




Research Article

# Gradient local auto-correlation features for depth human action recognition



Mohammad Farhad Bulbul<sup>1</sup>  · Hazrat Ali<sup>2</sup> 

Received: 3 April 2020 / Accepted: 20 March 2021 / Published online: 7 April 2021

© The Author(s) 2021 

## Abstract

Human action classification is a dynamic research topic in computer vision and has applications in video surveillance, human–computer interaction, and sign-language recognition. This paper aims to present an approach for the categorization of depth video oriented human action. In the approach, the enhanced motion and static history images are computed and a set of 2D auto-correlation gradient feature vectors is obtained from them to describe an action. Kernel-based Extreme Learning Machine is used with the extracted features to distinguish the diverse action types promisingly. The proposed approach is thoroughly assessed for the action datasets namely MSRAction3D, DHA, and UTD-MHAD. The approach achieves an accuracy of 97.44% for MSRAction3D, 99.13% for DHA, and 88.37% for UTD-MHAD. The experimental results and analysis demonstrate that the classification performance of the proposed method is considerable and surpasses the state-of-the-art human action classification methods. Besides, from the complexity analysis of the approach, it is turn out that our method is consistent for the real-time operation with low computational complexity.

## Article Highlights

- The work proposes to process depth action videos through 3D Motion Trail Model (3DMTM) to represent the video as a set of 2D motion and motionless images.
- This work improves the above action representation by configuring all the 2D motion and motionless images as binary-coded images with the help of the Local Binary Pattern (LBP) algorithm.
- This work evaluates the use of auto-correlation features extracted on binary-coded versions of the 2D action representations instead of extracting these features from the non-binary-coded versions of the early action illustration.

**Keywords** 3D action classification · Depth action sequences · Action feature extraction · Auto-correlation features · Extreme learning machine

---

✉ Mohammad Farhad Bulbul, farhad@just.edu.bd; Hazrat Ali, hazratiali@cuiatd.edu.pk | <sup>1</sup>Department of Mathematics, Jashore University of Science and Technology, Jashore 7408, Bangladesh. <sup>2</sup>Department of Electrical and Computer Engineering, COMSATS University Islamabad, Abbottabad Campus, Abbottabad, Pakistan.



SN Applied Sciences (2021) 3:535 | <https://doi.org/10.1007/s42452-021-04528-1>

## 1 Introduction

A large number of researchers have been attracted to human action classification problem due to its wide range of real-world applications. The notable implementations cover visual surveillance [1], smart homes [2], sports [3], entertainment [4], healthcare monitoring [5], patient monitoring [6], elderly care [7], Virtual-Reality [8], human–computer interaction [9], and so on.

Human actions refer to distinctive sorts of activities including walking, jumping, waving, etc. However, the vivid variations in human body sizes, appearances, postures, motions, clothing, camera motions, viewing angles, illumination changes make the action recognition task very challenging. Over few years, a large number of researchers introduced several action or activity recognition model by using data sensors like RGB video cameras [10], depth video cameras [2], and wearable sensors [11]. Among these two video data sources, action recognition research based on conventional RGB cameras (e.g. [12]) has achieved great progress in the last few decades. However, utilization of RGB cameras for action recognition raises significant impediments such as lighting variations and cluttered background [13]. On the contrary, depth cameras generate depth images, which are insensitive to lighting variations and make background subtraction and segmentation easier. In addition, we can obtain body shape and structure characteristics and the human skeleton information from depth images.

Many previous attempts can be listed for efficient recognition systems such as DMM [14], HON4D [15], Super Normal Vector [16], Skeletons Lie group [17] and etc. But, those existing methods still face some crucial challenges such as depth video processing, appropriate feature extraction and reliable performance of classification model. Considering the aforementioned challenges, this study focuses to build an effective and efficient human action recognition framework on depth action video sequences. The main objective of this work is to enhance the classification accuracy by proposing an efficient recognition framework, which can overcome the above challenges more effectively. More specifically, the action video is illustrated through three 2D motion and three 2D static segments oriented images of the action. In fact, the dynamic and motionless maps are derived from the implementation of 3DMTM [18] on a video. However, the obtained representations are then enhanced with the help of LBP [19] tool. The tool enriches the action illustration by encoding the motion and motionless maps into binary pattern. Eventually, the outputs of the LBP are treated as input of GLAC [20] to generate the auto-correlation gradient vectors. In fact, there are three feature vectors for

the action motion segments images and another three feature vectors for the action static segments images. The first three vectors are concatenated to construct a motion information based GLAC vector. Similarly, another single GLAC vector is gained by incorporating the above mentioned last three vectors. For more boosting the proposed method, the aforementioned two single action representation vectors are concatenated for building the final action description. Finally, the action is recognized by passing the vector to a supervised learning algorithm named Extreme Learning Machine with Kernel (KELM) [21].

The main contributions of this paper are:

- We enhance the auto-correlation features for the optimal description of an action. Besides, to observe the significance of the feature augmentation, the action is also presented with the ordinary auto-correlation features. In fact, our action representation technique addresses the intra-class variation and inter-class similarity problem significantly.
- We report recognition results on three benchmark datasets namely; MSRAAction3D [22], DHA [23] and UTD-MHAD [24]. The recognition results are compared with state-of-the-art handcrafted as well as deep learning methods.
- We compare the recognition results based on the enhanced auto-correlation features compared to recognition results using the auto-correlation features only. These comparisons are made for the same data sets to fairly evaluate and elaborate the effectiveness of enhanced auto-correlation features.
- Finally, we report the computational efficiency in terms of the running and computational requirements.

Based on three publicly available data sets - MSRAAction3D [22], DHA [23] and UTD-MHAD [24], the proposed method is compared with handcrafted and deep learning methods extensively. The computational efficiency assessment indicates that the proposed approach offer feasibility for the real-time implementation. The working flow of the system is illustrated in Fig. 1.

This paper is organized as follows: in Sect. 2, we present some related literature review. Section 3 describes research methodology. The results of experimental and discussions are presented in Sect. 4. Finally, Sect. 5 concludes the work.

## 2 Related work

Feature extraction is a key step in Computer Vision research problems like object localization, human gait recognition, face recognition, action recognition, text

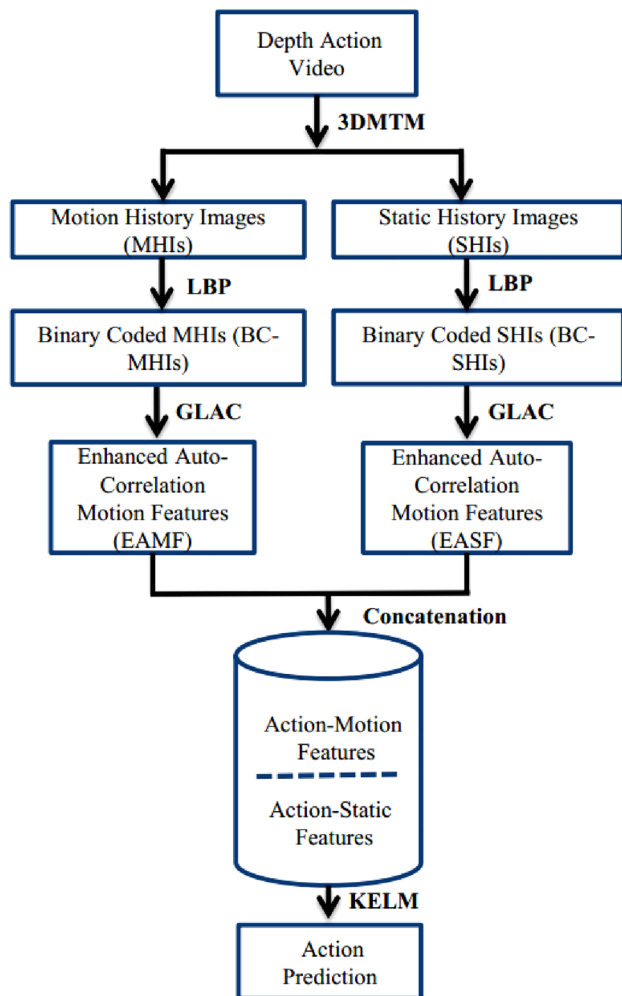


Fig. 1 Workflow illustration of our method

recognition and etc. As a result, researchers have given more attention to extract features effectively. For example, for object recognition, Ahmed et al. [25] introduced a Saliency map on RGB-D indoor data which had numerous applications such as vehicle monitoring system, violence detection, driverless driving system, etc. Hough voting and distinct features were used to measure the efficiency of that work. To explore silhouettes of humans from noisy backgrounds, Jalal et al. [26] applied embedded HMM for activity classifications where spatiotemporal body joints and depth silhouettes were fused to improve accuracy. In another work, to recognize online human action and activity, Jalal et al. [27] performed multi-features fusion along with skeleton joints and shape features of humans. For feature extractions in activity recognition, Tahir et al. [28] applied 1-D LBP and 1-D Hadamard wavelet transform along with Random Forest. On depth video sequences, Kamal et al. [29] utilized modified HMM to complete another fusion process of temporal joint features and

