Multimedia Tools and Applications
https://doi.org/10.1007/s11042-020-08704-0

Particle swarm optimization with deep learning for human action recognition

S. Jeba Berlin[1] · Mala John[1]

Received: 14 February 2019 / Revised: 4 December 2019 / Accepted: 28 January 2020
Published online: 16 February 2020

Abstract
A novel method for human action recognition using a deep learning network with features optimized using particle swarm optimization is proposed. The binary histogram, Harris corner points and wavelet coefficients are the features extracted from the spatiotemporal volume of the video sequence. In order to reduce the computational complexity of the system, the feature space is reduced by particle swarm optimization technique with the multi-objective fitness function. Finally, the performance of the system is evaluated using deep learning neural network (DLNN). Two autoencoders are trained independently and the knowledge embedded in the autoencoders are transferred to the proposed DLNN for human action recognition. The proposed framework achieves an average recognition rate of 91% on UT interaction set 1, 88% on UT interaction set 2, 91% on SBU interaction dataset and 94% on Weizmann dataset.

Keywords Video surveillance · Human action recognition · Autoencoder · Deep learning network · Particle swarm optimization

1 Introduction

Video classification is exhaustively explored in computer vision due to its wide range of applications including crowd scene analysis [9], video retrieval [59], health care, human-machine interaction [36] and human action recognition. Among this, human action recognition is an interesting topic that aims in recognizing a particular event of interest in the video sequences. On the other hand, the solution to this problem suffers several bottlenecks [21, 22], which include intra-pattern variation, execution time, camera motion, occlusion, variation in scales, cluttered background and variation in viewpoints. The major tasks in human action recognition involve feature extraction, representation and classification. However, the main

✉ S. Jeba Berlin
jebaberlin@gmail.com

[1] Department of Electronics Engineering, Madras Institute of Technology, Anna University, Chennai, India

challenge is finding the discriminate features from the video streams that characterize the human motion while eliminating the effect of camera and background motions.

In this work, the first level of processing is to extract the foreground of each frame from the video sequences. However, the focus of the paper is on the classification of action sequence and therefore existing methods of foreground segmentation are used. To obtain the foreground, background subtraction is carried out in case of the static background and Gaussian Mixture modelling (GMM) for dynamic background scenarios. The Gaussian mixture modelling [23] is a powerful method for separating foreground object and the method is also robust to shadow and other artefacts. The Harris corner points and the wavelet transform coefficients are the features computed in the spatiotemporal volume in order to detect target actions from the video sequence.

The irrelevant and redundant features obtained during feature extraction have large feature space which in turn increases the computational complexity of the system. Therefore, it is necessary to reduce the number of features to make the system simple. Evolutionary computation technique [53] is the recently used feature selection process due to its global search ability. Infact, Particle Swarm Optimization (PSO) is one of the powerful feature selection method used in the recent decades to overcome the problem of local optima. These methods use multi-objective fitness function to measure the quality of the feature and to select the optimal features required for classification.

Finally, the DLNN is used for classification of actions. Based on the theory of transfer learning, re-utilizing the knowledge from the base model is a great starting point to develop a successful recognition system. This increases the model performance with limited resource utilization and has the potent to solve the problem of insufficient training samples. Therefore, in this work, two autoencoders are trained independently in order to efficiently compress the input. Later, the knowledge embedded in the hidden layers of the autoencoders is transferred to the DLNN. Unlike conventional machine learning techniques that attempts to learn from scratch, the autoencoders transfer the hidden knowledge to the DLNN. Moreover, the cross-person recognition is employed in this work.

The contribution of the proposed work is summarized as follows

1) Detect Harris corner points from salient motion region of the region of interaction (RoI) inorder to reduce the computational burden for further analysis.
2) Develop the feature descriptor that posses compact and discriminate wavelet coefficients which is extracted from the temporal domain of pre-processed action sequence.
3) Employ an RBM based autoencoder (AE) to pre-train the DLNN, so as to reduce the computational cost.
4) Pop out the non significant wavelet coefficients using the popular global search evolutionary algorithm called particle swarm optimization with multi-objective fitness function to get the sparse representation of action sequence.

The rest of this paper is organized as follows. Section 2 describes the traditional and deep learning based human action recognition and PSO based feature reduction. Section 3 gives technical details of the proposed algorithm. Section 4 shows the experimental results and discussions. Finally, the conclusion of the paper is presented in Section 5.

2 Related work

Human action recognition in the last few decades has been dominated by different approaches including spatiotemporal interest points, spatiotemporal volume, trajectory based, skeleton based and transform based approaches. Human actions can also be recognized through deformable body parts, where everybody parts are detected separately and the detected results are then combined to develop the probabilistic or discriminative model [16] of an action. The bag of words [25, 32, 33] technique focuses more attention in human action recognition to represent the local features in the spatio-temporal environment. This representation is more robust to complex backgrounds and occlusions but it fails to capture the temporal information of the human actions. Besides, the skeleton based approach [19, 20, 29] relies on human pose to track body joints in the form of skeleton. The skeleton data obtained from single camera is more noisy and unreliable when the person is not facing the camera and under occlusions.

