Computers, Materials & Continua
DOI: 10.32604/cmc.2022.024760
Article

Tech Science Press

A Template Matching Based Feature Extraction for Activity Recognition

Muhammad Hameed Siddiqi[1,*], Helal Alshammari[1], Amjad Ali[2], Madallah Alruwaili[1], Yousef Alhwaiti[1], Saad Alanazi[1] and M. M. Kamruzzaman[1]

[1]College of Computer and Information Sciences, Jouf University, Sakaka, Aljouf, 2014, Kingdom of Saudi Arabia
[2]Department of Computer Science, COMSATS University Islamabad, Lahore Campus, Pakistan
*Corresponding Author: Muhammad Hameed Siddiqi. Email: mhsiddiqi@ju.edu.sa
Received: 30 October 2021; Accepted: 21 December 2021

Abstract: Human activity recognition (HAR) can play a vital role in the monitoring of human activities, particularly for healthcare conscious individuals. The accuracy of HAR systems is completely reliant on the extraction of prominent features. Existing methods find it very challenging to extract optimal features due to the dynamic nature of activities, thereby reducing recognition performance. In this paper, we propose a robust feature extraction method for HAR systems based on template matching. Essentially, in this method, we want to associate a template of an activity frame or sub-frame comprising the corresponding silhouette. In this regard, the template is placed on the frame pixels to calculate the equivalent number of pixels in the template correspondent those in the frame. This process is replicated for the whole frame, and the pixel is directed to the optimum match. The best count is estimated to be the pixel where the silhouette (provided via the template) presented inside the frame. In this way, the feature vector is generated. After feature vector generation, the hidden Markov model (HMM) has been utilized to label the incoming activity. We utilized different publicly available standard datasets for experiments. The proposed method achieved the best accuracy against existing state-of-the-art systems.

Keywords: Activity recognition; feature extraction; template matching; video surveillance

1 Introduction

Human activity recognition has a significant role in many applications such as telemedicine and healthcare, neuroscience, and crime detection. Most of these applications need additional grades of independence like rotation or orientation, scale or size and viewpoint distortions. Rotation might be felt by spinning the template or by utilizing Arctic coordinates; scale invariance might be attained using templates of various size. Having additional parameters of attention infers that the accumulator space becomes bigger; its dimensions rise through one for every extra parameter of attention. Position-invariant template matching infers a 2D parameter space; while, the enlargement of scale and position-invariant template matching needs 3D parameter space [1].

Human activity recognition (HAR) systems try to automatically recognize and examine human activities by acquiring data from different sensors [2]. HAR is frequently associated to the procedure of finding and naming actions using sensory annotations [3]. Generally, a human activity states the movement of one or many parts of the human body, which might be static or composed of numerous primitive actions accomplished in some successive order. Hence, HAR should permit classification the same activity with the similar label even when accomplished by various persons under various dynamic [2].

There are various types of audio and video sensors that can be employed in HAR systems. However, most of them have their own limitations. In audio sensors-based data collection, we may lose the data because of utilizing GPRS to transmit the data. This is one the main disadvantages of the audio-based data collection. Therefore, in this work, we will be using video-sensor (such as 2D RGB camera). HAR system has three basic stages. In the first stage, the noise and environmental distortion will be diminished from the video frame. Furthermore, in this stage, we also segment the human body. In the second stage, we extract the best and informative features from the segmented body. While, in the last stage, a classifier is employed to categorize the incoming activities as shown in Fig. 1.

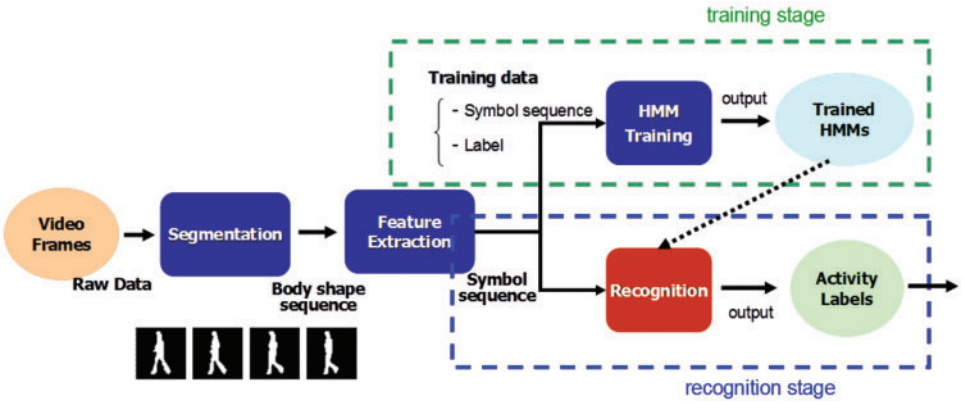

Figure 1: General flow diagram of a HAR system

Commonly, classification has two types: First is the frame-based classification; while, second is sequence-based classification. In frame-based classification, only the present frame is employed with or without a standard frame in order to categorize the human actions from the arriving videos. On the other hand, in the sequence-based classification, the symmetrical movement of the feature pixels is considered among the present frame and the preliminary frame. Therefore, the frame-based classification does not have the ability in such domains in order to classify human activities; hence, the concentration of this work is the sequence-based classification [4].

Accordingly, some latest works have been developed for the sequence-based HAR systems that showed significant performance in various dynamic scenarios. A state-of-the-art system was proposed by [5–8] that is based on the extraction of the individual persons' scene from the sequence of frames. Then, 3D convolutional neural network was utilized in order to detect and classify the corresponding activities of every sequence of frames. Activity-based video summarization is accomplished by saving every person's activity at every time of the incoming video. Similarly, another sequence-based HAR system is proposed by [9] for the identification of the human in healthcare domains. This system takes video frames of COVID-19 patients, then finds for a match inside the grip on frames. In this

system, the Gabor filter is utilized for feature extraction where the personal sample generation formula along with Gabor filter is utilized on input frame in order to collect the optimum and non-redundant Gabor features. Further, deep learning models are employed for matching the human activities with input frame. Furthermore, a robust sequence-based HAR system was proposed by [10] that was assessed on Weizmann and KTH actions datasets. In the pre-processing step of this system, the authors extracted the initial frames from input videos and resized. Then, frame by frame, the region of interest has been considered by employing Blob detection technique and tracing is done with the help of Kalman filter. Furthermore, an ensembled method (which is a group of various techniques such as bi-dimensional empirical mode decomposition, scale invariant feature transform, and wavelet transform) was employed for feature extraction, which extracts the features from moving object. Similarly, this method was also utilized on pre-processed frames in order to extract the best features from multi-scaled frames. Finally, convolution neural network was employed for activity classification. Most of these systems suffer from their own limitations such as the degradation of accuracy in dynamic and naturalistic environments.

Therefore, in this work, we have proposed an adoptive feature extraction method. Essentially in this method, we want to associate a template of an activity frame which will be the template like sub-frame which comprises the silhouette, we are going to search. Therefore, we focus the template on the frame pixels and calculate the equivalent number of pixels in the template correspondent those in the frame. This process is replicated for the whole frame, and the pixel that directed to the optimum match, the best count is estimated to be the pixel where the silhouette (provide via the template) presented inside the frame. For the experiments, we utilized various publicly available standard datasets such as Weizmann dataset [11], KTH action dataset [12], UCF sports dataset [13], and IXMAS action dataset [14] respectively. The proposed technique showed best performance against existing works.

The remaining article is ordered as: Section 2 provides some recent literature review about sequence-based human activity classification systems. The detailed description on the proposed feature extraction is presented in Section 3. The utilized action datasets are explained in Section 4. The Section 5 describes the experimental setup. While, in Section 6, the results along with the discussion are explained. Lastly, in Section 7, the proposed HAR system will be summarized along with little future directions.

