the motions in the video sequences; however, majority of the existing research and most of the conventional methodologies and classic neural networks either neglect or are not able to use temporal information for action recognition prediction in a video sequence. On the other hand, the computational cost of a proper and accurate human action recognition is high. In this paper, we address the challenges of the preprocessing phase, by an automated selection of representative frames from the input sequences. We extract the key features of the representative frame rather than the entire features. We propose a hierarchical technique using background subtraction and HOG, followed by application of a deep neural network and skeletal modelling method. The combination of a CNN and the LSTM recursive network is considered for feature selection and maintaining the previous information; and finally, a Softmax-KNN classifier is used for labelling the human activities. We name our model as *"Hierarchical Feature Reduction & Deep Learning"*-based action recognition method, or HFR-DL in short. To evaluate the proposed method, we use the UCF101 dataset for the benchmarking which is widely used among researchers in the action recognition research field. The dataset includes 101 complicated activities in the wild. Experimental results show a significant improvement in terms of accuracy and speed in comparison with eight state-of-the-art methods.

Keywords Human action recognition · Deep neural networks · Histogram of oriented gradients · HOG · Skeleton model · Feature extraction · Spatio-temporal information

Introduction

Although the Human Activity or Human Action Recognition (HAR) is an active field in the present era, there are still key aspects which should be taken into consideration in order to accurately realise how people interact with each other or while using digital devices [11, 12, 63]. Human activity recognition is a sequence of multiple and complex sub-actions. This has been recently investigated by many researchers around the world using different types of sensors. Automatic recognition of human activities using computer vision has

been more effective than the past few years, and as a result with rapidly growing demands in various industries. These include health care systems, activities monitoring in smart homes, Autonomous Vehicles and Driver Assistance Systems [35, 36], security and environmental monitoring to automatic detection of abnormal activities to inform relevant authorities about criminal or terrorist behaviours, services such as intelligent meeting rooms, home automation, personal digital assistants and entertainment environments for improving human interaction with computers, and even in the new challenges of social distancing monitoring during the COVID-19 pandemic [33].

In general, we can obtain the required information from a given subject using different types of sensors such as cameras and wearable sensors [1, 39, 58]. Cameras are more suitable sensors for security applications (such as intrusion detection) and other interactive applications. By examining video regions, activities in different directions can be identified as forward or backward, rotation, or

✉ Mahdi Rezaei
m.rezaei@leeds.ac.uk

1 Faculty of Computer and Information Technology Engineering, Qazvin Branch, Islamic Azad University, Qazvin, Iran

2 University of Leeds, Institute for Transport Studies, 34–40 University Road, Leeds LS2 9JT, UK

sitting positions. The concept of action and movement recognition in video sequences is very interesting and challenging research topics to many researchers. For example, in walking action recognition using computer vision and wearable devices, the challenges could be visual limitation of sensory devices. As a result, there may be a lack of available information to describe the movements of individuals or objects [39, 42, 56].

On the other hand, the complex action recognition using computer vision demands a very high computation cost, while video capturing itself can be heavily influenced by light, visibility, scale, and orientation [50]. Therefore, to reduce the computational cost [16], a system should be able to efficiently recognise the subject's activities based on minimal data, given that the action recognition system is mostly online and needs to be assessed in real time, as well. Accordingly, useful frames and frame index information can be exploited; in human pose estimation, the body pose is represented by a series of directional rectangles. Combination of rectangles' directions and positions defines a histogram to create a state descriptor for each frame. In the background subtraction methods (BGS), the background is considered as the offset and the methods such as histogram of oriented gradients (HOG), histogram of optical flow (HOF) and motion boundary histogram (MBH) can increase the efficiency of the video based action recognition systems [30, 53]. Skeletons model can capture the position of the body parts or the human hands/arms to be used for human activity classification [17, 30, 40]. Different machine learning methods have been proposed for action recognition and address the mentioned challenges, each of which has its own strengths, deficiencies and weaknesses.

Convolutional Neural Networks (CNNs) is a type of deep neural network that effectively classifies the objects using a combination of layers and filtering [51]. Recurrent Neural Network (RNN) can be used to address some of the challenges in activity recognition. In fact, RNNs include a recursive loop that retains the information obtained in the previous moments. The RNN only maintains a previous step that is considered as a disadvantage. Therefore, LSTM was introduced to maintain information of several sequential stages [8]. Theoretically, RNNs should be able to maintain long-term dependencies to solve two common problems of exploding and vanishing gradients, while LSTM deals with the above-mentioned issues more efficiently [1, 14, 23, 50].

In the following sections, we discuss in more details and provide further information about our ideas. The rest of this paper is organised as follows: the next section reviews some of the most related work in the field. The following section explains the proposed method and procedures. Before the final section, we will review the experimental and evaluation results and compare them with eight state-of-the-art methods. The final section concludes the paper and provides suggestions for future works.

Related Work

In the last decade, Human Action Recognition (HAR) has attracted the attention of many researchers from different disciplines and for various applications. Most of the existing methods use hand-crafted features, and thanks to the GPU and extended memory developments, the deep neural networks can also recognise the activities of subjects in the live videos. Human action recognition in a sequence of image frames is one of the research topics of the machine vision that focuses on correct recognition of human activities using single view images [44, 50]. In conventional hand-crafted approaches, the low-level features associated with a specific action were extracted from the video signal sequences, followed by labelling by a classifier, such as K-Nearest Neighbour (KNN), Support Vector Machine (SVM), decision tree, K-means, or Hidden Markov Models (HMMs) [50, 64]. Handcraft-based techniques require an expert to identify and define features, descriptors, and methods of making a dictionary to extract and display the features. Deep learning techniques for image classification, object detection, HAR, or sound recognition have also taken traditional hand-crafting techniques, but in a more automated manner than conventional approaches [38].

In [50], the authors performed an analytical study on every six frames of input video sequences and tried to extract relevant features for action recognition using a pre-trained AlexNet Network. The method uses deep LSTM with two forward and backward layers to learn and extract the relevant features from a sequence of video frames.

In [38], a pre-trained deep CNN is used to extract features, followed by the combination of SVM and KNN classifiers for action recognition. A pre-trained CNN on a large-scale annotation dataset can be transmitted for the action recognition with a small training dataset. So transfer learning using deep CNN would be a useful approach for training models where the dataset size is limited.

