Computing and Informatics, Vol. 42, 2023, 965–992, doi: 10.31577/cai_2023_4_965

ENSEMBLE BASED FEATURE EXTRACTION AND DEEP LEARNING CLASSIFICATION MODEL WITH DEPTH VISION

Kumari Priyanka Sinha*

Department of Computer Science and Engineering
Nalanda College of Engineering, Chandi
Bihar, India
e-mail: kumaripriyankas@outlook.com

Prabhat Kumar, Rajib Ghosh

Department of Computer Science and Engineering
National Institute of Technology Patna
Patna, India
e-mail: {Prabhat, rajib.ghosh}@nitp.ac.in

Abstract. It remains a challenging task to identify human activities from a video sequence or still image due to factors such as backdrop clutter, fractional occlusion, and changes in scale, point of view, appearance, and lighting. Different appliances, as well as video surveillance systems, human-computer interfaces, and robots used to study human behavior, require different activity classification systems. A four-stage framework for recognizing human activities is proposed in the paper. As part of the initial stages of pre-processing, video-to-frame conversion and adaptive histogram equalization (AHE) are performed. Additionally, watershed segmentation is performed and, from the segmented images, local texton XOR patterns (LTXOR), motion boundary scale-invariant feature transforms (MoBSIFT) and bag of visual words (BoW) based features are extracted. The Bidirectional gated recurrent unit (Bi-GRU) and the Bidirectional long short-term memory (Bi-LSTM) classifiers are used to detect human activity. In addition, the combined decisions of the Bi-GRU and Bi-LSTM classifiers are further fused, and their accuracy levels are determined. With this Dempster-Shafer theory (DST)

* Corresponding author

technique, it is more likely that the results obtained from the analysis are accurate. Various metrics are used to assess the effectiveness of the deployed approach.

Keywords: Human activities, improved LTXOR, BoW, Bi-LSTM, Bi-GRU classifier

Mathematics Subject Classification 2010: 46-T30

1 INTRODUCTION

Human action recognition (HAR) is a classification task in which the movement of a person is evaluated using data from different sources, such as sensors and cameras. It has several applications in the health care industry, primarily, in monitoring the actions of elderly people and detecting falls [1, 2, 3]. The system has the potential to support innovative appliances, such as enhanced realism, internet of things (IoT), interior localization, and smart building control systems to maintain a secure indoor environment with superior energy efficiency [4, 5]. Several studies have been conducted using computer vision (CV), smartphones, sensors, and ambient devices to develop HAR schemes. There are two types of monitoring schemes: passive and active [6, 7, 8]. Wearable cameras and sensors are used as part of active monitoring systems (AMS). A sensor-based HAR requires individuals to hold sensors, such as gyroscopes, accelerometers, and pedometers, which can be difficult to manage in several situations, particularly for senior citizens [9, 10, 11].

With a vision-oriented HAR [12, 13], humans can be recognized both indoors and outdoors using cameras and machine vision schemes. One of the major challenges that such schemes face is the presence of noise in the video and image streams [14, 15, 16]. Depending on the surrounding environment, the capturing device, and several other factors, there may be several types of noise, including multiplicative noise, additive noise, etc. If a vision-based recognition and classification scheme is to be effective, it must eliminate the noise before implementation. In recent studies, deep learning schemes, such as deep belief networks (DBNs), convolutional neural networks (CNNs), and deep convolutional neural networks (DCNNs) have produced competent results in image-oriented HAR [17, 18, 19]. Although these improvements have been made, certain challenges remain, including a lack of accuracy, that must be overcome by the application of appropriate technology in the future [20, 21, 22, 23]. The major contributions of this study are listed as follows.

1. The first step in this study is to propose a multi-feature fusion approach that will combine local texton XOR pattern (LTXOR), BoW, and motion boundary scale-invariant feature transform (MoBSIFT) features to investigate datasets at a deeper level.
2. To improve the accuracy of the HAR system, bidirectional gated recurrent units (Bi-GRU) have been combined with bidirectional long short term memory (Bi-LSTM) classifiers using Dempster-Shafer theory (DST).

This paper is organized as follows: Section 2 discusses the literature on HAR. An overview of datasets is provided in Section 3. In Section 4, the proposed HAR model is described. A discussion of the experimental results is provided in Section 5. Section 6 concludes this paper.

2 RELATED WORKS

Bokhari et al. [24] proposed a method called deep gated recurrent unit (DGRU) for non-obtrusive HAR. Further, a de-noising approach based on the empirical model decomposition (EMD) has been used, followed by a linear discriminant analysis (LDA) and a discrete wavelet transform (DWT), aimed at reducing the dimensionality and extracting features from the data. It is evident from the results of the investigation that the DGRU produces the highest level of classification accuracy. A faster networking approach was developed by Xu et al. [25]. To enhance the efficiency of the optical flow features, fusion methods of spatio-temporal features were investigated, in which the temporal and spatial information was combined to form a single feature. Further, CNN with OFF was projected in place of VGG16-network, which was used to achieve a high number of features. In terms of accuracy, the proposed method outperformed these conventional schemes. A deep neural network (DNN) for HAR has been developed by Qin et al. [26] based on several sensor data sets. The DNN encoded the time series of the sensor data as images and controlled these deformed images to preserve the important characteristics of the HAR. In addition, a deep residual network (DRN) with different layers was used to accommodate the different dataset sizes. According to the results, the DNN technique outperformed the previously demanding techniques in terms of F1-score and precision.

Explaining and reasoning with knowledge-oriented and data-driven models, Jia et al. [27] developed a hierarchical structure-oriented scheme and technique for HAR. The method consists primarily of creating a hierarchical representation of the compound action based on the semantic meaning. As a result, the hierarchical approach to symbolic analysis had been established as a useful methodology. As demonstrated by Liu et al. [28], the normalized dynamic graph convolutional network (MRDGCN) continuously updates the structure of the data until an optimal model is obtained. An optimal convolution layer was constructed to determine the structure of the data. Thus, the MRDGCN has learned higher

level sample features to improve its learning performance on the data representations.

Jung et al. [29], Sena et al. [30] and L'Yvonnet et al. [31] presented a sound recognition-oriented HAR method using recurrent neural networks (RNNs). This study collected sound data by analyzing ten groups of people who performed daily concerts in the internal environment. With the help of a Log Mel-filter bank energies technique, the features have been derived from aural data, and an RNN scheme with various layers has been trained based on the aural data. In comparison with the existing models, the RNN model provided enhanced recall scores.

As a result of the DCNNs, Sena et al. [30] derived patterns using data from a variety of chronological scales. It was appropriate to use this method as the data were presented prior to a temporal series and the derived scales provided valuable information regarding the activities carried out by the users. By using this scheme, it was possible to extract both simple and composite movement. Multitemporal and multimodal systems have been developed that outperform the previous studies using two diverse datasets. The HAR scheme presented by L'Yvonnet et al. [31] is based on the possible differences in human performance. A discrete-time Markov chains (DTMC) and a PRISM model were used to examine and express the motivating temporal reasons that pertain to the dynamic development of activities within this scheme. It was observed that the DTMC's scheme performed better compared to the other schemes.

The reviews on the HAR models are summarized in Table 1. The LDA is generally used for improving the accuracy and increasing the F1-scores [24]. The tentative scenarios, however, were not measured as a result of the study. To improve the speed and accuracy, Xu et al. [25] used the CNN method. Further, the DRN model was used to provide higher F1 values and greater accuracy than previous models [26]. To accomplish this, it is necessary to focus on datasets that are larger in size. Although it is a complex method, the HMM scheme was able to achieve high reliability, as well as improve the recognition rate despite its complexity [27]. The MRDGCN scheme has been used for increasing the accuracy of the recognition rate and for improving the recognition rate; however, it is important to investigate the local invariance as well [28]. In addition, the RCNN was found to provide greater precision and recall when compared to other methods; however, it is essential to consider a variety of classes of activity [29]. Generally, the DCNN incurs a negligible bias and condensed arrays, however, the kernel selection for the DCNN must be examined [30]. Additionally, the DTMCs provided better prediction accuracy and required a shorter processing time [31]. However, the temporal reasons characteristics are not taken into account.