2 Related Work

Human activity states the movement of one or many parts of the human body, which might be static or composed of numerous primitive actions accomplished in some successive order. There lots of state-of-the-art methods have been proposed for HAR systems. However, most of them their own limitations. The authors of [15] developed a state-of-the-art system that is based on the architecture of deep learning and V4 inception in order to classify the incoming activities. However, deep learning lacks mutual intelligence, which makes the corresponding systems flimsy and the errors might be very large if the errors are made [16]. Moreover, due to the larger number of layers, the step time of Inception-v4 is suggestively slower in practice [17].

Similarly, an HAR system was proposed by [18] that is based on dissimilarity in body shape, which has been divided into five parts that associate to five fractional occupancy regions. For every frame, the region ratios have been calculated that further be employed for classification purpose. For classification, they utilized the advantages of AdaBoost algorithm that has the greater acumen capacity. However, AdaBoost algorithm cannot be equivalent since every predictor might only be trained after the preceding one has been trained and assessed [19]. A novel ensembled model was

proposed by [20] for Har systems, where they utilized multimodal sensor dataset. They proposed a new data preprocessing method in order to permit context reliant feature extraction from the corresponding dataset to be employed through various machine learning techniques such as linear discriminant, decision trees, kNN, cubic SVM, DNN, and bagged tree. However, every of these algorithms has its own limitations, for instance, kNN, SVM and DNN are frame-based classifiers that do not have the ability to accurately recognize the human activities from incoming sequences of video frames [21].

A new HAR approach was introduced by [22] which is based on entropy-skewness and dimension reduction technique in order to get the condensed features. These features are then transformed into a codebook through serial-based fusion. In order to select the prominent and best features, a genetic algorithm is applied on the created feature codebooks, and for classification, a multi-class SVM has been employed. However, the well-known limitation of the genetic algorithm is that it does not guarantee any variety amongst the attained solutions [23]. Moreover, SVM does not have the capability to correctly classify the human activities from incoming sequences of video frames [21]. A naturalistic HAR system was proposed by [24] for which the human behavior is demonstrated as a stochastic sequence of activities. Activities are presented through a feature vector including both route data such as position and velocity, and a group of local movement descriptors. Activities are classified through probabilistic search of frames feature records on behalf of formerly seen activities. Hidden Markov Models (HMM) was employed for activity classification from incoming videos. However, the local descriptors have one of the main limitations, means that due to this algorithm the results might not be directly transferred to pixel descriptors which cannot be further utilized for classification [25].

A motion-based feature extraction was proposed by [26] for HAR systems. They employed the context information from various resources to enhance the recognition. So, for that purpose, they presented the scene context features which presents the situation of the subject at various levels. Then for classification, the structure of deep neural network was utilized in order to get the higher-level presentation of human actions, which further combined with context features and motion features. However, deep neural network has major limitations such as short transparency and interpretability, and requires huge amount of data [27]. Moreover, the motion features are very scant if human or background comprise non-discriminative features, and sometimes, the extracted features are defective and vanish in succeeding frames [28]. A very recent system was proposed by [29] that is based on various machine learning techniques such as Spatio-temporal interest point, histogram orient gradient, Gabor filter, Harris filter coupled with support vector machine, and they claimed best accuracy. However, the aforementioned techniques have major limitations such as the high-frequency response of Gabor filter produces ring effect closer to the edges which may degrade the accuracy [30]. Moreover, Harris filter requires much time for feature extraction and space to store them, which might not be suitable for naturalistic domains [31].

On the other hand, an automatic sequence-based HAR system was proposed by [32], which is based on group features along with high associations into category feature vectors. Then every action is classified through the amalgamation of Gaussian mixture models. However, Gaussian mixture model is a frame-based classifier which does not has the ability to accurately classify video-based activities. Another sequence-based HAR system was designed by [33] that was based on the neural network. The corresponding networks were created the features database of various activities that were extracted and selected from sequence of frames. Finally, multi-layer feed forward perceptron network was utilized used in order to classify the incoming activities. However, neural network is a vector-based classifier that has low performance against sequence of frames [21]. Similarly, a multi-viewpoint HAR systems was proposed by [34] that was based on two-stream convolutional neural networks integrated with temporal pooling scheme (that builds non-direct feature subspace depictions. However, their accuracy

was very low in naturalistic domains. Moreover, temporal pooling scheme receive the shortcomings in performance generalization as described in [35] that clearly make the benefit of trained features over handmade ones [36].

A multimodal scheme was proposed for human action recognition [37]. This system was based on ascribing importance to the semantic material of label texts instead of just mapping them into numbers. After this step, they modelled the learning framework that reinforces the video description with additional semantic language management and allows the proposed model to the activity recognition without additional required parameters. However, semantic information has some major issues like dimension detonation, data sparseness, incomplete generalization capacity [38].

Accordingly, this work presents an accurate, robust and dynamic feature extraction method that has the ability to extract the best features from the sequence of video frames. In this method, we want to associate a template of an activity frame which will be the template like sub-frame which comprises the silhouette, we are going to search. Therefore, we focus the template on the frame pixels and calculate the equivalent number of pixels in the template correspondent those in the frame. This process is replicated for the whole frame, and the pixel that directed to the optimum match, the best count is estimated to be the pixel where the silhouette (provide via the template) presented inside the frame. By this way, the feature vector is generated. After feature extraction, the hidden Markov model (HMM) has been utilized in order to label the incoming activities.

3 Proposed Feature Extraction Method

In a typical human activity recognition system, the accuracy is completely relying on the feature extraction module. Therefore, we proposed a robust and naturalistic method for feature extraction module. In this method, we want to associate a template of an activity frame which will be the template like sub-frame which comprises the corresponding silhouette. Therefore, we focus the template on the frame pixels and calculate the equivalent number of pixels in the template corresponding to those in the frame. This process is replicated for the whole frame, and the pixel that is directed to the optimum match, the best count is estimated to be the pixel where the silhouette (provided via the template) is inside the frame.

Generally, template matching might be explained as an algorithm of parameter calculation. The parameters describe the template location in the image, which might be defined as a distinct function $\boldsymbol{F}_{i,j}$ that accepts the values in a frame such as the coordinates of the pixels like $(i, j) \in \boldsymbol{S}$. For instance, a set points of 3×3 template may be defined as $\boldsymbol{S} = \{(0, 0, 0)\ (0, 0, 1)\ (0, 1, 0)\ (0, 1, 1)\ (1, 0, 0)\ (1, 0, 1)\ (1, 1, 0)\ (1, 1, 1)\}$.

Let assume that every pixel in the activity frame $Im_{i,j}$ is disturbed by the noise of additive Gaussian, and the corresponding noise is the mean of zero and the unidentified standard deviation that is represented by σ. Hence, the probability at a pixels' template positioned at the coordinates (x, y) ties the equivalent pixel at location $(i, j) \in \boldsymbol{S}$ that is shown by the general distribution

$$Q_{x,y}(i,j) = \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{1}{2}\left(\frac{Im_{i+x,\ j+y} - F_{i,j}}{\sigma}\right)^2} \tag{1}$$

where σ indicates the Gaussian distribution and π is the ratio between the edges and diameter of the image area. Meanwhile, the noise that affect every pixel is autonomous, the probability of the template

at location (x, j) is the fused probability of every pixel that is covered by the template, such as

$$M_{x,y} = \prod_{(i,j)\in \boldsymbol{S}} Q_{x,y}(i,j) \tag{2}$$

Put Eq. (1), then we have

$$M_{x,y} = \left(\frac{1}{\sqrt{2\pi\sigma}}\right)^{k} e^{-\frac{1}{2}\sum_{(i,j)\in \boldsymbol{S}} \left(\frac{Im_{i+x,\ j+y} - F_{i,j}}{\sigma}\right)^2} \tag{3}$$

where k represents the number of points in the corresponding template, which is known as the likelihood function. Commonly, for simpler analysis, this function is expressed in the form of logarithmic. It should be noticed that the scale of the logarithm function does not modify the location of the maximum likelihood. Hence, the updated likelihood function under the logarithm is shown as shown below

$$\log M_{x,y} = n \log\left(\frac{1}{\sqrt{2\pi\sigma}}\right) - \frac{1}{2}\sum_{(i,j)\in \boldsymbol{S}} \left(\frac{Im_{i+x,\ j+y} - F_{i,j}}{\sigma}\right)^2 \tag{4}$$