spatial depth shape features. On the other hand, to recognize facial expression Rizwan et al. [30] implemented local transform features where HOG and LBP were used for feature extraction. Again, skin joint features by using skin color and self-organized maps were used for activity recognition [31]. In another work, Kamal et al. [32] employed distance parameters features and motion features. Yaacob et al. [33] introduced a discrete cosine transform, particularly for gait action recognition.

In developing vision based handcrafted action recognition, researchers have also done struggle in feature extraction for optimal action representation. The motion features of an action through the simplified depth motion maps were extracted by works in DMM [14], DMM-CT-HOG [34], DLE [35]. The texture features extracted by LBP were utilized in [36]. Recently, Dhiman et al. [37] introduced Zernike moments and R-transform to create a powerful feature vector for abnormal action detection. A genetic algorithms based system was proposed by Chaaroui et al. [38] to improve the efficiency of the skeleton joint-based recognition system by optimizing skeleton-joint subset. Vemulapalli et al. [17] represented human actions as curves that contain skeletal action sequences. Gao et al. [39] proposed a model to recognize 3D actions where they constructed a difference motion history image for RGB and depth sequences. Then, they captured motions through multi-perspective projections. Next, they extracted the pyramid histogram of oriented gradients. Finally, human action was identified by combining multi-perspective and multi-modality discriminated and joint representation. In the work by Rahmani et al [40], the features obtained from depth images were combined with skeleton movements encoded by difference histogram and finally a Random Decision Forest (RDF) was applied to obtain discriminant features for action classifications. On the other hand, Luo et al. [41] represented features by sparse coding-based temporal pyramid matching approach (ScTPM). They also proposed a capturing technique for spatio-temporal features from RGB videos called Center-Symmetric Motion Local Ternary Pattern (CS-Mltp). Finally, they explored the feature-level fusion and classifier-level fusion applying the above mentioned features to improve recognition accuracy. Again, decisive pose features were used by imposing another two distinct transformations called Ridgelet and Gabor Wavelet Transform to detect human action [42]. Moreover, Wang et al. [43] studied ten Kinect-based methods for cross-view and cross-subject action recognition on six dissimilar datasets and finally concluded that skeletal-based recognition is superior to other processes for cross-view.

Deep learning models usually learn features automatically from raw depth sequences, which are then useful to compute high level semantic representations. For

example, 2D-CNN and 3D-CNN were employed by Yang and Yang to address the deep learning based depth action classification [44]. Wang et al. [45] used to improve the action representation, unlike DMM, Wang et al. proposed Weighted Hierarchical Depth Motion Maps (WHDMM). The WHDMM was fed into CNN along three CNN streams to recognize actions. In another concept, before passing to CNN, the depth video was described by Dynamic Depth Images (DDI), Dynamic Depth Normal Images (DDNI) and Dynamic Depth Motion Normal Images (DDMNI) [46]. In [47], a novel notion in action classification is introduced by using the RGB domain features as depth domain features by domain adaptation. Motion History Images (MHI) from RGB videos and DMM of depth videos are utilized together to generate a four-stream CNN architectures [48]. By using inertial sensor data and depth data, Ahmad et al. [11] expressed a multimodal  $M^2$  fusion process with the help of CNN and multi-class SVM. Very recently, Dhiman et al. [49] have merged shape and motion temporal dynamics by proposing a deep view-invariant human action system. To detect the human gesture and 3D action, Weng et al. [50] proposed pose traversal convolution Network which applied joint pattern features from the human body. They also represented human gesture and action as a sequence of 3D poses. A self-supervised alignment method was used for unsupervised domain adaptation (UDA) [51] to recognize human action. Busto et al. [52] expressed another concept for action recognition and image classification called open set domain adaptation which works for unsupervised and semi-supervised domain adaptation model.

### 3 Proposed system

Our proposed framework consists of feature extraction, action representation and action classification. In this section, we discuss the three parts respectively. Figure 2 shows the pipeline of the system.

#### 3.1 Feature extraction

For each action video, three motion and three static information images are firstly computed by applying the 3DMTM [18] on the video. The 3DMTM yields the set  $MHI_{XOY}, MHI_{YOZ}, MHI_{XOZ}$  of motion images and the set  $SHI_{XOY}, SHI_{YOZ}, SHI_{XOZ}$  of static images by simultaneously stacking all the moving and stationary body parts (along the front, side and top projection views) of an actor in a depth map sequence.

Now, the MHIs and SHIs are converted to the binary coded form by the LBP [19]. In fact, the later versions are more enhanced than the earlier version of those images. Figure 3 shows an MHI and the corresponding  $BC - MHI$  is

represented by Fig. 4. It is clear that the motion information of the action is improved in the  $BC - MHI$ .

The binary coded motion images ( $BC - MHIs$ ) are referred to as  $BC - MHI_{XOY}, BC - MHI_{YOZ}$  and  $BC - MHI_{XOZ}$  on three Euclidean faces whereas the binary coded static images on those faces are denoted as  $BC - SHI_{XOY}, BC - SHI_{YOZ}$  and  $BC - SHI_{XOZ}$ . The binary coded images thus obtained, are fed into the GLAC [20] descriptor to extract spatial and orientational auto-correlations for illustrating action. This paper extracts the 0th order and 1st order auto-correlation features to describe an action. In fact, the auto-correlation features are used to describe an action through the rich texture information from images. The texture information includes the image gradients and curvatures simultaneously. Overall, the auto-correlation features are more dominant over the standard histogram oriented features. Thus, we consider the auto-correlation features in our approach. For a spontaneous discussion of the GLAC utilization on  $BC - MHIs / BC - SHIs$ , let  $I$  be a binary coded motion/static image (i.e.,  $BC - MHI / BC - SHI$ ). For each pixel of  $I$ , we use image gradient operators to obtain a gradient vector. The magnitude and the orientation of the gradient vector are computed as follows:

$$m = \begin{cases} \sqrt{\left(\frac{\partial I}{\partial x}^2 + \frac{\partial I}{\partial y}^2\right)}, & \text{if } I = I(x, y) \\ \sqrt{\left(\frac{\partial I}{\partial y}^2 + \frac{\partial I}{\partial z}^2\right)}, & \text{if } I = I(y, z) , \\ \sqrt{\left(\frac{\partial I}{\partial x}^2 + \frac{\partial I}{\partial z}^2\right)}, & \text{if } I = I(x, z) \end{cases} \quad (1)$$

$$\theta = \begin{cases} \arctan\left(\frac{\partial I}{\partial x}, \frac{\partial I}{\partial y}\right), & \text{if } I = I(x, y) \\ \arctan\left(\frac{\partial I}{\partial y}, \frac{\partial I}{\partial z}\right), & \text{if } I = I(y, z) , \\ \arctan\left(\frac{\partial I}{\partial x}, \frac{\partial I}{\partial z}\right), & \text{if } I = I(x, z) \end{cases} \quad (2)$$

The above orientation  $\theta$  can be coded into D orientation bins by voting weights to the nearest bins to form a sparse gradient orientation vector  $\mathbf{g} \in \mathbb{R}^D$ .