The global techniques give the holistic information regarding the human actions considering both the spatial and temporal information. The motion energy image (MEI), motion history image (MHI), motion history volumes (MHV) and gait energy image (GEI) are some of the temporal templates [50] extracted from the silhouette of the video sequences. This gives poor performance in case of partial occlusion, noisy backgrounds and shadows. Later, spatio-temporal volume based approach [4, 55] is introduced where the 3D space-time volume representation of the human actions is considered. It preserves spatial and temporal information in the video sequences but they are sensitive to photometric and geometric distortions [7] and needs very precise background subtraction techniques to extract the silhouette of an image.

One of the successful space-time features is the trajectory that relies on optical flow [40] of the action sequence. It is computationally complex and expensive as it is difficult to track necessary points involved in an action. The dense trajectories [47] perform well on human action recognition, but it involves huge number of features. Moreover, the addition of motion boundary histogram eliminates the effect of camera motion and excludes the irrelevant trajectories in the background. However, in dense crowd scenarios, only the impact of trajectories is insufficient, hence it is necessary to include the velocity and physical interaction either between multiple humans [24] or human-object interactions [1].

Recently, the wavelet based approach is used in human action recognition because it generates the shortest descriptor which is extremely compact, discriminative and informative. It provides directional information of various actions thereby increases the classification accuracy. The 2D DWT is often combined with feature reduction techniques such as step wise linear discriminant analysis [43] and principal component analysis [17] to reproduce salient localized features and to classify the actions based on regression values. However, the 3D DWT with orientation sensitive wavelet sub bands [37] represent the global dynamic events present in the videos with low computational cost. In the work reported in [3], the spatiotemporal changes are detected using 3D stationary wavelet transform and then local binary pattern is used as texture descriptor. These features are then fused with moments to give both local and global features.

Further, the 3D stationary wavelet transform is applied to generate wavelet based energy image [2] of the video sequence. The consolidated foreground image is then generated by fusing those images obtained through different scales and sub bands of action sequence. Later, the seven Hu moments are used as the feature vector for classification. The method addressed in [60] doesn't include any external feature selection technique. However, the DWT is applied over three orthogonal planes which are intersected on the middle point of each of the cuboids

around the estimated interest points. They found that DWT generates the coefficients less than one thirtieth of the 3D gradient descriptor.

In recent decades, deep learning based human activity recognition is becoming popular because of its outstanding performance. The convolutional neural network is the most commonly used deep learning model which consists of several layers of trainable filters and pooling layers to yield the hierarchy of complex features. The 3D convolutional neural networks [18, 34] capture appearance and motion information embedded in spatiotemporal volume by performing 3D convolution in the spatio temporal domain. The factorized spatio-temporal convolutional networks used in [44] performs 2D spatial convolution followed by 1D temporal convolution instead of performing 3D convolution directly over the video sequence so as to reduce the computational complexity of the system.

In the work presented in [6], deep trajectory based features that gives motion information is combined with appearance based deep features extracted from the individual frames to deliver the spatiotemporal information. To share the merits of both handcrafted and deep features, the trajectory pooled convolutional descriptor is reported in [48]. Here, the discriminative convolutional features are initially extracted using deep architectures and then the trajectory constrained pooling is done to construct the effective descriptor. Moreover, these models has the potential to act as the primitive one to deploy it in human computer interfaces for content based video retrieval after applying binary hashing techniques [57, 58].

In recent years, transfer learning is becoming increasingly popular, as it alleviates the requirement for training deep learning models from scratch, giving rise to a huge saving in terms of training hours involved. Thereby, the classical machine learning techniques are now replaced by pre-trained models such as GoogleNet [45], VGG16 [49], AlexNet [39], Inspection-v3[46], Resnet [13] and L-CNN [35], either for classification or for feature extraction. Despite of the deployment of pre-trained models, some researchers exploit the use of transfer learning to the handcrafted features. In [31], both the feature and sample level adaptation is done to transfer the knowledge from the well labelled samples in the source domain to the poorly labelled samples in the target domain. The low rank discriminant learning is employed in [30] to share the common feature representation for different viewpoints and modalities which posses different probability distributions.

The particle swarm optimization [26] is one of the effective evolutionary computation techniques that work on population based mechanism to produce multiple solutions in a single iteration. The PSO-SVM model used in [15] optimizes the feature subset and SVM kernel parameters simultaneously. In catfish particle swarm optimization proposed in [8], the particle with the worst fitness is introduced, when the fitness of the global best is not improved for consecutive iterations. In [56], the cost based feature selection technique which involves multi-objective PSO is addressed. Further, to enhance the search capability, the probability based encoding, the crowd distance and the Pareto relationship are added with traditional PSO concepts.

