Penerbit Akademia Baru

Journal of Advanced Research in Computing and Applications

Journal homepage: www.akademiabaru.com/arca.html
ISSN: 2462-1927

A Hybrid Model for Human Action Recognition Based on Local Semantic Features

Sri Ganes Palaniapan[1,*], Sokchoo Ng[1]

[1] Faculty of Arts and Science, International University of Malaya-Wales, Kuala Lumpur, Malaysia

ARTICLE INFO

Article history:
Received 19 August 2023
Received in revised form 10 October 2023
Accepted 19 December 2023
Available online 28 December 2023

Keywords:
Illumination; Human action recognition; Encoder-Decoder; iSIFT

ABSTRACT

One of the challenging research in real-time applications like video surveillance, automated surveillance, real-time tracking, and rescue missions is Human Action Recognition (HAR). In HAR complex background of video, illumination, and variations of human actions are domain-challenging issues. Any research can address these issues then only it has a reputation. This domain is complicated by camera settings, viewpoints, and inter-class similarities. Uncontrolled environment challenges have reduced many well-designed models' performance. This paper aims to design an automated human action recognition system that overcomes these issues. Redundant features and excessive computing time for training and prediction are also issues. We propose hybrid model having four modules: the first one is Encoder-Decoder Network (EDNet) need to extract deep features. The second one is an Improved Scale-Invariant Feature Transform (iSIFT) needs to reduce feature redundancy. The third one is Quadratic Discriminant Analysis (QDA) algorithm to reduce feature redundancy. The fourth one is the Weighted Fusion strategy to merge properties of different essential features. The proposed technique is evaluated on two publicly available datasets, including KTH action dataset and UCF-101, and achieves average recognition accuracy of 94% and 90%, respectively.

1. Introduction

The video surveillance or closed-circuit television (CCTV) technology [1,2] is used for better video quality, lower cost and secure communications. Hence the number applications are increased using CCTV in monitoring and recognition of many human action recognitions (HAR). The CCTV collects huge size of raw data and target is ubiquitously [3]. Indeed, target human action recognition has significant and essential. HAR lets machines look at data from sensors and multimedia to figure out what people are doing. In the early 1990s, Foerster *et al.,* showed that HAR was accurate more than 95% of the time [4]. Researchers are trying to improve HAR systems because of how quickly smartphones, wearable devices, and CCTV 15 systems are getting better. HAR is used in surveillance systems [5,6], behaviour analysis [7], gesture recognition [8–10], patient monitoring systems [11,12], ambient assisted living (AAL) [13,14], and different healthcare systems [15,16]. When it comes to

* *Corresponding author.*
E-mail address: sriganes@sgacademy.edu.my (Sri Ganes Palaniapan)

patients, their daily activities need to be tracked so that clinicians can get up-to-date reports and give patients real-time feedback about their progress. Based on the type of data that is processed, HAR can be split into vision-based HAR and sensor-based HAR [17,18]. The first one looks at pictures or videos from optical sensors [6,19], and the second one looks at raw data from wearable and environmental sensors [8,15]. Optical sensors are different from other types because of the data type. Optical sensors can make data in 2D, 3D, and video, while wearable sensors can only make data in 1D. Sensor-based HAR includes things like wearable devices that can track activities like sitting, jogging, running, and sleeping [20]. A sensor doesn't work when a subject is out of range [21] or does something that can't be figured out [28]. Vision-based HAR systems have been using CCTVs for a long time [6]. Video analysis has been used to study systems that can recognise gestures and activities [12,23]. This topic is also helpful for security, surveillance, and interactive applications. In recent years, most research has been done on HAR that is based on vision because it is cheaper and easier to get than data from sensors. So, this study only looks at a small part of HAR studies that are based on vision.

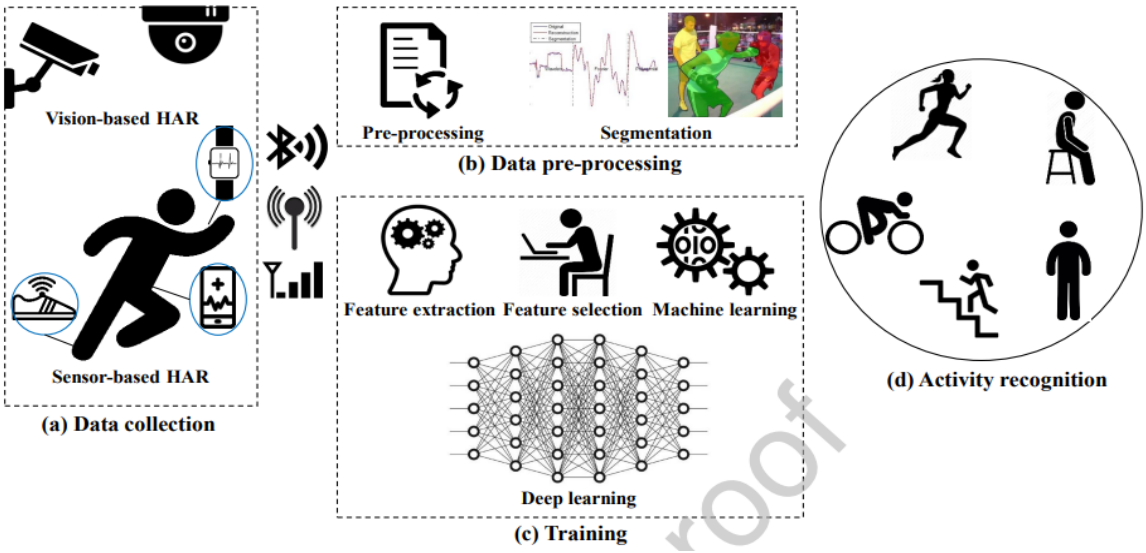

Fig. 1. Human activity recognition framework. A) Data collection B) video pre-processing C) extracting of the features using deep learning frameworks. D) Activity recognition

1.1 Real Time Applications

In the last decade, the number of HAR publications has grown significantly, and each study focuses on a specific activity or behavior.

i. Security and surveillance [24]: In this research, identifying of potential suspicious activity and rising of alarm.
ii. Healthcare [14] : Identifying of daily activities using smart watch and smart phones with high accuracy
iii. Autonomous driving [25] : Classification of driver activity using head and eyes status
iv. Human-robot interaction [26]: designing of algorithm for automatic record the video and classify the human expressions and make robot interaction more precisely.
v. Smart home [27]: In framing of smart home system with features like privacy, reusable, applicability and scalability
vi. Entertainment [28]: identifying of group activity with low features and achieve the good performance regardless of surrounding conditions.