Through the gradient orientation vector  $\mathbf{g}$  and the gradient magnitude  $m$ , the  $K$ th order auto-correlation function of local gradients can be written as

$$F(d_0, \dots, d_K, \mathbf{b}_1, \dots, \mathbf{b}_K) = \int w[(m(\mathbf{r}), m(\mathbf{r} + \mathbf{b}_1), \dots, m(\mathbf{r} + \mathbf{b}_K))] g_{d_0}(\mathbf{r}) g_{d_1}(\mathbf{r} + \mathbf{b}_1) \dots g_{d_K}(\mathbf{r} + \mathbf{b}_K) d\mathbf{r}, \quad (3)$$

where  $\mathbf{b}_1, \mathbf{b}_2, \dots, \mathbf{b}_K$  are the shifting vectors from the position vector  $\mathbf{r}$  (indicates the position of each pixel in image

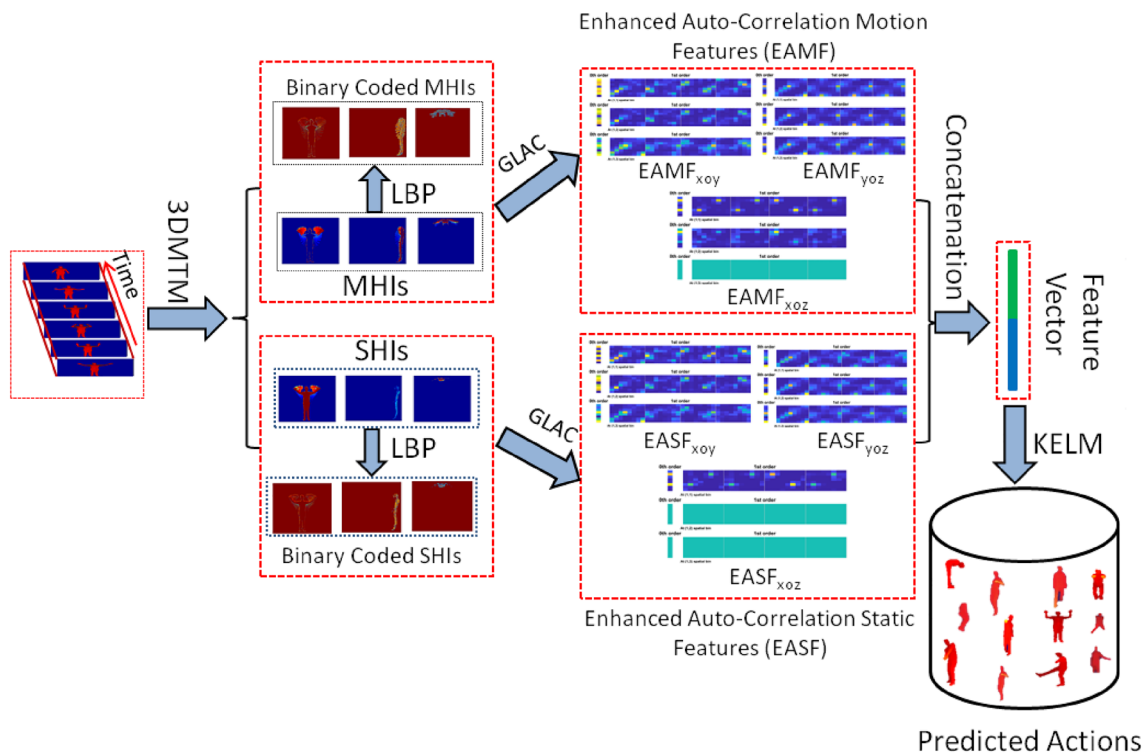


Fig. 2 Architecture visualization of our proposed framework

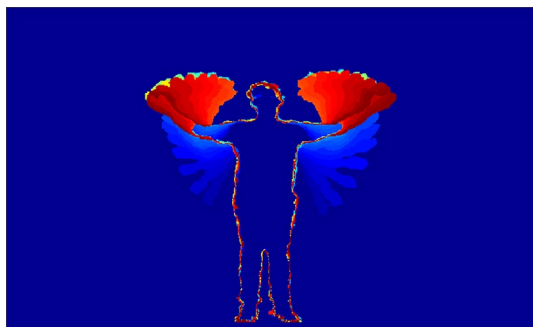


Fig. 3 Motion history image of two hand wave action

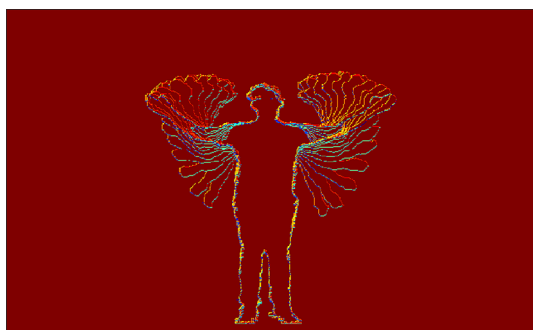
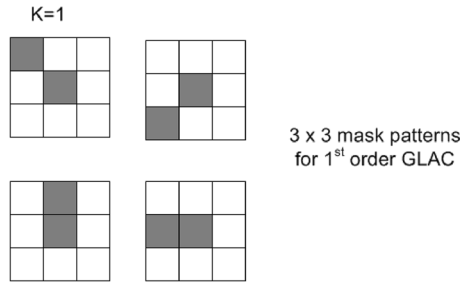


Fig. 4 Enhanced motion history image of two hand wave action

$l$ ).  $g_d$  indicates the  $d$ th element of  $\mathbf{g}$  and  $w(\cdot)$  is a weighting function of  $m$  functions. Indeed, the function  $w(\cdot)$  is used as the auto-correlation's weights. All the shifting vectors are restricted to local neighbours since the local neighbouring gradients might be immensely correlated. However, two types of correlations among gradients are obtained from Eq. (3): *spatial* gradient correlations gained with the vectors  $\mathbf{b}_i$  and *orientational* gradient correlations attained through the multiplications of the values  $g_{d_i}$ . By changing the values of  $K$ ,  $\mathbf{b}_i$ , and the weight  $w$ ; Eq. (3) may take various forms. The lower values of  $K$  assists to capture lower order auto-correlation features, which are rich geometric characteristics together with the shifting vectors  $\mathbf{b}_i$ . Because of image isotropic characteristic, the shifting intervals are kept identical along the horizontal and vertical directions. For  $w(\cdot)$ , the min is accepted for suppressing the impact of isolated noise around auto-correlations.

According to the suggestion by [20],  $K \in \{0, 1\}$ ,  $b_{1 \times y} \in \{\pm \Delta \mathbf{r}, 0\}$  and  $w(\cdot) \equiv \min(\cdot)$  are considered in this paper. The  $\Delta \mathbf{r}$  is the displacement interval in both horizontal and vertical directions. The interval is same for both directions due to the isotropic property of the image. Now, from Eq. (3), for  $K \in \{0, 1\}$  the 0th order ( $\mathbf{F}_0$ ) and the 1st order ( $\mathbf{F}_1$ ) GLAC features are as follows:

$$1^{th} \text{ order GLAC: } F_1 = \sum_{r \in I} \min[m(r), m(r + b_1)] g_{d_0}(r) g_{d_1}(r + b_1)$$



$$0^{th} \text{ order GLAC: } F_0 = \sum_{r \in I} m(r) g_{d_0}(r)$$

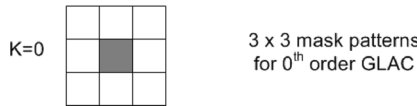


Fig. 5 Mask patterns for the 0th and 1st order auto-correlation

$$F_0 : F_{K=0}(d_0) = \sum_{r \in I} m(r) g_{d_0}(r) \tag{4}$$

$$F_1 : F_{K=1}(d_0, d_1, b_1) = \sum_{r \in I} \min [m(r), m(r + b_1)] g_{d_0}(r) g_{d_1}(r + b_1) \tag{5}$$

A single mask pattern can be used for Eq. (4), and there are four independent mask patterns for Eq. (5) for computing the auto-correlations. The mask/spatial auto-correlation patterns of  $(r, r + b_1)$  are depicted with Fig. 5). Since there is a single mask pattern for  $F_0$  and four mask patterns for  $F_1$  then the dimensionality of the above GLAC features ( $F_0$  and  $F_1$ ) is  $D + 4D^2$ . Although the dimensionality of the GLAC features is high, the computational cost is low due to the sparseness of  $g$ . It is worth noting that the computational cost is invariant to the number of bins,  $D$ , since the sparseness of  $g$  does not depend on  $D$ .