In contrast to the above mentioned methods, the proposed method utilizes 1D DWT which posses both directional and scale invariance property is applied along the temporal domain of the pre-processed action sequence. Building an efficient 1D DWT is simple and computationally less intensive as compared to other 3D DWT [2, 3, 17, 37]. Unlike other descriptors [4, 25, 47, 55] that relies on huge number of features, the proposed work posses only 19 features from a single spatio-temporal volume. Also, the shallow network called sparse autoencoder [14] is learned in an unsupervised manner to pre-train the DLNN so as to enhance the generalization capability of the deep learning neural network. In addition, the PSO with multi-objective

fitness function is used for the simultaneous optimization of feature space and the network performance.

3 Proposed method

In this section, the detailed description of the proposed human action recognition system is presented. Specifically, this section discuss on RoI extraction, wavelet based feature extraction and action classification using pre-trained DLNN. The overview of the proposed framework is depicted in Fig. 1.

3.1 Foreground extraction

The first step in the proposed work is the segmentation of foreground image. In static environment, the foreground image is obtained just by subtracting each frame with the background model. In case of complex and dynamic background scenario, the GMM that uses expectation maximization algorithm is applied for foreground segmentation. It is then followed by morphological operations such as erosion and dilation to remove unnecessary objects present in the foreground image.

The supporting RoI is determined under the following conditions [5]. If two objects in the segmented image are detected as spatially separated, then the region of interest is considered as the rectangular region with their vertical sides pass through the centroid of those two objects as in Fig. 2(a), (b). The region of interaction of the overlapped objects is considered to be the rectangle whose width is half that of the width of the overlapped region and its centroid coincides with the foreground object as shown in Fig. 2(c). The RoIs are represented with green rectangles in Fig. 2. Thus instead of extracting feature from the overall interaction region, the region is limited over certain area where the movement of arms takes place.

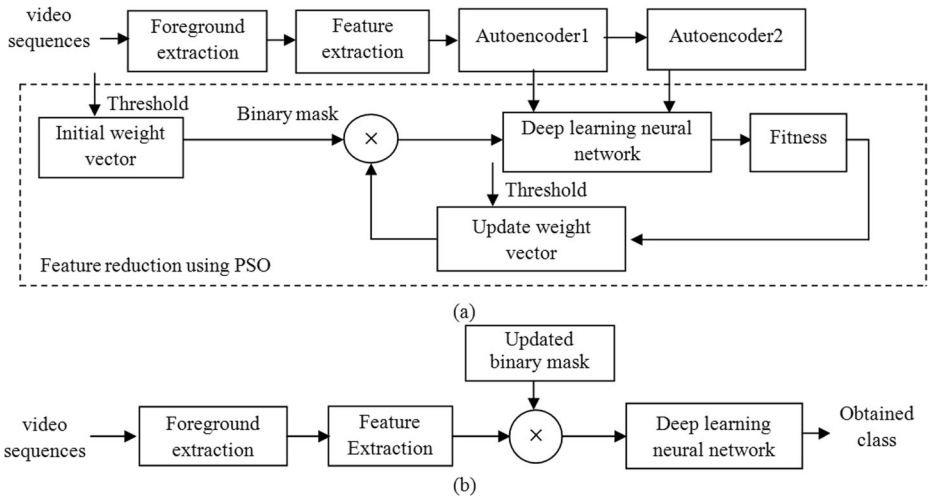


Fig. 1 Overview of the proposed approach (a) Training (b) Testing

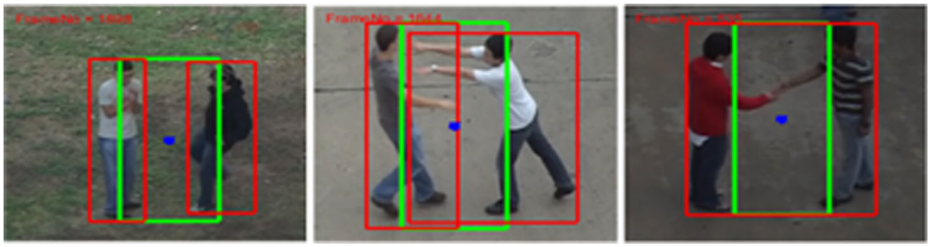


Fig. 2 Supporting RoI extraction (a)Non overlapping objects (b) overlapping objects without contact (c) overlapping objects with contact - Bounding box enclosing the interacting individuals; − RoI; − Center of RoI

3.2 Feature extraction

This section discuss on spatial and temporal feature extraction from the salient RoI. The spatial features are extracted from every frame of the action sequences to give the structural or appearance based information. However, the temporal features in the video sequences are computed across entire frames of corresponding blocks to deliver the motion information.