Machine learning (ML) approaches like random forest, Bayesian networks, Markov models, and support vector machine have been utilized to handle the HAR problem for a long time. Traditional machine learning techniques function well with limited data. Inefficient and time-consuming, they require many hand-made pre-processing stages [35]. Shallow characteristics make learning in tiny steps and independently difficult [35,36]. Deep learning has gained popularity because it can detect and recognize objects, classify photos, and interpret natural language [37–40]. Deep learning automatically extracts abstract features over numerous buried layers. This framework works for unsupervised and reinforcement learning [41, 42]. Deep learning-based HAR frameworks are new. Oyedotun *et al.,* taught CNN using static ACPI Source Language (ASL) hand movements [9]. Figure 2 depicts pre-processing and feature extraction.

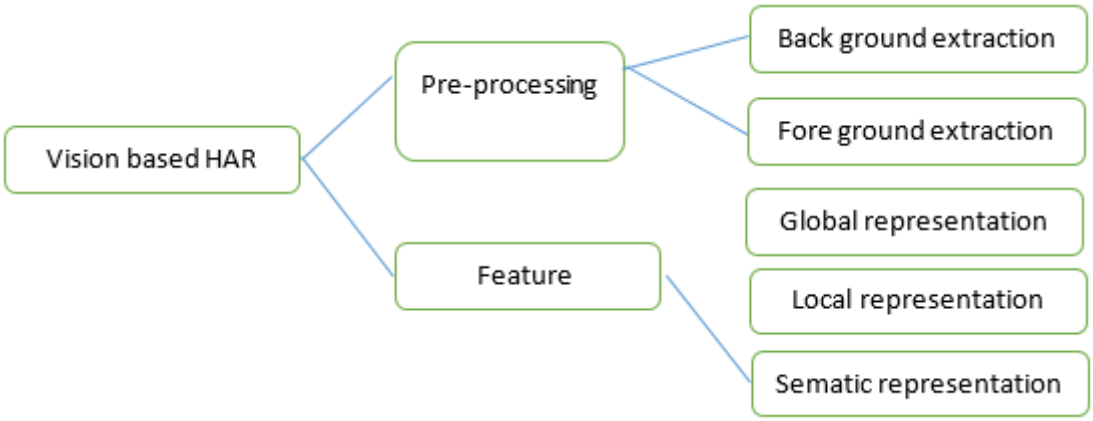

Fig. 2. Basic modules in human action recognition

The rest of the paper is organized as follows: the second section discuss about literature survey of human action recognition. The third section presents the proposed model of EDNet with iSIFT and QDA. And then in section four and five gives proposed model and respective conclusions are mentioned.

2. Literature Review

Visual Geometry Group (VGG 16), ResNet 152, Inception v3, and Inception ResNet v2 are Deep CNN models. Removing the top classification layer produces a vector of CNN model features. The CNN model's spatial features are given to Long Term Short Memory (LSTM) for action classification. LSTM [43] is Python Keras RNN [8]. Shuiwang Ji *et al.,* used 3D convolution to recognise video actions [7]. The architecture employs two adjacent video frames to create data channels. Each channel undergoes convolution and subsampling. The final feature representation was the sum of all channels. Y. Wan *et al.,* [44] suggested a 2-stream convolutional network for spatiotemporal features (LSF CNN). LT-Net gets layered RGB images. ST-Net uses two adjacent frames for optical flow. Linear Support Vector Machine (SVM's) fully connected layer combines spatial and temporal data. Y. Han *et al.,* [1] employed global spatial attention (GSA) to obtain human activity data. The accumulative learning curve (ALC) model weights each intermediate learning outcome to demonstrate which frames are most essential. LSTM's action recognition architecture uses human skeletal joints, GSA, and ALC. Andrej Karpathy [8] created context and fovea streams. Context frames have half the original resolution, while fovea frames have the full resolution. The paper analyses each movie as a bag of short, fixed-length segments to characterise Early Fusion, Late Fusion, and Slow Fusion. Single-frame CNN uses space-time patterns.

3. Proposed Model

The proposed model description having following modules and description as follows.

3.1 Feature Extraction by Encoder

Recurrent Convolution Nueral Network (R-CNN) can find objects, and frame content can be considered regions of interest (ROI). Each ROI has a fully convolutional neural network and a fixed-size bounding box. Final fixed-feature map has linked layers (FCs). The Decoder predicts saliency values using low rank and margin regularizations to increase classification accuracy. This research uses Fast R-CNN to extract saliency features. Super-pixel segmentation preserves image edges and eliminates hollow maps. Figure 1 depicts feature extraction. First, the ResNetnetwork calculates a feature map from an image; second, because the original size of the feature map is small, a De-convolution layer is used to make it bigger and help the Roi projection, which projects the bounding box of each super-pixel to the 100 feature map; finally, the cropped features are pooled into a fixed size by the Roi pooling layer; we then introduce a super-pixel attention mechanism. Some non-important body components may look important, while important body parts may look different and be missed. Human vision prevents mistakes by glancing around. Neural networks can replicate this by expanding each super-perception pixel's field and reducing confusion.
Here a_i output layer and it is calculated using following Eq. (1).

$$a_i = \sum_{j \in M(i)} \delta_{ij} . f_j \tag{1}$$

Where M(i) is neighbor of the i[th] super pixel. F_j is feature vector and δ_{ij} is coefficient which is calculated using

$$\delta_{ij} = \frac{e^{Sim_{ij}}}{\sum_{j \in M(i)} e^{Sim_{ij}}}, \quad Sim_{ij} = f_i^T . M f_i \tag{2}$$

Where M is the weight matrix of fully connected layer, Simij is the similarity between two feature vectors. The attention scope of each super-pixel is limited in its adjacent neighbors, resulting in a sharp reduction in the times of similarity comparison compared to the conventional global mode.

3.2 Action Recognition using Decoder

Structure of bidirectional LSTMs: the proposed bidirectional LSTMs is designed using basic LSTM. The cell structure and pre-requisites are as follows.

$$f_i = \sigma (W_f . h_t + U_f . y_{t-1} + b_f) \tag{3}$$
$$i_t = \sigma (W_i . h_t + U_i . y_{t-1} + b_i) \tag{4}$$
$$g_t = tanh (W_g . h_t + U_g . y_{t-1} + b_g) \tag{5}$$
$$o_t = \sigma (W_o . h_t + U_o . y_{t-1} + b_o) \tag{6}$$
$$c_i = f_t \circ c_{t-1} + i_t \circ g_t \tag{7}$$
$$h_t = o_t \circ tanh(c_{t)} \tag{8}$$

Where f is forget, I is the input, o is the output and g is the gates of data processing. σ is sigmoid function, c and h is the cell and hidden state.

In order to collect spatial information, this paper using two bidirectional LSTMs are used. One for segmented and another one is the skipping LSTMs. These two layers used for extracting the internal relation in different directions.

