HAR-Depth: A Novel Framework for Human Action Recognition using Sequential Learning and Depth Estimated History Images

Suraj Prakash Sahoo, Samit Ari, *Member, IEEE,* K.K. Mahapatra, *Senior Member, IEEE* and Saraju P. Mohanty, *Senior Member, IEEE*

***Abstract*—Human action recognition (HAR) is a challenging task due to the presence of the pose and temporal variations in the action videos. To address these challenges, HAR-Depth is proposed in this paper with sequential and shape learning along with the novel concept of depth history image (DHI). A deep bidirectional long short term memory (DBiLSTM) is constructed for sequential learning to model the temporal relationship existing between the action frames. Action information in each frame is extracted using pre-trained convolutional neural network (CNN). The depth information of each action frame is estimated and projected onto the X-Y plane to form the DHI. During shape learning, the shape information through DHI is used to train a deep pre-trained CNN network. By leveraging the trained knowledge of the pre-trained network, overfitting issue is handled. The finetuned network is used to recognize actions from query DHI images. Data augmentation is adopted to avoid overfitting of the network by virtually increasing the training set. The proposed work is evaluated on publicly available datasets like KTH, UCF sports, JHMDB, UCF101, and HMDB51 and achieves the performance accuracy of 97.67%, 95.00%, 73.13%, 92.97%, and 69.74% respectively. The results on these datasets suggest that the proposed work of this paper performs better in terms of overall accuracy, kappa parameter and precision compared to the other state-of-the-art algorithms present in the earlier reported literature.**

***Index Terms*—Action recognition, data augmentation, depth estimation, fine tuning, sequential learning.**

I. Introduction

RECOGNIZING human actions through computer vision is a trending research area. Human action recognition (HAR) is helpful in the field of the visual surveillance system as it is having applications in both indoor and outdoor environments. Now-a-days, research in this area attracts more number of researchers due to its vast application fields. Some of the applications of HAR are detection of abnormal activities in sensitive areas, patient's behavior recognition in hospitals, sports data analysis, video retrieval etc. However, the HAR paradigm faces challenges in recognizing the actions efficiently due to the presence of inter-class similarities and intra-class variations among the action classes. Inter-class similarity arises during the recognition of actions like 'run', 'jog' and 'walk', where the shape of the actions are nearly similar. Different persons perform a same action in a different way, which results in intra-class action variations. Similarly, appearance of an action also changes due to camera angle.

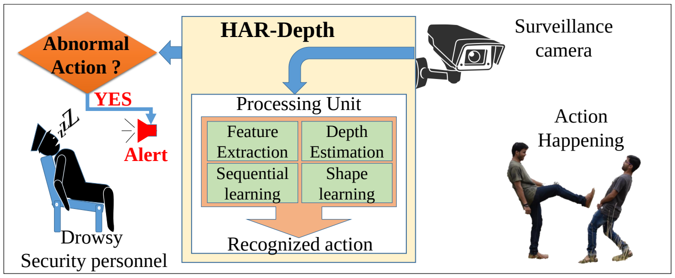


Fig. 1. A thematic diagram of HAR-Depth paradigm aiding the surveillance application.

Human actions are characterized by simple hand and leg motions. In complex scenarios, it is the combination of body motion and interaction with surroundings. Therefore, single image is not sufficient to perceive an action correctly. It is about learning the relation between informations extracted from each frame. The temporal relationship in an action video is an important aspect to recognize the action type. Various techniques are developed to recognize human actions, such as space-time features [1]–[4] and trajectory based methods [5]–[8]. Recently, sequential learning based long short term memory (LSTM) is used by the researchers for HAR. LSTM is made up of input units, output units and hidden units. The advantage of the LSTM lies in its memory cells, which can remember the learning parameters or past information. The modified version of LSTM is bidirectional LSTM or BiLSTM [9], which is a combination of two LSTMs in forward and backward directions. The BiLSTM network is capable of remembering the past information as well as future information. However, the success of the sequential learning is very much dependent on efficient features extracted from the action frames. Along with the sequential information, the shape information of the whole action needs to be studied. The HAR paradigm will certainly be more efficient if depth information is available. Unfortunately, most of the available datasets for HAR do not have depth information. Depth information acquired from depth cameras provides better shape

S.P. Sahoo and S. Ari are with the Pattern Recognition Laboratory, Department of Electronics and Communication Engineering, National Institute of Technology Rourkela, Odisha, India, 769008, e-mail: (surajprakashsahoo@gmail.com, samit@nitrkl.ac.in).

K.K. Mahapatra is with the Department of Electronics and Communication Engineering, National Institute of Technology Rourkela, Odisha, India, 769008, e-mail: (kkm@nitrkl.ac.in).

S. P. Mohanty is with the Department of Computer Science and Engineering, University of North Texas, e-mail: (saraju.mohanty@unt.edu).

TABLE I
THE HIERARCHY OF EXISTING TECHNIQUES AND THEIR EVOLUTION IN HUMAN ACTION RECOGNITION.

Approaches	Existing works	Propositions	Solutions
Holistic approach	Dalal, *et al*. [2]	Histogram of oriented gradient (HOG)	Shape through gradients for human detection
Sparse representation	Laptev [3]	Spatio-temporal interest points (STIP)	Detection of 3D local motional corners
	Sahoo, *et al*. [4]	STIP filtering through region of interest by MHI	Reduction of noisy interest points
	Lin, *et al*. [19]	Spatial-temporal histogram of gradients (SPHOG) feature	Representation of locally extracted regions
Feature indexing trees	Yu, *et al*. [18]	Random projection tree with median split	Indexing of local features extracted as STIPs
	Sahoo, *et al*. [10]	Random projection tree with overlapping split	Indexing of local features extracted from LMDI
Feature fusion	Li, *et al*. [16]	Fusion of features from STIP and 3DSURF	Combining feature for better human action recognition
	Yu, *et al*. [17]	Fusion of appearance features and motion features	Leveraging the motion and shape information for HAR
3DCNN	Ji, *et al*. [20]	3D convolution kernel based CNN network	Deep network for spatio-temporal information
Multi-stream CNN	Qin, *et al*. [22]	Fusion of handcrafted features and deep features	Leveraging different features for HAR
Sequential learning	Gammulle, *et al*. [25]	LSTM on extracted deep features from action frames	Modelling of temporal sequential information
	Xu, *et al*. [26]	Leverages LSTM and attention networks along with CNN	Extracting sequential and shape information for HAR
Proposed work	(HAR-Depth)	Two-stream deep network with depth history image	Shape representation by depth history image (DHI) and sequential learning for better HAR

representation, however the depth cameras are not cost effective. Thus, estimating depth from action frames became an open challenge to the research society. A thematic diagram of the proposed HAR-Depth is depicted in Fig. 1.

The rest of this paper is structured as follows: contributions of this paper are mentioned in section II. The state-of-the-art methods are discussed in Section III. The proposed work is explained thoroughly in Section IV, where the sequential and shape learning details are described along with their detailed parameter set ups. The algorithm is validated by conducting experiments on publicly available datasets and the results are provided in section V. Finally, section VI concludes the work of this paper.

II. Contributions of the Current Paper

The challenges of HAR lie in efficient representation of different actions. The novel contributions of this work which addresses this issue, are as follows.

- Depth information is estimated for RGB action frames in this work. The depth frames are projected onto a single frame for creation of the proposed DHI to represent shape of an action.
- Another aspect of representing an action is to model the temporal information lying in between the action frames. This is carried out by sequential learning through proposed DBiLSTM network. For DBiLSTM network, features are extracted from each action frame by pre-trained CNN network.
- During training of DHI images, the network undergoes overfitting due to limited training data. The techniques like transfer learning and data augmentation are adopted to handle overfitting problem.
- Finally, the scores from both the learning network (shape learning and sequential learning) are fused to provide the final recognition score.

III. Related Prior Research

In the literature, two different approaches are available for feature extraction of HAR: handcrafted features and deep learning based automatic feature extraction. A hierarchy of existing techniques and their evolution in human action recognition is presented in Table I. Hand crafted features are the only source for feature extraction before deep learning approaches. Handcrafted feature extraction technique includes histogram of optical flow (HOF) [1], histogram of oriented gradient (HOG) [2], spatio-temporal features like STIP [3], LMDI [10], HOG3D, HOF3D [4], and motion trajectories [5]–[7]. Initial development of human action recognition has started with the view based template matching [11]. A tangent space [12] is reported, which helps to recognize actions by projecting the action differences onto the tangent space. For compact representation of a video, STIP [3] is proposed to extract movable corners in the action videos. The work reported in [13] has concentrated on handling multi-view action classes, where a view invariant feature is proposed to represent the actions. In [14], the authors have proposed a human intention inference system for human target prediction. Haar wavelet transform is used for feature extraction in [15]. The handcrafted features like STIP and 3DSURF features are combined [16] and the combined feature is classified by a multi-class SVM. Fusion of appearance features and motion features is carried out in [17] to recognize human actions. 2D HOG and HOF local features are extracted and being indexed by random forest tree structures by Yu *et al.* [18]. Another type of spatial-temporal histogram of gradients (SPHOG) feature for HAR is proposed in [19].