3 DATASET DESCRIPTION

The proposed method is evaluated using multiple benchmark datasets, including UCF-ARG [32], UCF-101 [33], Hollywood2 [34] and HMDB51 [35].

Author	Deployed Schemes	Features	Challenges
Bokhari et al. [24]	LDA	Higher accuracy Improved F1-score.	No consideration on experimental scenarios.
Xu et al. [25]	CNN	Superior speed High exactness.	Optical flow was not produced.
Qin et al. [26]	DRN	Higher accuracy Higher F1 value.	Need spotlight on advanced datasets.
Jia et al. [27]	HMM	Highly consistent Better detection.	More multifaceted.
Liu et al. [28]	MRDGCN	Higher detection rate Improved accuracy.	Requires deliberation on local invariance.
Jung et al. [29]	RCNN	Improved precision Higher recall.	Need spotlight on diverse activity classes.
Sena et al. [30]	DCNN	Negligible bias Condensed count of array.	Require assessment on kernel election.
L'Yvonnet et al. [31]	DTMCs	Least time period Higher prediction accurateness.	Need spotlight on temporal reason characteristics.

Table 1. Study on existing HAR models

3.1 UCF-ARG Dataset

There are ten different types of human activities included in the UCF-ARG dataset, including *carrying*, *digging*, *boxing*, *jogging*, *opening-closing trunks*, *clapping*, *running*, *walking*, *throwing*, and *waving*. It is a multi-view dataset collected using aerial cameras mounted on ground cameras, helium expands, and roof cameras. The videos were recorded in high resolution throughout and were divided into three sets, namely training, testing, and validation, in a ratio of 6:3:1. In Figure 1, a few samples of the video frames from the UCF-ARG dataset are presented.

Figure 1. Video frame samples from UCF-ARG dataset

3.2 UCF-101 Dataset

There are 101 different realistic action videos in the UCF-101 dataset, which depict a variety of human activities. While the videos were being recorded, there were large variations in the camera's motion. There are five categories of data that can be found in this dataset: *human-human communication*, *human-object cooperation*, *playing instrument*, *body-motion only*, and *sports*. There is a wide range of human actions in this dataset. The collected videos have been divided into three sets: training, testing, and validation. In Figure 2, a few samples of the video frames from the UCF 101 dataset are presented.

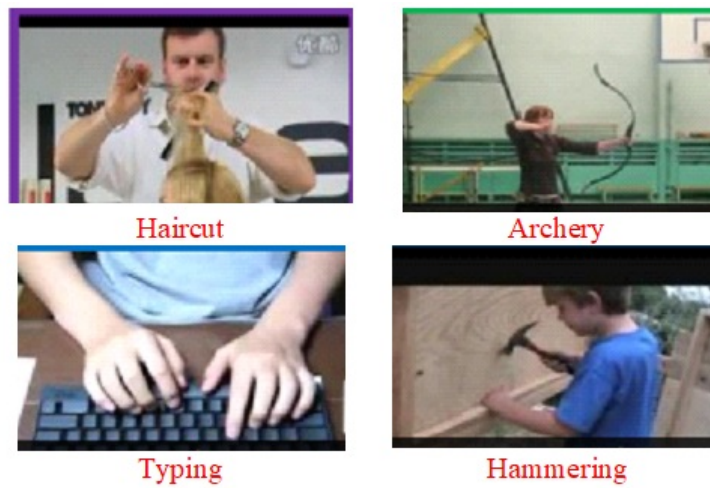

Figure 2. Video frame samples from UCF101 dataset

3.3 Hollywood2 Dataset

The Hollywood dataset contains twelve human activities. Datasets such as this one are extremely comprehensive and are considered benchmarks in the field of activity recognition. This repository contains 810 video clips in the AVI format, which were created from 69 Hollywood films. Figure 3 shows a few video frame samples from the Hollywood2 dataset.

3.4 HMDB51 Dataset

There are 51 different types of realistic action videos in the HMDB51 dataset. A variety of internet sources and digitized movies were used to collect the videos. There are five categories in this dataset: *body movements for human interaction*, *general facial actions*, *body movements with object interaction*, *general body movements*, and

Figure 3. Video frame samples from the Hollywood2 dataset

facial actions with object manipulation. The training and testing sets were divided in a 7:3 ratio between all the collected videos.

In Figure 4, a few samples of the video frames from the HMDB51 dataset are presented.

Figure 4. Video frame samples from HMDB51 dataset

4 METHODOLOGY

To classify human activity, the proposed method consists of four phases: preprocessing, segmentation, feature extraction, and classification. Figure 5 illustrates the overall architecture of the proposed work.

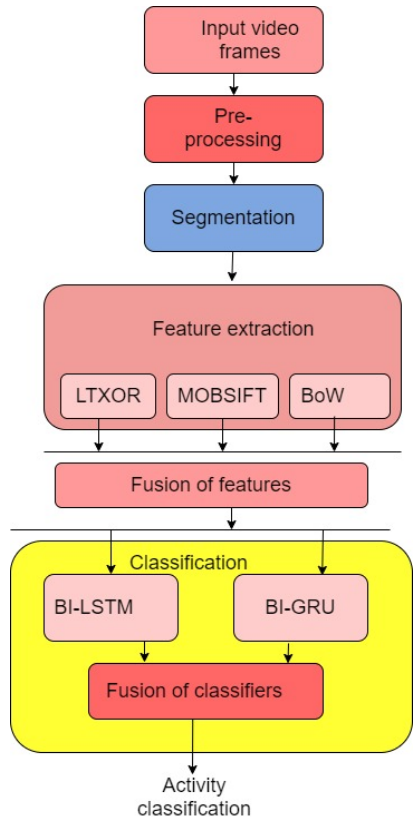

Figure 5. Overall architecture of the proposed model

4.1 Pre-Processing

As the frames have diverse backgrounds and resolutions, it becomes necessary to pre-process the images to enhance their quality.

4.1.1 Conversion of Video to Frame

Original video footage was collected which included human activities, such as *boxing*, *carrying*, *clapping*, *digging*, *jogging*, *running*, and *throwing*. Each frame of the videos was extracted using the video capture OpenCV function to extract the individual frames in the form of moving frames. These video frames are implied by *fr*, which will be used for further processing.

4.1.2 Adaptive Histogram Equalization

Adaptive histogram equalization (AHE) [36] is a digitalized image processing method that enhances the contrast of an image. It differs from the usual histogram equalization (HE) method in that the adaptive technique improves the local contrast. The HE is calculated for each division by dividing the image into separate blocks. Therefore, the AHE calculates plenty of different histograms, each of which is related to a distinct part of the image. In the different areas of the image, the contrast is improved locally and the edges are described better. However, the AHE exhibits certain noise disturbances. Therefore, in this study, a new modification is made to overcome this problem.

Conventionally, the AHE is evaluated as shown in Equation (1), where, r refers to the pixel with grey level value for new images, $P(r)$ refers to the probability density, x refers to the mean of the image and σ refers to the standard deviation of the image. As per the improved concept, the AHE is modelled as shown in Equation (2).

$$P_S(s)\,\mathrm{d}s = P_r(r)\,\mathrm{d}r, \tag{1}$$

$$P_S(s)\,\mathrm{d}s = P_r(r)\,\mathrm{d}r + G(x), \tag{2}$$

$$G(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{x^2}{2\sigma^2}}. \tag{3}$$

The pre-processed frames are implied as fr^{pr}.

4.2 Watershed Segmentation

When it comes to extracting regions from images, watershed segmentation [37] is a more effective and simpler method. As outlined below, there are several steps involved in the watershed segmentation process.

Step 1: Image simplification: This process helps smooth out the image and eliminates the noise interference.