To select the parameter which enlarges the likelihood function, we need to estimate the maximum likelihood. For instance, the location enlarges the rate of modification of the objective function.

$$\frac{\partial \log M_{x,y}}{\partial_x} = 0 \text{ and } \frac{\partial \log M_{x,y}}{\partial_y} = 0 \tag{5}$$

So,

$$\begin{aligned} \sum_{(i,j)\in \boldsymbol{S}} (Im_{i+x,\ j+y} - F_{i,j}) \left(\frac{\partial Im_{i+x,\ j+y}}{\partial_x}\right) = 0 \\ \sum_{(i,j)\in \boldsymbol{S}} (Im_{i+x,\ j+y} - F_{i,j}) \left(\frac{\partial Im_{i+x,\ j+y}}{\partial_j}\right) = 0 \end{aligned} \tag{6}$$

Hence, the aforementioned equations also provide the solution of the minimization issue, which is given as

$$min\ e = \sum_{(i,j)\in \boldsymbol{S}} (Im_{i+x,\ j+y} - F_{i,j})^2 \tag{7}$$

Here, the estimation of maximum likelihood is equal to picking the location of the template which diminishes the shaped errors. The location where the utmost matches of the frame template is the projected location of the template inside the frame. Hence, if the solution of maximum likelihood has been selected based on the measurement of the matching under the criteria of squared error. This indicates that the result attained via template matching is optimum for frames that are crooked through Gaussian noise. It should be noted that practically assessed noise might be presumed to be the Gaussian noise based on the recommendation of the algorithm of the central limit, though many frames seem to deny this presumption. Alternatively, other errors criteria like the complete difference, instead of the squared difference.

The alternative criteria of the squared error can be derived by substituting Eq. (7), which can be written as:

$$min\ e = \sum_{(i,j)\in \boldsymbol{S}} (Im_{i+x,\ j+y})^2 + (F_{i,j})^2 - 2(Im_{i+x,\ j+y})(F_{i,j}) \tag{8}$$

The final part of the Eq. (8) does not rely on the location of the template (x, y). Intrinsically, it is continuous and might not be diminished. Hence, the optimal in Eq. (8) might be gained through minimizing.

$$min\ e = \sum_{(i,j)\in \boldsymbol{S}} (Im_{i+x,\ j+y})^2 - 2\sum_{(i,j)\in \boldsymbol{S}} (Im_{i+x,\ j+y})(F_{i,j}) \tag{9}$$

If the initial term

$$\sum_{(i,j)\in \boldsymbol{S}} (Im_{i+x,\ j+y})^2 \tag{10}$$

is almost continuous, then the rest of the terms give a quantity of the likeness among the template and frame. Specifically, we might enlarge the cross correlation among the frame and template. Hence, the best location might be calculated as

$$max\ e = \sum_{(i,j)\in \boldsymbol{S}} (Im_{i+x,\ j+y})(F_{i,j}) \tag{11}$$

But, the term of square in Eq. (10) may be changed with location; so, the defined match through Eq. (11) might be poor. Similarly, the variety of the cross-correlation is reliant on the template size, which means that under various environmental conditions, it does not vary. Hence, it is more feasible to utilize either Eqs. (7) or (9) in implementation.

On the other hand, in order to normalize the cross-correlation, Eq. (8) can be defined as below

$$min\ e = 1 - 2\frac{\sum_{(i,j)\in \boldsymbol{S}} (Im_{i+x,\ j+y})(F_{i,j})}{\sum_{(i,j)\in \boldsymbol{S}} (Im_{i+x,\ j+y})^2} \tag{12}$$

Accordingly, the first part is consistent, and hence, the optimal value might be attained as

$$max\ e = \frac{\sum_{(i,j)\in \boldsymbol{S}} (Im_{i+x,\ j+y})(F_{i,j})}{\sum_{(i,j)\in \boldsymbol{S}} (Im_{i+x,\ j+y})^2} \tag{13}$$

Generally, it is feasible to stabilize the window for every activity frame against the template. So,

$$max\ e = \frac{\sum_{(i,j)\in \boldsymbol{S}} (Im_{i+x,\ j+y} - \overline{Im}_{x,y})(F_{i,j} - \bar{F})}{\sum_{(i,j)\in \boldsymbol{S}} (Im_{i+x,\ j+y} - \overline{Im}_{x,y})^2} \tag{14}$$

where $Im_{x,\ y}$ is the average of the pixels $Im_{i+x,\ j+y}$, which is utilized for points inside the window (such as $(i,\ j) \in \boldsymbol{S}$) and $\boldsymbol{F}$ indicates the is the average of the pixels in the corresponding template. Likewise, normalized cross-correlation is presented by Eq. (14), which does not modify the location of the optimal and provides a clarification as the vector of cross-correlation is normalized. Hence,

$$max\ e = \frac{\sum_{(i,j)\in \boldsymbol{S}} (Im_{i+x,\ j+y} - \overline{Im}_{x,y})(F_{i,j} - \bar{F})}{\sqrt{\sum_{(i,j)\in \boldsymbol{S}} (Im_{i+x,\ j+y} - \overline{Im}_{x,y})^2 (F_{i,j} - \bar{F})^2}} \tag{15}$$

If the activity frame and the corresponding template are binary, then such type of combination for template matching will be more beneficial, which might present the regions in the frame or it may comprise the edges. The overall flowchart of the proposed approach is presented in Fig. 2.

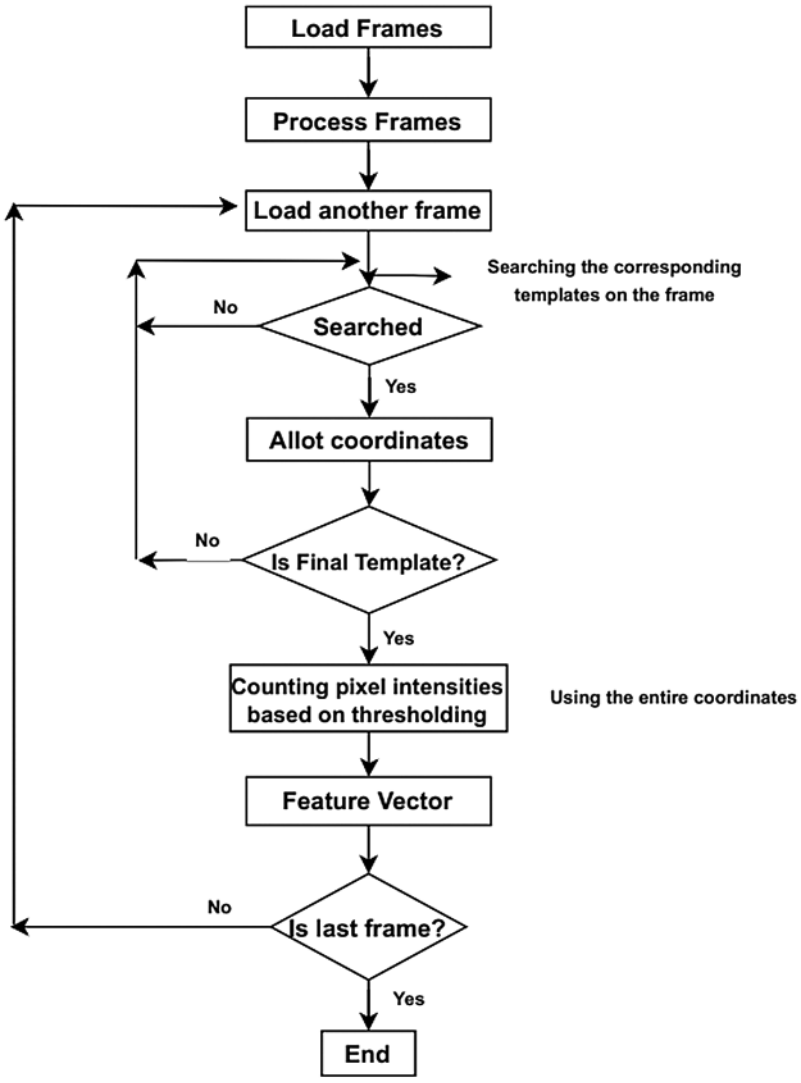

Figure 2: The flowchart of the proposed feature extraction approach

4 Utilized Action Datasets

The proposed feature extraction technique has been tested and validated on four publicly available standard action datasets such as Weizmann dataset, KTH action dataset, UCF sports dataset, and IXMAS action dataset respectively. Every action dataset is explained as below:

4.1 Weizmann Dataset

In this dataset, there are ten various activities which are performed by nine different subjects. The corresponding activities are skip, bend, walk, run, side changing, place jumping, forward jumping, one hand waving (Wave-1) and two hand waving (Wave-2) respectively. The dataset has total 90 activity clips having approximately 15 frames/activity. In order to normalize the entire frames of the dataset, we resized them to 280×340.

4.2 KTH Action Dataset

This dataset was created by 25 subjects who performed total six activities such as walk, boxing, run, clapping, jogging, and waving in various dynamic distinctive situations. This dataset was created under the setting of static camera against consistent background. The dataset has total 2391 sequences under the size of were taken with a frame size 280×320.

4.3 UCF Sports Dataset

This dataset contains 182 videos of total that were assessed through n-fold cross validation scheme from television channels. This dataset was created from various sports persons who were performing different sport matches. Moreover, the entire activities were collected under the settings of static camera. Some of the classes have high intra-class resemblances. There is total nine activities such as diving, run, lifting, skating, golf swimming, kick, walk, baseball swimming, and horse back riding. Each activity frame has a size 280×320.

4.4 IXMAS (INRIA Xmas Motion Acquisition Sequences) Action Dataset

In this dataset, there were total thirteen activities that were performed by eleven subjects. Each actor selected a free angle and location. For each subject, there were corresponding silhouettes in this dataset. We have chosen eight activity classes such as cross arm, walk, turn around, punch, wave, sit down, kick, and get up. This dataset has a view-invariant HAR where the size of each activity frame is size 280×320 (for our experiments). This dataset suffers from high occlusion which may reduce the performance of the proposed approach; therefore, we employed one of our previous methods [39] to normalize the occlusion concern.

5 Experiments Setup

The proposed method was assessed and validated against the following set of experiments.

5.1 First Experiment

This experiment presents the accuracy of the HAR system under the presence of the proposed feature extraction technique. So, for that purpose, we performed four sub-experiments against each dataset in order to show the significance and robustness of the proposed technique.

5.2 Second Experiment

This experiment indicates the role and importance of the designed approach in a typical HAR system. So, we utilized an inclusive set of sub-experiments for such persistence. For these experiments, we employed various state-of-the-art feature extraction methods instead of using the proposed technique.

5.3 Third Experiment

Finally, in this experiment, we compared the accuracy of the proposed method against state-of-the-art systems.

6 Results and Discussions

6.1 First Experiment

In this sub-experiment, we presented the performance of the proposed feature extraction technique against each dataset. For reach dataset, we utilized *n*-fold cross validation structure, which means that every activity is utilized for training and testing respectively. The overall result of the proposed method is shown in Tab. 1 (Weizmann dataset), Tab. 2 (KTH action dataset), Tab. 3 (UCF dataset), and Tab. 4 (IXMAS dataset) respectively.

Table 1: Analysis of the proposed approach on Weizmann dataset

Actions	Bending	Jacking	P-Jump	Running	Side	Skipping	Walk	Wave-1	Wave-2
Bending	**100**	0	0	0	0	0	0	0	0
Jacking	0	**99**	1	0	0	0	0	0	0
P-Jump	0	0	**98**	0	0	0	0	1	1
Running	0	0	0	**99**	0	0	1	0	0
Side	0	1	0	0	**99**	0	0	0	0
Skipping	0	0	0	1	0	**97**	2	0	0
Walk	0	0	0	1	0	0	**99**	0	0
Wave-1	0	0	0	0	0	0	0	**100**	0
Wave-2	0	0	1	0	0	0	0	1	**98**
Average					**98.8%**				

Table 2: Analysis of the proposed approach on KTH action dataset

Actions	Walking	Jogging	Running	Boxing	Waving	Clapping
Walking	**100**	0	0	0	0	0
Jogging	0	**98**	2	0	0	0
Running	2	1	**97**	0	0	0
Boxing	0	0	0	**98**	0	2
Waving	0	0	0	0	**100**	0
Clapping	0	0	0	2	0	**98**
Average				**98.5%**		

It should be noted from Tabs. 1–4 that the common HAR system along with the proposed feature extraction method achieved accuracy on every dataset. From these results, we observed that the proposed method is robust, which means the proposed feature extraction method did not achieve best accuracy only on one dataset but also showed significant performances on other datasets respectively. This is because the averaging intrinsic in the proposed feature extraction method is the reduction of the vulnerability to noise and the maximization stage diminishes defenselessness to occlusion.

Table 3: Analysis of the proposed approach on UCF dataset, where GoS is Golf Swimming, HoBR is Horse Back Riding, and BaS is Baseball Swimming

Actions	Diving	GoS	Kick	Lift	HoBR	Running	Skating	BaS	Walk
Diving	**100**	0	0	0	0	0	0	0	0
GoS	0	**98**	0	0	0	0	0	2	0
Kick	0	0	**97**	0	1	0	1	1	0
Lift	0	0	0	**100**	0	0	0	0	0
HoBR	0	0	0	1	**98**	0	0	0	1
Running	0	0	0	0	0	**98**	0	0	2
Skating	0	0	1	0	0	0	**99**	0	0
Bas	0	0	0	0	0	0	0	**99**	1
Walk	0	0	0	0	0	2	0	0	**98**
Average					**98.6%**				

Table 4: Analysis of the proposed approach on IXMAS dataset, where CrA is Cross Arm, SiD is Sit Down, GeU is Get Up, TuA is Turn Around

Actions	CrA	SiD	GeU	TuA	Kicking	Punching	Waving	Walking
CrA	**100**	0	0	0	0	0	0	0
SiD	0	**99**	0	0	1	0	0	0
GeU	0	1	**97**	0	0	2	0	0
TeA	0	0	2	**98**	0	0	0	0
Kicking	0	0	1	1	**98**	0	0	0
Punching	1	0	0	0	1	**98**	0	0
Waving	0	1	0	0	0	0	**99**	0
Walking	0	0	0	0	0	0	0	**100**
Average					**98.6%**			

6.2 Second Experiment

For this experiment, we performed a group of sub-experiments in order to show the performance of the proposed HAR system. The entire sub-experiments were performed on every dataset under the absence of the proposed feature extraction method. For these sub-experiments, we utilized recent well-known feature extraction techniques such as wavelet transform [4], Curvelet transform [40], local binary pattern (LBP) [41], local directional pattern (LDP) [42], and stepwise linear discriminant analysis (SWLDA) [43] respectively. The overall results of the sub-experiments are presented in Tabs. 5–24 against Weizmann dataset, KTH dataset, UCF dataset, and IXMAS dataset of various activities.