The proposed LSTM used in the decoder as first layer and its structure is formulated as follows;

$$f_i = \sigma\,(W_f.x_t + U_f.h_{t-1} + b_f) \quad (9)$$
$$i_t = \sigma\,(W_i.x_t + U_i.h_{t-1} + b_i) \quad (10)$$
$$s_t = f_{Multy}\,(\sigma\,(W_s.x_t + U_s.h_{t-1} + b_s) \quad (11)$$
$$g_t = tanh\,(W_g.x_t + U_g.h_{t-1} + b_g) \quad (12)$$
$$o_t = \sigma\,(W_o.x_t + U_o.h_{t-1} + b_o) \quad (13)$$
$$c_i = f_t \circ c_{t-1} \circ (1 - S_{t-1} + i_t \circ g_t \quad (14)$$
$$h_t = o_t \circ tanh\,tanh\,(c_t) \quad (15)$$

The symbols meaning is same as basic LSTM cell. Here segmented gate is introduced (st) to find image boundaries. F_{Multy} is used for classifying n human actions recognition.

Another LSTM is used to skip the states having similar features. The internal structure of this as follows:

$$f_i = \sigma\,(W_f.x_t + U_f.h_{t-1} + b_f) \quad (16)$$
$$i_t = \sigma\,(W_i.x_t + U_i.h_{t-1} + b_i) \quad (17)$$
$$p_t = f_{Multy}\,(\sigma\,(W_p.x_t + U_p.h_{t-1} + b_p) \quad (18)$$
$$g_t = tanh\,(W_g.x_t + U_g.h_{t-1} + b_g) \quad (19)$$
$$o_t = \sigma\,(W_o.x_t + U_o.h_{t-1} + b_o) \quad (20)$$
$$c_i = f_t \circ c_{t-1} \circ (1 - S_{t-1} + i_t \circ g_t \quad (11)$$
$$h_t = o_t \circ tanh(c_t) \quad (22)$$

All symbols meaning is same as previous. The additionally pt is added where this is the skipping gate used to determine whether state is update or copied from previous state.

There are four (4) steps, including a) deep feature extraction using Encoder-Decoder Network (EDNet); b) local feature extraction using improved Scale-Invariant Feature Transform (iSIFT), Quadratic Discriminant Analysis

The Scale Invariant Feature Transform (SIFT) has four steps namely:

i. Extreme localization
ii. Key points accurately positioning
iii. Orientation assignment
iv. Key point descriptor formation. (For more details, please refer to [1] and [2])

3.2.1 Extreme localization

The SIFT is basically depends on finding the extreme of scale-space of image. The scale-space of frame is defined as E(x,y,σ) which can be obtained by variable scale Gaussian kernel G(x,y,σ) of image I(x,y). DoG concept is used for extracting extreme of frame as follows:

$$D = E(x, y, k\sigma) - e(x, y, \sigma) \quad (23)$$

Here non-maxima suppression (NMS) used where need to compare 26 pixels.

3.2.2 Key point accurately-positioning

Here unstable extrema points will be rejected. The unstable extrema points are defined with low contrast and these points can be find using Taylor expansion of the scale-space function D(x,y,σ) at extrema point (x0,y0).

$$D(x, y, \sigma) = D(x_0, y_0, \sigma) + \frac{d\ D^T}{d\ X_0}\ X + \frac{d^2\ D^T}{d\ {X_0}^2} \tag{24}$$

Derivate above equation and equate to 0 then Xmax will be obtained.

$$D(x, y, \sigma) = D(x_0, y_0, \sigma) + \frac{d\ D^T}{d\ X_0}\ X + \frac{d^2\ D^T}{d\ {X_0}^2} \tag{25}$$

The D (Xmax) is less than certain threshold then accept it otherwise reject it.

3.2.3 Orientation assignment

The third step is orientation assignment. This can be done using gradient magnitude and orientation of the magnitude and these factors can be calculated using following equation below.

$$M(x, y) = \sqrt{E(x+1, y) - E(x-1, y)^2} - E(x, y+1) - E(x, y-1)^2 \tag{26}$$

$$\theta(x, y) = (\frac{E(x, y+1) - E(x, y-1)}{E(x+1, y) - E(x-1, y)} \tag{27}$$

3.2.4 Key point descriptor formation and matching

128 feature vectors are computed from 16*16 pixels. Find the Euclidean distance between each image's key points. If the nearest distance to second nearest distance ratio is below a threshold, accept the matching points.

3.3 Improved SIFT (iSIFT)

For each key point, find the gradient values in both vertical and horizontal directions in a 41x41 frame patch. So, a 3042 (39x39x2) dimensional vector is obtained. For example p is number of key points, then p*3042 dimensional matrix is formed. The edge effect is efficiently removed by multiplying weight value w (i,j).

$$w(i, j\ ) = exp(-(i - x)^2\ (i - y)^2 /\ \sigma^2 \tag{28}$$

i,j are horizontal and vertical coordinates of key point and σ scale parameter of the key point. To perform dimensionality reduction principle component analysis (PCA) is used which is computationally effective. Rather than Euclidean distance, in this paper used weighted values to find

high Eigen vectors for each descriptor. By taking of co-efficient 0.7 and 0.3, the d value is calculated using following equation.

$$d = \alpha \sqrt{\sum_{i} \quad (a_i - b_i)^2 + \beta \sum_{i} \quad (\sigma_1 - \sigma_2)^2} \tag{29}$$

The improved SIFT (iSIFT) is well performed in three aspects like: rotation, scale-transformation and robustness to noise.

3.4 Linear Discriminant Analysis (LDA)

Take any classification problem, it can be donoted by Bayes Probability distribution P(Y=k| X=x). LDA can model X distribution for given predicted class(Y)

$$P(Y = k\,|X = x\,) = \frac{P(Y = k\,|X = x\,).\ P(Y = k)}{P(X = x)} = \frac{P(X=x\,|Y=k\,).\ P(Y=k)}{\sum_{j=1}^{K} \quad P(X=x|Y=j\,).\ P(Y=j)} \tag{30}$$

In LDA, the multivariate Normal distribution that is given by Eq. (31)

$$f_k\,(x) = \frac{1}{(2\pi)^2\ |\vartheta|^{1/2}}\ e^{-\frac{1}{2}} \quad (x - \mu_k\,)^T\ \vartheta^{-1}\,(\,x - \mu_k\,) \tag{31}$$

Where μ_k is mean of examples of category k and ϑ covariance of all categories. The Eq. (31) can rewrite with some assumptions as follows:

$$P(Y = k\,|X = x\,) = \frac{\pi_k \quad \frac{1}{(2\pi)^{\,p/2}\ |\vartheta|^{1/2}}\ e^{-\frac{1}{2}} \quad (x - \mu_k\,)^T\ \vartheta^{-1}\,(\,x - \mu_k\,)}{\sum_{j=1}^{K} \quad \frac{1}{(2\pi)^{\,p/2}\ |\vartheta|^{1/2}} e^{-\frac{1}{2}} \quad (x - \mu_j\,)^T\ \vartheta^{-1}\,(\,x - \mu_j\,)} \tag{32}$$