Step 2: Calculation of morphological gradient image (MGI): The gray change in an image is reflected by the MGI *gr(f)* [38] that is formulated as in Equation (4), wherein s is a sign of the structuring component, Θ stands for the eroding conversion and $+$ is a sign of the dilating conversion.

$$gr(f) = (s + f) - (s\Theta f). \tag{4}$$

Step 3: This floating-point image $f^g(f)$ of activity is calculated according to Equation (5).

$$f^g(f) = \frac{gr(f) * gr(f)}{255.0}. \tag{5}$$

Step 4: Obtain the initial segmentation results similar to the watershed approach. The segmented images are implied as fr^{seg}.

4.3 Feature Extraction

From fr^{seg}, various features can be derived. The extracting features assist in identifying the most valuable characteristics and removing those that are no longer relevant. A description of the derived features is provided below.

4.3.1 LTXOR Features

The LTXOR pattern [39] uses seven different texton shapes to generate the texton images. As a preliminary step, the image is split into overlapping blocks of 2×2, which are referred to by the name B_1. The gray value positions are referred to as P, Q, R, and S for the purpose of examination. According to the shape of the texton, the subblocks are modeled as described in Equation (6).

$$Tx(Y,Z) = \begin{cases} 1, & B_1(P) = B_1(Q) \,\&\, B_1(R) \neq B_1(S), \\ 2, & B_1(Q) = B_1(S) \,\&\, B_1(P) \neq B_1(R), \\ 3, & B_1(R) = B_1(S) \,\&\, B_1(P) \neq B_1(Q), \\ 4, & B_1(P) = B_1(R) \,\&\, B_1(Q) \neq B_1(S), \\ 5, & B_1(P) = B_1(S) \,\&\, B_1(Q) \neq B_1(R), \\ 6, & B_1(Q) = B_1(R) \,\&\, B_1(P) \neq B_1(S), \\ 7, & B_1(P) = B_1(Q) \,\&\, B_1(R) = B_1(S), \\ 8, & B_1(P) \neq B_1(Q) \;\&\; B_1(R) \neq B_1(S). \end{cases} \tag{6}$$

The center of every pixel and its neighbors are collected on a texton image. After the texton image has been computed, an XOR function is applied between the center texton and its neighbors. LTXOR patterns are typically calculated as shown in Equation (7). To obtain more precise content-oriented outputs, certain modifications are made to the existing LTXOR. Based on the improved concept, the LTXOR is calculated in accordance with Equation (8), where, HM_w denotes the weighted harmonic mean formulated in Equation (9), N refers to the overall weight,

w_i refers to the weight randomly selected among 1 and 2.

$$LTXOR_{G,L} = \sum_{l=1}^{G} 2^{l-1} \times \tilde{f}_3(Tx(b_l) \otimes Tx(b_a)), \tag{7}$$

$$ILTXOR_{G,L} = \sum_{l=1}^{G} 2^{l-1} \times \frac{\tilde{f}_3(Tx(b_l) \otimes Tx(b_a))}{HM_w}, \tag{8}$$

$$HM_w = \frac{\sum_{l=1}^{n} w_i}{\sum_{l=1}^{n} \frac{w_i}{y_i}}, \tag{9}$$

$$\tilde{f}_3(y \otimes z) = \begin{cases} 1, & \text{if } y \neq z, \\ 0, & \text{else.} \end{cases} \tag{10}$$

In Equation (10), $\otimes$ points to the XOR function amid the variables, $Tx(b_a)$ points to the texton shape for the centre pixel b_a, and $Tx(b_l)$ points to the texton shape for neighbour pixel b_l. In addition, the particular image of the texton is malformed to the maps of the LTXOR within 0 to $2^{\tilde{p}-1}$. In the LTXOR calculation, it specifies the total map using the histogram construction as in Equation (11).

$$His_{LTXoR}(m) = \sum_{J=1}^{T_1} \sum_{k=1}^{T_2} \tilde{f}_2(LTXoR(\tilde{j}, \tilde{k}), \tilde{m});\ \tilde{m} \in [0, (2^{\tilde{p}} - 1)]. \tag{11}$$

The extracted LTXOR features are specified as fe^{ILT}.

4.3.2 BoW

Bag of words (BoW) [40] is most likely used as a feature representation scheme for the still images and the videos in HAR. The bag of visual words, also known as the BoW, is a symbolic scheme used to symbolize the documents for retrieval purposes. The scheme was implemented to retrieve videos and images. BoW features are indicated by the expression fe^{BOW}.

4.3.3 MoBSIFT

MoBSIFT interest point detection is similar to the working of a MoSIFT [41] detector. It behaves as a temporal expansion of the well accepted SIFT [42] model. Scale invariant feature transform (SIFT) established by Lowe is the most accepted scheme to find the feature descriptors and key points (interest points). In SIFT, the key points were the spatial interest points recognized by building the difference of Gaussian (DoG) pyramids and after that local extremes of the DoG imageries were found across the neighbouring scales. The interest point detection is mathematically

depicted as shown in Equations (12) and (13).

$$O(a, b, \sigma) = g(a, b, \sigma) * fr^{se}(a, b), \tag{12}$$

$$U(a, b, \sigma) = O(a, b, l\sigma) - O(a, b, \sigma). \tag{13}$$

In Equations (12) and (13), $O(a, b, \sigma)$ refers to the scale spacing of the input imagery $fr^{se}(a, b)$ attained by convolving it with the variable scale Gaussian, $g(a, b, \sigma)$ and $U(a, b, \sigma)$ refers to the DoG of the input imagery. The extracted MoBSIFT features are denoted by fe^{SLBT}.

4.3.4 Proposed Method for Fusion of Features

A Bi-LSTM and Bi-GRU variant of the RNN classifiers are trained using the extracted multi-features:

$$f = fe^{ILT}.fe^{BOW}.fe^{SIFT}. \tag{14}$$

4.4 Human Activity Classifcation

The Bi-GRU and Bi-LSTM classifiers have been used in the proposed work to classify human activities. The performance of the HAR was tested using both the individual classifiers as well as the combination of these classifiers. The Bi-LSTM and Bi-GRU variants of the RNN classifiers are used in this study to classify the feature vectors f from Equation (14). As the recurrently connected nodes are present in the hidden layers of the RNN, its internal states are capable of remembering inputs from several past timestamps.

4.4.1 Bi-GRU Variant of RNN Based Recognition

The Bi-GRU [40] uses exceptional gates (ut), known as reset and update gates for the declining gradient dispersion with smaller loss. The ut substitute input and forget gate of the LSTM, which depict the preservation degree of the preceding data. The proposed architecture of the Bi-GRU variant of the RNN for HAR is illustrated in Figure 6.

$$ut = \mu(W_u.(R_{t-1}, Fea_t) + f_u). \tag{15}$$

In Equation (15), μ points out the sigmoid activation function among 0 and 1, Fea_t stands for the input matrix at time step t, R_{t-1} stands for the hidden state at the prior time step $t - 1$. W_u stands for the weight matrix of ut and f_u stands for the bias matrix of ut. The rt regulates the amount of chronological data that has to be ignored which is revealed in Equation (16), wherein, W_r characterizes the weight matrix of rt and f_r symbolizes the bias matrix of rt.

$$rt = \mu(W_r.(R_{t-1}, Fea_t) + f_r). \tag{16}$$

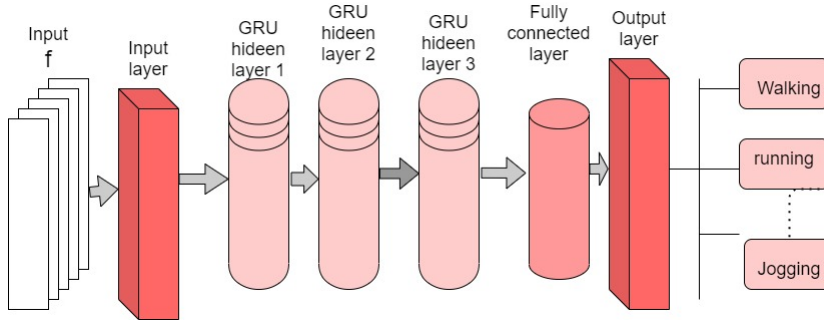

Figure 6. Proposed architecture of Bi-GRU variant of RNN for HAR in the present work