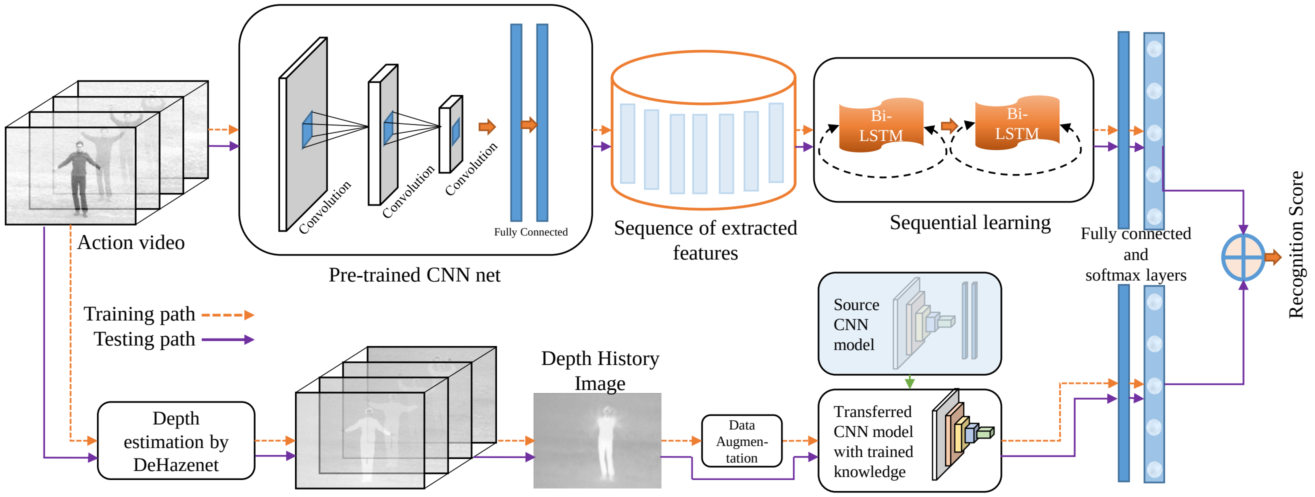


Fig. 2. Block diagram of the proposed human action recognition framework (HAR-Depth) containing sequential learning through DBiLSTM and shape learning using DHI images.

In contrast to handcrafted feature extraction techniques, the deep neural networks extract features automatically. CNN based techniques provide both feature extraction and classification in a single network. The convolution layer extracts feature in spatio-temporal direction. However, adjusting kernels to distinguish actions like ‘run’ and ‘jog’ is a tedious task. 3D convolutional neural network (3DCNN) is developed for human action recognition in [20]. CNN based feature extraction along with SVM based classification used for HAR in [21]. Fusion of deep features are proposed by [22] where traditional features are combined with 3DConvNet features to provide a better feature vector for HAR. In [4], the above mentioned closely related actions are better classified compared to earlier techniques using local 3D features along with semi-supervised random forest classifier. Stacked Fisher vector [23] is a deep architecture which relies on dense extraction of cuboids and trajectories. In [24], single shot multibox detector based action recognition is performed on each frame.

The recurrent neural network (RNN) is a recent trend in HAR as it can explore the sequential informations residing in an action video. Specifically, LSTM represents the sequential information present among the video frames. Sequential learning by LSTM on extracted deep features from action frames is reported in [25]. The work of [26] leverages LSTM networks and attention networks along with CNN to recognize human actions. Spatial-optical data organization along with sequential learning [5] is used to recognize different actions. During spatial-optical data organization, motion trajectories and optical flow are used on whole RGB video data. The concepts of faster R-CNN, two stream CNN and multi region CNNs are used by Peng *et al.* [27] for action recognition. Two LSTM layers are applied in opposite directions to form a bidirectional LSTM network for recognition of complex frame-to-frame hidden sequential patterns in [28]. Multi-layer LSTM network is applied on extracted optical flow features to learn long term action sequences for industrial surveillance applications in [29]. Bidirectional LSTM or BiLSTM network, which is a special kind of LSTM, is used to recognize actions in [9].

IV. Proposed Two-Stream Network for Human Action Recognition Technique

The block diagram of the proposed framework is shown in Fig. 2. The work is divided into two streams: sequential learning and shape learning streams. The shape of the action plays an important role for action representation. In the first stream, relation between the action frames is extracted using the DBiLSTM networks. Pre-trained CNN architecture like AlexNet [30], is used to extract frame level features. In second stream, depth information of each action frame is estimated and depth history image (DHI) is proposed to project the action onto a single X-Y plane. During testing, the recognition score is generated from both the streams and then fused to provide the final recognition score to each action class.

The intuition behind proposing the HAR-Depth network is to leverage maximum information from action videos and model them effectively to reduce recognition error. As frame-to-frame sequential information and the action shape are important to model an action, HAR-Depth is more advantageous.

A. A Deep Bidirectional LSTM (DBiLSTM) Network for HAR

Bidirectional LSTM or BiLSTM network [9] is a combination of two LSTM cells in forward and backward directions. It is required to learn the sequential information residing between frame-to-frame of an action video. Bidirectional nature helps to remember the past and future information during network training. In this work, a deep architecture of BiLSTM layers is created and trained on extracted features. In experiments, three BiLSTM layers are placed in series with a dropout in each layer. These dropout factor helps to reduce overfitting of the network by ignoring some neurons during training. The proposed two-stream network learns the long term sequences with the help of DBiLSTM stream. DBiLSTM stream is built with LSTM cells which is a special kind of RNN network designed to learn the long term sequences. LSTM layer learns

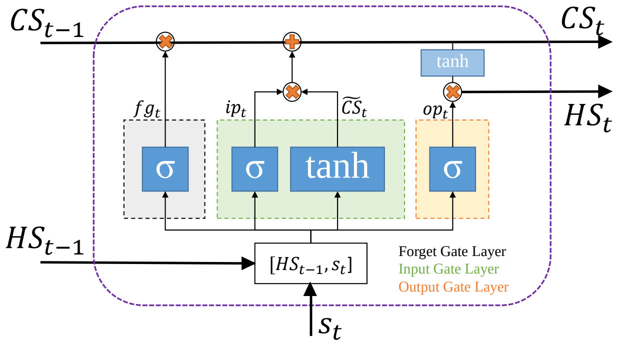


Fig. 3. Cell structure of LSTM showing input gate, hidden gate and output gate layers.

the long term sequences by (i) overcoming the vanishing gradient problem, and (ii) regulating the cells by non-linear gating units known as input gate, output gate, and forget gate.

Input to the DBiLSTM network is the collection of features extracted from each action frame. The recent success of CNN has made it as a better feature extractor. As action frames are simply images, any CNN trained on huge image data can be used for feature extraction. The simplest CNN trained on large scale ImageNet dataset is AlexNet [30], which comprises of 5 convolution layers and 3 fully connected layers. During feature extraction, the frames are passed through the CNN network and features are collected from last fully connected layer. The extracted features with a dimension of 1000 are fed to DBiLSTM network to learn the sequential temporal information among them.

The LSTM network operates with basic memory cell structures as shown in Fig. 3. Each memory cell is operated by gate structures, cell state updation, and hidden state calculation. The forget gate layer decides whether the previous information will be forgotten or not. The forget gate fg_t can be formulated using the previous hidden state HS_{t-1}, corresponding weight W_{fg}, and bias b_{fg} as below:

$$fg_t = \sigma(W_{fg}.[HS_{t-1}, s_t] + b_{fg}), \tag{1}$$

where s_t is the signal or feature vector at time t. Similarly, input and output gate layers are formulated as follows:

$$ip_t = \sigma(W_{ip}.[HS_{t-1}, s_t] + b_{ip}), \tag{2}$$

$$\tilde{CS_t} = tanh(W_{CS}.[HS_{t-1}, s_t] + b_{CS}), \tag{3}$$

$$op_t = \sigma(W_{op}.[HS_{t-1}, s_t] + b_{op}), \tag{4}$$

In the updation stage, the cell state and the hidden state are updated as in (5) and (6).

$$CS_t = CS_{t-1}.fg_t + ip_t.\tilde{CS_t}, \tag{5}$$

$$HS_t = tanh(CS_t).op_t. \tag{6}$$

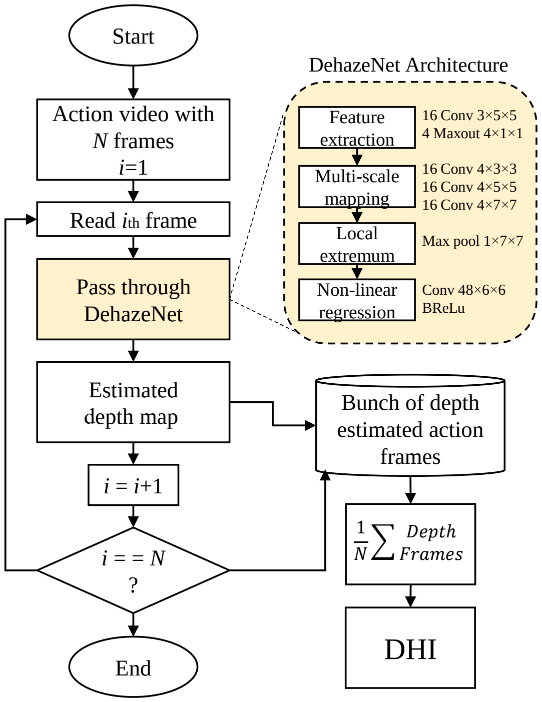


Fig. 4. Flowchart of the proposed depth history image (DHI) using deep learning based DehazeNet network.