In [25] an extended version of the LSTM units named C^2 LSTM is presented in which the motion data are perceived as well as the spatial features and temporal dependencies. They used both spatial and motion structure of the video data and developed a new deep network structure for HAR. The new network is evaluated on the UCF101 and HMDB51.

In [57], a novel Mutually Reinforced Spatio-Temporal Convolutional Tube (MRST) is represented for HAR. The model decomposes 3D inputs into spatial and temporal representations and mutually enhances them by exploiting the interaction of spatial and temporal information and selectively emphasising on informative spatial appearance

and temporal motion, while reducing the complexity of the structure.

By increasing the size of the dataset, the issue of overfitting will be eliminated; however, providing a large amount of annotated data is very difficult and expensive. In such conditions, the transfer learning is appropriate. The proposed technique in [38] aims to build a new architecture using a successful pre-trained model.

In some research works, the human activity and hand gesture recognition problems are investigated using 3-D data sequence of the entire body and skeletons. Also, a learning-based approach, which combines CNN and LSTM, is used for pose detection problems and 3-D temporal detection [50]. Singh et al. [44] proposed a framework for background subtraction (BGS) along with a feature extraction function, and ultimately they used HMMs for action recognition.

In [30], an action recognition system is presented using various feature extraction fusion techniques for UCF dataset. The paper presents six different fusion models inspired by the early fusion, late fusion, and intermediate fusion schemes [34]. In the first two models, the system utilises an early fusion technique. The third and fourth models exploit intermediate fusion techniques. In the fourth model, the system confront a kernel-based fusion scheme, which takes advantage of a kernel-based SVM classifier. In the fifth and sixth models, late fusion techniques have been demonstrated.

[64] has processed only one frame of a temporal neighbourhood efficiently with a 2-D Convolutional architecture to capture appearance features of the input frames. However, to capture the contextual relationships between distant frames, a simple aggregation of scores is insufficient. Therefore, they feed the feature representations of distant frames into a 3-D network that learns the temporal context between the frames, so, it can improve significantly over the belief obtained from a single frame especially for complex long-term activities.

[32] has proposed a Robust Non-linear Knowledge Transfer Model (R-NKTM) for human action recognition from unseen viewing angles. The proposed R-NKTM is a fully connected deep neural network that transfers knowledge of human actions from any unknown view to a shared high-level virtual view by finding a non-linear virtual path that interconnects different views together. The R-NKTM is trained by dense trajectories of synthetic 3-D human models fitted to capture real motion data, and then generalise them for real videos of human actions. The strength of the proposed technique is that it trains only one single R-NKTM for all action detections and all viewpoints for knowledge transfer of any human action video, without the requirement of re-training or fine-tuning the model.

In [21], a probabilistic framework is proposed to infer the dynamic information associated with a human pose. The model develops a data-driven approach, by estimating the density of the test samples. The statistical inference on the estimated density provides them with quantities of interests, such as the most probable future motion of the human and the amount of motion information conveyed by a pose.

[19] proposes a novel robust and efficient human activity recognition scheme called ReHAR which can be used to handle single person activities and group activity prediction. First, they generate an optical flow image for each video frame. Then both the original video frames and their corresponding optical flow images are fed into a single frame representation model to generate representations. Finally, an LSTM network is used to predict the forthcoming activities based on the generated representations.

Methodology

The methodology includes multiple stages and sub-modules; therefore, we divide this section into multiple subsections. First, the overall architecture of the model is described for an overall understanding, followed by the learning phase, and finally the transfer learning phase is explained.

Model Architecture

The architecture of the proposed method called Hierarchical Feature Reduction and Deep Learning (HFR-DL) is shown in Fig. 1. The proposed system consists of three main components: the input, the learning process, and the output.

The learning process module includes Background Subtraction, Histogram of Oriented Gradients, and Skeletons (BGS-HOG-SKE), where we also call it feature reduction module; then we develop the CNN-LSTM model as deep learning module; and finally the KNN and Softmax layer as the human action classification sub-modules. The UCF101 dataset and AlexNet are also utilized in the system. The former is a collection of large and complex video clips and the latter one is a pre-trained system designed to enhance the system action detection performance. We use AlexNet for transfer learning [50] and as the backbone of the network.

In the action recognition component, we have three sub-components: preprocessing, feature selection, and classifications. In the preprocessing step, the video clips are converted to a sequence of frames. However, the operations are performed only on selected frames which can have a positive impact on the cost and performance. Two deep CNN and LSTM neural networks are used to select the features with optimised weights. The parameters are trained on a variety of datasets and are adjusted more precisely comparing to previous non-deep learning based methods. Later in "Experimental Results", we will show the main advantage of RNNs and deep LSTM with a higher accuracy rate in complex action recognition, comparing
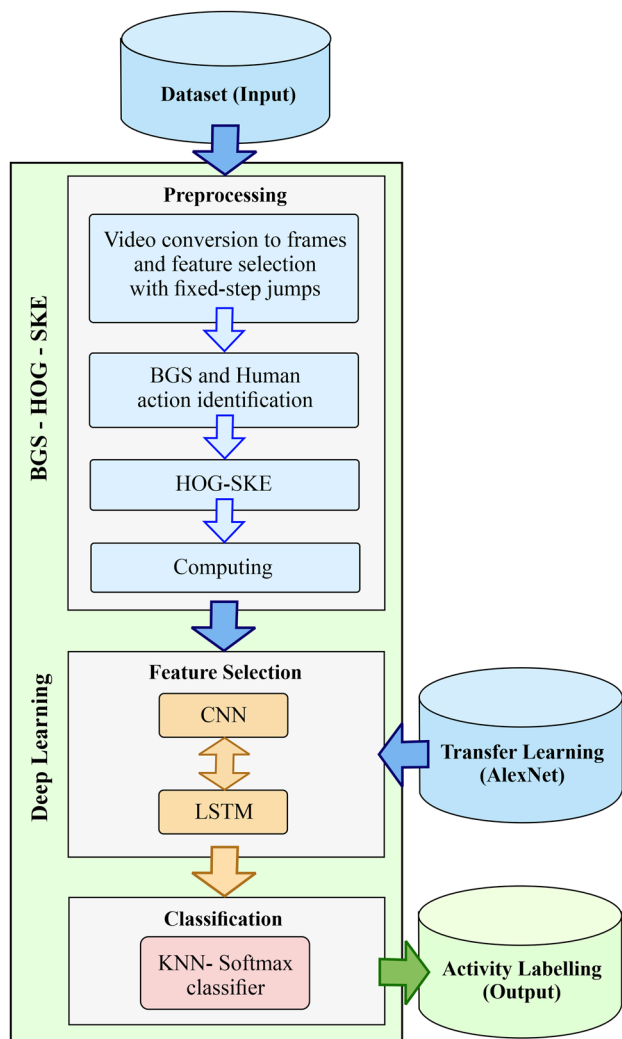
Fig. 1 The HAR system including three main modules: the preprocessing component, the deep neural network based feature extraction module, and classification module

to other deep neural network models. In "KNN-Softmax Classifier", two methods of Softmax and KNN are used to label and classify the output as an action.