The hidden candidate state is exposed in Equation (17), wherein, tanh stands for the activation function. f_R and W_R stand for the bias matrix and weight matrix of the new cell state, * stands for the dot multiplication function. As a result, the output (R_t) implies linear disruption amid R_{t-1} and $(\tilde{R}_t)$.

$$\tilde{R}_t = \tanh(W_R.(R_{t-1} * rg, Fea_t) + f_R), \tag{17}$$

$$R_t = (1 - ut) * R_{t-1} + ut * \tilde{R}_t. \tag{18}$$

The forward and backward GRUs capture the prior and forthcoming facts of the input data. The Bi-GRU is devised as shown in Equation (19), wherein, $\overleftarrow{R_t}$ and $\overrightarrow{R_t}$ correspond to the hidden state of backward and forward GRU in that order, Ct corresponds to combining technique of the outputs at two directions.

$$Yt = Ct(\overleftarrow{R_t}, \overrightarrow{R_t}). \tag{19}$$

4.4.2 Bi-LSTM Variant of RNN Based Recognition

The Bi-LSTM classifier [39] covers a series of recurrent LSTM cells. The Bi-LSTM cells include the "forget gate, input gate, and output gate". Let, the variables be the hidden and cell state. The proposed architecture of the Bi-LSTM variant of the RNN for the HAR is illustrated in Figure 7.

(X_t, D_{t-1}, Z_{t-1}) and (Z_t, D_t) designate the input and output layer. At certain times, the output, input and forget gate implies O_t, I_t, F_t. The Bi-LSTM primarily uses F_t to sort the information. F_t is formulated as shown by Equation (20).

$$F_t = \sigma(J_{IF}X_t + L_{IF} + J_{ZF}Z_{t-1} + L_{ZF}). \tag{20}$$

In Equation (20), (J_{ZF}, L_{ZF}) and (J_{IF}, L_{IF}) points out the weight and bias constraint to map the hidden and input layers to forget that the gate and activation function is signified by σ. The input gate is exploited by the Bi-LSTM as revealed in Equations (21), (22) and (23), wherein, the (J_{ZG}, L_{ZG}) and (Z_{II}, L_{II}) imply the

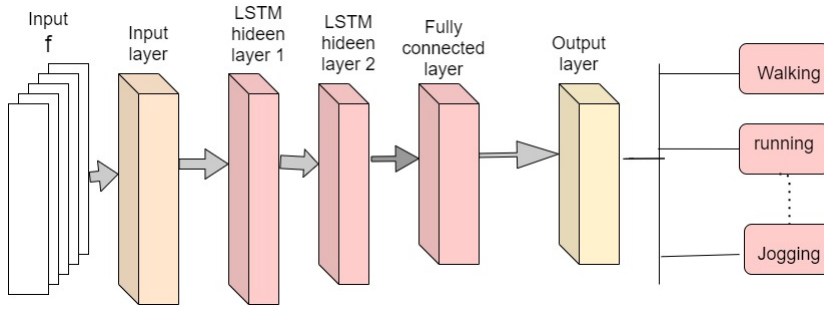

Figure 7. Proposed architecture of Bi-LSTM varient of RNN for HAR in the present work

weight and bias constraint to map the hidden and input layers to I_t.

$$G(t) = \tanh(J_{IG}X_t + L_{IG} + J_{ZG}Z_{t-1} + L_{ZG}), \tag{21}$$

$$I_t = \sigma(J_{II}X_t + L_{II} + J_{ZI}Z_{t-1} + L_{ZI}), \tag{22}$$

$$D_t = F_t D_{t-1} + I_t G_t, \tag{23}$$

$$O_t = \sigma(J_{IO}X_t + L_{IO} + J_{ZO}Z_{t-1} + L_{ZO}), \tag{24}$$

$$Z_t = O_t \tanh(D_t). \tag{25}$$

The Bi-LSTM cell obtains a hidden-layer from the output gate as shown in Equations (24) and (25), in which, (J_{ZO}, L_{ZO}) and (J_{IO}, L_{IO}) represent the weight and bias to map the hidden and the input layer to O_t.

4.4.3 Proposed Approach of Combining Bi-GRU and Bi-LSTM Variants of RNN Classifiers

The purpose of this section is to demonstrate how Bi-LSTMs and Bi-GRU variants of RNN can be combined to produce more accurate predictions from the individual class probabilities. In this process, both Bi-LSTMs and Bi-GRU variants of RNN are simultaneously trained with f and then combined with the probabilistic output of Bi-LSTMs and the probabilistic output of Bi-GRU using DST [43], as shown in Figure 8.

Fusion using DST. A DST of evidence is different from statistically based combination methods in that it is capable of representing lack of knowledge and uncertainty. This is quite important in the context of classifier combinations, since each classifier generally possesses a certain degree of uncertainty related to its performance. The DST method utilizes gradient descent learning to minimize the mean square error (MSE) between the output of a training set and the target output [43]. An illustration of a DST-based method is provided below.

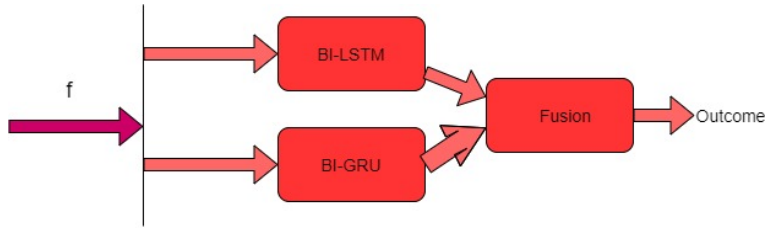

Figure 8. The proposed technique of fusing the Bi-LSTMs and Bi-GRU variants of RNN classifiers

Suppose $C = \{C_1, C_2, \ldots, C_n\}$ is a finite set of exclusive classes. There is a mass function M that is defined on the power set of C, represented by $P(C)$, which maps onto $[0, 1]$.

$$\sum M(x) = 1, x \subseteq C \text{ and } M(\phi) = 0, \tag{26}$$

$$P(C) = 2^{|C|}. \tag{27}$$

In Equation (27), $P(C)$ denotes the number of elements. The belief function *bel* is defined in terms of Equation (28), which refers to the probabilistic lower bound.

$$bel(x) = \sum M(y); \forall x \subseteq C, \text{where } y \subseteq x \text{ and } \ y \neq \phi. \tag{28}$$

In addition, the plausibility function *pf* is defined using Equation (29), which represents the probability that all the evidence does not contradict x.

$$pf(x) = \sum M(y); \forall x \subseteq C, \text{where } y \cap x \neq \phi. \tag{29}$$

$pf(x) - bel(x)$ denotes the imprecision associated with subset x of C.

Using Equation (30), let us assume that two mass functions M_1 and M_2 derived from two independent sources can be combined to form a consonant mass function *MC*.

$$MC(z) = \frac{\sum_{x \cap y = z} M_1(x) \times M_2(y)}{1 - \sum_{x \cap y = \phi} M_1(x) \times M_2(y)}. \tag{30}$$

5 EXPERIMENT RESULTS AND DISCUSSIONS

In this section, an evaluation of the proposed activity recognition approach based on different metrics including overall accuracy, precision, and recall is presented. The different benchmark datasets including the UCF-ARG, UCF-101, Hollywood2, and HMDB51 are used to evaluate the proposed approach. The performance of the proposed model is compared with the performance of the state-of-the-art models for activity recognition. It is implemented and evaluated in Python 3.9.7 on

a Windows 11 operating system (64 bit) using an Intel Core i7 processor (11^{th} generation) and a 12 GB GeForce Titan X graphics processing unit (GPU). For all the four datasets, the experiments were conducted using a stratified sample of 70 % for training and 30 % for testing. An explanation of the parameters evaluated, the results of each dataset, and a comparison with the current state-of-the-art techniques are provided below.