Figure 7 shows an example of 0th and 1st order GLAC features with 8 orientation bins (bins are shown in Fig. 6). Based on texture features, an action with motion images can be described as a vector  $EAMF = [EAMF_{XOY}, EAMF_{YOZ}, EAMF_{XOZ}]$ , where  $EAMF_{XOY}, EAMF_{YOZ}$  and  $EAMF_{XOZ}$  are vectors, which are obtained by conveying the set of binary coded motion information images to the 2D GLAC. In order to represent the static image action based on texture features, the vector  $EASF = [EASF_{XOY}, EASF_{YOZ}, EASF_{XOZ}]$  is obtained by linking the enhanced auto-correlation feature vectors extracted on multi-view static images. The  $EAMF$  is complementary to the  $EASF$ , therefore we fuse these two

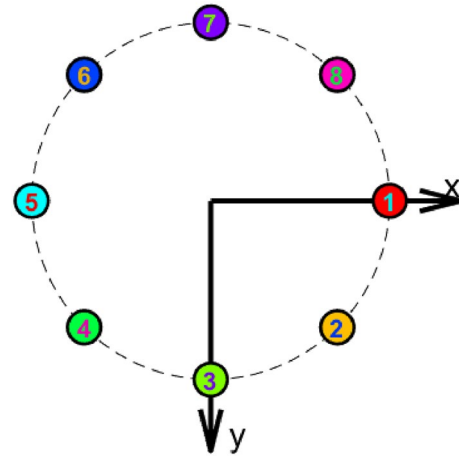


Fig. 6 Example of orientation bins in auto-correlation computation

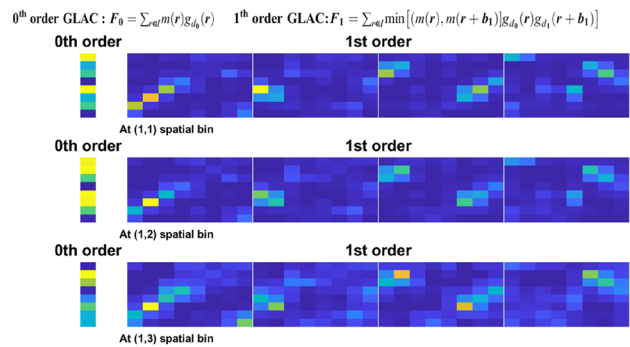


Fig. 7 Example of 0th and 1st order GLAC features

vectors in to a single vector to get optimal representation of an action. In our work (for all experiments), the dimension of the single feature vector is of 4752. It is flexible to compute the feature vector due to the sparse vector  $g$ . The work in [20] provides more detail on GLAC (Fig. 7).

### 3.2 Action classification

To gain the promising classification outcome, the fused version of  $EAMF$  and  $EASF$  is passed to Kernel based Extreme Learning Machine (KELM). The classification algorithm is discussed in detail in this section. The KELM [21] is an enhancement of Extreme Learning Machine (ELM) classifier [53]. By associating a suitable kernel with ELM, the KELM is derived to improve the discriminatory power of the classification algorithm. The Radial Basis Function (RBF) kernel is employed in our work. For an intuitive illustration, the classifier is described as a single algorithm in Algorithm 1.

**Algorithm 1** Algorithm of Kernel-based Extreme Learning Machine

**Input:** The training feature set  $\{\mathbf{x}_j, \mathbf{y}_j\}_{j=1}^m$ , ( $m$  is a number of training samples,  $y_i \in \{0, 1\}$  are class labels with  $i \in \{1, 2, \dots, C\}$ ,  $\mathbf{x}_j \in \mathbb{R}^{D'}$ ,  $\mathbf{y}_j \in \mathbb{R}^C$ ), and a test sample  $c$ .

**Steps:**

1. Construct a Feed-forward Neural Network (FFNN) as  $h_N(\mathbf{x}_j) = \sum_{k=1}^N \alpha_k f(\mathbf{w}_k \cdot \mathbf{x}_j + e_k) = \mathbf{y}_j$

2. Write **Step-1** in compact form as  $\mathbf{F}\alpha = \mathbf{Y}$ , where  $\alpha = [\alpha_1^T, \dots, \alpha_m^T]^T \in \mathbb{R}^{N \times C}$ ,  $\mathbf{Y} = [\mathbf{y}_1^T, \dots, \mathbf{y}_m^T]^T \in \mathbb{R}^{m \times C}$  and  $\mathbf{F} = \begin{bmatrix} f(\mathbf{w}_1 \cdot \mathbf{x}_1 + e_1) & \dots & f(\mathbf{w}_N \cdot \mathbf{x}_1 + e_N) \\ \vdots & \ddots & \vdots \\ f(\mathbf{w}_1 \cdot \mathbf{x}_m + e_1) & \dots & f(\mathbf{w}_N \cdot \mathbf{x}_m + e_N) \end{bmatrix}$ .

3. Solve  $\alpha$  with regularization coefficient ( $\rho > 0$ ) as  $\alpha = \mathbf{F}^T(\frac{\mathbf{I}}{\rho} + \mathbf{F}\mathbf{F}^T)^{-1}\mathbf{Y}$ ,

4. Update **Step-1** as:

$$h_N(\mathbf{x}_j) = \mathbf{f}(\mathbf{x}_j)\alpha = \mathbf{f}(\mathbf{x}_j)\mathbf{F}^T(\frac{\mathbf{I}}{\rho} + \mathbf{F}\mathbf{F}^T)^{-1}\mathbf{Y}.$$

5. Update **Step-4** by replacing  $\mathbf{F}\mathbf{F}^T$  with a kernel matrix  $\Omega_{ELM} : \Omega_{ELM_{j,s}} = \mathbf{f}(\mathbf{x}_j) \cdot \mathbf{f}(\mathbf{x}_s) = K(\mathbf{x}_j, \mathbf{x}_s)$  as

$$h_N(\mathbf{x}_j) = \begin{bmatrix} K(\mathbf{x}_j, \mathbf{x}_1) \\ \vdots \\ K(\mathbf{x}_j, \mathbf{x}_m) \end{bmatrix}^T \left(\frac{\mathbf{I}}{\rho} + \Omega_{ELM}\right)^{-1} \mathbf{Y}.$$

**Output:** The class label of  $c$ : the index of the output nodes  $h_N(\mathbf{x}_j)$  with the largest value.

**4 Experimental results and discussion**

We evaluate the proposed framework on MSRAction3D [22], DHA [23] and UTD-MHAD [24] datasets. Example depth images of each dataset are illustrated in Fig. 8. From the figure, it is clear that these datasets are ready for direct use without any segmentation process. Like other methods in [22–24], we straightforward input the depth map sequences in the proposed system without employing a preprocessing algorithm on the sequences.

**4.1 Experiments on MSRAction3D dataset**

MSRAction3D dataset [22] consists of 20 actions delivered by 10 diverse actors. The dataset includes inter-class similarity in different types of actions. The action examples generated by odd label oriented actors are utilized for model training and even label oriented samples are employed for the model testing [22]. The KELM uses

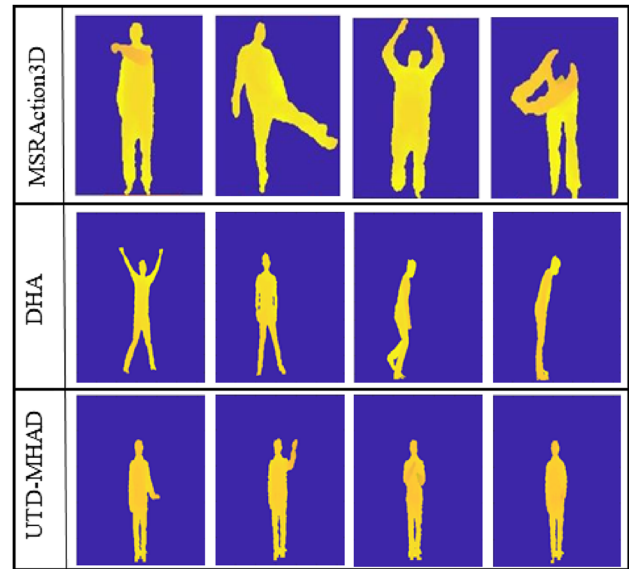


Fig. 8 Action snaps of MSRAction3D, DHA and UTD-MHAD dataset

$C = 10^4$  and  $\gamma = 0.03$  for training the classification model as optimal parameter values which are determined by 5-fold cross validation.

