Video augmentation technique for human action recognition using genetic algorithm

Nudrat Nida¹ | Muhammad Haroon Yousaf¹,² | Aun Irtaza³ | Sergio A. Velastin⁴,⁵

¹Department of Computer Engineering, UET, Taxila, Pakistan
²Swarm Robotics Lab, National Centre of Robotics and Automation, Islamabad, Pakistan
³Department of Computer Science, UET, Taxila, Pakistan
⁴Applied Artificial Intelligence Research Group, Department of Computer Science and Engineering, University Carlos III de Madrid, Madrid, Spain
⁵School of Electronic Engineering and Computer Science, Queen Mary University of London, London, UK

Correspondence
Muhammad Haroon Yousaf, Department of Computer Engineering, UET, Taxila, Pakistan.
Email: haroon.yousaf@uettaxila.edu.pk

Abstract
Classification models for human action recognition require robust features and large training sets for good generalization. However, data augmentation methods are employed for imbalanced training sets to achieve higher accuracy. These samples generated using data augmentation only reflect existing samples within the training set, their feature representations are less diverse and hence, contribute to less precise classification. This paper presents new data augmentation and action representation approaches to grow training sets. The proposed approach is based on two fundamental concepts: virtual video generation for augmentation and representation of the action videos through robust features. Virtual videos are generated from the motion history templates of action videos, which are convolved using a convolutional neural network, to generate deep features. Furthermore, by observing an objective function of the genetic algorithm, the spatiotemporal features of different samples are combined, to generate the representations of the virtual videos and then classified through an extreme learning machine classifier on MuHAVi-Uncut, iXMAS, and IAVID-1 datasets.

KEYWORDS
computer vision, evolutionary deep features augmentation, genetic algorithm, human action recognition, video augmentation

1 | INTRODUCTION

Human action recognition (HAR) are predicted by visual and motion metaphors, either using traditional handcrafted features with prior information (that is, dense trajectories [1], BoF [2], and holistic representation [3,4]) or deep learning models, such as two-stream network [5], 3D convolutional neural network (CNN) [6], and recurrent neural networks [7,8]. Despite their successes, deep learning models require large training sets. These two categories of HAR techniques potentially improve the accuracy of the learning model based on the quality and quantity of the training data. However, data skewness and irregularity degrade performance due to missing feature representation and class imbalance problems.

Synthetic data generation and data sampling are effective measures for mitigating data irregularity. Computer vision research has focused on synthetic data generation by applying data augmentation techniques to grow the training sets for stable HAR models.
Recently, synthetic data generation has been employed to learn models for action recognition [5], scene understanding [9], pose estimation [10], and object recognition [11]. Video augmentation techniques, including inverse-order frame [12], upsampling [13], frame mirroring [14], frame skipping [15], generative adversarial networks (GANs) [16–21], and long short-term memory (LSTM) networks [22,23], increase the video training data for model learning. The PredNet CNN model [17] was used to forecast future frames of videos. Here, every layer of PredNet computed a local prediction and passed it to the next layer, to encode spatiotemporal information using fewer training samples.

Tiny videos [20] were generated using GANs by convolved streams from two-stream inputs: foreground motion and background frames contextual stream. However, a stable input video involving Random Sample Consensus (RANSAC) and Scale-Invariant Feature Transform (SIFT) feature computation at each video frame is required. Similarly, small videos were generated by extending recurrent neural networks (RNNs) for temporal recurrence and spatial convolutions [22]. GAN-based methods for video generation have higher computational overhead and are limited by nonconverging gradients and mode collapse. Thus, they produce limited sample variation and loss imbalance between the generator and discriminator networks [24].

Because data augmentation with standard techniques generates translated replications of existing examples, it might not result in significant improvements in learning. Moreover, existing data augmentation approaches lack semantic diversity that improves classification results [18]. Therefore, robust data augmentation approaches ensure dynamic feature generation for more accurate HAR prediction models.