5.1 Hyperparameters Bi-GRU and Bi-LSTM Variants of RNN Classifiers

This study presents a Bi-LSTM model that has two hidden layers (optimal values). There are two hidden layers, the first of which processes the input sequence forward and the second of which processes it backward. A stochastic gradient descent (SGD) optimization technique has been used to determine the optimal number of hidden layers. In addition, there are 64 memory blocks that are recurrently connected in the hidden layers. A ReLU activation function has been applied to each memory block. The gates were activated using a sigmoid activation function. The softmax activation function has been used to activate the neurons in the output layer. There are three hidden layers in the Bi-GRU model, two of which process the input sequences forward and one of which processes the input sequences backward. This model uses the sigmoid functions for the control reset gate and the update gate, whereas the functions are used for the hidden state. The activation function at the hidden layer is the ReLU.

Bi-LSTM and Bi-GRU variants of the RNN accept a 128 by 128 dimensional feature vector as input. Based on the RMSprop optimizer algorithm with a learning rate of 0.0001, both networks have been optimized to minimize the categorical cross-entropy losses. A variety of the memory blocks, epochs, and batch sizes have been experimented with for each hidden layer of the Bi-LSTM and Bi-GRU. In this study, the optimal values of the various hyperparameters of the RNNs have been determined, as shown in Table 2.

5.2 Classification Results Using Bi-GRU Variant of RNN

Bi-GRU have been trained and tested on data using varying epochs, block sizes, and batch sizes. The accuracy of the proposed HAR system using the Bi-GRU variant of RNN classifiers is presented in Table 3. Based on Table 3, it can be seen that LSTMs with small batch sizes produce better recognition results. Additionally, Table 3 demonstrates that hidden layer 3 has improved recognition performance over hidden layer 1.

The performance of the HAR system using the Bi-GRU variant of the RNN classifier on UCF-101, Hollywood2, and HMDB51 dataset are given in Table 4.

Hyperparameters	**Search Space**	**Optimal Value**
Bi-LSTM hidden layer	1–3	2
Bi-GRU hidden layer	1–3	3
Memory blocks in the first Bi-LSTM hidden layer	32–64	64
Memory blocks in the second Bi-LSTM hidden layer	32–64	32
Memory blocks in the first Bi-GRU hidden layer	32–64	64
Memory blocks in the second Bi-GRU hidden layer	32–64	32
Memory blocks in the third Bi-GRU hidden layer	32–64	64
Batch size in Bi-LSTM hidden layer	20–40	30
Batch size in Bi-GRU hidden layer	10–20	10
Epochs	100–300	200
Learning rate	0.0001–0.0002	0.0001

Table 2. Optimal set of values of the various hyperparameters used in the Bi-LSTM and Bi-GRU variants of the RNN classifier

Hidden Layer	**Blocks**	**Epochs**	**Batch Size**	**Accuracy**
1	32	100	10	96.7 %
			20	96.3 %
		200	10	96.9 %
			20	96.7 %
	64	100	10	97.3 %
			20	96.5 %
		200	10	97.2 %
			20	96.8 %
2	32	100	10	97.6 %
			20	97.4 %
		200	10	98.1 %
			20	97.8 %
	64	100	10	98.1 %
			20	95.1 %
		200	10	97.4 %
			20	98.1 %
3	32	100	10	98.2 %
			20	97.8 %
		200	10	98.1 %
			20	97.9 %
	64	100	10	98.3 %
			20	98.1 %
		200	10	**98.9** %
			20	98.3 %

Table 3. Accuracy of the proposed HAR system using the Bi-GRU variant of the RNN classifier [UCF-ARG dataset]

Metric	UCF-101	Hollywood2	HMDB51
Accuracy	95.2 %	98.3 %	77.6 %
Precision	88.9 %	72.5 %	69.3 %
Recall	85.1 %	68.6 %	66.2 %

Table 4. Performance of the HAR system using the Bi-GRU variant of the RNN classifier

5.3 Classification Results Using Bi-LSTM Variant of RNN

Bi-LSTMs have been trained and tested on data using varying epochs, block sizes, and batch sizes. The accuracy of the proposed HAR system using the Bi-LSTM variant of RNN classifiers is presented in Table 5. Based on Table 5, it can be seen that hidden layer 2 has improved recognition performance over hidden layer 1.

The performance of the HAR system using the Bi-LSTM variant of the RNN classifier on UCF-101, Hollywood2, and HMDB51 dataset are given in Table 6.

5.4 Classification Results by Combining the Bi-GRU and the Bi-LSTM Variant of the RNN Classifier

Based on the proposed classifier combination discussed in Section 4.4.3, the activity recognition rates are presented in this section. On the UCF-ARG dataset, the performance of the Bi-GRU and Bi-LSTM variants of the RNN classifiers was evaluated. Using the Bi-GRU and Bi-LSTM variants of the RNN classifiers, the proposed HAR system was analysed, as shown in Table 7.

Three evaluation metrics were used to evaluate the proposed technique. The accuracy, precision, and recall for each dataset was calculated to assess the positive predictive value and sensitivity of the proposed technique. The results can be found in Figure 9.

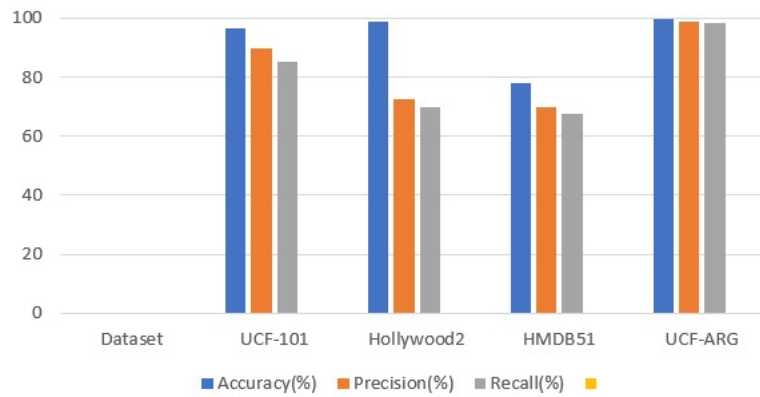

Figure 9. Evaluation of the proposed technique using accuracy, precision, and recall

Hidden Layer	Blocks	Epochs	Batch Size	Accuracy
1	32	100	20	95.7%
			30	96.1%
			40	95.6%
		200	20	95.9%
			30	96.3%
			40	95.7%
		300	20	95.5%
			30	95.9%
			40	95.4%
	64	100	20	96.3%
			30	96.5%
			40	96.1%
		200	20	96.5%
			30	96.7%
			40	96.3%
		300	20	96.1%
			30	96.3%
			40	95.9%
2	32	100	20	96.4%
			30	96.9%
			40	96.3%
		200	20	96.7%
			30	**97.4%**
			40	96.4%
		300	20	96.2%
			30	96.6%
			40	96.2%
	64	100	20	96.7%
			30	96.9%
			40	96.5%
		200	20	96.9%
			30	97.1%
			40	96.7%
		300	20	96.5%
			30	96.7%
			40	96.3%
3	32	100	20	96.2%
			30	96.7%
			40	96.1%
		200	20	96.5%
			30	97.2%
			40	96.1%
		300	20	95.9%
			30	96.3%
			40	96.0%
	64	100	20	96.4%
			30	96.6%
			40	96.2%
		200	20	96.6%
			30	96.9%
			40	96.5%
		300	20	96.2%
			30	96.4%
			40	96.0%

Table 5. Accuracy of the proposed HAR system using the Bi-LSTM variant of the RNN classifier [UCF-ARG dataset]

Metric	UCF-101	Hollywood2	HMDB51
Accuracy	94.9 %	97.9 %	77.2 %
Precision	88.1 %	72.1 %	68.6 %
Recall	84.6 %	69.2 %	67.2 %

Table 6. Performance of the HAR system using the Bi-LSTM variant of the RNN classifier

Metric	Bi-GRU	Bi-LSTM	Fusion (Bi-GRU + Bi-LSTM)
Accuracy	98.9 %	97.4 %	**99.8 %**
Precision	97.6 %	96.3 %	98.9 %
Recall	98.3 %	97.7 %	98.5 %

Table 7. Overall performance analysis of the proposed HAR system by combining the Bi-GRU and Bi-LSTM variant of the RNN classifier [UCF-ARG]

5.5 Comparison with the State-of-the-Art Results

As shown in the Table 8, some existing HAR systems available in the literature are compared to the proposed system. To evaluate the performance of the existing systems, the same datasets (UCF-ARG, UCF-101, Hollywood2 and HMDB51) were used similar to the state-of-the-art study.