After the training phase of the developed action recognition system, the second phase is the system test and performance analysis, which specifies the system error and its accuracy. We provide more details in "Experimental Results".

Before we dive into the further technical details we review on common symbols used in the following sections. Table 1 describes the symbols and notations used in this article.

Table 1 The description of the symbols

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
$p^l = [p_1, ..., p_I]$	Output feature vector of the deep network and input of classifications
V	Video activity
TF(.)	Conversion function

Learning Phase

In the learning phase, we have three stages; preprocessing, feature selection, and classification (see Fig. 1). The preprocessing stage is a very sensitive stage and the model performance highly depends on this stage and can lead to increased accuracy in the HAR output. In the following sub-sections, more steps and details will be described.

Preprocessing

As shown in Fig. 2, in the preprocessing stage, the input videos are converted into a sequence of frames. Then the representative frames will be selected from the given sequences of frames. In this study, we removed the background of representative frames using BGS technique (Fig. 2, bottom row). After that we apply the deep and skeletal method on the representative frames, where depth motion maps explicitly create the motion representations from the raw frames. Below we explain the details of the pre-processing phase in four stages, and in a step-by-step manner:

(1) Video to frame conversion and frame selection The input videos are first converted to a set of frames [2], each of which is represented by a matrix as shown in 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 is the kth representative frame, which has n rows and m columns. f_{ij} are the feature values (intensity of each
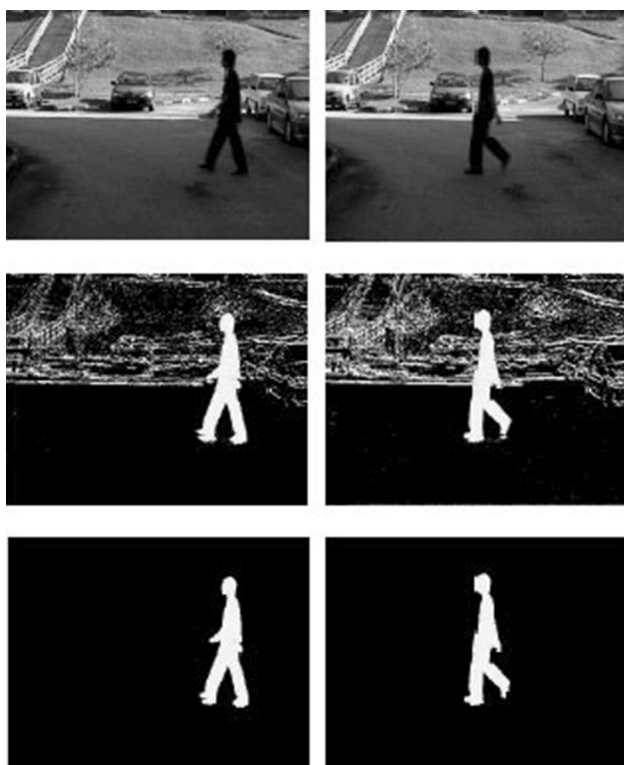
Fig. 2 A sample representation of BGS technique to preprocess the extracted frames from a video

pixel) for the corresponding frame k. After converting a video to frames we face a high volume of still images and frames that decrease the overall efficiency of the system due to high computational cost. To cope with the issue, we propose a simple yet effective solution to remove the redundant images. This can be done by fixed-step jumps J to eliminate similar sequential frames [52]. Based on our experiment, selecting one frame in every six frames will not significantly reduce the quality of the system, but speed it up significantly. We discuss this in more details later in "Experimental Results". Therefore, instead of extracting all features of all frames, only N_F frames [17, 20, 30, 40] were used. This makes our CNN network to perform more efficiently for the next steps.

(2) BGS and human action identification: The majority of the moving object recognition techniques include BGS, statistical methods, temporal differencing, and optical flow. We use background modelling like techniques to detect foreground objects [61]. Background subtraction-based methods have been used to detect moving objects in a video sequence; these methods enable the maintenance of a single background model constructed from previous frames [6].

The BGS scheme can be used indoors and outdoors, which is a popular method to separate moving parts of a scene by dividing it into background and foreground [2, 7, 48]. After separating the pixels from the static background of the scene, the regions can be classified into classes such as groups of humans. The classification algorithm depends on the comparison of the silhouettes of detected objects with pre-labelled templates in the database of an object silhouette. The template database is created by collecting samples of object silhouettes from samples of videos, labelled in appropriate categories. The silhouettes of the object regions are then extracted from the foreground pixel-map by using a contour tracing algorithm [2, 7]. In [55], the BGS steps are described, where f_k is the representative frame of the sequence of the video, assuming the neighbouring pixels share a similar temporal distribution. Given the pixel located in u, in the ith block of the image, the value and spatial neighbourhood are identified by $v_i(u)$ and $NG_i(u)$, respectively. Therefore, the value of the background sample of the pixel u, with $b_{i,j}(u)$ is determined to be equal to v, which is randomly chosen in $NG_i(u)$ (representative frame), as shown in Eq.2:

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

Then A_i the background model of the pixel u can be initialised by the background model of all pixels in the ith block:

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

This strategy can extract the foreground of selected frames from short video sequences or from embedded devices with limited memory and processing resources. Additionally, minimal but efficient size of data is preferred as too large data sizes may result in statistical correlation destruction within the pixels in different locations. Further information for tracking the foreground extraction steps can be found in [55]. This will also decrease the difference between the intensity of each pixel in the current image to the corresponding value in the reference background image.