We will get decision boundary by taking of log.

$$\delta_k\,(x) = log\ log\ \pi_k - \frac{1}{2}\ \mu_k^T\,\vartheta^{-1}\,\mu_k + \ x^T\ \vartheta\,.\mu_k \tag{33}$$

For multi-class classification, need to estimate mean, variance and prior proportions and $\left(\frac{p}{2}\right).K = \frac{p(p-1)}{2}\,K.$

3.5 Quadratic Discriminant Analysis (QDA)

It is similar to LDA, but we relaxed the assumption that all classes have equal mean and covariance. So we had to estimate it discretely. The covariance matrix for each class y is given by:

$$\vartheta_k = \frac{1}{N_y - 1} \sum_{y_i = y} \left(x_i - \mu_y\right)\left(x_i - \pi_y\right)^T \tag{34}$$

By taking log both sides

$$\vartheta_k = \frac{1}{N_y - 1} \sum_{y_i = y} \left(x_i - \mu_y\right)\left(x_i - \pi_y\right)^T \tag{35}$$

Figure 3 clearly demonstrates the complete module description of proposed model.

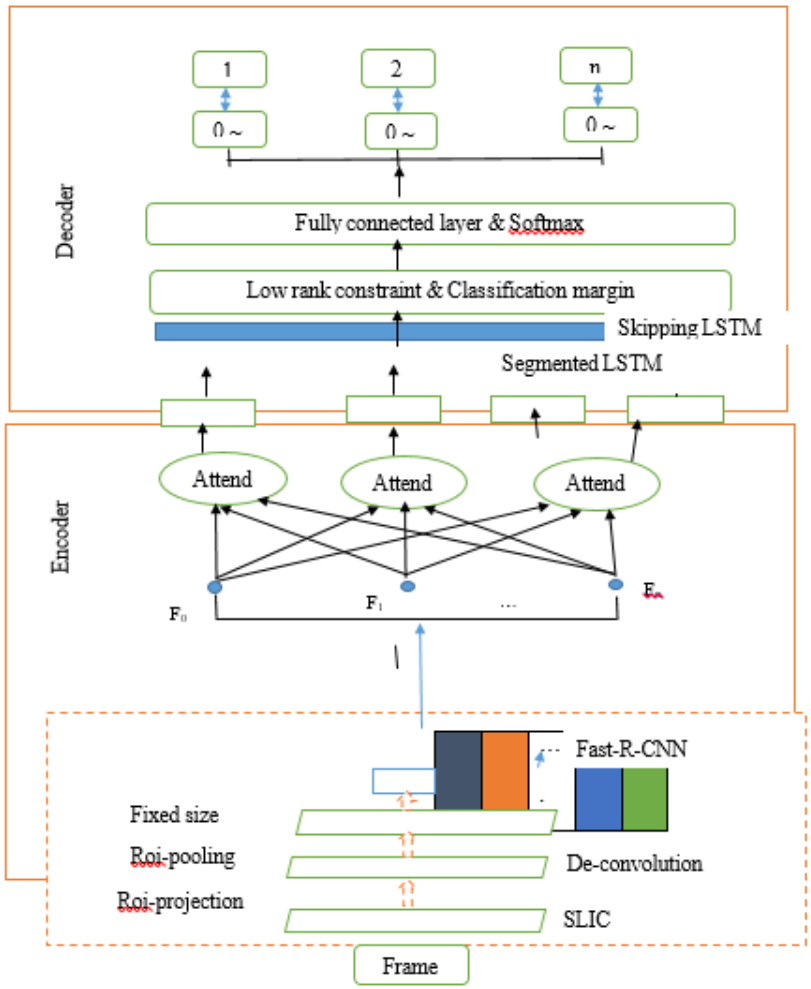

Fig. 3. The complete architecture of proposed model

4. Results and Discussions

Human action dataset KTH3 [45] is popular. The resolution is 160x120 and 6 video activities (walking, jogging, running, boxing, hand-waving and handclapping). Video classification uses 100 frames (Table 1). UCF Sports is an early actions recognition dataset [46]. It combines sports moves. The collection includes 150 720x480 films of 10 human behaviours. This dataset's frames vary by video. Each video has 30-130 frames. Maximum frames per input video, learning rate, frame size, batch size, and max. Table 1 lists Epoch numbers.

Table 1
The experimental setup of proposed method

Parameters	Values
No. of frames	N=100
Size of frame	224x224x3
Learning rate	0.01
The batch size	62
Maximum epochs	40

4.1 Performance Parameters

F1 score is the weighted average of Precision and Recall. Therefore, this score takes both false positive samples and false negative samples into account. The accuracy is calculated using Eq. (36).

The performance metrics are: sensitivity (St), specificity (Spt), precision (Pre), Accuracy (Ac), and F-score (FS). These metrics are derived from confusion matrix and respective equations are written below:

$$\text{Accuracy}=(\text{True Positive} + \text{True Negative}) / ((\text{True Positive} + \text{false negative})+(\text{false Positive}+ \text{True Negative})) \quad (36)$$

$$Se= (\text{True Positive})/(\text{True Positive}+\text{FalseNegative}) \quad (37)$$

$$Sp= (\text{True Negative})/(\text{True Negative} +\text{False Positive}) \quad (38)$$

$$Pr= (\text{True Positive})/(\text{True Positive}+FP) \quad (39)$$

$$F\_Score= (2*\text{True Positive})/(2*\text{True Positive} +\text{False Positive} +\text{False Negative}) \quad (40)$$

Where True positive (TP) reflects the classification of positives, such as cancer, and true negative (TN) represents the classification of negatives, such as infection. Furthermore, false positive (FP) reflects samples that have been erroneously identified, and false negative (FN) indicates cancer images that have been labelled as normal.

For model training, we used 100 input frames, 32 GRU hidden nodes, 64 batch size, 40 epochs, categorical cross-entropy as the loss function, and Stochastic Gradient Descent as the optimizer. KTH dataset video classification accuracy was 93% with Softmax activation in the output layers. Furthermore, to improve the performance of proposed method, two datasets are combined and form a 300 videos. From KTH dataset 150 plus UCF sports 150 videos. To extract potential features from the frame used iSIFT and multi-class GDA is used to reduce the feature dimensionality and classify the human action of the frame. To find accuracy and loss of training and validation sets used Softmax. The purpose of this paper is to classify the n human actions for each dataset used as input.