To overcome these limitations, this paper presents a virtual video generation using a genetic algorithm (GA) through evolving deep spatiotemporal feature representation. An initial training set contains a deep spatiotemporal feature representation of action classes. The GA increases the number of instances for each class to improve action classification and resolve data irregularity. Our contribution revolves around ways of achieving stronger predictive models rather than creating physical video data. Sequel to this, we evaluate diverse virtual video representations for HAR and compare our technique with state-of-the-art video augmentation techniques. GA introduces semantic diversity by combining the spatiotemporal video representation of the same class. These spatiotemporal representations are the semantic and temporal contextual information of each action video. The combination of spatiotemporal information of action video results in a diverse training set, extending the training set for HAR using virtual videos by inducing semantic diversity in training data to stimulate wider variations in scene contextual information (Figure 2) and improve the learning accuracy of HAR models. Third, our proposed approach overcomes data irregularity to generate stronger recognition models.

2 | THE PROPOSED APPROACH

Consider the training data \( (V_{x,y,t}^i, l^i) \), where \( V^i \) are the action videos and \( l^i \) are corresponding action labels, and \( i = \{1,2,3,...,m\} \) are the total number of training samples such that each video sample \( V_{x,y,t}^i \) consists of \( t \) video
frames. First, each video sequence $V_{i}^{\tau}$ is processed to obtain the actor’s silhouette $f(x,y,t)$ using graph-cut segmentation [27] (Figure 2). The parameter $x,y$ represents the pixel frame information and $t$ denotes the temporal shift of video frames. The actor’s silhouettes $f(x,y,t)$ of each video frame are aggregated to form a single spatiotemporal actor’s motion template $M_{t_{f}}(x,y,t)$ using motion history images [3]. Afterward, deep spatiotemporal features $D$ are computed for the motion template $M_{t_{f}}(x,y,t)$ using a CNN network. To extend the training set for HAR, we generate optimal deep spatiotemporal feature $D$ representations of virtual videos for $\psi$ number of generations of $\eta$ dimension using GA and select the elite deep spatiotemporal features $D^{+}$ representations that minimize the classification loss $J(y_{p})$ between the predicted $\hat{y}_{p}$ and actual action classes $l$. The extended training set is evaluated using an extreme learning machine (ELM) classifier to obtain optimal decision boundaries for HAR. However, our technique is neither limited to specific types of CNN models nor the input video, that is, RGB video, 2D frames, or silhouettes data, rather, our approach extends the training set irrespective of input data. The complete architecture of the proposed method is presented in Figure 2.

\section*{2.1 The deep spatiotemporal feature extraction}

For creating an initial population for generating virtual video, action videos are processed to segment the actor silhouettes $f(x,y,t)$ [27]. Then, the actor’s silhouettes $f(x,y,t)$ of the entire video form a single actor’s motion template $M_{t_{f}}(x,y,t)$ [3], to encode spatiotemporal movement information for $t$ video frames, as given by (1).

$$M_{t_{f}} = \begin{cases} \tau & \text{if } f(x,y,t) = 1 \\ \max(0, (0, f(x,y,t-1))) & \text{otherwise} \end{cases}, \quad (1)$$

where $\tau$ is the total silhouettes frames segmented to form $M_{t_{f}}$ for all action videos. Later, the actor’s motion templates $M_{t_{f}}$ are normalized using min–max normalization.

We further rescaled the $N \times N$ dimensions for presenting to deep CNN pretrained models. Next, the CNN model generates deep spatiotemporal representation $D$ by convolving 2D with a kernel $w$ of dimension $z \times z$, and the results are summed up onto the receptive field, (2).

$$D = \sum_{x=0}^{N-1} \sum_{y=0}^{N-1} w \ast M_{t_{f}}. \quad (2)$$

\section*{2.2 GA for generating virtual action videos}

For virtual video generation using GA, the set of features $D$ serve as the initial population $\phi_{m,k}$ of dimension $\eta$, where each row $m$ of $\phi_{m,k}$ represents a chromosome and each feature, a gene $\theta$ in the chromosome, $m \times k$ represents the feature dimension of $D$ with $m$ samples of $k$ features, and $k$ denote the total genes $\theta$ in one chromosome. The initial population $\phi_{m,k}$ was evolved for $\psi$ generations using GA’s selection and crossover operations to produce a new pair of offspring that fulfills the fitness criteria $J$ of the evolution.