Table 1 reports a notable accuracy of 97.44% of our approach. The table indicates that the proposed approach is able to achieve better classification accuracy over other existing methods considerably. It can be seen that our method overwhelms deep learning system described in [44] by 6.34% and by 11.34% (see Table 1). To clarify the effectiveness of the feature enhancement, the system based on only the auto-correlation features is also evaluated on this dataset. The enhanced auto-correlation feature based system improves the recognition accuracy by 5.5% over the system without feature enhancement on the same setup and parameters. Figure 9 shows the confusion matrix corresponding to the accurate and incorrect classification rates. Through Table 2, the failure cases of the approach are listed. The table shows that the “horizontal wave” is confused with “hammer” by 8.3%; “draw x” is confusion with “high wave” by 7.14% and “draw circle” by 21.43%. Also action named “draw tick” is confused with “draw x” by 16.67%. Overall, among 20 actions, 17 actions are classified correctly (i.e., 100% classification accuracy) and rest 3 actions are classified incorrectly being confused with other actions.

**4.2 Experiments on DHA dataset**

DHA dataset proposed by Lin et al. [23] with 23 action categories captured by 21 individuals. Due to having inter-class similarity of different types of action categories, such

**Table 1** Performance of our method compared to the existing systems in terms of the MSRAction3D dataset

Approach	Classification accuracy (%)
DCSF [54]	89.3
HON4D [15]	88.9
Super Normal Vector [16]	93.1
Skeletons Lie group [17]	89.5
DMM-LBP-DF [55]	93.0
2D-CNN on DMM-Pyramid [44]	91.1
3D-CNN on DMM-Cube [44]	86.1
HOG3D + LLC [56]	90.9
Hierarchical 3D Kernel [57]	92.7
GLAC on DMM [13]	89.4
DMM-GLAC-STACOG [13]	94.8
3DHoT + MBC [58]	95.2
Subspace encoding [59]	94.06
LSTM + trust gates [60]	94.8
Extended SNV [61]	93.5
Trust Gates [62]	94.8
ST-NBNN [63]	94.8
SSTKDes [64]	95.6
3D-CNN + DHI + relief + SVM [65]	92.8
WDMM + HOG [66]	91.9
WDMM + LBP [66]	91.6
WDMM + CNN [66]	90.0
Deep activations [67]	92.3
Deep activations + attributes [67]	93.4
Hierarchical Gaussian [68]	95.6
GMHI + GSHI + CRC [69]	94.5
MHF + SHF + KELM [36]	95.97
Spatiotemporal + HMM [70]	92.4
Only auto-correlation features	91.94
<b>Proposed method (enhanced auto-correlation features)</b>	<b>97.44</b>

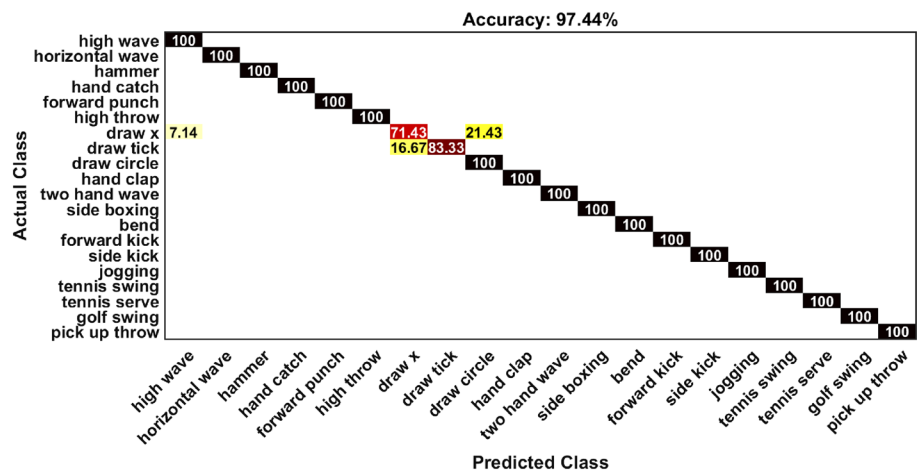
as *golf-swing* and *rod-swing* the dataset looks complex. The training and the test instances are separated according to the technique as discussed in the previous dataset [23]. The classification parameters  $C = 10^4$  and  $\gamma = 0.06$  are decided through the 5-fold cross validation technique.

Our approach gains a significant classification accuracy of %99.13 on this dataset. It can be seen from Table 3, our approach outperforms [23] by 12.33%, [55] by 7.83%, [71] by 3.69%, [58] by 2.44%, [39] by 6.73% and [39] by 4.13%. For this dataset, the enhanced auto-correlation method overcomes the auto-correlation method by an accuracy of 2.17%. The confusion matrix of the dataset is shown in Fig. 10. Furthermore a table is figured out to show the class based confusion information. Table 4 clarifies that the “skip” and “side-clap” are misclassified with low confusion rates and other 21 actions are classified with 100% accuracy. The wrong classification occurs when “skip” is confused with “jump” by 9.09% and “side-clap” is confused with “side-box” by 9.09%.

### 4.3 Experiments on UTD-MHAD dataset

The *UTD-MHAD* [24] is an action database constructed by a Microsoft Kinect camera and a wearable inertial sensor. In this dataset, 27 different actions are included and each action is performed four times by four females and four males. There are 861 depth sequences after removing 3 corrupted sequences. The 1st to 21st actions are obtained by positioning inertial sensor on the right wrist of performer, and the remaining actions are captured by inertial sensor placed on the subject’s right thigh. A comprehensive set of human actions is contained in the dataset, such as sport actions, daily activities, and training exercises. more detail on the dataset can be found in [24]. The entire database is split into training and test databases following the manner as in the last two databases [24]. The classification parameters  $C$  and  $\gamma$  are set to  $(10^4$  and 0.06 for promising recognition outcomes.

**Fig. 9** Confusion matrix obtained for the MSRAction3D dataset





**Table 2** Class oriented confusion on MSRAction3D dataset

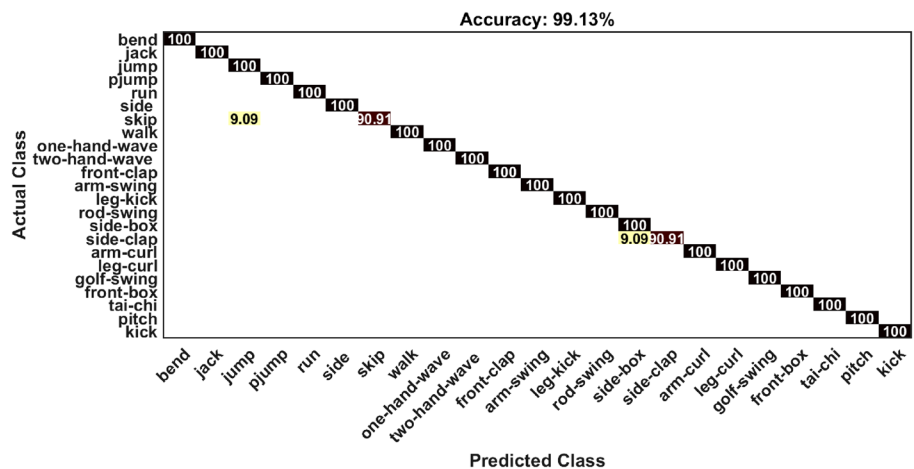
Action	Confusion (%)
Horizontal wave	Hammer (8.3)
Draw x	High wave (7.14), draw circle (21.43)
Draw tick	Draw x (16.67)