Table 5: Analysis of a common HAR system along with existing wavelet transform (without employing the proposed approach) on Weizmann dataset

Actions	Bending	Jacking	P-Jump	Running	Side	Skipping	Walk	Wave-1	Wave-2
Bending	**77**	2	4	3	2	4	2	4	2
Jacking	4	**81**	2	3	2	4	1	2	1
P-Jump	1	2	**82**	2	4	3	0	4	2
Running	3	1	3	**76**	3	4	5	2	3
Side	2	1	3	2	**86**	2	3	0	1
Skipping	3	0	3	4	3	**79**	3	1	4
Walk	2	1	2	4	2	3	**80**	4	2
Wave-1	2	3	1	1	3	1	0	**87**	2
Wave-2	0	2	1	4	2	0	3	4	**84**
Average					**81.3%**				

Table 6: Analysis of a common HAR system along with existing Curvelet transform (without employing the proposed approach) on Weizmann dataset

Actions	Bending	Jacking	P-Jump	Running	Side	Skipping	Walk	Wave-1	Wave-2
Bending	**80**	1	3	2	4	2	0	5	3
Jacking	1	**75**	5	4	2	3	4	3	3
P-Jump	2	1	**85**	2	3	1	0	2	4
Running	0	0	3	**88**	2	4	0	3	0
Side	1	4	2	3	**79**	0	3	5	3
Skipping	0	0	2	1	1	**89**	2	0	5
Walk	0	2	0	4	0	2	**90**	2	0
Wave-1	2	3	2	1	3	1	2	**83**	3
Wave-2	2	0	4	1	2	4	3	3	**81**
Average					**83.3%**				

Table 7: Analysis of a common HAR system along with existing local binary patter (LBP) (without employing the proposed approach) on Weizmann dataset

Actions	Bending	Jacking	P-Jump	Running	Side	Skipping	Walk	Wave-1	Wave-2
Bending	**83**	1	1	1	3	2	4	2	3
Jacking	1	**75**	4	3	4	2	3	5	3
P-Jump	4	3	**79**	1	3	4	3	1	2
Running	3	4	2	**77**	1	3	6	1	3

(Continued)

Table 7: Continued

Actions	Bending	Jacking	P-Jump	Running	Side	Skipping	Walk	Wave-1	Wave-2
Side	2	3	2	0	**80**	5	2	3	3
Skipping	2	5	2	3	0	**82**	1	4	1
Walk	3	2	3	6	3	2	**76**	2	3
Wave-1	5	2	3	1	5	4	2	**70**	8
Wave-2	2	1	2	1	2	1	2	5	**84**
Average					**78.4%**				

Table 8: Analysis of a common HAR system along with existing local directional pattern (LDP) (without employing the proposed approach) on Weizmann dataset

Actions	Bending	Jacking	P-Jump	Running	Side	Skipping	Walk	Wave-1	Wave-2
Bending	**70**	3	5	3	4	2	3	4	6
Jacking	4	**75**	3	6	1	3	2	1	5
P-Jump	4	3	**72**	4	3	4	3	3	3
Running	1	5	3	**80**	2	3	0	4	2
Side	2	4	0	3	**81**	2	3	5	1
Skipping	2	1	3	5	2	**77**	3	4	3
Walk	3	0	5	4	3	1	**79**	1	4
Wave-1	3	2	4	1	1	2	4	**81**	2
Wave-2	2	0	2	1	2	1	4	4	**84**
Average					**77.7%**				

Table 9: Analysis of a common HAR system along with existing stepwise linear discriminant analysis (SWLDA) (without employing the proposed approach) on Weizmann dataset

Actions	Bending	Jacking	P-Jump	Running	Side	Skipping	Walk	Wave-1	Wave-2
Bending	**69**	3	6	2	3	2	5	3	7
Jacking	4	**75**	1	3	5	4	3	1	4
P-Jump	2	3	**78**	2	3	3	3	3	3
Running	3	5	2	**70**	4	1	7	5	3
Side	2	3	5	3	**73**	4	2	3	5
Skipping	3	2	4	1	3	**80**	4	3	0
Walk	4	2	2	3	1	4	**79**	2	3
Wave-1	5	2	3	4	3	5	2	**71**	5
Wave-2	3	0	2	5	3	4	2	5	**76**
Average					**74.6%**				

Table 10: Analysis of a common HAR system along with existing wavelet transform (without employing the proposed approach) on KTH action dataset

Actions	Walking	Jogging	Running	Boxing	Waving	Clapping
Walking	**76**	6	7	4	5	2
Jogging	5	**80**	6	4	3	2
Running	6	5	**73**	6	3	7
Boxing	5	4	2	**79**	4	6
Waving	3	2	4	5	**82**	4
Clapping	4	2	3	6	1	**84**
Average				**79.0%**		

Table 11: Analysis of a common HAR system along with existing Curvelet transform (without employing the proposed approach) on KTH action dataset

Actions	Walking	Jogging	Running	Boxing	Waving	Clapping
Walking	**80**	5	6	2	4	3
Jogging	5	**79**	8	5	2	1
Running	3	5	**83**	4	3	2
Boxing	3	5	4	**77**	5	6
Waving	2	4	3	4	**83**	4
Clapping	4	3	5	10	2	**76**
Average				**79.7%**		

Table 12: Analysis of a common HAR system along with existing local binary pattern (LBP) (without employing the proposed approach) on KTH action dataset

Actions	Walking	Jogging	Running	Boxing	Waving	Clapping
Walking	**82**	3	5	2	5	3
Jogging	5	**85**	4	1	3	2
Running	4	5	**80**	3	5	3
Boxing	3	4	3	**84**	1	5
Waving	5	4	5	4	**78**	4
Clapping	4	3	2	6	4	**81**
Average				**81.7%**		

Table 13: Analysis of a common HAR system along with existing local directional pattern (LDP) (without employing the proposed approach) on KTH action dataset

Actions	Walking	Jogging	Running	Boxing	Waving	Clapping
Walking	**84**	5	4	2	3	2
Jogging	3	**88**	4	2	1	2
Running	4	3	**90**	0	1	2
Boxing	4	2	4	**81**	3	6
Waving	3	2	3	4	**86**	2
Clapping	1	1	1	5	3	**89**
Average				**86.3%**		

Table 14: Analysis of a common HAR system along with existing stepwise linear discriminant analysis (SWLDA) (without employing the proposed approach) on KTH action dataset

Actions	Walking	Jogging	Running	Boxing	Waving	Clapping
Walking	**90**	4	3	0	1	2
Jogging	2	**89**	4	3	0	2
Running	5	4	**86**	3	1	1
Boxing	4	3	4	**80**	4	6
Waving	2	3	2	3	**88**	2
Clapping	3	4	2	5	2	**84**
Average				**86.2%**		

Table 15: Analysis of a common HAR system along with existing wavelet transform (without employing the proposed approach) on UCF dataset, where GoS is Golf Swimming, HoBR is Horse Back Riding, and BaS is Baseball Swimming

Actions	Diving	GoS	Kick	Lift	HoBR	Running	Skating	BaS	Walk
Diving	**79**	1	3	2	4	2	4	3	2
GoS	4	**78**	5	2	3	2	1	2	3
Kick	3	4	**71**	4	3	4	3	4	4
Lift	4	3	5	**80**	3	0	2	1	2
HoBR	3	2	1	2	**81**	3	2	3	3
Running	2	3	4	0	2	**83**	2	1	3
Skating	4	3	3	4	2	3	**74**	4	3
Bas	2	3	2	3	2	3	2	**81**	2
Walk	2	1	3	2	1	2	3	0	**86**
Average					**79.2%**				

Table 16: Analysis of a common HAR system along with existing Curvelet transform (without employing the proposed approach) on UCF dataset, where GoS is Golf Swimming, HoBR is Horse Back Riding, and BaS is Baseball Swimming

Actions	Diving	GoS	Kick	Lift	HoBR	Running	Skating	BaS	Walk
Diving	**77**	3	2	3	4	3	2	1	5
GoS	3	**79**	3	3	0	3	4	3	2
Kick	3	4	**75**	3	2	4	1	5	3
Lift	1	3	2	**80**	5	2	3	1	3
HoBR	3	0	3	2	**82**	3	5	4	3
Running	3	4	2	5	3	**75**	3	2	3
Skating	3	4	3	2	1	3	**78**	2	4
Bas	3	2	3	0	3	1	3	**83**	2
Walk	1	2	1	4	2	2	1	3	**84**
Average					**79.2%**				