In our modified two-point discrete crossover, the two crossover points $[c_{1}, c_{2}]$ randomly opted from the real number $0 < \Re < k$ for interchanging the sections of two parents $\{P_{1}, P_{2}\} = \{\phi_{a,k}, \phi_{a+1,k}\}$, as expressed in (3) and (4). The crossover operation breeds new offsprings for inheriting the characteristics of parent chromosomes, and some evolved discriminative characteristics. We ensure that the generated new pair of offsprings are stronger and satisfy the fitness criteria.

$$\phi_{a,k} = \{\theta_{P_{1,1}} \leftarrow \theta_{P_{1,\alpha_{1}}}, \theta_{P_{1,\alpha_{1}}} \leftarrow \theta_{P_{2,\alpha_{1}}}, \theta_{P_{2,\alpha_{2}}} \leftarrow \theta_{P_{1,T}}\}. \quad (3)$$

$$\phi_{a+1,k} = \{\theta_{P_{1,1}} \leftarrow \theta_{P_{1,\alpha_{1}}}, \theta_{P_{2,\alpha_{1}}} \leftarrow \theta_{P_{1,\alpha_{2}}}, \theta_{P_{1,\alpha_{2}}} \leftarrow \theta_{P_{2,T}}\}. \quad (4)$$

The fitness function $J$ was gaged for the entire population in each generation $\psi$ to produce an elite deep spatiotemporal $D^{+}$ representation of virtual videos. In our work, the terminating condition for GA to evolved
chromosomes is to select elite deep spatiotemporal feature representation from generations of feature representation through observing minimum fitness measure. The generation of deep spatiotemporal \( D^+ \) representations is an evolved elite version of the initial feature ‘s chromosome population, which differs from the initial \( D \) computed using the human action videos. Our evolved representation extends the training set as an important function for model optimization. Here, we emphasize video analysis to recognize different human action classes and achieve better action prediction performance.

2.3 Classification using ELM

After computing the evolved \( D^+ \), the representations are passed to a feedforward ELM classifier for model generation. The ELM classifier predicts \( y_p(D^+) \) as a single output unit using hidden nodes \( h \), as described in (5):

\[
y_p(D^+) = \sum_{p=1}^{h} \gamma_p \times H_p(D^+),
\]

where \( \gamma_p = \{\gamma_1, \gamma_2, \gamma_3, \ldots, \gamma_h\} \) are the weights between the ELM’s \( h \) hidden nodes, and ELM’s output vectors \( H_p \). The ELM output vectors are \( y_p(D^+) = \{y_1, y_2, y_3, \ldots, y_h\} \), and ELM decision function for action recognition is a sigmoid activation function. The classification loss \( J(y_p^l) \) of ELM between the predicted and actual action class is shown in (6).

\[
J(y_p^l) = \min \left( \frac{1}{2} \|y^l - y_p^l\|_2 \right).
\]

3 EXPERIMENTS AND FINDINGS

3.1 Datasets and evaluation protocol

3.1.1 Datasets

The performance evaluation was conducted on the activity recognition datasets of INRIA Xmas (iXMAS) [28], MuHAVi-Uncut [29], and IAVID-1 [30]. The INRIA Xmas motion acquisition sequence (iXMAS) [28] is a silhouette-based activity dataset that comprises 12 activities performed by 12 actors. iXMAS, as multiview dataset, captures human actions from five viewpoints.

The MuHAVi-Uncut dataset is a multiview activity video dataset that contains 17 activities performed by 14 actors. It comprises 2898 single actor silhouette videos. The IAVID-1 dataset was developed for the instructor activity recognition and contains 100 videos with 8 action classes performed by 12 actors. The action video samples in MuHAVi-Uncut and IAVID-1 are unbalanced, as shown in Figure 3.

Generating stronger decision boundaries for learning models with imbalanced training datasets is difficult. We selected an imbalance dataset to evaluate the strength of the proposed augmentation approach against data irregularity due to the class imbalance problem. Our proposed augmentation approach introduced semantic diversity and increased the training data to overcome the data irregularity. Thus, the proposed approach is evaluated on MuHAVi-Uncut and iXMAS datasets because they constitute class imbalance samples. Our proposed video augmentation approach improved the generation of stronger predictive models by overcoming data irregularity.