3.2.1 Spatial feature extraction

The Harris corner detector is used to find the key points from the region of interest of every frame in the video sequences. Only 'M' key points (x_i,y_i), i = 0,1,...,M-1 nearer to the centroid of the region of interest (c_x,c_y) along the horizontal direction (i.e. the M lowest value of $|x_i - c_x|$) are retained as in Fig. 3. The polar coordinates $(r_j,\varnothing_j)$ of these 'M' key points are considered as one of the spatial features referred as **K** and is given by

$$\mathbf{K}_n = \left[(r_{0n}, \varnothing_{0n}), (r_{1n}, \varnothing_{1n}), \ldots., \left(r_{(M-1)n}, \varnothing_{(M-1)n}\right)\right] \tag{1}$$

The histogram of the region of interest is another spatial feature extracted along the frames. This is computed by dividing the region of interest into 'L' number of non-overlapping blocks. The number of foreground pixels falling within every block forms the histogram vector **H**. The vector **H** for the n^{th} frame is depicted as

$$\mathbf{H}_n = \left[h_{0n}, h_{1n}, , \ldots, h_{(L-1)n}\right] \tag{2}$$

Finally, the feature vector '**B**' is formed with the elements of the vectors **K** and **H** of the spatiotemporal volume comprised of 'T' frames of dimension (2 M + L)T is referred as given below.

Fig. 3 Keypoints extracted from a single frame within the RoI from different datasets (a) UT interaction set 1 (b)UT interaction set 2 (c) SBU interaction (d)Wiezmann dataset

$$\mathbf{B} = \begin{bmatrix} r_{00}, \emptyset_{00} & r_{10}, \emptyset_{10} & \cdots & r_{(M-1)0}, \emptyset_{(M-1)0} & H_0 \\ r_{01}, \emptyset_{01} & r_{11}, \emptyset_{11} & \cdots & r_{(M-1)1}, \emptyset_{(M-1)1} & H_1 \\ \vdots & \vdots & & \vdots & \vdots \\ r_{0(T-1)}, \emptyset_{0(T-1)} & r_{1(T-1)}, \emptyset_{1(T-1)} & \cdots & r_{(M-1)(T-1)}, \emptyset_{(M-1)(T-1)} & H_{(T-1)} \end{bmatrix} \quad (3)$$

3.2.2 Wavelet transform based feature extraction for temporal correlation

Wavelets are powerful in characterizing non-stationary signals and transients. Here, the human action sequence could be viewed as multiple temporal signals with each signal corresponding to a specific spatial location. The spatial locations pertaining to 'no change' would contribute to a high magnitude of approximation coefficients and low value of detail coefficients. However, those spatial locations which undergo significant changes during an action would contribute to a transient signal, thereby giving rise to high magnitude detail coefficients.

The algorithm for the computation of wavelet coefficient operates on the spatiotemporal volume comprised of 'T' frames containing the foreground of the interacting humans. The size of the region of interaction is normalized and divided into 'N' non-overlapping blocks and the histogram of each frame is computed. Considering only the p^{th} block of each frame in the spatiotemporal volume of 'T' frames and ordering them into a time sequence, $\mathbf{G}_p$ is obtained. $\mathbf{G}_p = [h_{p0}, h_{p1}, \ldots, h_{p(T-1)}]$. Formation of $\mathbf{G}_p$ and computation of wavelet coefficient is illustrated through Fig. 4. The wavelet transform of $\boldsymbol{G}_p$ is represented as $\mathbf{C}_p$ as given below.

$$\mathbf{C}_p = \left[C_p^{A3}, C_p^{D3}, C_p^{D2}, C_p^{D1} \right] \quad (4)$$

The dimension of $\mathbf{C}_p$ is $1 \times T$ and the notation C_p^{A3}, C_p^{D3}, C_p^{D2}, C_p^{D1} represents the approximate coefficients and detail coefficients at level 3, level 2 and level 1 respectively for the p^{th} block of video sequences. The procedure is repeated for all the blocks and the wavelet coefficients extracted from the spatiotemporal volume of 'T' frames are arranged to form the feature vector '**C**' of dimension $TN \times 1$ as shown below.

$$\mathbf{C} = \left[C_0^{A3}, C_1^{A3}, \ldots C_N^{A3}, C_0^{D3}, C_1^{D3}, \ldots, C_N^{D3}, C_0^{D2}, C_1^{D2}, \ldots, C_N^{D2}, C_0^{D1}, C_1^{D1}, \ldots C_N^{D1} \right] \quad (5)$$

The vectors **B** and **C** thus formed are concatenated to form the feature vector '**X**' of dimension $\lambda \times 1$ ($\lambda = T(2M + L + N)$).