Table 17: Analysis of a common HAR system along with existing local binary patter (LBP) (without employing the proposed approach) on UCF dataset, where GoS is Golf Swimming, HoBR is Horse Back Riding, and BaS is Baseball Swimming

Actions	Diving	GoS	Kick	Lift	HoBR	Running	Skating	BaS	Walk
Diving	**70**	4	3	4	3	4	5	3	4
GoS	3	**74**	3	5	4	3	4	3	1
Kick	3	2	**77**	3	1	4	3	2	5
Lift	4	3	4	**69**	4	3	4	5	4
HoBR	3	1	4	5	**79**	3	0	3	2
Running	2	1	4	3	2	**81**	2	0	5
Skating	3	5	3	2	4	5	**74**	1	3
Bas	3	2	4	3	4	3	4	**75**	2
Walk	4	3	2	3	1	5	2	4	**76**
Average					**75.0%**				

As can be seen from Tabs. 5–24 that under the absence of the proposed approach (like feature extraction technique), the HAR system did not achieved best accuracy. This is because the inattentiveness to noise and occlusion are the main benefits of the proposed feature extraction technique. Noise may happen in any frame of the incoming video. Similarly, there might be low noise in digital photographs; however, in image processing it is made inferior through edge detection by the quality of variation procedures. Furthermore, shapes might simply be obstructed or hidden, for instance, a person may walk behind a streetlamp, or illumination may be one of reasons to create occlusion. The

averaging intrinsic in the proposed feature extraction method is the reduction of the vulnerability to noise and the maximization stage diminishes defenselessness to occlusion.

Table 18: Analysis of a common HAR system along with existing local directional patter (LDP) (without employing the proposed approach) on UCF dataset, where GoS is Golf Swimming, HoBR is Horse Back Riding, and BaS is Baseball Swimming

Actions	Diving	GoS	Kick	Lift	HoBR	Running	Skating	BaS	Walk
Diving	**80**	0	4	2	4	2	3	1	4
GoS	3	**78**	3	1	3	1	3	5	3
Kick	3	2	**77**	1	3	5	0	6	3
Lift	3	2	4	**75**	3	2	4	3	4
HoBR	2	3	1	2	**83**	2	3	1	3
Running	3	2	3	1	3	**79**	3	1	5
Skating	3	4	3	4	3	3	**74**	3	3
Bas	2	1	2	3	2	1	4	**85**	0
Walk	2	3	2	3	2	5	2	0	**81**
Average					**79.1%**				

Table 19: Analysis of a common HAR system along with existing stepwise linear discriminant analysis (SWLDA) (without employing the proposed approach) on UCF dataset, where GoS is Golf Swimming, HoBR is Horse Back Riding, and BaS is Baseball Swimming

Actions	Diving	GoS	Kick	Lift	HoBR	Running	Skating	BaS	Walk
Diving	**72**	3	2	5	4	2	4	3	5
GoS	3	**77**	3	2	1	3	5	2	4
Kick	2	3	**80**	1	4	2	5	0	3
Lift	3	2	2	**83**	2	2	2	2	2
HoBR	3	2	3	4	**74**	3	4	5	2
Running	2	3	1	3	2	**80**	3	1	5
Skating	4	3	4	1	4	3	**76**	2	3
Bas	3	5	3	2	1	2	1	**79**	4
Walk	1	3	2	3	0	5	2	3	**81**
Average					**78.0%**				

Table 20: Analysis of a common HAR system along with existing wavelet transform (without employing the proposed approach) on IXMAS dataset, where CrA is Cross Arm, SiD is Sit Down, GeU is Get Up, TuA is Turn Around

Actions	CrA	SiD	GeU	TuA	Kicking	Punching	Waving	Walking
CrA	**75**	4	5	3	4	2	4	3
SiD	2	**78**	1	4	5	3	4	3
GeU	3	2	**80**	3	1	4	2	5
TeA	3	4	2	**77**	3	5	4	2
Kicking	2	3	0	3	**82**	2	5	3
Punching	2	4	2	0	4	**84**	1	3
Waving	3	5	3	4	2	4	**76**	3
Walking	5	1	2	3	5	2	3	**79**
Average					**78.9%**			

Table 21: Analysis of a common HAR system along with existing Curvelet transform (without employing the proposed approach) on IXMAS dataset, where CrA is Cross Arm, SiD is Sit Down, GeU is Get Up, TuA is Turn Around

Actions	CrA	SiD	GeU	TuA	Kicking	Punching	Waving	Walking
CrA	**82**	3	1	4	2	3	0	5
SiD	2	**85**	2	3	4	1	1	2
GeU	3	4	**79**	4	2	3	4	1
TeA	2	3	1	**83**	4	0	3	4
Kicking	2	1	3	0	**86**	1	4	3
Punching	3	2	3	2	4	**82**	1	3
Waving	3	5	1	4	1	3	**81**	2
Walking	4	3	4	2	4	5	4	**74**
Average					**81.5%**			

Table 22: Analysis of a common HAR system along with existing local binary pattern (LBP) (without employing the proposed approach) on IXMAS dataset, where CrA is Cross Arm, SiD is Sit Down, GeU is Get Up, TuA is Turn Around

Actions	CrA	SiD	GeU	TuA	Kicking	Punching	Waving	Walking
CrA	**79**	3	3	3	3	3	3	3
SiD	3	**82**	1	2	4	3	1	4
GeU	2	3	**84**	3	1	2	3	2
TeA	3	4	1	**77**	5	4	2	4

(Continued)

Table 22: Continued

Actions	CrA	SiD	GeU	TuA	Kicking	Punching	Waving	Walking
Kicking	3	4	2	3	**80**	4	1	3
Punching	2	3	2	1	4	**85**	3	0
Waving	4	2	4	5	4	1	**75**	5
Walking	4	2	3	1	3	4	2	**81**
Average					**80.4%**			

Table 23: Analysis of a common HAR system along with existing local directional pattern (LDP) (without employing the proposed approach) on IXMAS dataset, where CrA is Cross Arm, SiD is Sit Down, GeU is Get Up, TuA is Turn Around

Actions	CrA	SiD	GeU	TuA	Kicking	Punching	Waving	Walking
CrA	**86**	2	0	3	4	2	3	0
SiD	3	**81**	5	0	3	1	3	4
GeU	4	3	**79**	3	4	2	1	4
TeA	2	1	2	**87**	0	4	1	3
Kicking	4	3	0	3	**83**	2	4	1
Punching	3	5	2	4	5	**77**	1	3
Waving	3	1	5	1	4	2	**80**	4
Walking	2	4	1	4	2	0	3	**84**
Average					**82.1%**			

Table 24: Analysis of a common HAR system along with existing stepwise linear discriminant analysis (SWLDA) (without employing the proposed approach) on IXMAS dataset, where CrA is Cross Arm, SiD is Sit Down, GeU is Get Up, TuA is Turn Around

Actions	CrA	SiD	GeU	TuA	Kicking	Punching	Waving	Walking
CrA	**76**	3	5	2	4	3	5	2
SiD	3	**81**	3	1	4	2	4	2
GeU	4	3	**85**	2	0	2	1	3
TeA	2	1	4	**88**	0	1	3	1
Kicking	2	4	2	1	**80**	3	5	3
Punching	3	4	1	3	5	**78**	2	4
Waving	2	3	0	5	1	2	**84**	3
Walking	3	2	5	0	3	4	1	**82**
Average					**81.8%**			

6.3 Third Experiment

Finally, in this group of experiments, we have compared the recognition rate of the proposed approach against latest HAR systems. For some system, we have borrowed their implementation code; while, for the remaining system, we have presented their accuracies as described in their respective articles. The entire systems were implemented under the exact settings as indicated in their respective articles. For comparison, we also utilized, UCF50 dataset [44] and HMDB51 dataset [45]. The comparison results are accordingly presented in Tab. 25.