3.1.2 Experimental setup

We employed Alexnet [22], VGG19 [18], ResNet [31], Google Net [32], and DenseNet201 [33] CNN models for feature extraction. The features extracted from the best performing layers are reported in the results. For simplicity, the deep spatiotemporal features extracted from the FC6 layers of Alexnet [22] and VGG19 [18], FC1000 layers of ResNet [31] and DenseNet201 [33], and AVPool5 layer of Google Net [32] are denoted as \( DA, DV, DR, DD, \) and \( DG \), respectively. These deep spatiotemporal features \( DA, DV, DR, DD, \) and \( DG \) were augmented with the proposed augmentation approach to introduce feature diversity. The ELM classifier for the MuHAVi-Uncut dataset contains 129 060 hidden nodes and that of iXMAS and IAVID-1 datasets contains 2100 hidden nodes [30].
3.1.3 Evaluation criteria

We evaluated our method using the leave one actor out (LOAO) and leave one camera out (LOCO) validation schemes [30]. The performance is measured by evaluating the computing accuracy (Ac) and the F score (F*) over the feature set with and without the proposed augmentation.

3.2 Evaluation of actor invariance (LOAO validation)

LOAO validation scheme allows all actors to evaluate the performance of HAR with the proposed augmentation technique using MuHAVi-Uncut, iXMAS, and IAVID-1 datasets. The averaged action prediction accuracy is reported in Table 1. To understand the impact of semantic diversity introduced using the proposed augmentation approach, we extracted deep spatiotemporal features from AlexNet, VGG, ResNet18, ResNet101, GoogleNet, and DenseNet, represented by $D_A$, $D_V$, $D_{R18}$, $D_R$, $D_G$, and $D_D$, and trained the feedforward ELM classifier with and without the proposed augmented approach for HAR. The confusion matrices in Figures 4 and 5 show the perclass prediction accuracies. Those on the left represent the perclass prediction rates without the proposed augmentation technique, whereas those on the right represent the perclass prediction rates with the proposed augmentation technique, using MuHAVi-Uncut and iXMAS datasets. Our proposed augmentation approach improves the action prediction rate in all classes, as illustrated by the confusion matrices.

From Table 1, although the augmented deep spatiotemporal features extracted from AlexNet showed the highest accuracy on all three datasets, most significant improvements were observed for the augmented deep spatiotemporal features extracted from VGG19.

| D          | MuHAVi-Uncut | iXMAS | IAVID-1 |
|------------|--------------|-------|---------|
|            | Without aug  | With aug | Without aug | With aug | Without aug | With aug |
|            | Ac   | F*   | Ac   | F*   | Ac   | F*   | Ac   | F*   |
| $D_A$      | 91.1 | 92.1 | 93.6 | 94.1 | 73.1 | 80.1 | 79.4 | 84.2 |
| $D_V$      | 88.0 | 89.1 | 91.8 | 92.6 | 69.4 | 70.9 | 74.2 | 74.9 |
| $D_R$      | 80.9 | 81.5 | 85.9 | 86.6 | 65.8 | 65.3 | 70.8 | 70.5 |
| $D_{R18}$  | 92.1 | 93.0 | 93.4 | 94.2 | 67.6 | 66.2 | 72.7 | 72.1 |
| $D_R$      | 84.4 | 85.1 | 88.5 | 89.6 | 70.7 | 70.4 | 75.4 | 76.2 |
| $D_D$      | 83.3 | 84.0 | 88.1 | 89.0 | 72.1 | 71.7 | 78.2 | 77.7 |

**Table 1** Evaluation of the proposed augmentation approach with leave one actor out (LOAO) validation scheme on MuHAVi-Uncut, iXMAS, and IAVID-1 for person invariant human action recognition (HAR)

![Confusion Matrix](image)