B. Depth History Image (DHI)-The Novel Concept of this Work

The shape of an action plays an important role in differentiating it from other actions. The DHI is proposed in this work to describe the shape of an action through depth estimation. The detailed procedure is explained in Fig. 4. Generally the foreground and background objects exist at different depth levels. By using the estimated depth map, the foreground human action can be separated from the background. The advantage of depth based imaging lies in the availability of more information about the appearance of an object. As depth data is not available for all datasets, depth estimation technique can be helpful. In this work, depth information is extracted through medium transmission map as used in DehazeNet [31] for haze removal. DehazeNet uses a deep learning network to estimate the depth map. The details of the architecture are provided in Fig. 4. Let, *H* be the hazy image, *I* be the haze free image, then the relation between them can be represented as:

$$H(x, y) = I(x, y).T_r(x, y) + \mu(1 - T_r(x, y)), \tag{7}$$

where T_r is the medium transmission map and μ is the global atmospheric light. To recover *I*, T_r should be estimated properly. T_r depends on the depth of the scene *i.e.* distance of the scene from the camera as defined below.

$$T_r(x, y) = e^{-\eta d(x,y)}, \tag{8}$$

where η is the scattering coefficient of the atmosphere and *d* is the distance of the scene from the camera. This concept

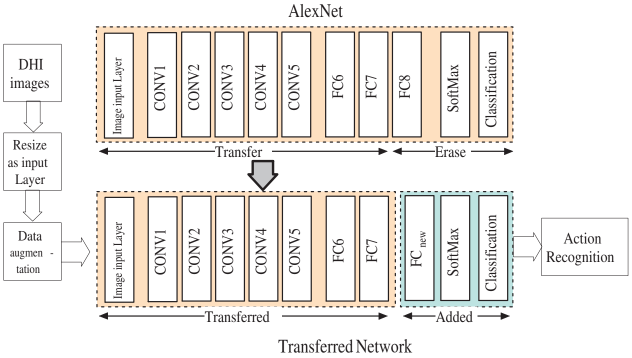


Fig. 5. Training of DHI images through transfer learning on AlexNet.

is utilized in the proposed work to estimate depth maps for action scenes.

The next step is to prepare the DHI. All the estimated depth frames are projected onto a single X-Y plane to calculate the DHI for its corresponding action video as shown in Fig. 4. The DHI is calculated by averaging the depth estimated frames as shown in (9). Let, the depth estimated frames be d_i^{BS}, for i =1,...,N and then, the depth history image for an action is represented as follows.

$$I_{DHI} = \frac{1}{N}\sum_{i=1}^{N} d_i^{BS} \tag{9}$$

C. Proposed Methods for Transfer Learning and Data Augmentation for DHI

Training of any deep neural network from scratch requires huge training data. As number of extracted DHI images are not sufficient, the training procedure faces overfitting. To handle this problem, two techniques such as transfer learning [32] and data augmentation [32] are adapted in this work.

1) Transfer Learning: It is the process of transferring learned weights or knowledge from a pre-trained network and retraining them. In this work, the pre-trained weights of AlexNet are taken as the initial weights while training the DHI images. The detailed procedure is shown in Fig. 5. The layers except the final fully connected layer (fc8), softmax layer and the classification layer of AlexNet are transferred. The final three layers of the AlexNet are replaced according to the requirement. Number of channels in the final fully connected layer (fc8) is decided by the number of action classes present in the dataset. Feature maps for DHIs after first convolution layer through transfer learning are depicted in Fig. 6.

2) Data Augmentation: Data augmentation is a technique to increase the number of samples internally when training data is very less. The network will look a single image as different images in each loop through specific data variations. The procedure increases the efficiency of the network by reducing the overfitting effect upto some extent. In this work, combination of four types of augmentation techniques is used. The details are given in Table II. The subjective representation of output of data augmentation is provided in Fig. 7.

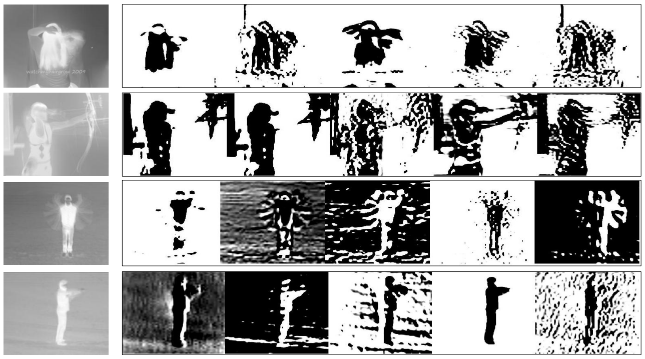
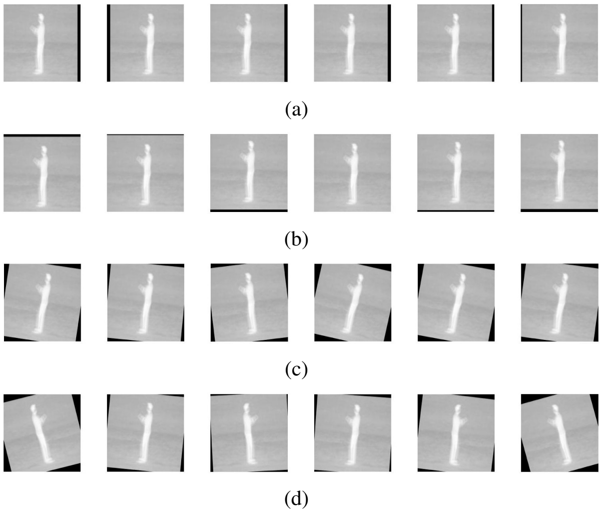


Fig. 6. Extracted feature maps for DHI of example actions from first convolution layer.


(a)

(b)

(c)

(d)

Fig. 7. Data augmentation on DHI images. (a) various X-Translation (b) various Y-Translation (c) various rotation (d) random reflection combined with (a)-(c). (DHI of boxing action of KTH dataset is used for representation purpose. The translation and rotation values are increased for this figure to demonstrate the changes).

D. Score Fusion

Let the scores generated from sequential learning be S_{sq} and scores generated from shape learning be S_{sp}. Then, final score is calculated by taking α probability of S_{sq} score and (1-α) probability of S_{sp}. Here, α is the fusion parameter.

$$S_{final}(i) = \alpha S_{sp}(i) + (1-\alpha)S_{sq}(i), \tag{10}$$

where, i=1,...,N, and N is the number of action classes present in the dataset.

The effect of α on performance accuracy is studied to decide the optimal value. As an example, the results on two small-scale (KTH, JHMDB) and one mid-scale (UCF101) datasets are depicted in Fig. 8. It can be observed that the performance of the proposed technique is better at α equal to 0.5. From experiments, it is found that by giving equal weightage to both the representations, the performance is better. This observation concludes to choose the α equals to 0.5 throughout this work.

E. Training Procedure

This section discusses about the effect of the number of BiLSTM layers in the DBiLSTM stream on the performance, the

TABLE II
DETAILS OF THE DATA AUGMENTATION PARAMETERS USED DURING TRAINING.

Sl No	Type	Range
1	X-Translation	random shifting in [-5, 5] pixels
2	Y-Translation	random shifting in [-5, 5] pixels
3	Rotation	random rotation in [-10, 10] degrees
4	X-Reflection	-

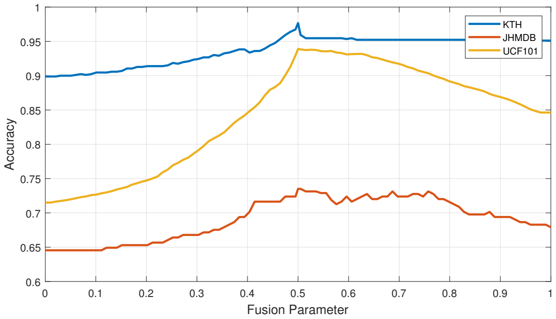


Fig. 8. The effect of fusion parameter on overall accuracy for KTH, JHMDB, and UCF101 datasets.

optimization technique in the training procedure and the loss function used to calculate the loss. A detailed training/testing step is provided in the Algorithm 1.