An example of a sequence of BGS steps for walking is shown in Fig. 2. Human shape plays an important role in recognising human action, which can extract blobs from BGS as shown in Fig. 2, middle and bottom rows. Several methods based on global features, boundary, and skeletal descriptors have been proposed to illustrate the human shape in a scene [48]. After applying the BGS, a series of noise may disappear; however, some other noise may arise in other regions [15, 46]. To remove such artefacts we use erosion and dilation morphological operators, with the structural elements of 3×3. The feature extraction step determines the diagnostic information needed to describe the human silhouette. In general, we can say that BGS extracts useful features from an object that increases the performance of our model by decreasing the size of the initial raw data,

while maintaining the important parts of the embedded information.

(3) HOG-SKE: histogram of oriented gradients and skeleton: In our proposed method, four different methods are used to evaluate the performance of the position descriptor: frame voting, global histogram, SVM classification, and dynamic time deviation. After that, the human body is extracted using complex screws or volumetric models such as cones, elliptical cylinders, and spheres. HOG is a well-known feature extraction technique and HOG descriptors from each training/testing video into a fixed-sized vector is known as a histogram of words. Histogram of words shows the frequency of each visual word that is present in a video sequence [9].

HOG features can be extracted from the silhouette we made from the BGS stage, as also shown in Fig. 3 [30]. The technique is a window-based descriptor used to compute points of interest, where the window is divided into an $n \times n$ frequency grid of the histograms. The frequency histogram is generated from each grid cell to indicate the magnitude and direction of the edge for every individual cell [3].

The cells are interconnected and HOG calculates the derivative of each cell (or sub-image), I, with respect to X and Y as shown in Eqs.4 and 5:

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

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

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

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

I_X and I_Y, are the derivative of the image with respect to X and Y, respectively. To obtain these derivatives, horizontal and vertical Sobel filters (i.e. DX and DY) are convolved on the image.

Normally, every video consists of hundreds of frames, and using the HOG will lead to an elongated vector and, therefore, a higher computational cost. For resolving these challenges, an overlap and 6-step frame jumps are used.

Then magnitude and the angle of each cell is calculated as per the Eqs.6 and 7, receptively. Finally histograms of cells will be normalised.
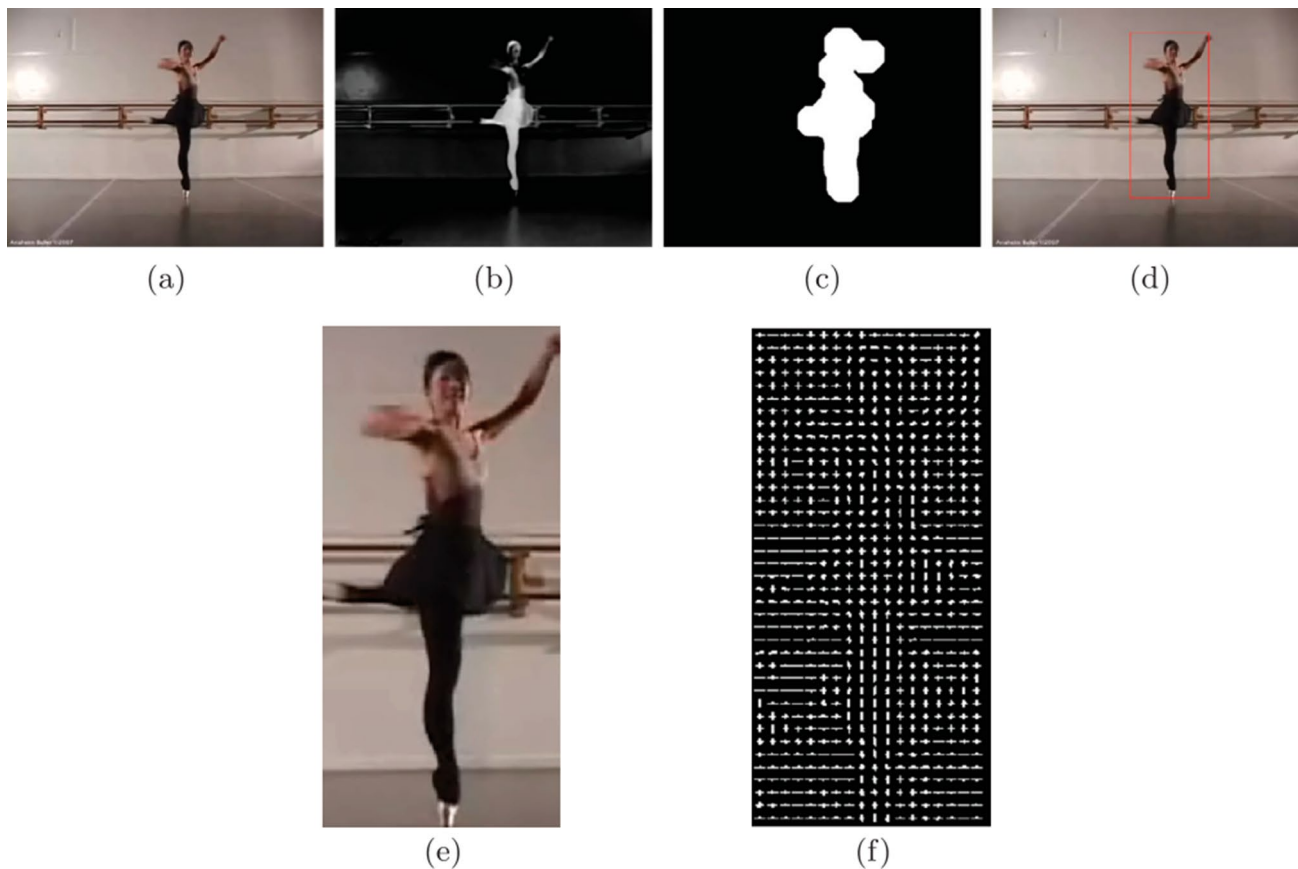

Fig. 3 HOG steps for a sample *"dancing"* action recognition [30]

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

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

In this paper, in addition to the HOG method a simple skeleton view is also used for action recognition. Real-time skeleton estimation algorithms are used in commercial deep integrate cameras. This technology allows the fast and easy joints extraction of the human body [3]. Some studies, only use part of the body in a skeleton method, such as hands. However, in this research, the whole body is used to increase the overall accuracy. Figure 4-left illustrates a skeletal method on three activities of sitting, standing, and raising hand and Fig. 4-right focuses more on hand activity recognition.

One of the advantages of deep data and skeletal data, as compared with traditional RGB data, is that they are less sensitive to changes in lighting conditions [17]. We use Skeleton and inertia data at both levels of feature extraction and decision making to improve the accuracy of our action recognition model.