**Figure 4** Confusion matrix of MuHAVi-Uncut dataset for action recognition evaluated without and with the proposed augmentation approach for person-invariant human action recognition (HAR) (leave one actor out [LOAO]). The impact of the proposed augmentation approach is shown in the confusion matrix using the perclass action recognition rate. The confusion matrix on the left represents the HAR without the proposed augmentation approach and that on the right represents the HAR with the proposed augmentation approach.
DenseNet, and Googlenet where HAR performance increased by 5% to 7%, whereas, for augmented deep spatiotemporal features extracted from Alexnet, the performance gain was 2%. Overall, LOAO performance improved by 2% to 7% showing that the proposed augmentation approach improves the semantic diversity of the spatiotemporal feature representation of human actions. Furthermore, our proposed feature augmentation approach enhances the discriminative ability of features extracted from any CNN network, including sequential or directed acyclic graph architecture.

Moreover, the MuHAVi-Uncut, iXMAS, and IAVIS-1 datasets for HAR consist of imbalance action class samples that disturb the decision boundaries of the recognition model, as illustrated in Figure 3. To overcome data irregularity, GA was employed to increase the number of spatiotemporal feature instances for each class to balance the class samples and remove the skewness from the data. Our proposed augmentation approach evolved and combined deep spatiotemporal features and generated a new deep spatiotemporal representation of actions per class while keeping the same class labels. This implies that our proposed feature augmentation approach improves action classification and resolves data irregularity by balancing the class samples. Therefore, augmented features are effective in establishing strong decision boundaries and improved predictive models for HAR.

### 3.3 Evaluation of multiview HAR (LOCO validation)

To understand the impact of semantic diversity introduced by the proposed augmentation approach in a multiview setting, we conducted a LOCO validation scheme for HAR. The LOCO scheme examines the view invariance attribute of the proposed augmentation approach for HAR using eight viewpoints for MuHAVi-Uncut and five viewpoints for iXMAS video datasets. The deep spatiotemporal features \( D_A, D_V, D_{R15}, D_R, D_G \), and \( D_D \) were used for training the feedforward ELM classifier with and without the proposed augmented approach for LOCO scheme. From Table 2, accuracy and \( F \) score improved by 4% to 5% on the MuHAVi-Uncut dataset (accuracy improved from 77.78% to 81.76%) and 3% to 7% on iXMAS dataset (71.9% to 77.4%). The perclass prediction accuracies are represented in Figures 6 and 7. Our proposed augmentation approach improves the action prediction rate in all classes for view-invariant HAR, as shown by the confusion matrices.

Like, LOAO validation scheme, our proposed augmentation approach improves the prediction accuracy for multiview HAR using the LOCO validation scheme. Therefore, the feature space augmented using the proposed approach proved as a more robust set for view invariance HAR.

### 3.4 Comparison against video augmentation techniques

The extended training set was generated and compared against the state-of-the-art video augmentation methods for model learning (Table 3). The inputted training videos of IAVIS-1 were augmented with upsampling [13], adding perturbations to video frames (Gaussian and salt & pepper noise) [13], rotate & crop [34], inverting frame order [12], frame skipping [15], frame mirroring [14],...
TABLE 2 Evaluation of leave one camera out (LOCO) validation scheme on MuHAVi and IAVID-1 datasets for multiview human action recognition (HAR)

|            | MuHAVi-Uncut |            | iXMAS     |
|------------|--------------|------------|-----------|
|            | Without aug  | With aug   | Without aug | With aug   | Without aug | With aug   |
|            | Ac           | Ac         | F*          | F*          | Ac           | Ac         | F*          | F*          |
| DA         |             |            |             |             |             |            |             |             |
| D_A        | 77.78       | 81.76      | 76.80       | 80.98       | 79.3         | 83.8       | 80.1        | 84.2        |
| D_V        | 76.67       | 81.76      | 75.95       | 81.67       | 72.4         | 78.2       | 72.9        | 78.9        |
| D_D        | 75.39       | 80.23      | 75.16       | 80.11       | 68.8         | 73.8       | 69.3        | 73.8        |
| D_R18      | 72.65       | 77.85      | 72.43       | 77.24       | 69.6         | 73.7       | 68.2        | 74.4        |
| D_R        | 73.83       | 78.43      | 73.56       | 78.46       | 71.9         | 77.4       | 71.4        | 78.2        |
| D_D        | 76.22       | 80.81      | 76.05       | 80.28       | 74.1         | 81.2       | 73.7        | 82.0        |

FIGURE 6 Confusion matrix of MuHAVi-Uncut dataset for the action recognition scheme evaluated without and with the proposed augmentation approach for view-invariant human action recognition (HAR) (LOCO). The impact of the proposed augmentation approach is shown in the confusion matrix using the perclass action recognition rate. The confusion matrix on the left represents the HAR without the proposed augmentation approach and that on the right represents the HAR with the proposed augmentation approach.