Table 25: Performance of the proposed approach against recent HAR systems

Latest HAR systems	Performance accuracies	Std Dev
[46]	89.3% (Wezmann + KTH + UCF + IXMAS)	±3.1
	82.1% (on UCF50)	±4.7
	80.4% (on HMDB51)	±1.9
[47]	85.9% (Wezmann + KTH + UCF + IXMAS)	±4.7
	79.4% (on UCF50)	±1.5
	80.5% (on HMDB51)	±2.0
[48]	79.1% (Wezmann + KTH + UCF + IXMAS)	±2.9
	71.9% (on UCF50)	±3.7
	69.7% (on HMDB51)	±2.4
[49]	92.8% (Wezmann + KTH + UCF + IXMAS)	±0.9
	87.4% (on UCF50)	±4.1
	89.8% (on HMDB51)	±2.2
[50]	89.7% (Wezmann + KTH + UCF + IXMAS)	±2.6
	77.4% (on UCF50)	±3.3
	75.6% (on HMDB51)	±1.2
[51]	71.9% (Wezmann + KTH + UCF + IXMAS)	±1.8
	65.5% (on UCF50)	±4.4
	66.4% (on HMDB51)	±3.9
Proposed approach	**98.4%** (Wezmann + KTH + UCF + IXMAS)	±1.4
	95.5% (on UCF50)	±2.8
	94.9% (on HMDB51)	±3.8

It is vibrant from Tab. 25 that the proposed approach achieved best weighted average classification accuracy against state-of-the-art works. The reason is that, the proposed technique has the capacity to extract the prominent features from the action frames under the presence of occlusion, illumination and background disorder and scale changes. Moreover, the proposed approach extracts the best features from various resources such as shapes, textures, and colors in order to build the feature vector that will be input for a classifier.

7 Conclusions

Human activity recognition (HAR) has a fascinating role in our daily life. HAR can be applied for healthcare domains to check the patients' daily routines. Also, HAR has a significant role in other applications such as crime control, sports, defense etc. There are many resources for HAR systems. Among them, video-camera is one of the best candidates for HAR systems. The accuracy of such systems completely depends upon the extraction and selection of the best features from the activity frames. Accordingly, in this work, we have proposed a new feature extraction technique that is based on template matching. In the proposed approach, we matched a template of an image which will be the template like sub-frame which comprises the silhouette. Therefore, we focus the template on the frame pixels and calculate the equivalent number of pixels in the template correspondent those in the frame. The proposed approach was assessed against four publicly available standard datasets of activities, which sowed showed the best performance against existing recent HAR systems. The averaging intrinsic in the proposed approach is the reduction of the vulnerability to noise and the maximization stage diminishes defenselessness to occlusion. Moreover, the proposed algorithm has the capacity to extract the prominent features from the activity frames under the presence of occlusion, illumination and background disorder and scale changes. Also, the proposed approach extracts the best features from various resources such as shapes, textures, and colors for building the feature vector that will be input for a classifier.

In the future, we will implement and deploy the proposed HAR system under the presence of the proposed feature extraction in healthcare, which will facilitate the physicians to remotely check the daily exercises of the patients through which they might easily recommend the corresponding recommendations for the patients. This approach may also help the patients sufficiently improve the quality of their lives in healthcare and telemedicine.

Funding Statement: The authors extend their appreciation to the Deputyship for Research & Innovation, Ministry of Education in Saudi Arabia for funding this work through the Project Number "375213500". Also, the authors would like to extend their sincere appreciation to the central laboratory at Jouf University to support this study.

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

References

[1] M. Nixon and A. Aguado, "Feature extraction by shape matching," in *Feature Extraction and Image Processing for Computer Vision*, 2nd ed., London, United Kingdom, Academic press, Chapter No. 5, Section No. 5.3.1, pp. 191, 2008.

[2] D. R. Beddiar, B. Nini, M. Sabokrou and A. Hadid, "Vision-based human activity recognition: A survey," *Multimedia Tools and Applications*, vol. 79, no. 41, pp. 30509–30555, 2020.

[3] D. Weinland, R. Ronfard and E. Boyer, "A survey of vision-based methods for action representation, segmentation and recognition," *Computer Vision and Image Understanding*, vol. 115, no. 2, pp. 224–241, 2011.

[4] M. H. Siddiqi, R. Ali, M. Rana, E. K. Hong, E. S. Kim *et al.,* "Video-based human activity recognition using multilevel wavelet decomposition and stepwise linear discriminant analysis," *Sensors*, vol. 14, no. 4, pp. 6370–6392, 2014.

[5] N. Almaadeed, O. Elharrouss, S. Al-Maadeed, A. Bouridane and A. Beghdadi, "A novel approach for robust multi human action recognition and summarization based on 3D convolutional neural networks,"

Computer Vision and Pattern Recognition, pp. 1–14, arXiv:1907.11272, 2019. [Online]. Available: http://arxiv.org/abs/1907.11272.

[6] A. Ullah, K. Muhammad, T. Hussain and S. W. Baik, "Conflux LSTMs network: A novel approach for multi-view action recognition," *Neurocomputing*, vol. 435, pp. 321–329, 2021.

[7] A. Ullah, K. Muhammad, W. Ding, V. Palade, I. U. Haq *et al.,* "Efficient activity recognition using lightweight CNN and DS-GRU network for surveillance applications," *Applied Soft Computing*, vol. 103, pp. 107102, 2021.

[8] A. Ullah, K. Muhammad, K. Haydarov, I. U. Haq, M. Lee *et al.,* "One-shot learning for surveillance anomaly recognition using siamese 3d cnn," in *Int. Joint Conf. on Neural Networks (IJCNN)*, Glasgow, Scotland, UK, pp. 1–8, 2020.

[9] V. Parameswari and S. Pushpalatha, "Human activity recognition using SVM and deep learning," *European Journal of Molecular & Clinical Medicine*, vol. 7, no. 4, pp. 1984–1990, 2020.

[10] J. Basavaiah and C. Patil, "Robust feature extraction and classification based automated human action recognition system for multiple datasets," *International Journal of Intelligent Engineering and Systems*, vol. 13, no. 1, pp. 13–24, 2020.

[11] L. Gorelick, M. Blank, E. Shechtman, M. Irani and R. Basri, "Actions as space-time shapes," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 29, no. 12, pp. 2247–2253, 2007.

[12] I. Laptev, M. Marszalek, C. Schmid and B. Rozenfeld, "Learning realistic human actions from movies," in *IEEE Int. Conf. on Computer Vision and Pattern Recognition*, Anchorage, Alaska, pp. 1–8, 2008.

[13] K. Soomro and A. R. Zamir, "Action recognition in realistic sports videos," in *Computer Vision in Sports*, 1st ed., Castle Donington, United Kingdom: Springer, Chapter No. 9, Section No. 9.2, pp. 181–208, 2014.

[14] D. Weinland, E. Boyer and R. Ronfard, "Action recognition from arbitrary views using 3D exemplars," in *11th Int. Conf. on Computer Vision*, Janeiro, Brazil, pp. 1–7, 2007.

[15] M. Ahmed, M. Ramzan, H. U. Khan, S. Iqbal, M. A. Khan *et al.,* "Real-time violent action recognition using key frames extraction and deep learning," *Computers, Materials & Continua*, vol. 69, no. 2, pp. 2217–2230, 2021.

[16] B. Zohuri and M. Moghaddam, "Deep learning limitations and flaws," *Modern Approaches Mater. Sci. J.*, vol. 2, no. 3, pp. 241–250, 2020.

[17] C. Szegedy, S. Ioffe, V. Vanhoucke and A. A. Alemi, "Inception-v4, inception-resnet and the impact of residual connections on learning," in *31st Int. Conf. on Artificial Intelligence*, California, USA, pp. 4278–4284, 2017.

[18] N. Zerrouki, F. Harrou, Y. Sun and A. Houacine, "Vision-based human action classification using adaptive boosting algorithm," *IEEE Sensors Journal*, vol. 18, no. 12, pp. 5115–5121, 2018.