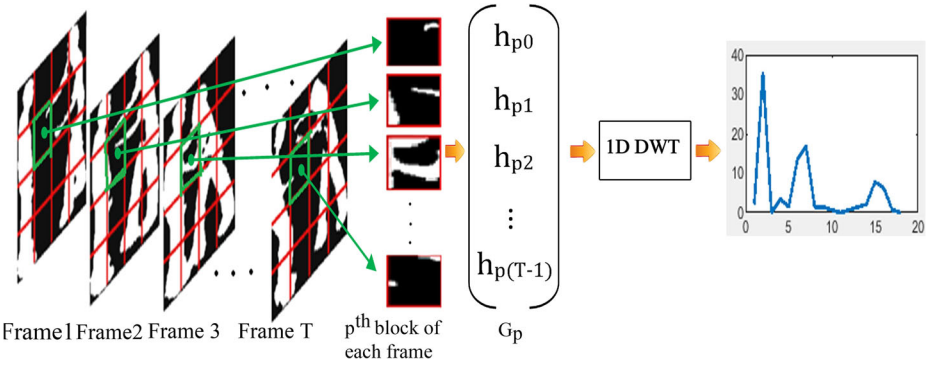


Fig. 4 Computation of wavelet coefficients from a single spatio temporal volume

3.3 Particle swam optimization (PSO) for optimal feature space selection

Though the orthogonal wavelet coefficients facilitate discrimination between various action sequences, the wavelet decomposition by itself does not give rise to dimensionality reduction. Therefore, a PSO based technique is used to extract optimized wavelet coefficients while achieving dimensionality reduction. Feature reduction techniques choose a subset of features that have the ability to achieve similar or better performance than the complete set of features. But, it has the conflicting objectives of maximizing the classification performance (minimizing the classification error) and minimizing the number of features. To solve this issue, the proposed work employs a multi-objective fitness based PSO technique.

The PSO is the optimization technique used herein for feature selection. Given the set of features $\mathbf{X} = \{x_1, x_2, \ldots x_\lambda\}$, the PSO will generate the set of weights $\theta = \{\theta_1, \theta_2, \ldots, \theta_\lambda\}$ after it passes through the number of generations. The weight vectors generated through PSO are thresholded to obtain the binary mask to perform the task of feature reduction. The position vector of the i^{th} swarm $\boldsymbol{\theta_i} = (\theta_{1i}, \theta_{2i}, \ldots, \theta_{\lambda i})$, where λ is the number of features in the feature space. The corresponding velocity vector defined by $\mathbf{v_i} = (v_{1i}, v_{2i}, \ldots, v_{\lambda i})$ are randomly initialized using uniform distribution. Then the vector $\boldsymbol{\theta_d}$ is used to generate reduced feature space according to the following condition.

$$x_{kd} = \begin{cases} x_{kd} & \text{if } \theta_{kd} > \delta \\ 0 & \text{if } \theta_{kd} < \delta \end{cases} \tag{6}$$

The value of '1' indicates the feature value at that particular location is selected. The reduced set of features is fed to the deep learning neural network to find the fitness value. The multi-objective fitness function is used to give importance to both the reduced feature set and the classification performance [51, 10]. The fitness function is expressed as.

$$f = a * \frac{\text{No. of features in the selected set}}{\lambda} + (1-a) * \frac{\text{MSE of selected feature set}}{\text{MSE of all available features}} \tag{7}$$

The position and velocity of the particles are updated for each iteration 't' according to its own fitness or fitness of its neighbours as

$$\boldsymbol{\theta}_i^{t+1} = \boldsymbol{\theta}_i^t + \boldsymbol{v}_i^{t+1} \tag{8}$$

$$\boldsymbol{v}_i^{t+1} = \mu * \boldsymbol{v}_i^t + \partial_1 * \sigma_1 * \left(\boldsymbol{p}_i - \boldsymbol{\theta}_i^t\right) + \partial_2 * \sigma_2 * \left(\boldsymbol{p}_g - \boldsymbol{\theta}_i^t\right) \tag{9}$$

where μ is inertia weight to control the impact of the previous velocities on the current velocity. ∂_1 and ∂_2 are acceleration constants. σ_1 and σ_2 are random values uniformly distributed in [0, 1]. Let $\mathbf{p}_i$ and $\mathbf{p}_g$ represent the local best and global best particles and p_{id} and p_{gd} denote the elements of local best and global best particles. The velocity is limited within the interval [0 1].To give importance to both number of features and classification performance [52], the local best is updated as given below.