FIGURE 7 Confusion matrix of iXMAS dataset for the action recognition scheme evaluated without and with the proposed augmentation approach for view-invariant human action recognition (HAR) (LOCO). The impact of the proposed augmentation approach is shown in the confusion matrix using the perclass action recognition rate. The confusion matrix on the left represents the HAR without the proposed augmentation approach and that on the right represents the HAR with the proposed augmentation approach.
and generating videos using adversarial networks [20]. The HAR method for all augmentation techniques remained the same for a fair comparison. We augmented the human action videos with state-of-the-art augmentation methods and generated motion templates, encoding the spatiotemporal representation of human actions and extracting deep spatiotemporal features $D_A$ to train the ELM classifier.

Our semantically diverse video augmentation technique improves the prediction accuracy by approximately 15% than frame skipping [15] and adding perturbations to video frames [13] on 70 to 30 training–testing splits, whereas 58.22% improvement is observed in contrast video augmented using the adversarial network [20]. Compared with Vondrick et al. [20], augmented videos generated using GANs have lower resolution with computational overhead, whereas frame skipping [15], mirroring [14], or inverting frame order [12] reflect existing training data; therefore, these methods generate lower data to improve the classification model. Table 3 shows the performance gain of our method from augmenting the feature space, inducing genetic diversity in feature set computed from AlexNet, and model relearning for performance optimization. From Table 3, our proposed method is computationally faster contrary to the state-of-the-art method due to single-pass feedforward learning of ELM regardless of backpropagation hyper-parameter tuning [35].

The analysis of the state-of-the-art video augmentation approaches using the MuHAVi-uncut dataset was not reported because the dataset consists of approximately 3000 action videos for 17 action classes. These video samples in each class cover almost all actions. IAVID-1 consists of 100 action samples for 8 action classes and required video augmentation to increase the training sample. Therefore, we evaluated all video augmentation approaches on IAVID-1 to compare the strength of the proposed video augmentation with standard video augmentation techniques.

### 3.5 Comparison against state-of-the-art HAR methods

To further evaluate the performance of the proposed augmentation approach, we implemented the techniques [5,27,33,34,36–40] on IAVID-1, MuHAVI-Uncut, and iXMAS datasets.

Table 4 provides the average prediction rate of the state-of-the-art HAR methods on MuHAVi [29], iXMAS, and IAVID-1 [30] datasets. In Howe and Deschamps [27], the aggregated actor silhouettes created motion templates, which were later used to learn deep action representations for training ELM. In Murtaza et al. [41], the actor silhouettes generated motion templates, which were used to generate Histogram of Oriented Gradients (HOG) representations of action to train the nearest neighbor classifier. Similarly, in [6], the C3D features were extracted from RGB videos and were used to train the support-vector machine (SVM) classifier. The CNN features with ELM [27] and SVM [6] classifiers yielded 78.89% and 48.77% on IAVID-1, 91% and 84.56% (LOAO), and 77.78% and 72.98% (LOCO) on MuHAVi-Uncut dataset. However, our proposed feature augmentation approach enhanced the CNN feature representation and improved the action prediction accuracy by 81.25% (2/3 split) on IAVID-1, 93.62% (LOAO), and 80.92% (LOCO) on MuHAVi-Uncut, and 83.3% (LOCO) on iXMAS dataset.

Our method improves prediction accuracies using virtual video generation by approximately 30% than MHI’s HOG [41] representation on LOCO and approximately 10% at LOAO on the MuHAVi dataset. Compared with C3D features using the SVM approach, an improvement of 32.66% was observed on IAVID-1 dataset [6].