The sequences s_k of the skeleton with N_F frames are shown as: $s_k = \{f_1, f_2, ... f_{N_F}\}$. We use same notations as in [31].

To represent spatial and temporal information, the coordinate skeleton sequence (X_i, Y_i, Z_i) is considered. For each f_i skeleton, $i \in [1, N_F]$, in the range [0, 255], and the normalisation operation is performed according to Eq.8 with the TF(.) conversion function:

$$(X_i', Y_i', Z_i') = TF(X_i, Y_i, Z_i)$$

$$X_i' = 255 \times \frac{X_i - \min\{C\}}{\max\{C\} - \min\{C\}},$$

$$Y_i' = 255 \times \frac{Y_i - \min\{C\}}{\max\{C\} - \min\{C\}}, \tag{8}$$

$$Z_i' = 255 \times \frac{Z_i - \min\{C\}}{\max\{C\} - \min\{C\}},$$

where $\min\{C\}$ and $\max\{C\}$ are minima and maxima of all coordinate values.

The new coordinate space is quantified to integral image representation and three coordinates (X_i', Y_i', Z_i') are considered as the three components R, G, B of a colour-pixel:

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

(X_i', Y_i', Z_i') is the new coordinate of the image display. The steps are shown in Fig. 5. Following the above steps and conversions, the raw data of the skeleton sequence changes into 3-D tensors and then is injected into the learning model as inputs. In Fig. 5, F_N denotes the number of frames in each skeleton sequence. K denotes the number of joints in each frame and it depends on the deep sensors and data acquisition settings.

(4) ROI calculation: During the process of feature extraction to display action, a combination of contour-based distance signal features, flow-based motion features [50, 53], and uniform rotation local binary patterns can be used to define region of interest for feature extraction [17, 30, 40, 44]. Therefore, at this stage, suitable regions for extraction of the features are determined. Depending on the nature of the dataset, the input videos may include certain multi-view activities, which
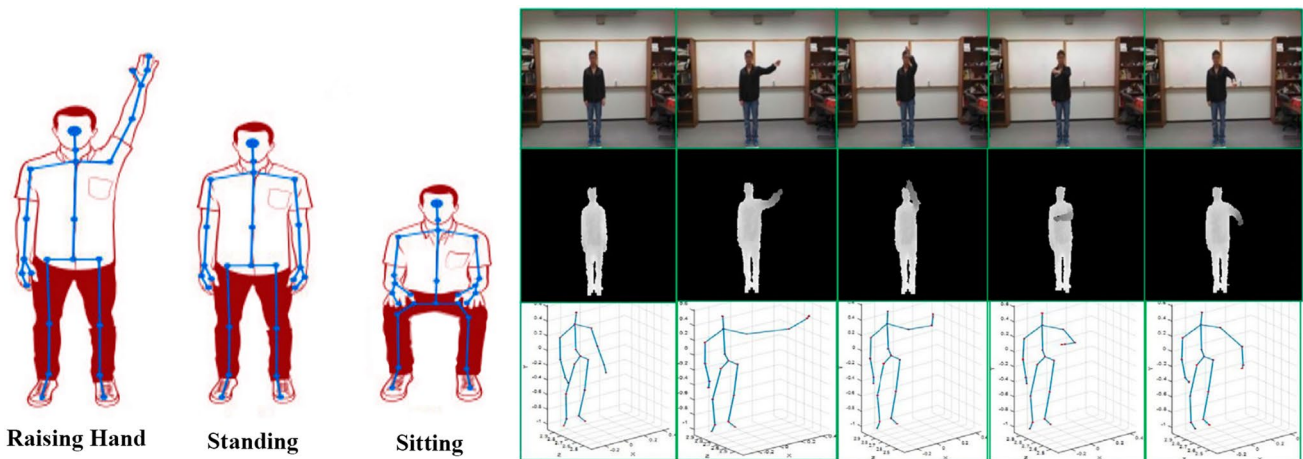

Raising Hand   **Standing**   **Sitting**

Fig. 4 The steps of the appropriate frame region selection and extraction of the skeleton motion [17, 31]
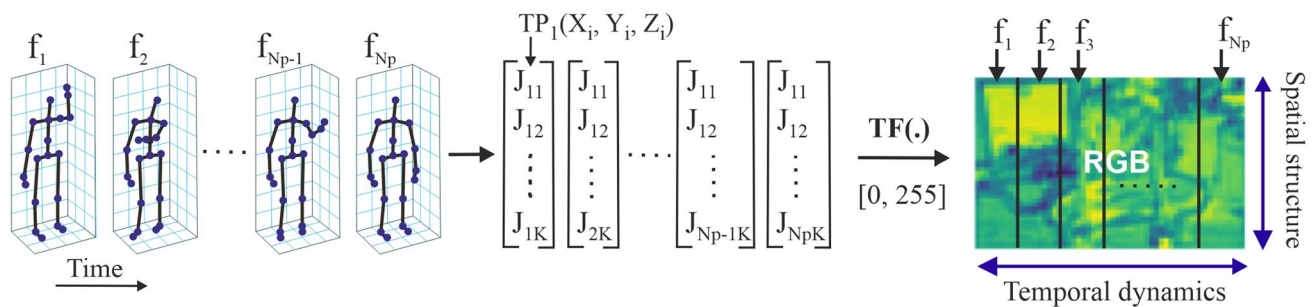
Fig. 5 The stages of converting skeletal sequences to spatio-temporal information to train the model

increase the accuracy of the classification. A similar method is presented in [18, 32] for extraction of entropy-based silhouettes.

Feature Selection

Given that in every movie an action is represented by a sequence of frames, we can perform the action recognition by analysing the contents of multiple frames in a sequence. We propose a series of techniques and methods to find out activities that are close to human perceptions of activities in real life.

One of the human abilities is to predict the upcoming actions based on the previous action sequences. Therefore, to enable a system with such characteristics, deep neural networks, inspired from natural human neural networks is very appropriate. These networks include but not limited to CNN, RNN, and LSTM.

In many research works, the CNN streams are fused with RGB frames and skeletal sequences at feature level and decision level. Classification at decision-making level is also done through voting strategy. As already mentioned, the existence of multidimensional visual data encourages us to combine all vision cues, such as depth and skeletal as in [17]. Many studies focus on the improved skeletal display of CNN architecture.