**Table 3** Performance of our method compared to the existing systems in terms of the DHA dataset

Approach	Classification accuracy (%)
D-STV/AS [23]	86.8
D-DMHI-PHOG [39]	92.4
DMPP-PHOG [39]	95.0
DMM-LBP-DF [55]	91.3
Multi-temporal DMM [71]	95.44
3DHoT + MBC [58]	96.69
Hierarchical Gaussian [68]	97.96
MHF + SHF + MSVM [36]	96.09
MHF + SHF + KELM [36]	98.26
Only auto-correlation features	96.96
<b>Proposed method (enhanced auto-correlation features)</b>	<b>99.13</b>

Experimental evaluation of our approach on UTD-MHAD dataset is represented by Table 5. The approach is able to acquire 88.37% overall classification accuracy on the dataset. The comparison of our method with other existing methods is also shown in the table. Our method outperforms [24] (Kinect) by 22.27%, [24] (Kinect+Inertial) by 9.27%, [58] by 3.97%, [72] by 2.57% and [68] by 6.87%. The enhanced auto-correlation system overwhelms the auto-correlation system by 0.93%. The confusion matrix

**Fig. 10** Confusion matrix obtained for the DHA dataset



**Table 4** Class oriented confusion on DHA dataset

Action	Confusion (%)
Skip	Jump (9.09)
Side-clap	Side-box (9.09)

is shown by Fig. 11. The figure describes, the approach has misclassified action classes although the overall classification rate for this dataset is of 88.37%. Due to inter-class similarity, 16 action classes show confusion among 27 action classes. Table 6 is extracted from the confusion matrix to furthermore analyze the experimental results. The table mentions that the action swipe-right is confused with the action “swipe-left”, and the confusion/misclassification rate is 15.79%. For the “wave” action this rate is 20.0% for having confusion with the action “draw-circle-CW”. Similarly, the confusion rate for the action class “clap” and “wave” is of 20.0% and for “wave” and “clap” is of 15.79%. Also, all the samples of the action classes “basketball shoot”, “draw-x”, “draw-circle-CW”, “draw-circle-CCW” “draw-triangle”, “baseball-swing”, “pus”, “knock”, “catch”, “jog”, “stand2sit” and “lunge” are not classified perfectly. Those misclassified samples are confused with samples of similar body postures of subjects.

### 4.4 System competency

The computational time and the complexity of key factors are considered to examine the system’s efficiency .

#### 4.4.1 Computational time

The system is evaluated on a Desktop whose configuration includes an Intel i5-7500 Quad-core processor of 3.41 GHz frequency and a 16 GB RAM. There are six major components in the system-i.e., MHI/SHI construction, binary coded MHI generation, binary coded SHI generation, EAMF

**Table 5** Performance of our method compared to the existing systems in terms of the UTD-MHAD dataset

Approach	Classification accuracy (%)
Microsoft kinect [24]	66.1
Wearable inertial [24]	67.2
Microsoft kinect + wearable inertial [24]	79.1
3DHoT + MBC [58]	84.4
Joint trajectory + CNN [72]	85.8
Hierarchical Gaussian [68]	81.5
MHF + SHF + MSVM [36]	83.26
Only auto-correlation features	87.44
<b>Proposed method (enhanced auto-correlation features)</b>	<b>88.37</b>

generation, EASF generation, and KELM classification. The operation time of those components are figured out to measure the time competency of the recognition system. The computational time (in millisecond) per action sample (with 40 depth frames on average) for all the components is represented in Table 7. Note that the system needs less than one second (i.e.,  $731.43 \pm 48.83$  ms) to process 40 depth frames. Consequently, it can be claimed that our recognition method can be utilized as real-time recognition system.

#### 4.4.2 Computational complexity

In fact, the PCA and the KELM are the key components in the computational complexity calculation of the introduced system. The PCA has complexity of  $O(m^3 + m^2r)$  [14] and the KELM has complexity of  $O(r^3)$  [73]. As a result, the complexity of the system can be revealed as  $O(m^3 + m^2r) + O(r^3)$ . Table 8 represents the calculated

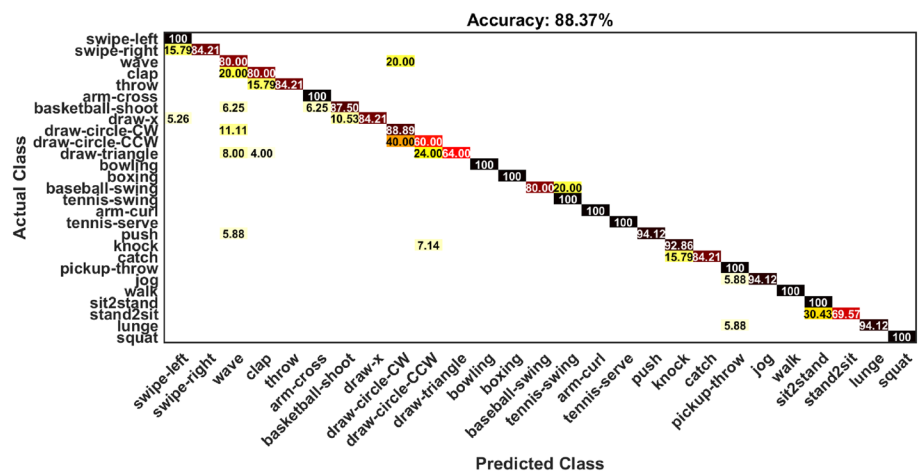
**Table 6** Class oriented confusion on UTD-MHAD dataset

Action	Confusion (%)
Swipe-right	Swipe-left (15.79)
Wave	Draw-circle-CW (20.00)
Clap	Wave (20.00)
Throw	Clap (15.79)
Basketball-shoot	Wave (6.25), arm-cross(6.25)
Draw-x	Swipe-left (5.26), basketball-shoot (10.53)
Draw-circle-CW	Wave (11.11)
Draw-circle-CCW	Draw-circle-CW (40.00)
Draw-triangle	Wave (8.00), clap (4.00), draw-circle-CCW (24.00)
Baseball-swing	Tennis-swing (20.00)
Push	Wave(5.88)
Knock	Draw-circle-CCW (7.14)
Catch	Knock (15.79)
Jog	Pickup-throw (5.88)
Stand2sit	Sit2stand (30.43)
Lunge	Pickup-throw (5.88)

**Table 7** Computational time (mean  $\pm$  std) of the key factors of the system

Key factors	Computational time (ms)/ action sample (40 frames)
3DMTM based MHI/SHI generation	606.2 $\pm$ 40.2
Binary coded MHI	50.4 $\pm$ 2.9
Binary coded SHI	50.8 $\pm$ 2.7
GLAC on binary coded MHI	12.2 $\pm$ 1.4
GLAC binary coded SHI	10.7 $\pm$ 0.8
KELM ensemble	1.1 $\pm$ 0.8
<b>Total running time</b>	<b>731.4 <math>\pm</math> 48.8</b>

**Fig. 11** Confusion matrix obtained for the UTD-MHAD dataset



**Table 8** Computational complexity comparison of the proposed approach with other existing approaches

Method	Components	Complexity	Total complexity
DMM [14]	Principal component analysis (PCA), l2-regularized collaborative representation classifier (l <sub>2</sub> -CRC)	$O(m^3 + m^2r)$ , $O(n_c \times r)$ m= feature vector dimension, r = training instance number, $n_c$ = class number	$O(m^3 + m^2r) + O(n_c \times r)$
DMM-LBP-DF [55]	Principal component analysis (PCA), kernel based extreme learning machine (KELM)	$O(m^3 + m^2r)$ , $O(r^3)$ m= feature vector dimension, r = training instance number	$O(m^3 + m^2r) + 3 * O(r^3)$
MHF + SHF + KELM [36]	Principal component analysis (PCA), kernel based extreme learning machine (KELM)	$O(m^3 + m^2r)$ , $O(r^3)$ m= feature vector dimension, r = training instance number	$O(m^3 + m^2r) + 2 * O(r^3)$
GMSHI + GSHI + CRC [69]	Principal component analysis (PCA), l2-regularized collaborative representation classifier	$O(m^3 + m^2r)$ , $O(n_c \times r)$ m= feature vector dimension, r = training instance number, $n_c$ = class number	$O(m^3 + m^2r) + O(n_c \times r)$
<b>Proposed method</b>	Principal component analysis (PCA), kernel based extreme learning machine (KELM) ensemble	$O(m^3 + m^2r)$ , $O(r^3)$ m= feature vector dimension, r = training instance number	$O(m^3 + m^2r) + O(r^3)$

complexity and compares with complexities of other existing methods. It can be seen that our method has lower computational complexity than other methods listed in the table. our method is also superior over them from the recognition perspective. Thus, our approach is superior in terms of recognition accuracy as well as computational efficiency.