Similarly, we extracted the spatiotemporal representation of action videos using I3D [42] and I3D with attention [43,44], for action recognition using iXMAS dataset. The RGB and optical-flow frames were provided at the input of the I3D model that convolved using unsymmetrical cubic filters to encode the feature representation, whereas for action recognition using I3D with attention, the empirical weight factor with I3D features was assigned to each RGB and optical-flow frames and was effective for action representation. However, the proposed video augmentation approach outperformed the I3D and I3D with attention, due to the genetic diversity introduced within the spatiotemporal features. Our technique outperforms state-of-the-

### Table 3: Performance comparison of proposed approach with state-of-the-art video augmentation techniques

| Technique                  | Acc | Time   |
|----------------------------|-----|--------|
| Proposed                   | 81.25 | 0.528 ms |
| Upsampling frame [13]      | 61.70 | 0.1013 s |
| Rotate & crop frame [34]   | 61.67 | 0.0255 s |
| Adding Gaussian noise [13] | 53.33 | 0.0166 s |
| Adding salt & pepper noise [13] | 65.87 | 0.0179 s |
| Inverse frame order [12]   | 54.39 | 0.0458 s |
| Frame skipping [15]       | 65.38 | 0.1219 s |
| Frame mirroring [14]      | 62.10 | 0.0133 s |
| Adversarial [20]          | 23.03 | 4 h   |
The semantic diversity of the evolved virtual video representation is beneficial for HAR. The benefits and performance of the proposed augmentation approach can be attributed to the stronger decision boundaries established by evolving deep spatiotemporal features and sustaining the elite virtual video representation. The proposed augmentation approach improves the baseline action prediction accuracy on MuHAVI-Uncut and IAVID-1 datasets.

### 3.6 Significance of virtual videos for HAR

For an efficient HAR model, a robust and discriminative feature representation of training data holds a significant role in learning because it ensures stronger decision boundaries, fast and cost-effective predictions, as validated in Tables 1 and 2. Therefore, robust feature representation of data is more beneficial than the physical existence of data for accurate predictive models, because they simplify interpretation by defining reliable predictive boundaries and enhance model generalization. Our proposed video augmentation technique provides significant improvement in model learning by inducing semantic diversity within the visual data. It augments human action video features to improve the decision boundaries of learning models for HAR. Contrary to standard augmentation techniques, our proposed approach augments the features, despite data irregularities.

Moreover, our feature augmentation approach extends the training samples in each class to balance the representation of each action sample and hence, improves the model learning. Thus, the proposed evolved video-set representation is responsible for efficient HAR through diverse deep spatiotemporal feature representations, regardless of the physical existence of video data.

### 4 Conclusion

In this paper, we proposed a video augmentation technique to improve the deep learning models for HAR. We showed that model optimization stems from inducing semantic diversity in the feature sets lacking using common data augmentation methods. Furthermore, we achieved semantic diversity by introducing virtual video generation using GA. Our proposed video augmentation scheme improved HAR accuracy by 2% to 7% on MuHAVi-Uncut, iXMAS, and IAVID-1 datasets. Also, the robust performance of the proposed approach was attributed to the following factors: First, augmenting feature space is better than augmenting the dataspace to counter data irregularities. Second, inducing semantic diversity overcomes classifier instability. Third, the model relearning using diverse feature sets generates stronger decision boundaries. Moreover, our augmentation technique can be induced for HAR using RGB data, and in other domains including object...
CONFLICT OF INTEREST
The authors declare no potential conflict of interests.