The CNN features have strong activations values on the human region rather than the background when the network is trained to discriminate between different pedestrians. Benefiting from such attention mechanism, a pedestrian of human can be relocated and aligned within a bounding box [61].

One of the major challenges in exploiting CNN-based methods for detecting skeletal-based action is how to display a temporal skeleton sequence effectively and feed them into a CNN for feature learning and classifications. To overcome this challenge, we encode the temporal and spatial dynamics of skeleton sequences in 2-D image structures. CNN is used to learn the features of the image and its classification

to identify the original skeleton sequences [28]. CNN generally consists of convolutional layers, pooling layers and fully connected layers. In the convolutional layer, filters are very useful for detecting the edges in the images [5, 37, 52, 58, 60]. The pooling layers are generally used in the Max-type, which is intended to reduce the dimension, and the fully-connected layers are used to convert a cubic dimensional data in to a 1-D vector [27].

Based on a stack of N_F input frames, this convolutional network learns to optimise the filters weight; however, it may not be capable of detecting complicated video sequences with complex activities, such as eating or jumping over obstacles. RNNs can resolve this problem [24, 26, 50], by storing only the previous step and consequently avoiding the exploding and vanishing gradient issue. It can be said that the LSTM network is a kind of RNN, which solves the aforementioned issues by holding up a short memory for a long time. In our research, we combine CNN and LSTM for feature selection and accurate action recognition due to their high performance in visual and sequential data. AlexNet is also injected into feature selection for identifying hidden patterns of the visual data. The feature selection operation is performed in parallel in order to speed up the processing, namely parallel duplex LSTMs. A similar approach is considered in [13, 19, 50, 64], and [29]. In other words, we use LSTM for two main reasons:

1. As each frame plays an important role in a video, maintaining the important information of successive frames for a long time will make the system more efficient. The "LSTM" method is appropriate for this purpose.
2. Artificial neural networks and LSTM have greatly gained success in the processing of sequential multimedia data and have obtained advanced results in speech recognition, digital signal processing, image processing, and text data analysis [28, 50, 62].

Figure 6 describes the structure of the proposed deep learning model using a CNN and dual LSTM networks. According to research conducted in [27, 31, 49, 50, 54], LSTM is
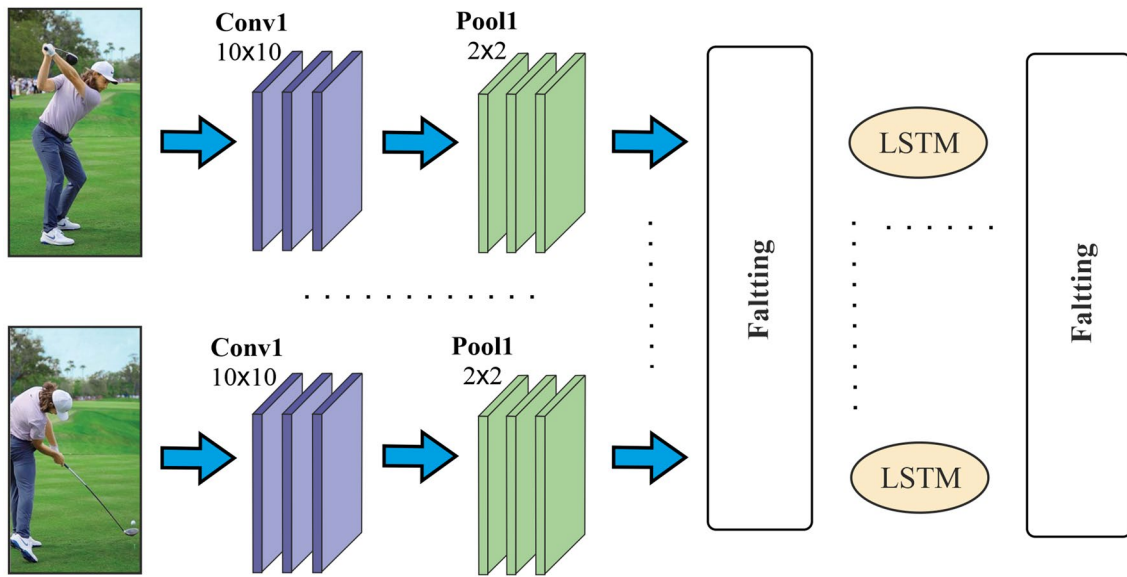
Fig. 6 The simplified architecture of the proposed Deep Learning model based on hybrid CNN and parallel LSTM to select the deep features of a given frame set (e.g. Golf swing action recognition using spatio-temporal information)

capable of learning long-term dependencies, and its special structure includes inputs, outputs and forget gates, which controls long-term sequence recognition. The gates are set by the Sigmoid unit opened and closed during the training. Each LSTM unit is calculated as Eqs. 10, 11, 12, 13, 14, 15, 16:

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

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

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

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

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

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

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

where x_t is the input at time t, f_t is the forget gate at time t which clears the information from the memory cell, if needed, and holds a record of the previous frame. Output gate o_t holds the information about the next step, g is the return unit and has the *tanh* activation function which is computed using the current frame input and the previous s_{t-1} frame status. s_t is the RNN output from the current mode. The hidden mode is calculated from one RNN stage

by activating *tanh* and c_t memory cells. W^i is the input gate weight, W^o is the output gate weight, W^f is the forget gate weight and W^g is the returning unit weight from the LSTM cell. b_i, b_o, b_f and b_g are the biases for input, output, forget and the returning unit gates, respectively.

As the action recognition does not need the intermediate output of the LSTM, we made a final decision making by applying a Softmax classifier on the final state of the RNN network. Training large data with complex sequence patterns (such as video data) can not be identified by a single LSTM cell, so we use stacking multiple LSTM cells to learn long term dependencies in video data.

Transfer Learning: AlexNet

AlexNet is an architecture for solving the challenges of the human action recognition system, trained on the large ImageNet dataset with more than 15 million images. The model is able to identify hidden patterns in visual data more accurately than many other CNN based architectures [38, 50]. Action recognition system requires high training data and computing ability. AlexNet is embedded in the architecture of our model to extract the higher performing features because the pre-trained AlexNet does not have any negative impacts on the performance of the system.

The AlexNet architectural parameters are presented in Table 2. It has six layers of convolution, three layers of pooling and three fully connected layers. Each layer is followed by a non-linear ReLU activation function and the vector of extracted features from the FC8 layer is 1000-dimensional.