1) Depth of DBiLSTM Layers: To study the effect of the number of BiLSTM layers in the DBiLSTM stream, an experiment is carried out for the analysis between classification accuracy and training time. As an example, the results are shown for the KTH dataset in Table III. From the tabulation, it is observed that the training time has increased by increasing the depth of the DBiLSTM layers. DBiLSTM network with 3 layers shows the maximum classification performance. By increasing the depth further, there is a little improvement or no improvement in the performance compared to large increment in training time. Therefore, to keep the balance between performance and training time, three number of BiLSTM layers are chosen in the DBiLSTM stream.

2) Steepest Descent Gradient with Momentum (SGDM): For network optimization, steepest descent gradient with momentum (SGDM) optimizer is applied. Let W, b, η be the weights, bias and learning parameter for the network. Similarly, dW and db are the derivatives of the cost function with respect to weight and bias respectively. v_{dW}, v_{db} are the velocities with respect to weight and bias. Then, SGDM is defined as:

$$\begin{aligned} W_{new} &= W_{prev} - \eta.v_{dW}, \\ b_{new} &= b_{prev} - \eta.v_{db}, \end{aligned} \tag{11}$$

where,

$$\begin{aligned} v_{dW} &= \beta.v_{dW} + (1-\beta).dW_{prev}. \\ v_{db} &= \beta.v_{db} + (1-\beta).db_{prev}. \end{aligned} \tag{12}$$

The momentum parameter β is chosen as 0.9 or above. By doing so, the vertical oscillation of the weight updation path is reduced resulting in a faster training procedure.

Algorithm 1: Training/testing steps of the proposed two stream HAR-Depth framework

```
Initialize training action video set $V_{train}$ with class
  labels $Y_{train}$, testing action video set $V_{test}$, number
  of action classes $C$, batch size $B$, number of epochs
  $E$, Fusion parameter $\alpha$, $X_B$ variable to store feature
  vectors, DBiLSTM layers $L_{seq}$ and transfer learning
  layers from AlexNet for depth learning layers $L_{depth}$.
Function Training ($V_{train}, Y_{train}, L_{seq}, L_{depth}, B, E$)
for epochs=1 to E do
    $V_B$ = $B$ number of videos from $V_{train}$
    $Y_B$ = $B$ corresponding labels from $Y_{train}$
    Sequential learning stream
        Load AlexNet                                  #Pre-trained CNN network
        $[X_B(i)]_{i=1}^{B} = [Pass\,(V_B(i),\ AlexNet)]_{i=1}^{B}$
        $S_{Seq} = TrainNetwork(X_B, Y_B, L_{Seq})$
        Clear $X_B$
    Shape learning stream
        Load $D$ =DehazeNet                           #For depth estimation
        $N_i$ = Number of frames in $V_B(i)$
        $[X_B(i)]_{i=1}^{B} = \left[\frac{1}{N}\sum_{j=1}^{N} Pass\,(V_B(i,j),\ D)\right]_{i=1}^{B}$
       $S_{Depth} = TrainNetwork(X_B, Y_B, L_{Depth})$
    Save $S_{Seq}, S_{Depth}$
    Update $L_{Seq}, L_{Depth}$
end
End Function

Function Testing ($V_{test}, S_{Seq}, S_{Depth}$)
for $V \in V_{test}$ do
                                   #$V$ is a single query video from $V_{test}$
    $X_{Seq}$ = $Pass\,(V, AlexNet)$;
    $X_{Depth}$ = $\sum_{j=1}^{N} Pass\,(V(j),\ D)$ ;    #N: Number of frames in $V$
    $Score1$ = Classify ($X_{Seq}$, $S_{Seq}$) ;
    $Score2$ = Classify ($X_{Depth}$, $S_{Depth}$) ;
    $Score_{Final} = \alpha.Score1 + (1-\alpha).Score2$ ;
    Recognized action class = arg $\max_{C}(Score_{Final})$
end
End Function
```

3) Cross Entropy Loss: Loss is calculated at the classification layer after each feed forward pass. To calculate the loss, cross entropy technique is applied in this paper. Let, C be the total number of action classes, p be the generated score in the softmax layer, and l represents the classification layer. Then, the cross entropy loss is defined as follows.

$$Loss = -\sum_{i=1}^{C} l_i \log(p_i) \tag{13}$$

Here, l_i is either 0 or 1, and the negative sign in (13) is to counter the negative value of the log function.

TABLE III
EFFECT OF DEPTH OF THE DBILSTM NETWORK ON CLASSIFICATION ACCURACY AND TRAINING TIME.

BiLSTM layers	1	2	3	4	5
Classification Accuracy (%)	82.33	82.85	85.46	84.53	84.19
Training Time (in Second)	701.77	821.12	957.8	1155.59	1343.63

V. EXPERIMENTAL RESULTS AND DISCUSSION

This section presents the results obtained from proposed HAR-Depth algorithm on different small-scale and mid-scale datasets and discusses the overall performance.

A. Datasets Used in Our Experiments

Three small-scale datasets such as KTH [33], UCF sports [34], JHMDB [35] and two mid-scale datasets like UCF101 [36], HMDB51 [37] are used to evaluate the proposed work.

KTH dataset: The dataset contains six types of actions, which are collected from indoor and outdoor environments. The dataset is having more inter-class similarity as it contains similar actions such as 'running', 'jogging', 'walking', 'boxing', 'hand waving', and 'hand clapping'. Nearly, 100 action videos are present per each class and the actions are performed by 25 different persons. In our experiments, actions of 16 persons are grouped as training set and the rest are utilized for testing as reported in [18].

UCF Sports dataset: The UCF sports dataset is more realistic and is collected from broadcast television. The dataset consists of 150 action videos of 480×720 resolution in 10 action classes. Due to the presence of less number of action videos per class, leave one out cross validation (LOOCV) method is adopted as mentioned in [34]. The actions present in this dataset are : 'diving', 'golf swing', 'kicking', 'lifting', 'riding horse', 'running', 'skate boarding', 'swinging bench', 'swinging side' and 'walking'.

JHMDB dataset: The JHMDB dataset is a relatively complex dataset having 21 action classes with 36-55 action videos in each class. All the actions are single human actions and are named as 'brush hair', 'catch', 'clap', 'climb stairs', 'golf', 'jump', 'kick ball', 'pick', 'pour', 'pull-up', 'push', 'run', 'shoot ball', 'shoot bow', 'shoot gun', 'sit', 'stand', 'swing baseball', 'throw', 'walk', 'wave'. The dataset is divided into training and testing set as mentioned in [35].

UCF101 dataset: The dataset is having 101 number of action classes with a total of 13320 number of action videos. The actions are broadly classified into five types such as 1) human-object interaction 2) body-motion only 3) human-human interaction 4) playing musical instruments 5) sports.

HMDB51 dataset: This complex dataset contains nearly 7000 videos in 51 action classes. Broadly, the dataset contains the action types such as real time basic actions, facial actions, human-object interaction, and human-human interaction.

B. Experimental Setup

- During the experiments, training is very much important as DHI images are new to AlexNet. Training of DHI images on AlexNet will update the network in accordance to depth knowledge which in turn produces better features during feature extraction. As the number of videos in UCF sports and JHMDB datasets are comparatively less than KTH dataset, the KTH dataset is first trained through transfer learning. The weights of the trained dataset is used for UCF sports and JHMDB datasets. The mid-scale datasets are trained independently as the number of videos are larger compared to small-scale datasets.
- In BiLSTM network, three BiLSTM units are used in sequence. The number of hidden units are 150 in first layer, 125 in second layer and 100 in third layer. Other chosen parameters are: mini-batch size as 8, maximum number of epochs as 300 and initial learning rate as 0.001. All the values are chosen empirically.
- The shape learning network, which uses the layers of AlexNet, has the following parameters: maximum number of epochs as 300 and initial learning rate as 0.0001.