REFERENCES
1. H. Wang et al., Dense trajectories and motion boundary descriptors for action recognition, Int. J. Comput. Vis. 103 (2013), no. 1, 60–79.
2. H. Wang et al., Evaluation of local spatio-temporal features for action recognition, in Proc. British Mach. Vis. Conf. (BMVC), (London, UK), Sept. 2009, pp. 124.1–124.11.
3. A. F. Bobick and J. W. Davis, The recognition of human movement using temporal templates, IEEE Trans. Pattern Anal. Mach. Intell. 3 (2001), 257–267.
4. W. Zhu et al., Co-occurrence feature learning for skeleton based action recognition using regularized deep lstm networks, Proc. AAAI Conf. Artif. Intell. 30 (2016), no. 1, 3697–3703.
5. K. Simonyan and A. Zisserman, Two-stream convolutional networks for action recognition in videos, in Proc. Annu. Conf. Neural Inf. Process. Syst. (Montreal, Canada), Dec. 2014, pp. 568–576.
6. S. Ji et al., 3D convolutional neural networks for human action recognition, IEEE Trans. Pattern Anal. Mach. Intell. 35 (2012), no. 1, 221–231.
7. Z. Li et al., Violestm convolves, attends and flows for action recognition, Comput. Vis. Image Underst. 166 (2018), 41–50.
8. B. Singh et al., A multi-stream bi-directional recurrent neural network for fine-grained action detection, in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (Las Vegas, NV, USA), June 2016, pp. 1961–1970.
9. J. Marin et al., Learning appearance in virtual scenarios for pedestrian detection, in Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (San Francisco, CA, USA), June 2010, pp. 137–144.
10. G. Varol et al., Learning from synthetic humans, in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (Honolulu, HI, USA), July 2017, pp. 109–117.
11. D. Vazquez et al., Virtual and real world adaptation for pedestrian detection, IEEE Trans. Pattern Anal. Mach. Intell. 36 (2013), no. 4, 797–809.
12. M. Hoai and A. Zisserman, Improving human action recognition using score distribution and ranking, in Asian Conference on Computer Vision, vol. 9007, Springer, Cham, Switzerland, 2014, pp. 3–20.
AUTHOR BIOGRAPHIES

**Nudrat Nida** completed her PhD in Computer Engineering from UET, Taxila, Pakistan. She received her BSc and MSc degrees in Computer Engineering from University of Engineering and Technology, Lahore, Pakistan. She was awarded a fellowship for PhD from UET, Taxila, in 2016. Before joining UET, she was working in COMSAT University, as a lecturer and Alkhawarizmi Institute UET Lahore as a research associate. Since 2016, she has been working for Center for Computer Vision Research group, UET, Taxila. Currently, Dr Nudrat is an artificial intelligence consultant at R&D organization. Her research interests include video analytics, human action recognition, behavior analysis, object detection, and biomedical image analysis.

**Muhammad Haroon Yousaf** received his BSc, MSc, and PhD degrees in Computer Engineering from University of Engineering and Technology, Taxila, Pakistan and Director Swarm Robotics Lab under National Centre for Robotics and Automation (NCRA). His research interests lie in image processing/computer vision. He has published number of papers in international conferences and journals. He is also working as editor and reviewer for different journals. He has been the recipient of Best University Teacher Award (2012–2013) given by Higher Education Commission of Pakistan.

**Aun Irtaza** completed his PhD in 2016 from FAST-nu, Islamabad Pakistan. During his PhD, he remained working as a research scientist in the Gwangju Institute of Science and Technology (GIST), South Korea. He became an associate professor in 2017 and chair of the Department of Computer Science at the University of Engineering and Technology (UET), Taxila, Pakistan, in 2018. He also remained working as
a visiting associate professor in the University of Michigan-Dearborn in 2019–2020. His research areas include computer vision, multimedia forensics, audio-signal processing, medical image processing, and big data analytics. He has more than 40 publications in IEEE, Springer, and Elsevier Journals.

Sergio A. Velastin received the BSc and MSc (Research) degrees in Electronics and the PhD degree from the University of Manchester, Manchester, UK, in 1978, 1979, and 1982, respectively, for research on vision systems for pedestrian tracking and road-traffic analysis. He worked in industrial R&D before joining King College London (UK) in 1991 and then Kingston University London where he became director of its Digital Imaging Research Centre and full professor of applied computer vision. In 2015, he moved to the University Carlos III of Madrid, Spain, where he was a Marie Curie Professor. He is a visiting professor at Universidad Carlos III in Madrid and at Queen Mary University of London, UK.

How to cite this article: N. Nida, M. H. Yousaf, A. Irtaza, and S. A. Velastin, Video augmentation technique for human action recognition using genetic algorithm, ETRI Journal 44 (2022), 327–338. https://doi.org/10.4218/etrij.2019-0510