Table 2 AlexNet architecture specifications

Layers	Conv1	Pool1	Conv2	Pool2	Conv3	Conv4	Conv5	Pool5	FC6	FC7	FC8
Kernel	11×11	3×3	5×5	3×3	3×3	3×3	3×3	3×3	-	-	-
Stride	4	2	1	2	1	1	1	2	-	-	-
Channels	96	96	256	256	384	384	256	256	4096	4096	1000

KNN-Softmax Classifier

Classification is usually done in deep neural networks based on Softmax function. The Softmax classifier practically is placed after the last layer in the deep neural network. In fact, the result of the convolutional and pooling layers (a feature vector $p^l = [p_1, ..., p_I]$) is the input of the Softmax [5, 60]. After forward propagation, weighs are updated, and errors are minimised through an stochastic gradient descent (SGD) optimisation on several training examples and iterations. Back-propagation balances the weights by calculating the the gradient of convolution weights. In case of large number of classes the Softmax does not perform very well. This is normally due to two main reasons: when the number of parameters is large, the last layer fails to increase the forward–backward speed; furthermore, syncing GPUs will be difficult as well [52, 60].

In this article, we use KNN when the number of classes is high and Softmax fails to perform well. After classifying by Sofmax, if it fails (that is the probability of closeness of action to two classes or several classes), then KNN should be used. KNN uses Euclidean distance [41] and Hamming distance to detect the similarity between two feature vectors. As previously mentioned, $p^l = [p_1, ..., p_I]$ is a classifier input which holds

$$p_I = \{(x_i, y_j), \ i = 1, 2, ..., n_I\},$$

where x_i is the number of extracted features and y_j is the equivalent label for each feature set of x_i. We use Euclidean distance in the KNN classifier, with $k = 10$ and squared inverse distance weights [41]. The Euclidean distance is formulated as Eq. 17.

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

Assuming u is a new instance with a label y_j, in order to find $v + 1$ and closest neighbour to u, the distance formula with $d(u, x_i)$ can be determined as Eq. 18.

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

We normalise $d(u, x_i)$ by the kernel function and weighing according to Eq. 19.

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

The final membership function of weighted K-nearest neighbour (W-KNN) is formulated as follows:

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

Experimental Results

In this section, we evaluate the proposed method on the UCF101 dataset as a common benchmarking dataset based on the accuracy criterion, followed by discussion on the experimental results. The dataset is divided into three parts: training, testing, and validation, based on 60%, 20% and 20% splits, respectively. In Fig. 7, examples of dataset are shown. To implement the proposed model we use Python 3 and TensorFlow deep learning framework.

In our evaluations, we compare the proposed method with eight state-of-the-art methods using the accuracy criterion.

Applying Lipstick Playing Guitar Golf Swing

Fig. 7 Sample frames and activities from the UCF101 video dataset

UCF101 Dataset

The UCF101 dataset is a complex dataset due to many action categories. Some categories include variety of actions, such as sport related actions. The videos are captured in different lighting conditions, gestures, and viewpoint. One of the major challenges in this dataset is the mixture of natural realistic actions and the actions played by various individuals and actors, while in other datasets, the activities and actions are usually performed by one actor only [29, 38, 41, 50]. The UCF101 dataset includes 101 action classes, over 13,000 video clips and 27 hours of video data. This dataset contains realistic uploaded videos with camera motion and custom backgrounds. Therefore, UCF101 is considered as a very comprehensive dataset. The action categories in this dataset can generally be considered as five major types: interaction between human and object, body movement, human-to-human interaction, musical instruments, and sport [45, 47, 50]. Figure 7 shows one sample frame of three different video clips and actions from the UCF101 dataset. Sport category is the largest category of the UCF101 dataset and plays an important role in benchmarking. The sport sub-dataset contains 150 videos of sports broadcasts that are captured in cluttered, and dynamic environments. There are 10 action classes and each video corresponds to one action [47]. In some research works such as [4] which are based on temporal template matching, the UCF Sports action has been used for benchmarking purposes. This category is also useful for actions that are related to human body motion such as "Walk" or to human-object interaction such as "Horse-Riding" [47].

In the next sub-sections, we discuss about three types of tests and evaluations that we conducted in this research:

Performance Evaluation

Table 3 presents the outcome of our experiments for the proposed HFR-DL method in terms of classification accuracy. The proposed method shows an improvement rate of 0.8% to 4.47% comparing to eight other state-of-the art method.

As one of the main contributions in the proposed HFR-DL method we conduct an effective use of BGS, HOG and Skeleton methods to improve the results right from the early stage of the preprocessing, and to extract the most informative features in a customised DNN platform that played a major role in accurate action recognition in the wild. The combination of convolution, pooling, fully connected and LSTM units are used to achieve a better feature learning, feature selection, and classification. Therefore, the probability of error in the classification stage is greatly reduced; and furthermore, complex activities are recognised with higher accuracy rate, as well.

Optimum Frame Jumping

As the second experiment, we also evaluated the optimum jump length for the proposed HFR-DL method. Every video is considered as a single input, and then the features of the frames are extracted using one frame out of every x frames. Table 4 shows the evaluation of the proposed method, based on different frame jumps of 4, 6, and 8 and their impact on the performance of the system. Using the frame jump of $J = 6$ we achieved nearly 50% improvement in speed and computational cost of the system in comparison with $J = 4$, while we approximately lost only 1.5% in accuracy rate. Therefore, considering the speed-accuracy trade-off, we selected the frame jump of 6 as the optimum value for our intended application while it still outperforms the similar state-of-the-art method (DB-LSTM) [50].

Confusion Matrix

A confusion matrix contains visualised and quantised information about multiple classifiers using a reference classification system [13, 50]. Figure 8 shows the details of the results on the UCF Sports dataset for the proposed HFR-DL method. Each row represents the predicted class, and each column represents instances of the ground truth classes.

The confusion matrix results confirms that in overall the HFR-DL provides a more consistent confusion matrix comparing the ReHAR method [50], even for the *Golf*, *Run*, and *Walk* actions which are our weakest results with the accuracy rate of 82.6%, 83.42%, and 72.30%, respectively, in contrast to 83.33%, 75.00%, and 57.14% for the ReHAR method.