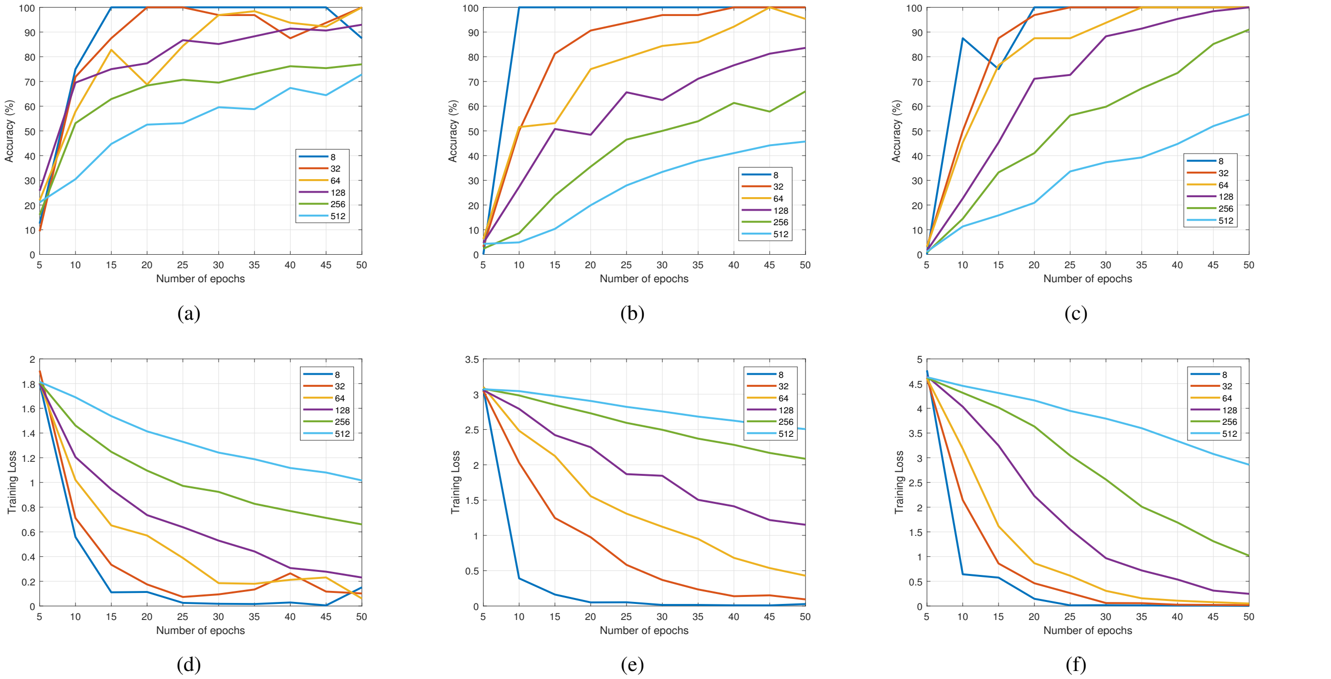
All the implementations are carried out in MATLAB platform (Version: 9.4.0.813654 / R2018a) with Intel Xeon E5-2630 v4 (10 Core, 2.2 GHz, 32GB RAM) processor and NVIDIA Quadro M4000 8GB GPU. For quantitative assessment, the parameters used to compare the proposed method with the previously reported techniques are: classification accuracy (CA), kappa parameter (k) and precision (P).

C. Evaluation of the Proposed HAR-Depth Network

1) Effect of Batch Size: Batch size is defined as the number of training samples passed through the learning network before calculation of the loss function. The lower the batch size, the faster is the convergence of the training procedure. A detailed experiment is carried out on two small-scale datasets (KTH, JHMDB) and one mid-scale dataset (UCF101) to explain the effect of batch size on training loss and performance accuracy. The same is depicted in the Fig. 9.

2) Sample Complexity: The performance of the proposed network is analyzed by varying the number of samples per each class and this is termed as sample complexity analysis. In the experiment, the amount of training samples are increased in a 20% increment strategy and the effect on the performance accuracy is analyzed. It is found that, as the number of training samples have increased, the performance of the system increases accordingly as depicted in Fig. 10.

3) Class Complexity: The performance of the proposed network is also analyzed by varying the number of action classes in the training dataset and this is termed as class complexity analysis. The analysis is carried out on shape learning stream to analyze the effect of DHI on the whole network. Similar to sample complexity analysis, the training classes are increased in a 20% increment set up and the performance accuracy is analyzed. As number of actions in the dataset increases, the performance is reduced due to increase in inter-action complexity as depicted in Fig. 11.



(a) (b) (c)

(d) (e) (f)

Fig. 9. The effect of batch size on training accuracy and training loss, (a)-(c) Training accuracy, (d)-(f) Training loss for KTH, JHMDB and UCF101 datasets.

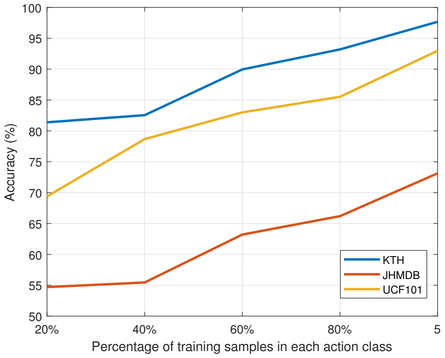


Fig. 10. Sample complexity analysis of KTH, JHMDB and UCF101 datasets.

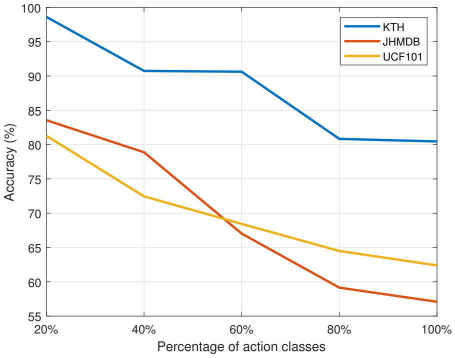


Fig. 11. Class complexity analysis for DHI stream with KTH, JHMDB and UCF101 datasets.

4) Ablation Study: As the proposed HAR-Depth is a two-stream network, the effect of each stream on HAR is analyzed. The performance for DBiLSTM stream and DHI based shape stream are provided separately in Table IV. The overall performance accuracy is also provided for comparison purpose. To evaluate the stability of the network, standard deviation of performance accuracy for all the datasets are also provided along with the overall accuracy.

The second stream of the HAR-Depth network consists of training of DHI images with data augmentation (DA) and transfer learning (TL) techniques. Therefore, an ablation study is carried out to analyze the effect of different parts of the network on the overall performance as depicted in Table V. From the table, it is observed that the transfer learning technique is very much effective to provide better performance on a small-scale dataset. The data augmentation has helped to improve the performance of the network.

TABLE IV
ABLATION STUDY OF DIFFERENT STREAMS OF HAR-DEPTH NETWORK FOR ALL THE EVALUATING DATASETS.

Datasets	DBiLSTM stream Accuracy (%)	DHI stream Accuracy (%)	Overall Accuracy (%)
KTH	85.46	80.46	97.67 ± 1.61
UCF sports	90.91	84.55	95.00 ± 0.45
JHMDB	60.82	57.09	73.13 ± 2.05
UCF101	85.25	62.39	92.97 ± 0.95
HMDB51	62.16	46.41	69.74 ± 2.46

D. Comparison of Results with the Earlier Reported Methods for Small-scale Datasets

To evaluate the proposed work, three datasets namely KTH, UCF sports and JHMDB are used. Performance of each dataset with earlier reported techniques is discussed one after another in this section. The KTH dataset is challenging to differentiate the actions like running, jogging and walking. The efficiency

TABLE V
ABLATIVE STUDY OF THE PROPOSED HAR-DEPTH NETWORK IN TERMS OF PERFORMANCE ACCURACY (IN %) ON KTH DATASET. ABLATIONS INCLUDE THE DBILSTM STREAM, BASELINE DHI STREAM, DHI WITH DATA AUGMENTATION (DA) ONLY, DHI WITH TRANSFER LEARNING (TL) ONLY, DHI WITH DL AND TL

Epochs	20	40	60	80	100
DBiLSTM	81.49	81.72	83.92	84.15	84.85
DHI	48.03	53.70	62.73	65.16	68.06
DHI+DA	49.31	57.29	64.70	70.02	70.72
DHI+TL	73.74	75.47	75.70	76.63	77.90
DHI+DA+TL	75.24	76.17	77.21	77.44	78.37

TABLE VI
COMPARISON OF THE PROPOSED TECHNIQUE WITH EXISTING STATE-OF-THE-ART TECHNIQUES FOR SMALL-SCALE DATASETS

Dataset	Method	CA (%)	*k*	P
KTH Dataset	Chou, *et al.* [13]	90.58	0.66	90.90
	Yu, *et al.* [18]	91.80	0.70	92.78
	Samanta, *et al.* [15]	94.91	0.81	94.82
	Megrhi, *et al.* [6]	94.90	-	-
	Qin, *et al.* [22]	95.10	-	-
	Sahoo, *et al.* [4]	96.29	0.87	96.55
	Shi, *et al.* [7]	96.80	0.89	97.08
	Proposed method	**97.67**	**0.92**	**97.73**
UCF Sports Dataset	Song, *et al.* [38]	73.67	0.25	77.10
	Lui, *et al.* [12]	88.00	0.35	89.10
	Lin, *et al.* [19]	89.80	-	-
	Xu, *et al.* [26]	91.89	-	-
	Sahoo, *et al.* [4]	92.67	0.54	92.64
	Proposed method	**95.00**	**0.72**	**95.00**
JHMDB Dataset	Gkioxari, *et al.* [21]	62.50	0.75	64.23
	Peng, *et al.* [23]	69.03	-	-
	Gammulle, *et al.* [25]	69.00	0.71	68.48
	Singh, *et al.* [24]	72.00	-	-
	Peng, *et al.* [27]	73.10	-	-
	Proposed method	**73.13**	0.66	**78.31**