$$\boldsymbol{p_i} = \begin{cases} \boldsymbol{\theta}_i & \text{if } f(\boldsymbol{\theta}_i) > f\left(\boldsymbol{p}_i^{Prev}\right) \text{ and } |\boldsymbol{\theta}_i| \leq \left|\boldsymbol{p}_i^{Prev}\right| \\ \boldsymbol{\theta}_i & \text{if } f(\boldsymbol{\theta}_i) = f\left(\boldsymbol{p}_i^{Prev}\right) \text{ and } |\boldsymbol{\theta}_i| < \left|\boldsymbol{p}_i^{Prev}\right| \\ \boldsymbol{p}_i^{Prev} & \text{elsewhere} \end{cases} \tag{10}$$

The best local particle obtained so far is considered to be the best global particle $\boldsymbol{p}_g$. It is updated iteratively if the present local particle is better than the existing global particle. The particles and velocities are updated iteratively based on its own experience and experience from its neighbours. The global minimum achieved at the final iteration is considered as the optimal feature weight vector. Finally, the binary mask is generated to indicate the feature is selected at the location of binary value 1 and the feature is rejected elsewhere.

3.4 Deep learning neural network

The autoencoder is the basic model present in the deep learning neural network depicted in Fig. 5. The deep learning neural network in the proposed work includes two AE's stacked together as given in [11]. The two AE's are trained independently and sequentially. The input feature vector '**X**' described elsewhere in this paper is fed as the input to the first AE and trained according to (11) & (12). Followed by this, the second AE is trained with the hidden layer of the first AE fed as the input. In Fig. 5(a), g_i, i = 1,2....γ and s_i, i = 1,2....Γ represent the hidden layers of the first and the second AE respectively.

In order to configure the deep learning network, the feature vector '**X**' is used as the input layer and the output layer is used as the classification layer. The hidden layers of the two (trained) AE's form the hidden layers of the deep learning network. The training methodology of the autoencoder is outlined below.

In the autoencoder [42] the input vector $\mathbf{X} \in \mathbb{R}^{\lambda \times 1}$ is encoded into the vector $\mathbf{\Psi}$, $\mathbf{\Psi} = [r_1, r_2, ..r_\gamma]^T$, through the equation.

$$\mathbf{\Psi} = \Phi(\mathbf{WX} + \mathbf{b_{IN}}) \tag{11}$$

Here, $\mathbf{W} \in \mathbb{R}^{\gamma \times \lambda}$ is the weight vector connecting the input layer and the hidden layer, $\mathbf{b_{IN}}$ is the bias vector and $\Phi(z)$ is the sigmoid function defined as $\Phi(z) = 1/(1 + \exp.(-z))$. The output of the AE, $(\widehat{\mathbf{X}})$, is reconstructed from $\mathbf{\Psi}$ using the equation.

$$\widehat{\mathbf{X}} = \Phi\left(\mathbf{V^T}\mathbf{\Psi} + \mathbf{b_{OUT}}\right) \tag{12}$$

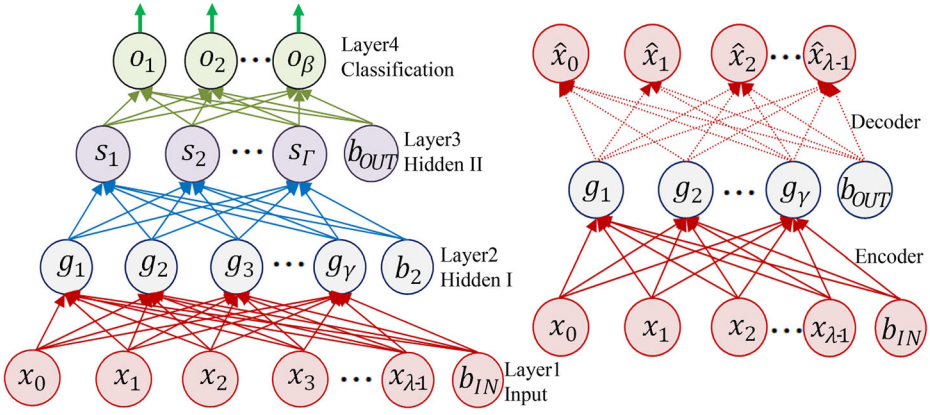


Fig. 5 Structure of (a) Deep learning neural network with stacked autoencoders (b) Autoencoder

where $\mathbf{V} \in \mathbb{R}^{\gamma \times \lambda}$ is the weight vector connecting the hidden layer and the output layer and $\mathbf{b_{OUT}}$ is the bias vector. The weights and the bias of the AE are learned by minimizing the sparse mean square error.

$$\xi = \left\| X - \widehat{X} \right\|_2 + \Lambda_1 \Omega_{\text{weights}} + \Lambda_2 \Omega_{\text{sparsity}} \tag{13}$$

The term Ω_{weights} and Ω_{sparsity} in the above equation corresponds to L_2 regularization and sparsity regularization respectively. The Λ_1 and Λ_2 represents the coefficient of L_2 regularization and sparsity regularization respectively. The higher degree of sparsity is achieved for the lower values of sparsity proportion (Λ_2). Similarly, increasing the coefficient (Λ_1) strengthens the regularization effect by making negligible weight values. Thus, the use of regularizer favours the sparse representation and avoids the problem of over fitting (underfitting).