6 CONCLUSION AND FUTURE WORK

In this study, an HAR scheme was developed that enabled video-to-frame conversion and AHE could be accomplished through a pre-processing step. In addition, watershed segmentation was performed, and features such as LTXOR, MoBSIFT, and BoW were extracted from the segmented images. A motion, shape, and texture feature was used to represent an activity in a selected shot. A novel deep learning model was developed to classify the human activities that combine Bi-GRU and Bi-LSTM variants of RNNs. To verify its effectiveness, the proposed system was extensively tested using different accuracy matrices for four benchmark activity recognition datasets. Due to its ability to process video streams in near real time, its low time complexity, and high detection accuracy, the system was considered suitable for industrial applications. Based on the empirical results, the proposed approach appears to be robust in the context of an HAR. As a result of this approach, it is possible to identify the activity of a single individual within a video. Further research will be conducted on individual and group activities for the HAR in the future. In addition, the multi-view datasets and complex datasets will be examined to recognize the activities.

Reference	**Classifier(s) Used**	**UCF-101**	**Hollywood2**	**HMDB51**	**UCF-ARG**
Mliki et al. [44]	LSTM	–	–	–	99.5 %
Subramanian et al. [45]	Deep genetic model	–	98.42 %	–	81.40 %
AlDahoul et al. [46]	Stochastic gradient descent	–	–	–	98 %
Burghouts et al. [47]	SVM	–	–	–	93 %
Ullah et al. [48]	LSTM	94.45 %	69.5 %	72.21 %	–
Ullah et al. [3]	DS-GRU	95.5 %	71.3 %	71.3 %	–
Xin et al. [49]	LSTM	85.3 %	63.1 %	58.2 %	–
Li et al. [50]	Bi-LSTM	94.2 %	–	70.4 %	–
Yang et al. [51]	3D-CNNs and bi-directional hierarchical LSTM	94.8 %	–	71.9 %	–
Wang et al. [52]	temporal segment network	94.2 %	–	69.4 %	–
Mahasseni et al. [53]	RLSTM	86.9 %	–	55.3 %	–
Liu et al. [54]	Hierarchical clustering	76.3 %	–	51.4 %	–
Ke et al. [55]	descriptor approaches	–	64.60 %	–	–
Lan et al. [56]	Multi-skIp Feature Stacking	89.1 %	68 %	65.4 %	–
Islam et al. [57]	SVM	–	87 %	–	–
Hou et al. [58]	FASNet, MIFS, SVM	–	78.1 %	–	–
Proposed system	**Fusion of Bi-GRU and Bi-LSTM**	**96.8 %**	**98.9 %**	**78.2 %**	**99.8 %**

Table 8. An analysis of the comparative performance with a limited number of studies already available

REFERENCES

[1] Tang, S.—Roberts, D.—Golparvar-Fard, M.: Human-Object Interaction Recognition for Automatic Construction Site Safety Inspection. Automation in Construction, Vol. 120, 2020, Art. No. 103356, doi: 10.1016/j.autcon.2020.103356.

[2] Janarthanan, R.—Doss, S.—Baskar, S.: Optimized Unsupervised Deep Learning Assisted Reconstructed Coder in the On-Nodule Wearable Sensor for Human Activity Recognition. Measurement, Vol. 164, 2020, Art. No. 108050, doi: 10.1016/j.measurement.2020.108050.

[3] Ullah, A.—Muhammad, K.—Ding, W.—Palade, V.—Haq, I. U.—Baik, S. W.: Efficient Activity Recognition Using Lightweight CNN and DS-GRU Network for Surveillance Applications. Applied Soft Computing, Vol. 103, 2021, Art. No. 107102, doi: 10.1016/j.asoc.2021.107102.

[4] Moayedi, F.—Azimifar, Z.—Boostani, R.: Human Action Recognition: Learning Sparse Basis Units from Trajectory Subspace. Applied Artificial Intelligence, Vol. 30, 2016, No. 4, pp. 297–317, doi: 10.1080/08839514.2016.1169094.

[5] Yurtman, A.—Barshan, B.: Human Activity Recognition Using Tag-Based Radio Frequency Localization. Applied Artificial Intelligence, Vol. 30, 2016, No. 2, pp. 153–179, doi: 10.1080/08839514.2016.1138787.

[6] Zhang, H.—Parker, L. E.: CoDe4D: Color-Depth Local Spatio-Temporal Features for Human Activity Recognition from RGB-D Videos. IEEE Transactions on Circuits and Systems for Video Technology, Vol. 26, 2016, No. 3, pp. 541–555, doi: 10.1109/TCSVT.2014.2376139.

[7] Yan, Y.—Ricci, E.—Liu, G.—Sebe, N.: Egocentric Daily Activity Recognition via Multitask Clustering. IEEE Transactions on Image Processing, Vol. 24, 2015, No. 10, pp. 2984–2995, doi: 10.1109/TIP.2015.2438540.

[8] Yousefzadeh, A.—Orchard, G.—Serrano-Gotarredona, T.—Linares-Barranco, B.: Active Perception with Dynamic Vision Sensors. Minimum Saccades with Optimum Recognition. IEEE Transactions on Biomedical Circuits and Systems, Vol. 12, 2018, No. 4, pp. 927–939, doi: 10.1109/TBCAS.2018.2834428.

[9] Popescu, A. C.—Mocanu, I.—Cramariuc, B.: Fusion Mechanisms for Human Activity Recognition Using Automated Machine Learning. IEEE Access, Vol. 8, 2020, pp. 143996–144014, doi: 10.1109/ACCESS.2020.3013406.

[10] Osayamwen, F.—Tapamo, J. R.: Deep Learning Class Discrimination Based on Prior Probability for Human Activity Recognition. IEEE Access, Vol. 7, 2019, pp. 14747–14756, doi: 10.1109/ACCESS.2019.2892118.

[11] Ehatisham-Ul-Haq, M.—Javed, A.—Azam, M. A.—Malik, H. M. A.—Irtaza, A.—Lee, I. H.—Mahmood, M. T.: Robust Human Activity Recognition Using Multimodal Feature-Level Fusion. IEEE Access, Vol. 7, 2019, pp. 60736–60751, doi: 10.1109/ACCESS.2019.2913393.

[12] Wu, X.—Chu, Z.—Yang, P.—Xiang, C.—Zheng, X.—Huang, W.: TW-See: Human Activity Recognition Through the Wall with Commodity Wi-Fi Devices. IEEE Transactions on Vehicular Technology, Vol. 68, 2019, No. 1, pp. 306–319, doi: 10.1109/TVT.2018.2878754.

[13] Muaaz, M.—Chelli, A.—Abdelgawwad, A. A.—Mallofré, A. C.—Pätzold, M.: WiWeHAR: Multimodal Human Activity Recognition Using Wi-Fi and Wearable Sensing Modalities. IEEE Access, Vol. 8, 2020, pp. 164453–164470, doi: 10.1109/ACCESS.2020.3022287.