TABLE VII
COMPARATIVE PERFORMANCES OF ALL CLASSES OF KTH DATASET USING PROPOSED TECHNIQUE

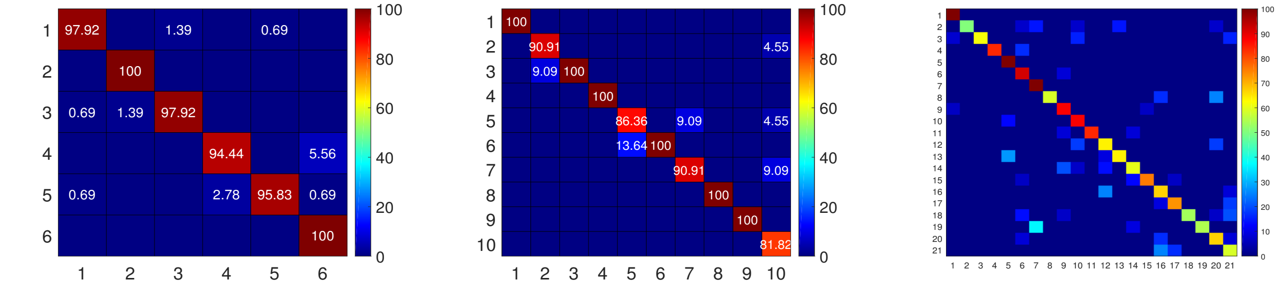
Actions	Sensitivity	Specificity	Positive predictivity
Boxing	0.98	1.00	0.99
Waving	1.00	1.00	0.99
Clapping	0.98	1.00	0.99
Jogging	0.94	0.99	0.97
Running	0.96	1.00	0.99
Walking	1.00	0.99	0.94

of the algorithm depends on the higher accuracy of these three related actions. The actions are difficult to differentiate since the shape of the actions are nearly similar. It is the action execution time which differentiates them from each other. The frequency of action pattern in running is higher than that of jogging and walking. In literature, the KTH dataset is tested and evaluated by the methods proposed by [4], [6], [7], [13], [15], [18], [22]. 2D features [18] and view invariant features [13] do not have any special approach to distinguish the actions differentiated by action time. In contrary to this, the trajectory based approach [6] and 3DCNN [22] based approaches can handle the problem upto some extent. In our previous work [4], 3D HOG and HOF features are combined with semi-supervised random forest which handles the closely related actions more accurately. CNN and trajectory based approaches are combined by Shi *et al.* [7] for which the overall accuracy is 96.8%. In this work, the sequential information among action frames are combined with depth estimated action shape to distinguish the actions more accurately. The confusion matrix for the proposed work is presented in Fig. 12(a). The performance accuracy of 97.67% is achieved by the proposed technique which is 0.87% and 1.38% better than [7] and [4] respectively as depicted in Table VI.

Statistical indices such as sensitivity, specificity and positive predictivity are presented for all the action classes in Table VII. The ‘waving’ action is better classified as its statistical index values (sensitivity of 1, specificity of 1 and positive predictivity of 0.99) are better than others. The better result of the proposed method is obtained by leveraging sequential learning, deep neural networks and combining them with action shape through depth estimation. As recognizing the actions like running, jogging, walking is the challenge for KTH dataset, average accuracy of these three classes is calculated for analysis. In this work, the closely related actions are classified with an average accuracy of 96.76%. Similarly, the average accuracy of other three classes and overall accuracy are found to be 98.61% and 97.67% respectively.

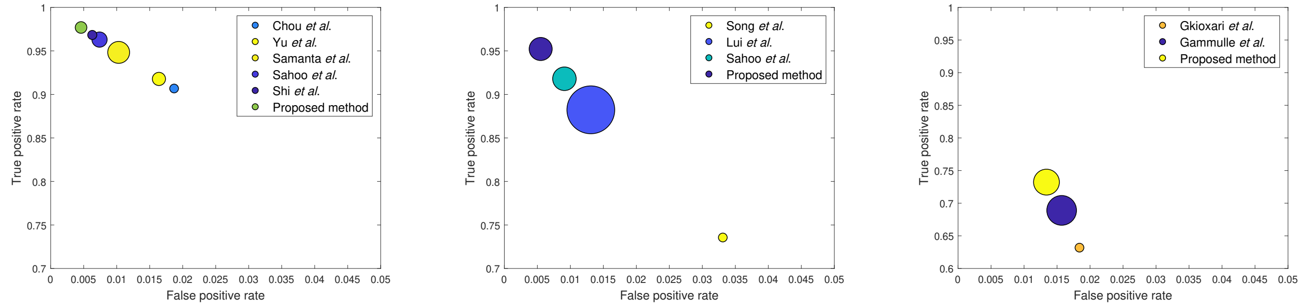
The UCF sports dataset is a complex dataset compared to KTH dataset as it is collected from real time broadcast televisions. Therefore, the performance on this dataset plays a vital role on efficiency of the proposed method. The UCF sports dataset is tested and evaluated by the methods proposed by [4], [12], [19], [26], [38] and our proposed method. The confusion matrix is shown in Fig. 12(b) and comparison with state-of-the-art methods is shown in Table VI. The work of [12] requires action localization for better action recognition. Lui *et al.* [12] have extended HOG to SPHOG for better temporal representation of an action. Recently sequential learning [26] is used to learn the sequential information present between frames. In this proposed work, sequential learning is aided with depth based shape information to make the algorithm more efficient. As a result, an accuracy of 95% is found which is better than the earlier techniques. Similar to KTH dataset evaluation, statistical indices are calculated for UCF sports actions to analyze the recognition efficiency of different actions, which is shown in Table VIII. The best recognized actions are found to be ‘diving’, ‘lifting’, and ‘swinging’ as all the parameters of these classes are better than other classes. The recognition performance of ‘walking’ class is the lowest as it is confused with ‘riding horse’, ‘golf swing’ and ‘skate boarding’.

The JHMDB dataset is tested and evaluated by the methods proposed by [21], [23]–[25], [27]. The confusion matrix for the dataset is given in Fig. 12(c). The overall accuracy is found to be 73.13% by the proposed method. Comparison of the performance accuracy with state-of-the-method is presented in Table VI. The methods reported in [21], [23], [24], [27] have not considered the long term temporal relationship residing in an action video. LSTM technique is a better algorithm to extract



(a) (b) (c)

Fig. 12. The confusion matrices for small-scale datasets using the proposed technique. The details of action names are reported in datasets subsection in sequence, (a) confusion matrix for KTH dataset, (b) confusion matrix for UCF sports dataset, (c) confusion matrix for JHMDB dataset



(a) (b) (c)

Fig. 13. Comparison of the discrete ROC graphs for small-scale datasets with state-of-the-art techniques (The better technique resides more towards (0,1) corner). (a) KTH dataset (b) UCF sports dataset (c) JHMDB dataset. (The size of circle is a scaled representation of amount of improvement from previous technique)

TABLE VIII
COMPARATIVE PERFORMANCES OF ALL CLASSES OF UCF SPORTS DATASET USING PROPOSED TECHNIQUE

Actions	Sensitivity	Specificity	Positive predictivity
Diving	1.00	1.00	1.00
Golf swing	0.95	0.99	0.91
Kicking	0.92	1.00	1.00
Lifting	1.00	1.00	1.00
Riding horse	0.86	0.98	0.86
Running	0.88	1.00	1.00
Skate boarding	0.91	0.99	0.91
Swinging bench	1.00	1.00	1.00
swinging side	1.00	1.00	1.00
walking	1.00	0.98	0.82

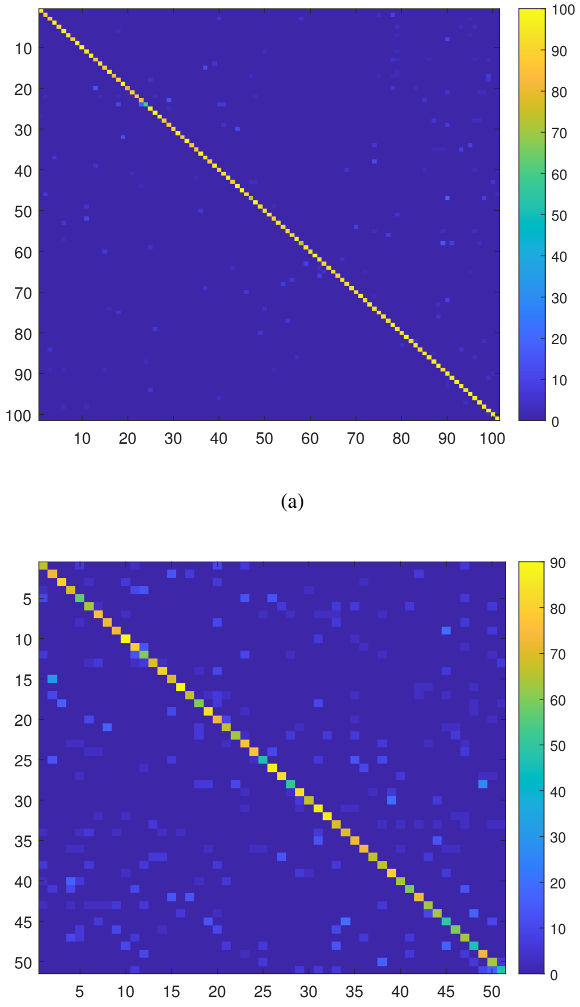
relation between action frames by sequential learning. Feature extraction by pre-trained CNN networks and sequence learning by LSTM are performed by [25] and achieved comparable results with existing techniques. Since, the proposed work uses BiLSTM networks and depth estimated shape information combiningly, it can extract temporal and shape information more accurately, which results in better performance compared to the earlier reported techniques.