Table 2
Classification rate of proposed model for each class of human action in KTH dataset

Models	Walking	Jogging	Running	Boxing	Hand-waving	Handclapping
Standard CNN [47]	0.73	0.77	0.75	0.74	0.76	0.80
VGG-16 [48]	0.76	0.72	0.73	0.80	0.80	0.78
ResNet-152 [49]	0.76	0.73	0.69	0.70	0.70	0.73
Inception_v3 [50]	0.79	0.69	0.70	0.69	0.69	0.71
Inception_ResNet_v2 [51]	0.84	0.79	0.82	0.84	0.79	0.85
Improved EDNet+iSIFT	0.86	0.81	0.87	0.86	0.89	0.89

This paper tested many CNN models for human activity classification. The Table 2 shows respective classification rate of proposed model and other models. This table values indicates classification rate of each and every human action of KTH dataset namely: Walking, Jogging, Running, Boxing, Hand-waving and Handclapping. The models experimented are: Standard CNN, VGG-16, ResNet-152, Inception_v3, and Inception_ResNet_v2. The proposed model used local features and deep features of EDNet and then classified each human action. The Table 2 shows respective classification rates of each model with respect to each action. The proposed model shown high classification rates for simple human actins like walking, jogging and running. For complex human activities like boxing, hand waving and hand clapping obtained less classification rates. Out of the all actions walking activity showing high classification rate. For boxing action got less classification rate.

Table 3 show classification rate of proposed model about UCF sports dataset. This dataset having of 10 classes of videos namely: Driving- Side, Golf-Swing, Kicking-Front, Lifting, Riding-Horse, Run-side, Skateboarding-front, Swing-Bench, Swing-side angel, Walk-front and Driving- Side. The proposed model showing high classification rate for human actions like walking-front and lifting. It is clear that if the human action is simple in the frame the proposed model sowing high performance. For complex activities also the proposed model showing satisfactory results.

Table 3
Classification rate of proposed model for each class of human action in UCF dataset

Models	Driving-side	Golf-swing	Kicking-front	Lifting	Riding-horse	Run-side	Skateboarding-front	Swing-bench	Swing-sideangel	Walk-front
Standard CNN [47]	0.76	0.88	0.75	0.74	0.76	0.8	0.72	0.76	0.69	0.73
VGG-16 [48]	0.79	0.83	0.73	0.8	0.8	0.78	0.78	0.79	0.69	0.65
ResNet-152 [49]	0.79	0.84	0.69	0.7	0.7	0.73	0.69	0.69	0.7	0.72
Inception_v3 [50]	0.82	0.8	0.7	0.69	0.69	0.71	0.71	0.73	0.77	0.78
Inception_ResNet_v2 [51]	0.87	0.9	0.82	0.84	0.79	0.85	0.83	0.85	0.85	0.89
Improved EDNet+iSIFT	0.89	0.91	0.86	0.87	0.85	0.89	0.88	0.89	0.89	0.92

The average accuracy of proposed model is given in Table 4. For Table 4, it is clear that the proposed modified EDNet model showing high accuracy on KTH than UCF. It is the reason that KTH dataset having very simple human actions but whereas UCF dataset having some complex human actions along with complex back grounds.

Table 4
The average accuracy of proposed model on each dataset

	Standard CNN	VGG-16	ResNet-152	Inception_v3	Inception_ResNet_v2	Improved EDNet+iSIFT
KTH	0.83	0.83	0.85	0.86	0.89	0.94
UCF	0.79	0.76	0.81	0.82	0.91	0.90

Figure 4 and Figure 5 shows confusion matrix of proposed model with respect to KTH and UCF datasets respectively.