[14] Wang, L.—Zhao, X.—Si, Y.—Cao, L.—Liu, Y.: Context-Associative Hierarchical Memory Model for Human Activity Recognition and Prediction. IEEE Transactions on Multimedia, Vol. 19, 2017, No. 3, pp. 646–659, doi: 10.1109/TMM.2016.2617079.

[15] Liu, W.—Zha, Z. J.—Wang, Y.—Lu, K.—Tao, D.: p-Laplacian Regularized Sparse Coding for Human Activity Recognition. IEEE Transactions on Industrial Electronics, Vol. 63, 2016, No. 8, pp. 5120–5129, doi: 10.1109/TIE.2016.2552147.

[16] Tu, Z.—Li, H.—Zhang, D.—Dauwels, J.—Li, B.—Yuan, J.: Action-Stage Emphasized Spatiotemporal VLAD for Video Action Recognition. IEEE Transactions on Image Processing, Vol. 28, 2019, No. 6, pp. 2799–2812, doi: 10.1109/TIP.2018.2890749.

[17] Chen, Z.—Zhang, L.—Cao, Z.—Guo, J.: Distilling the Knowledge from Handcrafted Features for Human Activity Recognition. IEEE Transactions on Industrial Informatics, Vol. 14, 2018, No. 10, pp. 4334–4342, doi: 10.1109/TII.2018.2789925.

[18] Yao, Y.—Liu, Y.—Liu, Z.—Chen, H.: Human Activity Recognition with Posture Tendency Descriptors on Action Snippets. IEEE Transactions on Big Data, Vol. 4, 2018, No. 4, pp. 530–541, doi: 10.1109/TBDATA.2018.2803838.

[19] Cai, L.—Liu, X.—Ding, H.—Chen, F.: Human Action Recognition Using Improved Sparse Gaussian Process Latent Variable Model and Hidden Conditional Random Filed. IEEE Access, Vol. 6, 2018, pp. 20047–20057, doi: 10.1109/ACCESS.2018.2822713.

[20] Zerrouki, N.—Harrou, F.—Sun, Y.—Houacine, A.: Vision-Based Human Action Classification Using Adaptive Boosting Algorithm. IEEE Sensors Journal, Vol. 18, 2018, No. 12, pp. 5115–5121, doi: 10.1109/JSEN.2018.2830743.

[21] Vishwakarma, D. K.—Kapoor, R.: Integrated Approach for Human Action Recognition Using Edge Spatial Distribution, Direction Pixel and R-Transform. Advanced Robotics, Vol. 29, 2015, No. 23, pp. 1553–1562, doi: 10.1080/01691864.2015.1061701.

[22] Najar, F.—Bourouis, S.—Bouguila, N.—Belghith, S.: Unsupervised Learning of Finite Full Covariance Multivariate Generalized Gaussian Mixture Models for Human Activity Recognition. Multimedia Tools and Applications, Vol. 78, 2019, No. 13, pp. 18669–18691, doi: 10.1007/s11042-018-7116-9.

[23] Uddin, M. Z.: Human Activity Recognition Using Segmented Body Part and Body Joint Features with Hidden Markov Models. Multimedia Tools and Applications, Vol. 76, 2017, No. 11, pp. 13585–13614, doi: 10.1007/s11042-016-3742-2.

[24] Bokhari, S. M.—Sohaib, S.—Khan, A. R.—Shafi, M.—Khan, A. u. R.: DGRU Based Human Activity Recognition Using Channel State Information. Measurement, Vol. 167, 2021, Art. No. 108245, doi: 10.1016/j.measurement.2020.108245.

[25] Xu, J.—Song, R.—Wei, H.—Guo, J.—Zhou, Y.—Huang, X.: A Fast Human Action Recognition Network Based on Spatio-Temporal Features. Neurocomputing,

Vol. 441, 2021, pp. 350–358, doi: 10.1016/j.neucom.2020.04.150.

[26] Qin, Z.—Zhang, Y.—Meng, S.—Qin, Z.—Choo, K. K. R.: Imaging and Fusing Time Series for Wearable Sensor-Based Human Activity Recognition. Information Fusion, Vol. 53, 2020, pp. 80–87, doi: 10.1016/j.inffus.2019.06.014.

[27] Jia, H.—Chen, S.: Integrated Data and Knowledge Driven Methodology for Human Activity Recognition. Information Sciences, Vol. 536, 2020, pp. 409–430, doi: 10.1016/j.ins.2020.03.081.

[28] Liu, W.—Fu, S.—Zhou, Y.—Zha, Z. J.—Nie, L.: Human Activity Recognition by Manifold Regularization Based Dynamic Graph Convolutional Networks. Neurocomputing, Vol. 444, 2021, pp. 217–225, doi: 10.1016/j.neucom.2019.12.150.

[29] Jung, M.—Chi, S.: Human Activity Classification Based on Sound Recognition and Residual Convolutional Neural Network. Automation in Construction, Vol. 114, 2020, Art. No. 103177, doi: 10.1016/j.autcon.2020.103177.

[30] Sena, J.—Barreto, J.—Caetano, C.—Cramer, G.—Schwartz, W. R.: Human Activity Recognition Based on Smartphone and Wearable Sensors Using Multiscale DCNN Ensemble. Neurocomputing, Vol. 444, 2021, pp. 226–243, doi: 10.1016/j.neucom.2020.04.151.

[31] L'Yvonnet, T.—De Maria, E.—Moisan, S.—Rigault, J. P.: Probabilistic Model Checking for Human Activity Recognition in Medical Serious Games. Science of Computer Programming, Vol. 206, 2021, Art. No. 102629, doi: 10.1016/j.scico.2021.102629.

[32] UCF-ARG Data Set. 2011, `http://Crcv.Ucf.Edu/Data/UCF-ARG.Php`.

[33] Soomro, K.—Zamir, A. R.—Shah, M.: UCF101: A Dataset of 101 Human Actions Classes from Videos in the Wild. CoRR, 2012, doi: 10.48550/arXiv.1212.0402.

[34] Marszalek, M.—Laptev, I.—Schmid, C.: Actions in Context. 2009 IEEE Conference on Computer Vision and Pattern Recognition, 2009, pp. 2929–2936, doi: 10.1109/CVPR.2009.5206557.

[35] Kuehne, H.—Jhuang, H.—Garrote, E.—Poggio, T.—Serre, T.: HMDB: A Large Video Database for Human Motion Recognition. 2011, pp. 2556–2563, doi: 10.1109/ICCV.2011.6126543.

[36] Pizer, S. M.—Amburn, E. P.—Austin, J. D.—Cromartie, R.—Geselowitz, A.—Greer, T.—ter Haar Romeny, B.—Zimmerman, J. B.—Zuiderveld, K.: Adaptive Histogram Equalization and Its Variations. Computer Vision, Graphics, and Image Processing, Vol. 39, 1987, No. 3, pp. 355–368, doi: 10.1016/S0734-189X(87)80186-X.

[37] Li, Y.—Shi, H.—Jiao, L.—Liu, R.: Quantum Evolutionary Clustering Algorithm Based on Watershed Applied to SAR Image Segmentation. Neurocomputing, Vol. 87, 2012, pp. 90–98, doi: 10.1016/j.neucom.2012.02.008.

[38] Weickert, J.: Efficient Image Segmentation Using Partial Differential Equations and Morphology. Pattern Recognition, Vol. 34, 2001, No. 9, pp. 1813–1824, doi: 10.1016/S0031-3203(00)00109-6.

[39] Bala, A.—Kaur, T.: Local Texton XOR Patterns: A New Feature Descriptor for Content-Based Image Retrieval. Engineering Science and Technology, an International Journal, Vol. 19, 2016, No. 1, pp. 101–112, doi: 10.1016/j.jestch.2015.06.008.

[40] Agusti, P.—Traver, V. J.—Pla, F.: Bag-of-Words with Aggregated Temporal Pair-Wise Word Co-Occurrence for Human Action Recognition. Pattern Recognition Letters, Vol. 49, 2014, pp. 224–230, doi: 10.1016/j.patrec.2014.07.014.