The proposed algorithm is evaluated through discrete ROC graph for all the mentioned small-scale datasets and presented in Fig. 13. ROC graph is a spatial presentation of true positive rate vs false positive rate. A perfect classifier will have the upper right corner (0,1) of the ROC graph. An algorithm is better compared to other if it is more north west positioned on ROC graph. As shown in Fig. 13, for KTH, UCF sports, and JHMDB datasets, the proposed technique performs better compared to state-of-the-art techniques as it is closest to (0,1) on ROC graph.

E. Comparison of Results with the Earlier Reported Methods for Mid-scale Datasets

The proposed network is evaluated on two mid-scale datasets such as UCF101 and HMDB51. The confusion matrix for both the datasets are depicted in Fig. 14. The performance accuracy of UCF101 dataset is compared with the techniques reported in [5], [7], [28], [39]–[43] and the performance accuracy of HMDB51 dataset is compared with the techniques reported in [5], [7], [29], [39]–[46]. Comparison of the performance accuracy of the proposed technique with state-of-the-art methods are depicted in Table IX for UCF101 dataset and in Table X for HMDB51 dataset. The performance accuracy of 92.97% and 69.74% are achieved for UCF101 and HMDB51 datasets respectively, which are better compared to all the reported state-of-the-art methods. From the confusion matrices, it can be visualized that for all the action classes, the diagonal cells or true positives are brighter which represents the effectiveness of the proposed technique. Similar to the parameters presented in Table VI, a k value of 0.72 and 0.87, a P value of 93.44 and 71.21 are reported for UCF101 and HMDB51 datasets respectively.

Finally, it is observed that the proposed HAR-depth technique faces challenge when two closely related faster actions



(a)

(b)

Fig. 14. The confusion matrices for mid-scale datasets by the proposed technique, (a) confusion matrix for UCF101 dataset, (b) confusion matrix for HMDB51 dataset.

TABLE IX
Comparison of the proposed technique with existing state-of-the-art techniques for UCF101 dataset

Methods	**CA (%)**
Liu *et al.* [39]	76.30
Lu *et al.* [40]	90.10
Wang *et al.* [41]	90.30
Feichtenhofer *et al.* [42]	90.80
Yuan *et al.* [5]	90.90
Ullah *et al.* [28]	91.21
Zhao *et al.* [43]	91.70
Shi *et al.* [7]	92.20
Proposed HAR-Depth	**92.97**

are recognized at the same time e.g. ‘riding horse’ and ‘running’ actions of UCF sports dataset. Similarly, the close temporal pattern of ‘kicking’ and ‘golf swing’ has also become troublesome during the recognition HAR-Depth technique. In KTH dataset, the closely related actions are recognized with an average performance accuracy of 96.67%.

TABLE X
Comparison of the proposed technique with existing state-of-the-art techniques for HMDB51 dataset

Methods	**CA (%)**
Liu *et al.* [39]	51.40
Yang *et al.* [44]	60.80
Xin *et al.* [45]	61.10
Feichtenhofer *et al.* [42]	62.10
Wang *et al.* [41]	63.20
Lu *et al.* [40]	64.50
Zhao *et al.* [43]	64.80
Shi *et al.* [7]	65.20
Yuan *et al.* [5]	65.70
Sekma *et al.* [46]	68.50
Proposed HAR-Depth	**69.74**

VI. Conclusions

This paper has introduced a two-stream HAR-Depth network for HAR. The proposed network is capable of learning the long-term sequences to recognize different action classes. The DBiLSTM stream learns the sequential information and the DHI stream learns the shape information of an action. The DHIs are constructed to provide better shape representation by estimating depth information from RGB action frames. The problem of network overfitting due to less training data is overcome by the transfer learning and data augmentation techniques. It is observed that the proposed HAR-Depth network performs better in terms of performance accuracy compared to the state-of-the-art techniques for five different publicly available datasets like KTH, UCF sports, JHMDB, UCF101, and HMDB51. Various ablation studies, parameter sensitivity, sample complexity, class complexity analysis suggest that the proposed HAR-Depth performs well and provides promising performance for HAR.

Acknowledgment

This publication is an outcome of the R&D work undertaken in the project under the Visvesvaraya PhD Scheme of Ministry of Electronics & Information Technology, Government of India, being implemented by Digital India Corporation (formerly Media Lab Asia), [grant number PhD-MLA/4(13)/2015-16].

References

[1] J. L. Barron and N. A. Thacker, “Tutorial: Computing 2D and 3D optical flow,” *Imaging Science and Biomedical Engineering Division, Medical School, University of Manchester*, vol. 1, 2005.

[2] N. Dalal and B. Triggs, “Histograms of oriented gradients for human detection,” in *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, vol. 1, 2005, pp. 886–893.

[3] I. Laptev, “On space-time interest points,” *International journal of computer vision*, vol. 64, no. 2-3, pp. 107–123, 2005.

[4] S. Sahoo, U. Srinivasu, and S. Ari, “3D features for human action recognition with semi-supervised learning,” *IET Image Processing*, vol. 13, no. 6, pp. 983–990, 2019.

[5] Y. Yuan, Y. Zhao, and Q. Wang, “Action recognition using spatial-optical data organization and sequential learning framework,” *Neurocomputing*, vol. 315, pp. 221–233, 2018.

[6] S. Megrhi, M. Jmal, W. Souidene, and A. Beghdadi, “Spatio-temporal action localization and detection for human action recognition in big dataset,” *Journal of Visual Communication and Image Representation*, vol. 41, pp. 375–390, 2016.

[7] Y. Shi, Y. Tian, Y. Wang, and T. Huang, "Sequential deep trajectory descriptor for action recognition with three-stream cnn," *IEEE Transactions on Multimedia*, vol. 19, no. 7, pp. 1510–1520, 2017.

[8] S. P. Sahoo and S. Ari, "Automated human tracking using advanced mean shift algorithm," in *Proc. IEEE International Conference on Communications and Signal Processing (ICCSP)*, 2015, pp. 0789–0793.

[9] Y. Bin, Y. Yang, F. Shen, N. Xie, H. T. Shen, and X. Li, "Describing video with attention-based bidirectional LSTM," *IEEE Transactions on Cybernetics*, vol. 49, no. 7, pp. 2631–2641, 2018.

[10] S. P. Sahoo and S. Ari, "On an algorithm for human action recognition," *Expert Systems with Applications*, vol. 115, pp. 524–534, 2019.

[11] A. F. Bobick and J. W. Davis, "The recognition of human movement using temporal templates," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 23, no. 3, pp. 257–267, 2001.

[12] Y. M. Lui and J. R. Beveridge, "Tangent bundle for human action recognition," in *Proc. IEEE International Conference on Automatic Face & Gesture Recognition and Workshops*, 2011, pp. 97–102.

[13] K. P. Chou, M. Prasad, D. Wu, N. Sharma, D. L. Li, Y. F. Lin, M. Blumenstein, W. C. Lin, and C. T. Lin, "Robust feature-based automated multi-view human action recognition system," *IEEE Access*, vol. 6, pp. 15 283–15 296, 2018.

[14] D. Das and C. Lee, "Cross-scene trajectory level intention inference using gaussian process regression and naive registration," 2018. [Online]. Available: https://docs.lib.purdue.edu/ecetr/491/

[15] S. Samanta and B. Chanda, "Space-time facet model for human activity classification," *IEEE Transactions on Multimedia*, vol. 16, no. 6, pp. 1525–1535, 2014.

[16] C. Li, B. Su, Y. Liu, H. Wang, and J. Wang, "Human action recognition using spatio-temoporal descriptor," in *Proc. International Congress on Image and Signal Processing (CISP)*, vol. 1, 2013, pp. 107–111.

[17] S. Yu, Y. Cheng, L. Xie, and S.-Z. Li, "Fully convolutional networks for action recognition," *IET Computer Vision*, vol. 11, no. 8, pp. 744–749, 2017.

[18] G. Yu, N. A. Goussies, J. Yuan, and Z. Liu, "Fast action detection via discriminative random forest voting and top-k subvolume search," *IEEE Transactions on Multimedia*, vol. 13, no. 3, pp. 507–517, 2011.

[19] B. Lin and B. Fang, "Spatial-temporal histograms of gradients and HOD-VLAD encoding for human action recognition," in *Proc. IEEE International Conference on Security, Pattern Analysis, and Cybernetics (SPAC)*, 2017, pp. 678–683.