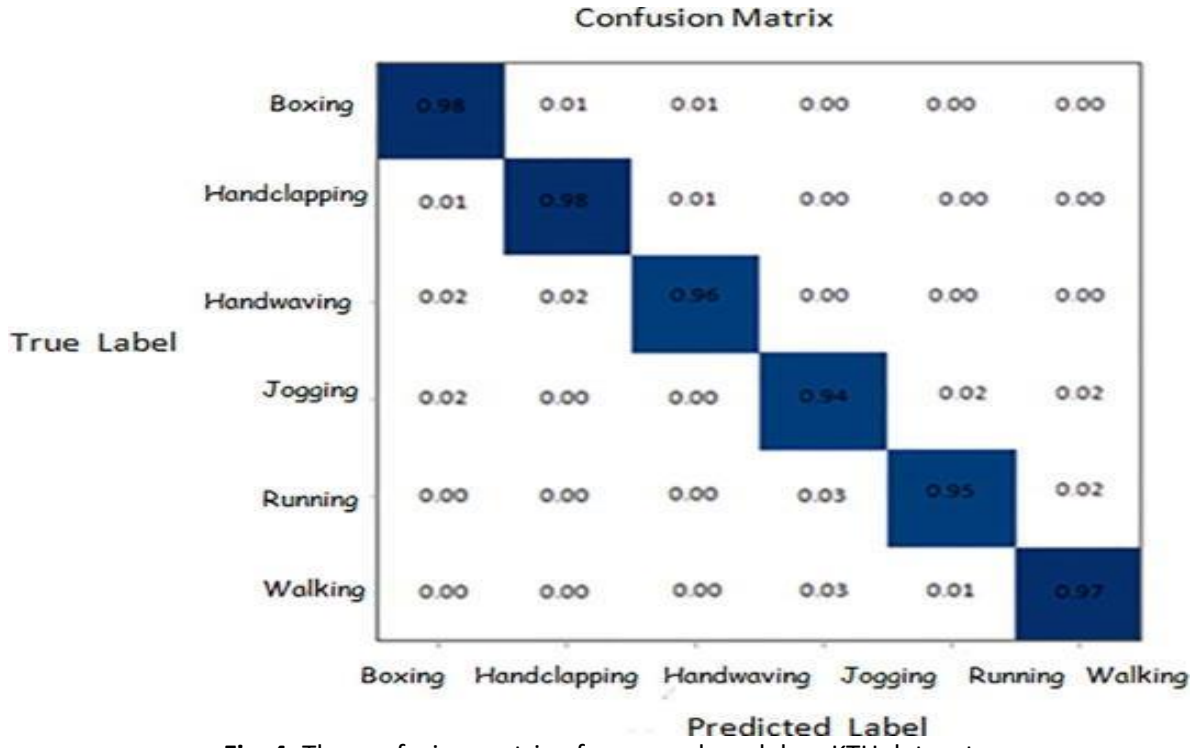

Fig. 4. The confusion matrix of proposed model on KTH dataset

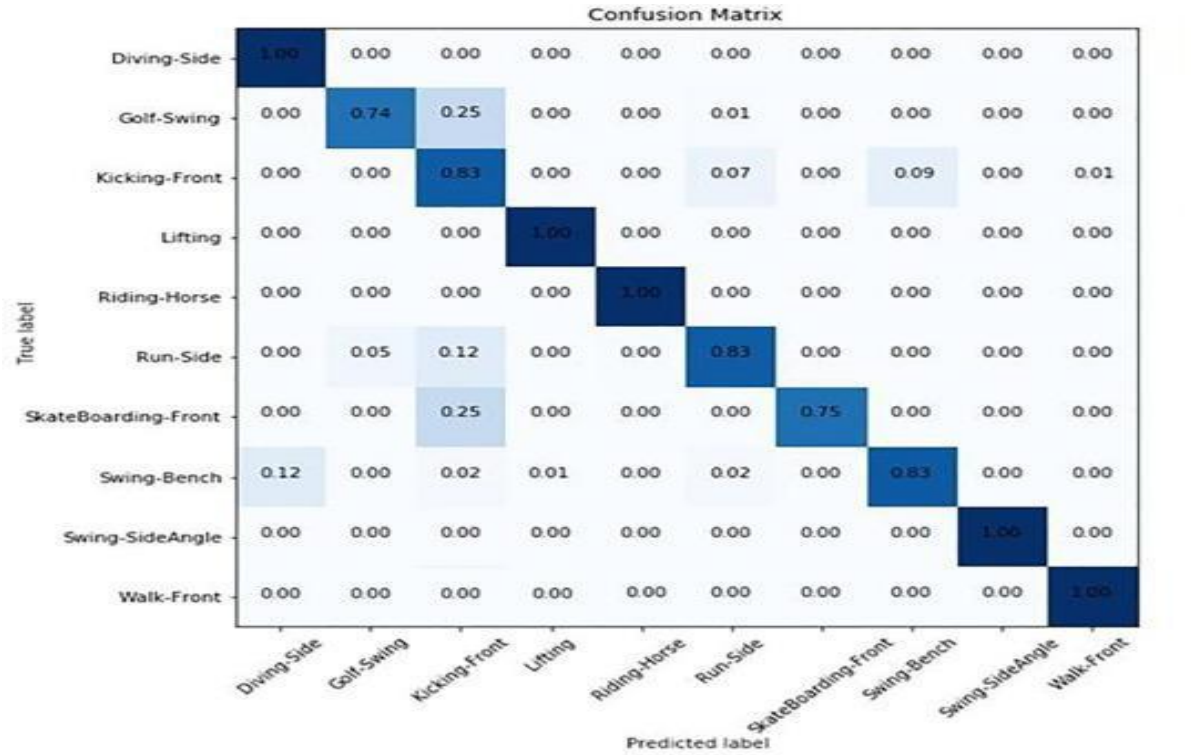

Fig. 5. The confusion matrix of proposed model on UCF dataset

Table 5 shows comparison results of proposed model with state-of-art methods. This table is witness for superiority of proposed model in terms of accuracy.

Table 5
Accuracy values of proposed method and existing state-of-art methods.

KTH	Accuracy (%)	UCF Sport	Accuracy (%)
Latah [52]	90.34	Wang *et al.,* [55]	88
Grushin *et al.,* [53]	90.70	de Oliveira Silva *et al.,* [56]	78.46
Veeriah *et al.,* [54]	93.96	Yeffet and Wolf, [57]	79.2
Proposed modified EdNet+iSIFT	94.00	Proposed modified EdNet+iSIFT	90

5. Conclusions

This work presents a hybrid model for human action recognition in complex videos. This approach is basically depends on potential feature extraction and selecting of optimal features of video content. The deep features are extracted from EDNet with BLSTM and combined with local features extracted from iSIFT feature descriptor. All these features are used for classifying human action in video sequences. The proposed model used GDA for accurate feature reduction and classification. The advantage of this model is combining deep features with local features in each time and each frame of video. The experimental results shows that the proposed model showing high classification rate in all actions of human in normal actions and sports actions. Indeed, the proposed model used for many diverse applications like human activity analysis and body movement of patients in daily analysis will be used tracking the personal health system. If there is an exact tracking of health system, we can suggest any medical precautions and alert with healthy suggestions. For future work, this work is extended to vision-based video analysis with the goal of short-time analysis on more challenging different datasets.

References

[1] Bux, Allah, Plamen Angelov, and Zulfiqar Habib. "Vision based human activity recognition: a review." In *Advances in Computational Intelligence Systems: Contributions Presented at the 16th UK Workshop on Computational Intelligence, September 7–9, 2016, Lancaster, UK*, pp. 341-371. Springer International Publishing, 2017.

[2] Dang, L. Minh, Syed Ibrahim Hassan, Suhyeon Im, and Hyeonjoon Moon. "Face image manipulation detection based on a convolutional neural network." *Expert Systems with Applications* 129 (2019): 156-168. https://doi.org/10.1016/j.eswa.2019.04.005

[3] Muñoz-Cristóbal, Juan A., María Jesús Rodríguez-Triana, Vanesa Gallego-Lema, Higinio F. Arribas-Cubero, Juan I. Asensio-Pérez, and Alejandra Martínez-Monés. "Monitoring for awareness and reflection in ubiquitous learning environments." *International Journal of Human–Computer Interaction* 34, no. 2 (2018): 146-165. https://doi.org/10.1080/10447318.2017.1331536

[4] Foerster, F., and Manfred Smeja. "Joint amplitude and frequency analysis of tremor activity." *Electromyography and clinical neurophysiology* 39, no. 1 (1999): 11-19.

[5] Ji, Xiaopeng, Jun Cheng, Wei Feng, and Dapeng Tao. "Skeleton embedded motion body partition for human action recognition using depth sequences." *Signal Processing* 143 (2018): 56-68. https://doi.org/10.1016/j.sigpro.2017.08.016

[6] Jalal, Ahmad, Yeon-Ho Kim, Yong-Joong Kim, Shaharyar Kamal, and Daijin Kim. "Robust human activity recognition from depth video using spatiotemporal multi-fused features." *Pattern recognition* 61 (2017): 295-308. https://doi.org/10.1016/j.patcog.2016.08.003

[7] Batchuluun, Ganbayar, Jong Hyun Kim, Hyung Gil Hong, Jin Kyu Kang, and Kang Ryoung Park. "Fuzzy system based human behavior recognition by combining behavior prediction and recognition." *Expert Systems with Applications* 81 (2017): 108-133. https://doi.org/10.1016/j.eswa.2017.03.052

[8] Xu, Chi, Lakshmi Narasimhan Govindarajan, and Li Cheng. "Hand action detection from ego-centric depth sequences with error-correcting Hough transform." *Pattern Recognition* 72 (2017): 494-503. https://doi.org/10.1016/j.patcog.2017.08.009

[9] Oyedotun, Oyebade K., and Adnan Khashman. "Deep learning in vision-based static hand gesture recognition." *Neural Computing and Applications* 28, no. 12 (2017): 3941-3951. https://doi.org/10.1007/s00521-016-2294-8

[10] Pigou, Lionel, Aäron Van Den Oord, Sander Dieleman, Mieke Van Herreweghe, and Joni Dambre. "Beyond temporal pooling: Recurrence and temporal convolutions for gesture recognition in video." *International Journal of Computer Vision* 126 (2018): 430-439. https://doi.org/10.1007/s11263-016-0957-7

[11] Capela, Nicole A., Edward D. Lemaire, and Natalie Baddour. "Feature selection for wearable smartphone-based human activity recognition with able bodied, elderly, and stroke patients." *PloS one* 10, no. 4 (2015): e0124414. https://doi.org/10.1371/journal.pone.0124414

[12] Prati, Andrea, Caifeng Shan, and Kevin I-Kai Wang. "Sensors, vision and networks: From video surveillance to activity recognition and health monitoring." *Journal of Ambient Intelligence and Smart Environments* 11, no. 1 (2019): 5-22.