## 5 Conclusion

This paper has introduced an efficacious and efficient human action framework based on enhanced auto-correlation features. The system uses the 3DMTM to derive three motion information images and three motionless information images from an action video. Those motion and static information-oriented maps are improved by engaging the LBP algorithm on them. The outputs of the LBP are fed into GLAC descriptor to get an action description vector. With the obtained feature vectors from GLAC, the action classes are distinguished through the KELM classification model. The approach is extensively assessed in terms of the MSRAction3D, DHA, and UTD-MHAD datasets. Because of our action representation strategy, the proposed algorithm exhibits considerable performance over the existing handcrafted as well as deep learning methods. It is also obvious that the enhanced auto-correlation features-based method overwhelms the simple auto-correlation features-based method successfully. Thus, the improvement of the features is significant to enhance the system. Furthermore, the computational efficiency of the method specifies its suitability for real-time operation.

It is worth mentioning, some misclassifications are observed in our method. Note that the proposed method did not remove noise to improve the performance. The system only employed the LBP as preprocessing method for edge enhancing. Besides the LBP, a noise removing algorithm can be utilized to address the miss-classification issues of the proposed approach and thus for further improvement of the overall recognition accuracy. The descriptor can be more improved to increase the discriminatory power of the approach.

In our future work, we aim to build a deep model using the obtained 2D motion and static images. Besides, the current approach is not evaluated on large and complex RGB datasets like UCF101 and HMDB51. With a proper modification, the approach would be tested on these datasets in the future. Furthermore, we have a plan to build a new recognition framework using the GLAC descriptor on RGB and depth datasets jointly..

**Acknowledgements** This work is partially supported by University Grants Commission (UGC), Bangladesh.

### Declarations

**Conflict of interest** The authors declare that they have no conflict of interest.

**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless

indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

## References

- Sabokrou M, Fayyaz M, Fathy M, Moayed Z, Klette R (2018) Deep-anomaly: fully convolutional neural network for fast anomaly detection in crowded scenes. *Comput Vis Image Underst* 172:88–97
- Kim K, Jalal A, Mahmood M (2019) Vision-based human activity recognition system using depth silhouettes: a smart home system for monitoring the residents. *J Electr Eng Technol* 14(6):2567–2573
- Zhuang Z, Xue Y (2019) Sport-related human activity detection and recognition using a smartwatch. *Sensors* 19(22):5001
- Hendry D, Chai K, Campbell A, Hopper L, O'Sullivan P, Straker L (2020) Development of a human activity recognition system for ballet tasks. *Sports Med-Open* 6(1):10
- Ogbuabor G, La R (2018) Human activity recognition for health-care using smartphones. In: *Proceedings of the 2018 10th international conference on machine learning and computing*, pp 41–46
- Gul MA, Yousaf MH, Nawaz S, Ur Rehman Z, Kim H (2020) Patient monitoring by abnormal human activity recognition based on CNN architecture. *Electronics* 9(12):1993
- Sebestyen G, Stoica I, Hangan A (2016) Human activity recognition and monitoring for elderly people. In: *2016 IEEE 12th international conference on intelligent computer communication and processing (ICCP)*. IEEE, pp 341–347
- Sagayam KM, Hemanth DJ (2017) Hand posture and gesture recognition techniques for virtual reality applications: a survey. *Virtual Real* 21(2):91–107
- Haria A, Subramanian A, Asokkumar N, Poddar S, Nayak JS (2017) Hand gesture recognition for human computer interaction. *Proc Comput Sci* 115:367–374
- Wu C-Y, Zaheer M, Hu H, Manmatha R, Smola AJ, Krähenbühl P (2018) Compressed video action recognition. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp 6026–6035
- Ahmad Z, Khan N (2019) Human action recognition using deep multilevel multimodal  $M^2$  fusion of depth and inertial sensors. *IEEE Sens J* 20(3):1445–1455
- Wang H, Schmid C (2013) Action recognition with improved trajectories. In: *Proceedings of the IEEE international conference on computer vision*, pp 3551–3558
- Chen C, Zhang B, Hou Z, Jiang J, Liu M, Yang Y (2017) Action recognition from depth sequences using weighted fusion of 2D and 3D auto-correlation of gradients features. *Multimed Tools Appl* 76(3):4651–4669
- Chen C, Liu K, Kehtarnavaz N (2016) Real-time human action recognition based on depth motion maps. *J Real-Time Image Process* 12(1):155–163
- Oreifej O, Liu Z (2013) Hon4d: histogram of oriented 4D normals for activity recognition from depth sequences. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp 716–723
- Yang X, Tian Y (2014) Super normal vector for activity recognition using depth sequences. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp 804–811
- Vemulapalli R, Arrate F, Chellappa R (2014) Human action recognition by representing 3d skeletons as points in a Lie group. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp 588–595
- Liang B, Zheng L (2013) Three dimensional motion trail model for gesture recognition. In: *Proceedings of the IEEE international conference on computer vision workshops*, pp 684–691
- Ojala T, Pietikäinen M, Mäenpää T (2002) Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE Trans Pattern Anal Mach Intell* 7:971–987
- Kobayashi T, Otsu N (2008) Image feature extraction using gradient local auto-correlations. In: *European conference on computer vision*. Springer, pp 346–358
- Huang G-B, Zhou H, Ding X, Zhang R (2012) Extreme learning machine for regression and multiclass classification. *IEEE Trans Syst Man Cybern Part B (Cybern)* 42(2):513–529
- Li W, Zhang Z, Liu Z (2010) Action recognition based on a bag of 3D points. In: *2010 IEEE computer society conference on computer vision and pattern recognition-workshops*. IEEE, pp 9–14
- Lin Y-C, Hu M-C, Cheng W-H, Hsieh Y-H, Chen H-M (2012) Human action recognition and retrieval using sole depth information. In: *Proceedings of the 20th ACM international conference on multimedia*. ACM, pp 1053–1056
- Chen C, Jafari R, Kehtarnavaz N (2015) Utd-mhad: a multimodal dataset for human action recognition utilizing a depth camera and a wearable inertial sensor. In: *IEEE international conference on image processing (ICIP)*. IEEE, pp 168–172
- Ahmed A, Jalal A, Kim K (2020) RGB-D images for object segmentation, localization and recognition in indoor scenes using feature descriptor and Hough voting. In: *2020 17th international Bhurban conference on applied sciences and technology (IBCAST)*. IEEE, pp 290–295
- Jalal A, Kamal S, Kim D (2015) Depth silhouettes context: a new robust feature for human tracking and activity recognition based on embedded HMMs. In: *2015 12th international conference on ubiquitous robots and ambient intelligence (URAI)*. IEEE, pp 294–299
- Jalal A, Kim Y-H, Kim Y-J, Kamal S, Kim D (2017) Robust human activity recognition from depth video using spatiotemporal multi-fused features. *Pattern Recognit* 61:295–308
- ud din Tahir SB, Jalal A, Batool M (2020) Wearable sensors for activity analysis using SMO-based random forest over smart home and sports datasets. In: *2020 3rd international conference on advancements in computational sciences (ICACS)*. IEEE, pp 1–6
- Kamal S, Jalal A, Kim D (2016) Depth images-based human detection, tracking and activity recognition using spatiotemporal features and modified HMM. *J Electr Eng Technol* 11(6):1857–1862
- Rizwan SA, Jalal A, Kim K (2020) An accurate facial expression detector using multi-landmarks selection and local transform features. In: *2020 3rd international conference on advancements in computational sciences (ICACS)*. IEEE, pp 1–6
- Farooq A, Jalal A, Kamal S (2015) Dense RGB-D map-based human tracking and activity recognition using skin joints features and self-organizing map. *KSII Trans Internet Inf Syst* 9(5):1856–1869
- Kamal S, Jalal A (2016) A hybrid feature extraction approach for human detection, tracking and activity recognition using depth sensors. *Arab J Sci Eng* 41(3):1043–1051
- Yaacob NI, Tahir NM (2012) Feature selection for gait recognition. In: *2012 IEEE symposium on humanities, science and engineering research*. IEEE, pp 379–383
- Bulbul MF, Jiang Y, Ma J (2015) Human action recognition based on DMMS, hogs and contourlet transform. In: *2015 IEEE international conference on multimedia big data*. IEEE, pp 389–394