[20] S. Ji, W. Xu, M. Yang, and K. Yu, "3D convolutional neural networks for human action recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 1, pp. 221–231, 2013.

[21] G. Gkioxari and J. Malik, "Finding action tubes," in *Proc. IEEE conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 759–768.

[22] Y. Qin, L. Mo, and B. Xie, "Feature fusion for human action recognition based on classical descriptors and 3D convolutional networks," in *Proc. IEEE International Conference on Sensing Technology (ICST)*, 2017, pp. 1–5.

[23] X. Peng, C. Zou, Y. Qiao, and Q. Peng, "Action recognition with stacked fisher vectors," in *Proc. European Conference on Computer Vision (ECCV)*, 2014, pp. 581–595.

[24] G. Singh, S. Saha, M. Sapienza, P. H. Torr, and F. Cuzzolin, "Online real-time multiple spatiotemporal action localisation and prediction," in *Proc. IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 3637–3646.

[25] H. Gammulle, S. Denman, S. Sridharan, and C. Fookes, "Two stream LSTM: A deep fusion framework for human action recognition," in *Proc. IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2017, pp. 177–186.

[26] Y. Xu, L. Wang, J. Cheng, H. Xia, and J. Yin, "DTA: Double LSTM with temporal-wise attention network for action recognition," in *Proc. IEEE International Conference on Computer and Communications (ICCC)*, 2017, pp. 1676–1680.

[27] X. Peng and C. Schmid, "Multi-region two-stream R-CNN for action detection," in *Proc. European Conference on Computer Vision (ECCV)*. Springer, 2016, pp. 744–759.

[28] A. Ullah, J. Ahmad, K. Muhammad, M. Sajjad, and S. W. Baik, "Action recognition in video sequences using deep bi-directional LSTM with CNN features," *IEEE Access*, vol. 6, pp. 1155–1166, 2017.

[29] A. Ullah, K. Muhammad, J. Del Ser, S. W. Baik, and V. H. C. de Albuquerque, "Activity recognition using temporal optical flow convolutional features and multilayer lstm," *IEEE Transactions on Industrial Electronics*, vol. 66, no. 12, pp. 9692–9702, 2018.

[30] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Proc. Advances in Neural Information Processing Systems*, 2012, pp. 1097–1105.

[31] B. Cai, X. Xu, K. Jia, C. Qing, and D. Tao, "Dehazenet: An end-to-end system for single image haze removal," *IEEE Transactions on Image Processing*, vol. 25, no. 11, pp. 5187–5198, 2016.

[32] D. Han, Q. Liu, and W. Fan, "A new image classification method using cnn transfer learning and web data augmentation," *Expert Systems with Applications*, vol. 95, pp. 43–56, 2018.

[33] C. Schuldt, I. Laptev, and B. Caputo, "Recognizing human actions: a local SVM approach," in *Proc. IEEE International Conference on Pattern Recognition (ICPR)*, vol. 3, 2004, pp. 32–36.

[34] M. D. Rodriguez, J. Ahmed, and M. Shah, "Action mach a spatio-temporal maximum average correlation height filter for action recognition," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2008, pp. 1–8.

[35] H. Jhuang, J. Gall, S. Zuffi, C. Schmid, and M. J. Black, "Towards understanding action recognition," in *Proc. IEEE International Conference on Computer Vision (ICCV)*, 2013, pp. 3192–3199.

[36] K. Soomro, A. R. Zamir, and M. Shah, "UCF101: A dataset of 101 human actions classes from videos in the wild," *arXiv preprint arXiv:1212.0402*, 2012.

[37] H. Kuehne, H. Jhuang, E. Garrote, T. Poggio, and T. Serre, "Hmdb: a large video database for human motion recognition," in *Proc. IEEE International Conference on Computer Vision*, 2011, pp. 2556–2563.

[38] Y. Song, S. Tang, Y. T. Zheng, T. S. Chua, Y. Zhang, and S. Lin, "A distribution based video representation for human action recognition," in *Proc. IEEE International Conference on Multimedia and Expo (ICME)*, 2010, pp. 772–777.

[39] A.-A. Liu, Y.-T. Su, W.-Z. Nie, and M. Kankanhalli, "Hierarchical clustering multi-task learning for joint human action grouping and recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 1, pp. 102–114, 2016.

[40] X. Lu, H. Yao, S. Zhao, X. Sun, and S. Zhang, "Action recognition with multi-scale trajectory-pooled 3d convolutional descriptors," *Multimedia Tools and Applications*, vol. 78, no. 1, pp. 507–523, 2019.

[41] L. Wang, Y. Qiao, and X. Tang, "Action recognition with trajectory-pooled deep-convolutional descriptors," in *Proc. IEEE conference on Computer Vision and Pattern Recognition*, 2015, pp. 4305–4314.

[42] C. Feichtenhofer, A. Pinz, and A. Zisserman, "Convolutional two-stream network fusion for video action recognition," in *Proc. IEEE conference on Computer Vision and Pattern Recognition*, 2016, pp. 1933–1941.

[43] S. Zhao, Y. Liu, Y. Han, R. Hong, Q. Hu, and Q. Tian, "Pooling the convolutional layers in deep convnets for video action recognition," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 28, no. 8, pp. 1839–1849, 2017.

[44] Y. Yang, R. Liu, C. Deng, and X. Gao, "Multi-task human action recognition via exploring super-category," *Signal Processing*, vol. 124, pp. 36–44, 2016.

[45] M. Xin, H. Zhang, H. Wang, M. Sun, and D. Yuan, "Arch: Adaptive recurrent-convolutional hybrid networks for long-term action recognition," *Neurocomputing*, vol. 178, pp. 87–102, 2016.

[46] M. Sekma, M. Mejdoub, and C. B. Amar, "Human action recognition based on multi-layer fisher vector encoding method," *Pattern Recognition Letters*, vol. 65, pp. 37–43, 2015.

Suraj Prakash Sahoo received his Bachelor of Technology degree in Electronics and Communication Engineering from Biju Patnaik University of Technology (BPUT) in the year 2012. He received his Master of Technology degree in Signal and Image processing specialization from National Institute of Technology, Rourkela in the year 2015. He is currently pursuing his PhD degree in the Department of Electronics and Communication Engineering at National Institute of Technology, Rourkela. His research interests include human action recognition, deep learning, machine learning and computer vision.

Samit Ari (M'10) received Bachelor of Technology degree in Electronics and Tele-Communication Engineering from University of Kalyani, West Bengal, India in 2001, Master of Technology degree in Instrumentation Engineering from University of Calcutta, West Bengal, India in 2003, and Ph.D. degree in Electronics and Electrical Communication Engineering from Indian Institute of Technology (IIT), Kharagpur, India in 2009. He joined National Institute of Technology (NIT), Rourkela as a faculty member in 2009, where he presently holds the position of Associate Professor in the Department of Electronics and Communication Engineering. His research interests include Signal Processing, Image processing, Pattern Recognition and Machine Learning. Dr. Ari is a member of IEEE and at present, he is also serving as an Associate Editor of IET Image Processing Journal. He has published more than 65 research articles and his Google Scholar h-index is 16 with 1000+ citations. He was awarded as Young Faculty Research Fellow under the Visvesvaraya PhD scheme for Electronics & IT, MeitY, for the year of 2015-16.

Kamalakanta Mahapatra (M'12) obtained his B. Tech degree with Honours from Regional Engineering College, Calicut in 1985, M. Tech from Regional Engineering College, Rourkela in 1989 and Ph. D. from IIT Kanpur in 2000. He is currently a Professor in Electronics and Communication Engineering Department of National Institute of Technology (NIT), Rourkela. He assumed this position since February 2004. He is a fellow of the Institution of Engineers (India) in ECE Division. He has published several research papers in National and International Journals. His research interests include Embedded Computing Systems, VLSI Design, Hardware Security and Industrial Electronics.

Saraju P. Mohanty (SM'08) received the bachelor's degree (Honors) in electrical engineering from the Orissa University of Agriculture and Technology, Bhubaneswar, in 1995, the master's degree in Systems Science and Automation from the Indian Institute of Science, Bengaluru, in 1999, and the Ph.D. degree in Computer Science and Engineering from the University of South Florida, Tampa, in 2003. He is a Professor with the University of North Texas. His research is in "Smart Electronic Systems" which has been funded by National Science Foundations (NSF), Semiconductor Research Corporation (SRC), U.S. Air Force, IUSSTF, and Mission Innovation. He has authored 350+ research articles, 4 books, and invented 4 U.S. patents. His Google Scholar h-index is 36 and i10-index is 134 with 5800+ citations. He is a recipient of 12 best paper awards, Fulbright Specialist Award in 2020, IEEE Consumer Electronics Society Outstanding Service Award in 2020, the IEEE-CS-TCVLSI Distinguished Leadership Award in 2018, and the PROSE Award for Best Textbook in Physical Sciences and Mathematics category in 2016. He has delivered 9 keynotes and served on 5 panels at various International Conferences. He is the Editor-in-Chief (EiC) of the IEEE Consumer Electronics Magazine (MCE).