As per Figure 8, it can also be interpreted that we have examples of "walking", "running" and "Golf" activities which are mistakenly identified as "Kicking", "Skateboarding", and "walking", respectively. These are expectable, as some of these actions have common features that lead to a misclassification. Furthermore, extra objects and people in the background of the scene are among the factors that also lead to a wrong classification. For example, in one of the examined videos "walk-front/006RF1-13902-70016.avi", there is a person who walks on a golf field with a golf pole. The environment is related to golf field and the motion of the golf pole in the background looks like a person is swinging the pole in front of him [13]. This was an examples of misclassifications by the proposed HFR-DL method.

Conclusion

According to the summarised performance report in Table 3, the proposed HFR-DL method led to an improved human action recognition using spatio-temporal information, which were hidden in sequential patterns and features.
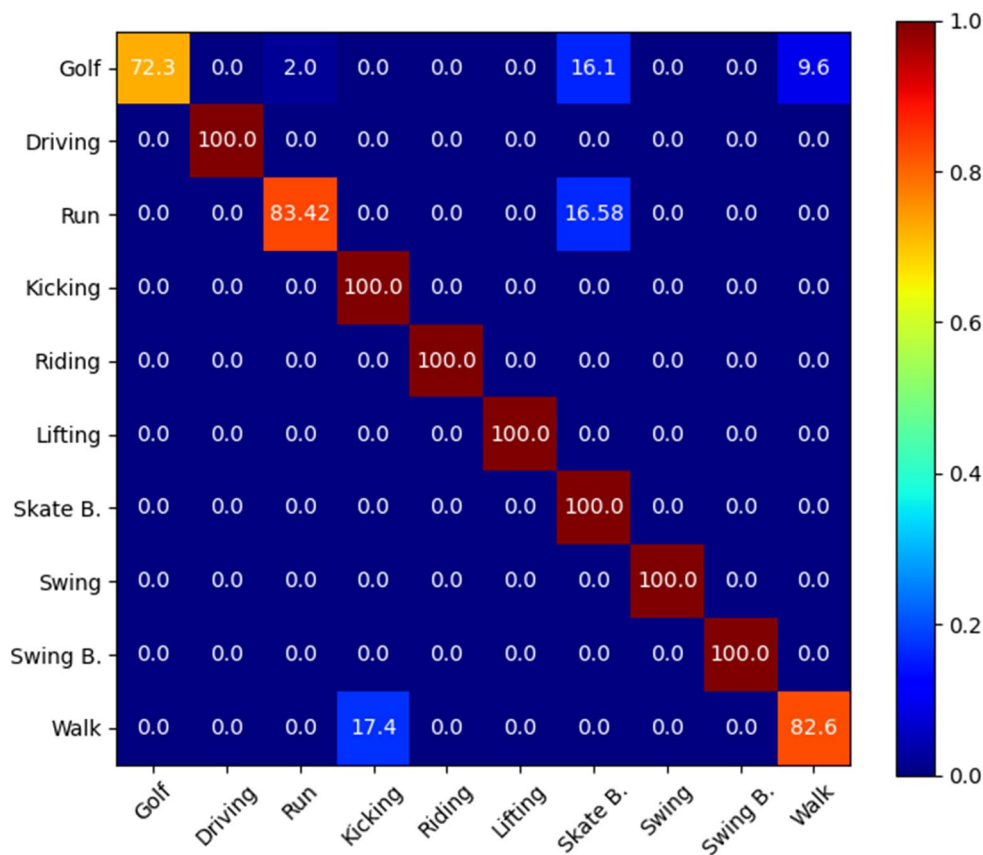
Fig. 8 Confusion matrix of the proposed HFR-DL method for sport actions

Table 3 Performance evaluation on the proposed method compared to eight other methods

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^{2}LSTM [25]	2020	92.80
ECO [64]	2019	93.10
HFR-DL (Proposed method)	2020	**93.90**

Table 4 Evaluation of the proposed method based on different jumps

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%

The combination of BGS, HOG and Skeletal was utilised to analyse and describe the appropriate frames for the preprocessing phase of the model. Then an efficient combination of deep CNN and LSTM were implemented for the feature selection. In the training phase, we initialised the weights randomly and trained the network by reiterating the training stage until getting the minimum errors [10, 22, 31,

43, 59]. The proposed system can reduce the effects of the degradation phenomenon for both training and test phases. It should be noted that degradation phenomena considerably depends on the size of the datasets. This is the reason why the networks with too many layers have higher errors than medium-size networks.

We also extended the skeleton encoding method by exploiting the Euclidean distance and the orientation relationship between the joints. According to Table 4, in both methods, the accuracy level is slightly higher with $J = 4$; however, the time complexity of Jump 6 is significantly less

than the Jump 4. Therefore, the Jump 6 is considered as an optimum trade-off in terms of accuracy and time complexity.

Finally a hybrid Softmax-KNN technique was utilised for the action recognition/classification. The experiment was performed on the commonly used UCF dataset which includes 101 different human actions. The accuracy metric and confusion matrix were assessed and analysed, and the overall results showed the proposed method outperforms in human action recognition compared with eight other state-of-the-art research in the field (Table 3).

Regarding the speed and computational costs, as Table 4 shows, the DB-LSTM method is slightly faster than the proposed HFR-DL method, but less accurate. In general, the suggested method of jumping 6 requires 1.6 second in a medium range Core i7 PC to process a 1-second HD video clip [50]. Depending on the nature of the application in terms of speed and accuracy requirements, this can be simply converted to a 1-1 real-time action recognition solution either by increasing the jump step, or by improving the CPU speed, or by reducing the input video resolution, or by considering a combination of all three factors.

In terms of application, the developed methodology can be applied in various domains, thanks to the diversity of the training dataset. Some of the real-world applications include but not limited to elderly and baby monitoring at home, accident detection, surveillance systems for crime detection and recognition, abnormal human behaviours detection, human-computer interaction, and sports analysis.

As on of the possible future research, we suggest extending this research to improve the current architecture and to predict the future actions of a subject based on the spatio-temporal information, current action, and semantic scene segmentation and understanding.

Compliance with Ethical Standards

Conflict of Interest The authors confirm that there is no conflict of interest in this research.

Human and Animal Rights The article does not contain any studies with human or animal subjects and no ethical concerns involves in this research.

Informed Consent No additional information consent was required from any parties included in this article.

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. Comput Vis 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. Silhouette-based 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 video-based 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. arXiv:1608.08128.

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 multisensor 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.2020.100005.

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;18:1473–88.

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-take-all. 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.