[13] Sankar, S., P. Srinivasan, and R. Saravanakumar. "Internet of things based ambient assisted living for elderly people health monitoring." Research Journal of Pharmacy and Technology 11, no. 9 (2018): 3900-3904. https://doi.org/10.5958/0974-360X.2018.00715.1

[14] Zdravevski, Eftim, Petre Lameski, Vladimir Trajkovik, Andrea Kulakov, Ivan Chorbev, Rossitza Goleva, Nuno Pombo, and Nuno Garcia. "Improving activity recognition accuracy in ambient-assisted living systems by automated feature engineering." *Ieee Access* 5 (2017): 5262-5280. https://doi.org/10.1109/ACCESS.2017.2684913

[15] Qi, Jun, Po Yang, Martin Hanneghan, Stephen Tang, and Bo Zhou. "A hybrid hierarchical framework for gym physical activity recognition and measurement using wearable sensors." *IEEE Internet of Things Journal* 6, no. 2 (2018): 1384-1393. https://doi.org/10.1109/JIOT.2018.2846359

[16] Aviles-Cruz, Carlos, Eduardo Rodriguez-Martinez, Juan Villegas-Cortez, and Andrés Ferreyra-Ramirez. "Granger-causality: An efficient single user movement recognition using a smartphone accelerometer sensor." *Pattern Recognition Letters* 125 (2019): 576-583. https://doi.org/10.1016/j.patrec.2019.06.029

[17] Abdallah, Zahraa S., Mohamed Medhat Gaber, Bala Srinivasan, and Shonali Krishnaswamy. "Activity recognition with evolving data streams: A review." *ACM Computing Surveys (CSUR)* 51, no. 4 (2018): 1-36. https://doi.org/10.1145/3158645

[18] Herath, Samitha, Mehrtash Harandi, and Fatih Porikli. "Going deeper into action recognition: A survey." *Image and vision computing* 60 (2017): 4-21. https://doi.org/10.1016/j.imavis.2017.01.010

[19] Yang, Xiaodong, and YingLi Tian. "Super normal vector for human activity recognition with depth cameras." *IEEE transactions on pattern analysis and machine intelligence* 39, no. 5 (2016): 1028-1039. https://doi.org/10.1109/TPAMI.2016.2565479
[20] Alsinglawi, Belal, Quang Vinh Nguyen, Upul Gunawardana, Anthony Maeder, and Simeon J. Simoff. "RFID systems in healthcare settings and activity of daily living in smart homes: A review." *E-Health Telecommunication Systems and Networks* (2017): 1-17. https://doi.org/10.4236/etsn.2017.61001
[21] Lara, Oscar D., and Miguel A. Labrador. "A survey on human activity recognition using wearable sensors." *IEEE communications surveys & tutorials* 15, no. 3 (2012): 1192-1209. https://doi.org/10.1109/SURV.2012.110112.00192
[22] Cornacchia, Maria, Koray Ozcan, Yu Zheng, and Senem Velipasalar. "A survey on activity detection and classification using wearable sensors." *IEEE Sensors Journal* 17, no. 2 (2016): 386-403. https://doi.org/10.1109/JSEN.2016.2628346
[23] Sanal Kumar, K. P., and R. Bhavani. "Human activity recognition in egocentric video using HOG, GiST and color features." *Multimedia Tools and Applications* 79 (2020): 3543-3559. https://doi.org/10.1007/s11042-018-6034-1
[24] Roy, Produte Kumar, and Hari Om. "Suspicious and violent activity detection of humans using HOG features and SVM classifier in surveillance videos." *Advances in Soft Computing and Machine Learning in Image Processing* (2018): 277-294. https://doi.org/10.1007/978-3-319-63754-9_13
[25] Billah, Tashrif, SM Mahbubur Rahman, M. Omair Ahmad, and M. N. S. Swamy. "Recognizing distractions for assistive driving by tracking body parts." *IEEE Transactions on Circuits and Systems for Video Technology* 29, no. 4 (2018): 1048-1062. https://doi.org/10.1109/TCSVT.2018.2818407
[26] Mojarad, Roghayeh, Ferhat Attal, Abdelghani Chibani, Sandro Rama Fiorini, and Yacine Amirat. "Hybrid approach for human activity recognition by ubiquitous robots." In *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 5660-5665. IEEE, 2018. https://doi.org/10.1109/IROS.2018.8594173
[27] Rafferty, Joseph, Chris D. Nugent, Jun Liu, and Liming Chen. "From activity recognition to intention recognition for assisted living within smart homes." *IEEE Transactions on Human-Machine Systems* 47, no. 3 (2017): 368-379. https://doi.org/10.1109/THMS.2016.2641388
[28] Wateosot, Chonthisa, and Nikom Suvonvorn. "Group activity recognition with an interaction force based on low-level features." *IEEJ Transactions on Electrical and Electronic Engineering* 14, no. 7 (2019): 1061-1073. https://doi.org/10.1002/tee.22901
[29] Hu, Chunyu, Yiqiang Chen, Lisha Hu, and Xiaohui Peng. "A novel random forests based class incremental learning method for activity recognition." *Pattern Recognition* 78 (2018): 277-290. https://doi.org/10.1016/j.patcog.2018.01.025
[30] Xiao, Qinkun, and Ren Song. "Action recognition based on hierarchical dynamic Bayesian network." *Multimedia Tools and Applications* 77 (2018): 6955-6968. https://doi.org/10.1007/s11042-017-4614-0
[31] Ronao, Charissa Ann, and Sung-Bae Cho. "Recognizing human activities from smartphone sensors using hierarchical continuous hidden Markov models." *International Journal of Distributed Sensor Networks* 13, no. 1 (2017): 1550147716683687. https://doi.org/10.1177/1550147716683687
[32] Sok, Pichleap, Ting Xiao, Yohannes Azeze, Arun Jayaraman, and Mark V. Albert. "Activity recognition for incomplete spinal cord injury subjects using hidden Markov models." *IEEE Sensors Journal* 18, no. 15 (2018): 6369-6374. https://doi.org/10.1109/JSEN.2018.2845749
[33] Abidine, Bilal M'hamed, Lamya Fergani, Belkacem Fergani, and Mourad Oussalah. "The joint use of sequence features combination and modified weighted SVM for improving daily activity recognition." *Pattern Analysis and Applications* 21, no. 1 (2018): 119-138. https://doi.org/10.1007/s10044-016-0570-y
[34] Chen, Zhangjie, and Ya Wang. "Infrared–ultrasonic sensor fusion for support vector machine–based fall detection." *Journal of Intelligent Material Systems and Structures* 29, no. 9 (2018): 2027-2039. https://doi.org/10.1177/1045389X18758183
[35] Portugal, Ivens, Paulo Alencar, and Donald Cowan. "The use of machine learning algorithms in recommender systems: A systematic review." *Expert Systems with Applications* 97 (2018): 205-227. https://doi.org/10.1016/j.eswa.2017.12.020
[36] Nguyen, Tan N., H. Nguyen-Xuan, and Jaehong Lee. "A novel data-driven nonlinear solver for solid mechanics using time series forecasting." *Finite Elements in Analysis and Design* 171 (2020): 103377. https://doi.org/10.1016/j.finel.2019.103377
[37] Ijjina, Earnest Paul, and Krishna Mohan Chalavadi. "Human action recognition in RGB-D videos using motion sequence information and deep learning." *Pattern Recognition* 72 (2017): 504-516. https://doi.org/10.1016/j.patcog.2017.07.013