[19] T. Chengsheng, L. Huacheng and X. Bing, "AdaBoost typical Algorithm and its application research," in *Int. Conf. on MATEC Web of Conf.*, Taichung, Taiwan, pp. 1–6, 2017.

[20] M. Moencks, V. De Silva, J. Roche and A. Kondoz, "Adaptive feature processing for robust human activity recognition on a novel multi-modal dataset," *Robotics and Autonomous Systems*, pp. 1–14, arXiv:1901.02858, 2019. [Online]. Available: http://arxiv.org/abs/1901.02858.

[21] M. H. Siddiqi, M. Alruwaili, A. Ali, S. Alanazi and F. Zeshan, "Human activity recognition using Gaussian mixture hidden conditional random fields," *Computational Intelligence and Neuroscience*, vol. 2019, pp. 1–14, 2019.

[22] M. A. Khan, M. Alhaisoni, A. Armghan, F. Alenezi, U. Tariq *et al.,* "Video analytics framework for human action recognition," *Computers, Materials & Continua*, vol. 68, no. 3, pp. 3841–3859, 2021.

[23] S. Bhargava, "A note on evolutionary algorithms and its applications," *Adults Learning Mathematics*, vol. 8, no. 1, pp. 31–45, 2013.

[24] N. Robertson and I. Reid, "A general method for human activity recognition in video," *Computer Vision and Image Understanding*, vol. 104, no. 2, pp. 232–248, 2006.

[25] K. Mikolajczyk and C. Schmid, "A performance evaluation of local descriptors," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, no. 10, pp. 1615–1630, 2005.

[26] L. Wei and S. K. Shah, "Human activity recognition using deep neural network with contextual information," in *12th Int. Joint Conf. on Computer Vision, Imaging and Computer Graphics Theory and Applications*, Porto, Portugal, pp. 34–43, 2017.
[27] D. Camilleri and T. Prescott, "Analyzing the limitations of deep learning for developmental robotics," in *Int. Conf. on Biomimetic and Biohybrid Systems*, California, USA, pp. 86–94, 2017.
[28] T. Alhersh, "From motion to human activity recognition," Ph.D. Dissertation, School of Business Informatics and Mathematics, University of Mannheim, Mannheim, Germany, 2021.
[29] K. Yashwanth, M. N. Sunay and S. Srinivas, "STIP based activity recognition," *International Journal of Engineering Research & Technology*, vol. 8, no. 11, pp. 229–234, 2020.
[30] L. Moraru, C. D. Obreja, N. Dey and A. S. Ashour, "Dempster-shafer fusion for effective retinal vessels' diameter measurement," in *Soft Computing Based Medical Image Analysis*, 1st ed., Kolkata, India: Evaluating Academic Research, Chapter No. 9, Section No. 2.2, pp. 149–160, 2018.
[31] E. F. Nasser, "Improvement of corner detection algorithms (Harris, FAST and SUSAN) based on reduction of features space and complexity time," *Engineering & Technology Journal*, vol. 35, no. 2, pp. 112–118, 2017.
[32] W. Lin, M. T. Sun, R. Poovandran and Z. Zhang, "Human activity recognition for video surveillance," in *IEEE Int. Symp. on Circuits and Systems*, Washington, USA, pp. 2737–2740, 2008.
[33] M. Babiker, O. O. Khalifa, K. K. Htike, A. Hassan and M. Zaharadeen, "Automated daily human activity recognition for video surveillance using neural network," in *IEEE 4th Int. Conf. on Smart Instrumentation, Measurement and Application (ICSIMA)*, Kuala Lumpur, Malaysia, pp. 1–5, 2017.
[34] A. G. Perera, Y. W. Law, T. T. Ogunwa and J. Chahl, "A multi-viewpoint outdoor dataset for human action recognition," *IEEE Transactions on Human-Machine Systems*, vol. 50, no. 5, pp. 405–413, 2020.
[35] Q. V. Le, W. Y. Zou, S. Y. Yeung and A. Y. Ng, "Learning hierarchical invariant spatio-temporal features for action recognition with independent subspace analysis," in *24th IEEE Int. Conf. on Computer Vision and Pattern Recognition*, Colorado, USA, pp. 3361–3368, 2011.
[36] F. Husain, B. Dellen and C. Torras, "Action recognition based on efficient deep feature learning in the spatio-temporal domain," *IEEE Robotics and Automation Letters*, vol. 1, no. 2, pp. 984–991, 2016.
[37] M. Wang, J. Xing and Y. Liu, "ActionCLIP: A new paradigm for video action recognition," *Computing Research Repository (CoRR)*, pp. 1–11, arXiv:2109.08472, 2021. [Online]. Available. https://arxiv.org/abs/2109.08472.
[38] H. Wu, Y. Liu and J. Wang, "Review of text classification methods on deep learning," *Computers, Materials & Continua*, vol. 63, no. 3, pp. 1309–1321, 2020.
[39] M. H. Siddiqi, S. Lee, Y. K. Lee, A. M. Khan and P. T. H. Truc, "Hierarchical recognition scheme for human facial expression recognition systems," *Sensors*, vol. 13, no. 12, pp. 16682–16713, 2013.
[40] M. H. Siddiqi, M. Alruwaili and A. Ali, "A novel feature selection method for video-based human activity recognition systems," *IEEE Access*, vol. 7, pp. 119593–119602, 2019.
[41] F. Kuncan, Y. Kaya and M. Kuncan, "A novel approach for activity recognition with down-sampling 1D local binary pattern," *Advances in Electrical and Computer Engineering*, vol. 19, no. 1, pp. 35–44, 2019.
[42] T. Jabid, M. H. Kabir and O. Chae, "Gender classification using local directional pattern (LDP)," in *20th Int. Conf. on Pattern Recognition*, Istanbul, Turkey, pp. 2162–2165, 2010.
[43] M. H. Siddiqi, "Accurate and robust facial expression recognition system using real-time YouTube-based datasets," *Applied Intelligence*, vol. 48, no. 9, pp. 2912–2929, 2018.
[44] K. K. Reddy and M. Shah, "Recognizing 50 human action categories of web videos," *Machine Vision and Applications*, vol. 24, no. 5, pp. 971–981, 2013.
[45] H. Kuehne, H. Jhuang, E. Garrote, T. Poggio and T. Serre, "HMDB: A large video database for human motion recognition," in *Int. Conf. on Computer Vision*, Barcelona, Spain, pp. 2556–2563, 2011.
[46] S. H. Basha, V. Pulabaigari and S. Mukherjee, "An information-rich sampling technique over spatio-temporal CNN for classification of human actions in videos," *Computing Research Repository (CoRR)*, pp. 1–7, arXiv:2002.02100, 2020. [Online]. Available. https://arxiv.org/abs/2002.02100.
[47] M. J. Roshtkhari and M. D. Levine, "Human activity recognition in videos using a single example," *Image and Vision Computing*, vol. 31, no. 11, pp. 864–876, 2013.

[48] L. Shiripova and E. Myasnikov, "Human action recognition using dimensionality reduction and support vector machine," in *CEUR Workshop Proc.*, Illinois, USA, pp. 48–53, 2019.
[49] J. Kim and D. Lee, "Activity recognition with combination of deeply learned visual attention and pose estimation," *Applied Sciences*, vol. 11, no. 9, pp. 1–18, 2021.
[50] A. B. Sargano, P. Angelov and Z. Habib, "Human action recognition from multiple views based on view-invariant feature descriptor using support vector machines," *Applied Sciences*, vol. 6, no. 10, pp. 1–14, 2016.
[51] N. Nida, M. H. Yousaf, A. Irtaza and S. A. Velastin, "Deep temporal motion descriptor (DTMD) for human action recognition," *Turkish Journal of Electrical Engineering & Computer Sciences*, vol. 28, no. 3, pp. 1371–1385, 2020.