The first AE is trained as per the procedure described above. On completion of the training process, the hidden layer of the first AE corresponding to every '**X**' is used as the input to the second AE and the weights and the bias of the second AE are learned by minimising the reconstruction error. The deep learning neural network (DLNN) is fine tuned through the minimization of the cross entropy error introduced at the classification layer using back propagation approach.

4 Experimental results and discussions

In this section, the performance of the proposed method is evaluated on three datasets that include UT interaction dataset [38], SBU Kinect RGB-D [54] video sequences and Weizmann dataset [12]. The leave one out cross validation is adopted to evaluate the performance of the proposed method over all the datasets. The proposed algorithm is implemented in MATLAB R2016a on Windows7 on core i5 processor.

4.1 Datasets

The UT interaction dataset [38] contains two sets of continuous videos with six types of two-person interactions, viz., shake hands (SH), point (PT), kick (KK), punch (PC), hug (HG) and push (PS). Each set has ten videos, one set of videos are taken from the parking lot with almost static background and the other set from the lawn with a dynamic background. The SBU Kinect Interaction dataset [54] contains eight two-person interactions: approach (AP), depart (DP), exchange something (EX), hug, push, kick, punch and shake hands taken in the same laboratory environment. The Weizmann (WZN) dataset [12] includes ten actions: bend (BD), jack (JK), jump (JP), pjump (PP), run (RN), side (SD), skip (SP), walk (WK), wave1 (W1) and wave2 (W2) with the known background.

4.2 Visualization of within and between-class variations

The RoI is restricted to an area where only the motion of arms in the human action takes place. After that, M = 40 strongest Harris points which are closer to the centre of normalised RoI are extracted from T = 16 overlapping frames. Further, the normalised RoI is divided into $N = 64$ blocks to generate the binary histogram of each frame. Finally, to compute the wavelet transform coefficients, the RoI is normalised to the size of 128×64 and each frame is divided

into non overlapping blocks of size 16 × 8. Later, the binary histogram is computed across the frames of each block, followed by 3-level 1D DWT.

The wavelet coefficients obtained for different action sequences are depicted in Figs. 6 and 7. The Fig. 6 shows the wavelet coefficients that are extracted from a spatio temporal volume for different actions such as hug, punch and shake hand of two different persons. For the action shake hand, the amplitude varies from 0 to 8 but in a wavery manner and for hug the amplitude is from 0-1which is comparatively very low because of negligible motions between consecutive frames. This ensures the within-class compactness of action sequence. Moreover, the camera placement and the scale of two action sequences are different. The change in the feature values of the different actions are depicted in Fig. 7. From this plot, it is inferred that the actions such as punch and kick has similar amplitude variation. It is also clear that there is strong between-class separation in the action sequence.

4.3 Impact of hyper-parameters

Two sparse AE's are trained separately to determine the initial weights of the deep learning neural network. The weights in the encoder side of the first AE are initialized as the weights between the input layer and first hidden node of DLNN. Similarly, the weights in the encoder side of the second AE are initialized as the weights between the first and second hidden nodes of the DLNN. The transfer function used in the AE is logsig as it gives better performance compared to satlin and purelin.

The training algorithm and loss function used for learning the AE is trainscg and msesparse respectively, whereas, the activation function and loss function used in the output neurons of DLNN is softmax and cross entropy respectively. The number of output neurons in the DLNN varies depending on the number of target actions present in the database. The autoencoders and DLNN are trained for 200 and 500 epochs respectively. In PSO, the initial particles and the initial velocities of size 20 × 3328 are randomly generated using uniform distribution in the range of 0 to 1. However, the inertia weight is iteratively varied and the values of acceleration constants are set to be 2.

The hyper-parameters such as L_2 weight regularization, sparsity regularization and sparsity proportion are selected based on the minimization of the discrepancy between input feature values and its reconstruction through autoencoders. In order to optimize the hyper-parameters, the different combinations of three parameters are considered and are fed either into the single layer network or to the two-layer network with different number of hidden nodes. The effect of the hyper-parameters on accuracy for the DLNN with single AE and two stacked AE's are illustrated in Figs. 8, 9.