35. Bulbul MF, Jiang Y, Ma J (2015) Real-time human action recognition using DMMs-based LBP and EOH features. In: International conference on intelligent computing. Springer, pp 271–282
36. Bulbul MF, Islam S, Zhou Y, Ali H (2019) Improving human action recognition using hierarchical features and multiple classifier ensembles. *Comput J* bxz123. <https://doi.org/10.1093/comjnl/bxz123>
37. Dhiman C, Vishwakarma DK (2019) A robust framework for abnormal human action recognition using *R*-transform and Zernike moments in depth videos. *IEEE Sens J* 19(13):5195–5203
38. Chaaraoui AA, Padilla-López JR, Climent-Pérez P, Flórez-Revuelta F (2014) Evolutionary joint selection to improve human action recognition with RGB-D devices. *Expert Syst Appl* 41(3):786–794
39. Gao Z, Zhang H, Xu G, Xue Y (2015) Multi-perspective and multi-modality joint representation and recognition model for 3D action recognition. *Neurocomputing* 151:554–564
40. Rahmani H, Mahmood A, Huynh DQ, Mian A (2014) Real time action recognition using histograms of depth gradients and random decision forests. In: IEEE winter conference on applications of computer vision. IEEE, pp 626–633
41. Luo J, Wang W, Qi H (2014) Spatio-temporal feature extraction and representation for RGB-D human action recognition. *Pattern Recognit Lett* 50:139–148
42. Vishwakarma DK (2020) A two-fold transformation model for human action recognition using decisive pose. *Cogn Syst Res* 61:1–13
43. Wang L, Huynh DQ, Koniusz P (2019) A comparative review of recent kinect-based action recognition algorithms. *IEEE Trans Image Process* 29:15–28
44. Yang R, Yang R (2014) DMM-pyramid based deep architectures for action recognition with depth cameras. In: Asian conference on computer vision. Springer, pp 37–49
45. Wang P, Li W, Gao Z, Zhang J, Tang C, Ogunbona PO (2015) Action recognition from depth maps using deep convolutional neural networks. *IEEE Trans Hum Mach Syst* 46(4):498–509
46. Wang P, Li W, Gao Z, Tang C, Ogunbona PO (2018) Depth pooling based large-scale 3-D action recognition with convolutional neural networks. *IEEE Trans Multimed* 20(5):1051–1061
47. Chen J, Xiao Y, Cao Z, Fang Z (2018) Action recognition in depth video from RGB perspective: a knowledge transfer manner. In: MIPPR 2017: pattern recognition and computer vision, vol 10609. International Society for Optics and Photonics, p 1060916
48. Imran J, Kumar P (2016) Human action recognition using RGB-D sensor and deep convolutional neural networks. In: 2016 international conference on advances in computing, communications and informatics (ICACCI). IEEE, pp 144–148
49. Dhiman C, Vishwakarma DK (2020) View-invariant deep architecture for human action recognition using two-stream motion and shape temporal dynamics. *IEEE Trans Image Process* 29:3835–3844
50. Weng J, Liu M, Jiang X, Yuan J (2018) Deformable pose traversal convolution for 3D action and gesture recognition. In: Proceedings of the European conference on computer vision (ECCV), pp 136–152
51. Munro J, Damen D (2020) Multi-modal domain adaptation for fine-grained action recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 122–132
52. Busto PP, Iqbal A, Gall J (2018) Open set domain adaptation for image and action recognition. *IEEE Trans Pattern Anal Mach Intell* 42(2):413–429
53. Huang G-B, Zhu Q-Y, Siew C-K (2006) Extreme learning machine: theory and applications. *Neurocomputing* 70(1–3):489–501
54. Xia L, Aggarwal J (2013) Spatio-temporal depth cuboid similarity feature for activity recognition using depth camera. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 2834–2841
55. Chen C, Jafari R, Kehtarnavaz N (2015) Action recognition from depth sequences using depth motion maps-based local binary patterns. In: 2015 IEEE winter conference on applications of computer vision (WACV). IEEE, pp 1092–1099
56. Rahmani H, Huynh DQ, Mahmood A, Mian A (2016) Discriminative human action classification using locality-constrained linear coding. *Pattern Recognit Lett* 72:62–71
57. Kong Y, Satarboroujeni B, Fu Y (2015) Hierarchical 3D kernel descriptors for action recognition using depth sequences. In: 2015 11th IEEE international conference and workshops on automatic face and gesture recognition (FG), vol 1. IEEE, pp 1–6
58. Zhang B, Yang Y, Chen C, Yang L, Han J, Shao L (2017) Action recognition using 3D histograms of texture and a multi-class boosting classifier. *IEEE Trans Image Process* 26(10):4648–4660
59. Liang C, Chen E, Qi L, Guan L (2016) 3D action recognition using depth-based feature and locality-constrained affine subspace coding. In: 2016 IEEE international symposium on multimedia (ISM). IEEE, pp 261–266
60. Liu J, Shahroudy A, Xu D, Wang G (2016) Spatio-temporal LSTM with trust gates for 3D human action recognition. In: European conference on computer vision. Springer, pp 816–833
61. Yang X, Tian Y (2017) Super normal vector for human activity recognition with depth cameras. *IEEE Trans Pattern Anal Mach Intell* 39(5):1028–1039
62. Liu J, Shahroudy A, Xu D, Kot AC, Wang G (2018) Skeleton-based action recognition using spatio-temporal LSTM network with trust gates. *IEEE Trans Pattern Anal Mach Intell* 40(12):3007–3021
63. Weng J, Weng C, Yuan J (2017) Spatio-temporal Naive-Bayes nearest-neighbor (ST-NBNN) for skeleton-based action recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 4171–4180
64. Asadi-Aghbolaghi M, Kasaei S (2018) Supervised spatio-temporal kernel descriptor for human action recognition from RGB-depth videos. *Multimed Tools Appl* 77(11):14115–14135
65. Keçeli AS, Kaya A, Can AB (2018) Combining 2D and 3D deep models for action recognition with depth information. *Signal Image Video Process* 12(6):1197–1205
66. Azad R, Asadi-Aghbolaghi M, Kasaei S, Escalera S (2018) Dynamic 3D hand gesture recognition by learning weighted depth motion maps. *IEEE Trans Circuits Syst Video Technol* 29(6):1729–1740
67. Zhang C, Tian Y, Guo X, Liu J (2018) Daal: deep activation-based attribute learning for action recognition in depth videos. *Comput Vis Image Underst* 167:37–49
68. Nguyen XS, Mouaddib A-I, Nguyen TP, Jeanpierre L (2018) Action recognition in depth videos using hierarchical Gaussian descriptor. *Multimed Tools Appl* 77(16):21617–21652
69. Bulbul MF, Islam S, Ali H (2019) Human action recognition using MHI and SHI based GLAC features and collaborative representation classifier. *Multimed Tools Appl* 78(15):21085–21111
70. Jalal A, Kamal S, Kim D (2016) Human depth sensors-based activity recognition using spatiotemporal features and hidden Markov model for smart environments. *J Comput Netw Commun* 2016(17):1–11
71. Chen C, Liu M, Zhang B, Han J, Jiang J, Liu H (2016) 3D action recognition using multi-temporal depth motion maps and Fisher vector. In: IJCAI, pp 3331–3337
72. Wang P, Li W, Li C, Hou Y (2018) Action recognition based on joint trajectory maps with convolutional neural networks. *Knowl-Based Syst* 158:43–53
73. Iosifidis A, Tefas A, Pitas I (2015) On the kernel extreme learning machine classifier. *Pattern Recognit Lett* 54:11–17

**Publisher's Note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.