[38] Tan, Tan-Hsu, Munkhjargal Gochoo, Shih-Chia Huang, Yi-Hung Liu, Shing-Hong Liu, and Yun-Fa Huang. "Multi-resident activity recognition in a smart home using RGB activity image and DCNN." *IEEE Sensors Journal* 18, no. 23 (2018): 9718-9727. https://doi.org/10.1109/JSEN.2018.2866806

[39] Young, Tom, Devamanyu Hazarika, Soujanya Poria, and Erik Cambria. "Recent trends in deep learning based natural language processing." *ieee Computational intelligenCe magazine* 13, no. 3 (2018): 55-75. https://doi.org/10.1109/MCI.2018.2840738

[40] Angeleas, Anargyros, and N. Bourbakis. "A two formal languages based model for representing human activities." In *2016 IEEE 28th International Conference on Tools with Artificial Intelligence (ICTAI)*, pp. 779-783. IEEE, 2016. https://doi.org/10.1109/ICTAI.2016.0122

[41] Seyfioğlu, Mehmet Saygın, Ahmet Murat Özbayoğlu, and Sevgi Zubeyde Gürbüz. "Deep convolutional autoencoder for radar-based classification of similar aided and unaided human activities." *IEEE Transactions on Aerospace and Electronic Systems* 54, no. 4 (2018): 1709-1723. https://doi.org/10.1109/TAES.2018.2799758

[42] Nguyen, Tan N., Seunghye Lee, H. Nguyen-Xuan, and Jaehong Lee. "A novel analysis-prediction approach for geometrically nonlinear problems using group method of data handling." *Computer Methods in Applied Mechanics and Engineering* 354 (2019): 506-526. https://doi.org/10.1016/j.cma.2019.05.052

[43] Jagadeesh, B., and Chandrashekar M. Patil. "Video based human activity detection, recognition and classification of actions using SVM." *Transactions on Machine Learning and Artificial Intelligence* 6, no. 6 (2019): 22. https://doi.org/10.14738/tmlai.66.5287

[44] Liu, Ye, Liqiang Nie, Li Liu, and David S. Rosenblum. "From action to activity: sensor-based activity recognition." *Neurocomputing* 181 (2016): 108-115. https://doi.org/10.1016/j.neucom.2015.08.096

[45] http://www.nada.kth.se/cvap/actions/

[46] Soomro, Khurram, Amir Roshan Zamir, and Mubarak Shah. "UCF101: A dataset of 101 human actions classes from videos in the wild." *arXiv preprint arXiv:1212.0402* (2012).

[47] Karpathy, Andrej, George Toderici, Sanketh Shetty, Thomas Leung, Rahul Sukthankar, and Li Fei-Fei. "Large-scale video classification with convolutional neural networks." In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pp. 1725-1732. 2014. https://doi.org/10.1109/CVPR.2014.223

[48] Simonyan, Karen, and Andrew Zisserman. "Very deep convolutional networks for large-scale image recognition." *arXiv preprint arXiv:1409.1556* (2014).

[49] He, Kaiming, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. "Deep residual learning for image recognition." In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770-778. 2016. https://doi.org/10.1109/CVPR.2016.90

[50] Szegedy, Christian, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. "Rethinking the inception architecture for computer vision." In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2818-2826. 2016. https://doi.org/10.1109/CVPR.2016.308

[51] Szegedy, Christian, Sergey Ioffe, Vincent Vanhoucke, and Alexander Alemi. "Inception-v4, inception-resnet and the impact of residual connections on learning." In *Proceedings of the AAAI conference on artificial intelligence*, vol. 31, no. 1. 2017. https://doi.org/10.1609/aaai.v31i1.11231

[52] Latah, Majd. "Human action recognition using support vector machines and 3D convolutional neural networks." *Int. J. Adv. Intell. Informatics* 3, no. 1 (2017): 47-55. https://doi.org/10.26555/ijain.v3i1.89

[53] Grushin, Alexander, Derek D. Monner, James A. Reggia, and Ajay Mishra. "Robust human action recognition via long short-term memory." In *The 2013 International Joint Conference on Neural Networks (IJCNN)*, pp. 1-8. IEEE, 2013. https://doi.org/10.1109/IJCNN.2013.6706797

[54] Veeriah, Vivek, Naifan Zhuang, and Guo-Jun Qi. "Differential recurrent neural networks for action recognition." In *Proceedings of the IEEE international conference on computer vision*, pp. 4041-4049. 2015. https://doi.org/10.1109/ICCV.2015.460

[55] Wang, Limin, Yu Qiao, and Xiaoou Tang. "Action recognition with trajectory-pooled deep-convolutional descriptors." In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4305-4314. 2015. https://doi.org/10.1109/CVPR.2011.5995407

[56] de Oliveira Silva, Vinicius, Flavio de Barros Vidal, and Alexandre Ricardo Soares Romariz. "Human action recognition based on a two-stream convolutional network classifier." In *2017 16th IEEE International Conference on Machine Learning and Applications (ICMLA)*, pp. 774-778. IEEE, 2017. https://doi.org/10.1109/ICMLA.2017.00-64

[57] Yeffet, Lahav, and Lior Wolf. "Local trinary patterns for human action recognition." In *2009 IEEE 12th international conference on computer vision*, pp. 492-497. IEEE, 2009. https://doi.org/10.1109/ICCV.2009.5459201