[41] Febin, I. P.—Jayasree, K.—Joy, P. T.: Violence Detection in Videos for an Intelligent Surveillance System Using MoBSIFT and Movement Filtering Algorithm. Pattern Analysis and Applications, Vol. 23, 2020, No. 2, pp. 611–623, doi: 10.1007/s10044-019-00821-3.

[42] Lowe, D. G.: Distinctive Image Features from Scale-Invariant Keypoints. International Journal of Computer Vision, Vol. 60, 2004, pp. 91–110, doi: 10.1023/B:VISI.0000029664.99615.94.

[43] Al-Ani, A.—Deriche, M.: A New Technique for Combining Multiple Classifiers Using the Dempster-Shafer Theory of Evidence. Journal of Artificial Intelligence Research, Vol. 17, 2002, pp. 333–361, doi: 10.1613/jair.1026.

[44] Mliki, H.—Bouhlel, F.—Hammami, M.: Human Activity Recognition from UAV-Captured Video Sequences. Pattern Recognition, Vol. 100, 2020, Art. No. 107140, doi: 10.1016/j.patcog.2019.107140.

[45] Subramanian, R. R.—Vasudevan, V.: A Deep Genetic Algorithm for Human Activity Recognition Leveraging Fog Computing Frameworks. Journal of Visual Communication and Image Representation, Vol. 77, 2021, Art. No. 103132, doi: 10.1016/j.jvcir.2021.103132.

[46] AlDahoul, N.—Sabri, A. Q. M.—Mansoor, A. M.: Real-Time Human Detection for Aerial Captured Video Sequences via Deep Models. Computational Intelligence and Neuroscience, Vol. 2018, 2018, Art. No. 1639561, doi: 10.1155/2018/1639561.

[47] Burghouts, G. J.—van Eekeren, A. W. M.—Dijk, J.: Focus-of-Attention for Human Activity Recognition from UAVs. In: Huckridge, D. A., Ebert, R. (Eds.): Electro-Optical and Infrared Systems: Technology and Applications XI. SPIE, Proceedings of SPIE, Vol. 9249, 2014, doi: 10.1117/12.2067569.

[48] Ullah, A.—Muhammad, K.—Del Ser, J.—Baik, S. W.—de Albuquerque, V. H. C.: Activity Recognition Using Temporal Optical Flow Convolutional Features and Multilayer LSTM. IEEE Transactions on Industrial Electronics, Vol. 66, 2019, No. 12, pp. 9692–9702, doi: 10.1109/TIE.2018.2881943.

[49] Xin, M.—Zhang, H.—Wang, H.—Sun, M.—Yuan, D.: ARCH: Adaptive Recurrent-Convolutional Hybrid Networks for Long-Term Action Recognition. Neurocomputing, Vol. 178, 2016, pp. 87–102, doi: 10.1016/j.neucom.2015.09.112.

[50] Li, W.—Nie, W.—Su, Y.: Human Action Recognition Based on Selected Spatio-Temporal Features via Bidirectional LSTM. IEEE Access, Vol. 6, 2018, pp. 44211–44220, doi: 10.1109/ACCESS.2018.2863943.

[51] Yang, H.—Zhang, J.—Li, S.—Luo, T.: Bi-Direction Hierarchical LSTM with Spatial-Temporal Attention for Action Recognition. Journal of Intelligent and Fuzzy Systems, Vol. 36, 2019, No. 1, pp. 775–786, doi: 10.3233/JIFS-18209.

[52] Wang, L.—Xiong, Y.—Wang, Z.—Qiao, Y.—Lin, D.—Tang, X.—Van Gool, L.: Temporal Segment Networks: Towards Good Practices for Deep Action Recognition. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (Eds.): Com-

puter Vision – ECCV 2016. Springer, Cham, Lecture Notes in Computer Science, Vol. 9912, 2016, pp. 20–36, doi: 10.1007/978-3-319-46484-8_2.

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

[54] Liu, A. A.—Su, Y. T.—Nie, W. Z.—Kankanhalli, M.: Hierarchical Clustering Multi-Task Learning for Joint Human Action Grouping and Recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 39, 2016, pp. 102–114, doi: 10.1109/TPAMI.2016.2537337.

[55] Ke, Y.—Sukthankar, R.—Hebert, M.: Efficient Visual Event Detection Using Volumetric Features. Vol. 1, 2005, pp. 166–173, doi: 10.1109/ICCV.2005.85.

[56] Lan, Z.—Lin, M.—Li, X.—Hauptmann, A. G.—Raj, B.: Beyond Gaussian Pyramid: Multi-Skip Feature Stacking for Action Recognition. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 204–212, doi: 10.1109/CVPR.2015.7298616.

[57] Islam, N.—Faheem, Y.—Din, I. U.—Talha, M.—Guizani, M.—Khalil, M.: A Blockchain-Based Fog Computing Framework for Activity Recognition as an Application to E-Healthcare Services. Future Generation Computer Systems, Vol. 100, 2019, pp. 569–578, doi: 10.1016/j.future.2019.05.059.

[58] Hou, J.—Wu, X.—Sun, Y.—Jia, Y.: Content-Attention Representation by Factorized Action-Scene Network for Action Recognition. IEEE Transactions on Multimedia, Vol. 20, 2018, No. 6, pp. 1537–1547, doi: 10.1109/TMM.2017.2771462.

Kumari Priyanka Sinha is pursuing her Ph.D. in the National Institute of Technology Patna. She is working as Assistant Professor in the Nalanda College of Engineering, Chandi, Department of Computer Science and Engineering. Her current research interests include computer vision, artificial intelligence and machine learning.

Prabhat Kumar is Professor in Computer Science and Engineering Department at the National Institute of Technology Patna, India. He is also the Professor-in-Charge of the IT Services and Chairman of Computer & IT Committee of NIT Patna. He is a member of NWG-13 (National Working Group 13) corresponding to ITU-T Study Group 13 "Future Networks, with focus on IMT-2020, cloud computing and trusted network infrastructures". He is a former HOD, CSE Department, NIT Patna and former State Student Coordinator of Bihar for Computer Society of India. He holds his Ph.D. in computer science and his M.Tech. in information technology. He has more than 100 publications in various reputed journals and international conferences. He has served as guest editor of special issues in international journals and has also edited several books published by reputed international publishers. He is also in the reviewing panel of multiple reputed SCI indexed journals. He has chaired sessions at several international conferences held in India and abroad. He is a senior member of IEEE, professional member of ACM, life member of CSI, International Association of Engineers (IAENG), Indian Society for Technical Education (ISTE) and global member of Internet Society. His research area includes wireless sensor networks, internet of things, data science, software engineering, e-governance, etc.

Rajib Ghosh is currently working as Assistant Professor in Computer Science and Engineering Department at the National Institute of Technology (NIT) Patna, India. He has more than 20 years of experiences of teaching in different engineering colleges. He completed his Ph.D. (computer science and engineering) degree at the National Institute of Technology (NIT) Patna, India. He also holds his M.Tech. degree in information technology and B.E. degree in computer science and engineering. His broader research domains are pattern recognition, machine learning and computer vision. His research areas of interest are document analysis and recognition, object detection, object tracking, human movement tracking, video surveillance, etc. His Ph.D. work explored pattern recognition methods using machine learning techniques for recognizing online handwritten text of different Indic scripts such as Bengali, Devanagari, Telugu, Tamil, etc. He has over 30 research publications in various reputed SCI-indexed as well as SCOPUS-indexed international journals and conferences of repute. He is also the reviewer of several reputed journals indexed in SCI, SCIE and SCOPUS. He is one of General Chairs in one international conference. He has also chaired sessions at several international conferences. He is a member of IEEE, life member of IUPRAI, life member of ISTE, and nominee member of CSI. He has delivered expert talks and guest lectures at various prestigious institutes.