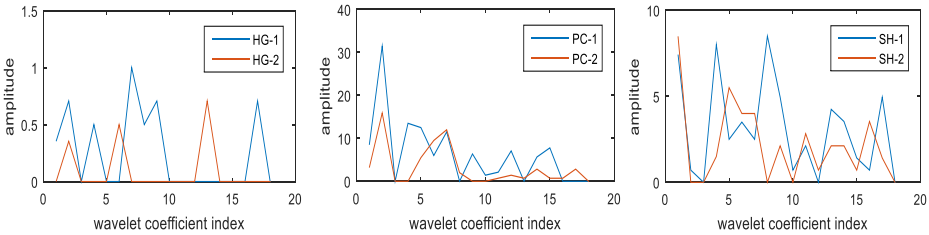


Fig. 6 Feature values obtained by the proposed algorithm for two action sequences of same class

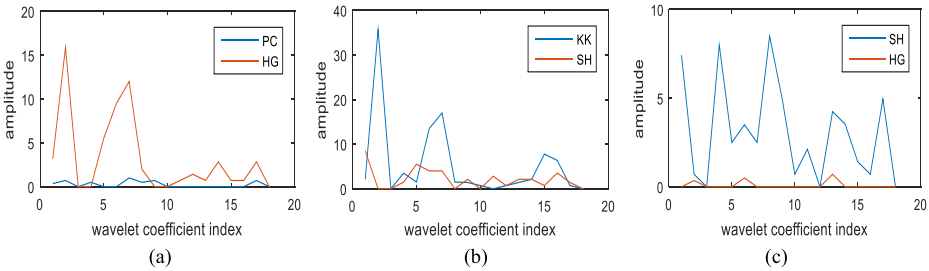


Fig. 7 Feature values obtained by the proposed algorithm for two action sequences of different classes

The sparse proportion and sparse regularization are set from 0.02 to .1 with a step of 0.02 and from 1 to 6 with a step of 1 respectively. The coefficient of L_2 regularization is varied as 0.001, 0.01 and 0. It is inferred from the figure that the optimal performance is achieved when L_2 regularization, sparse proportion and sparse regularization are set to be 0.01, 0.02 and 1 respectively. Further, the time required for training and testing the different combinations of hidden nodes are analyzed and are depicted in Fig. 10. From the figure, it is clear that as the

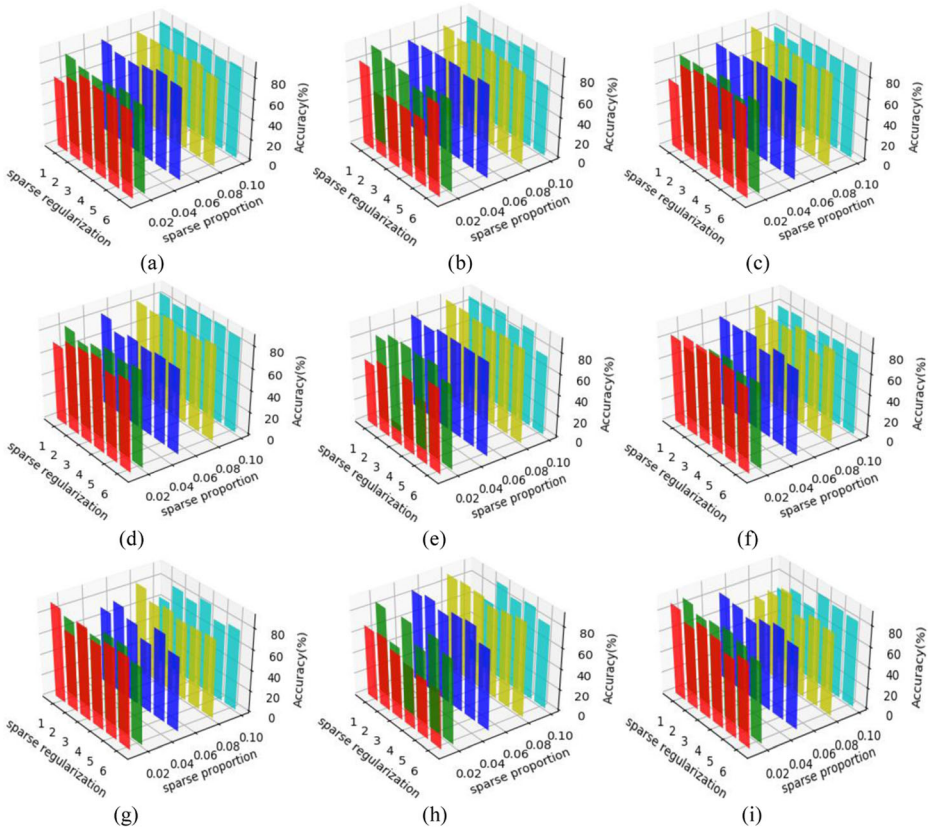


Fig. 8 Effect of sparse proportion, sparse regularization (Λ_2) and L_2 regularization (Λ_1) on accuracy for different number of hidden nodes in a DLNN with single AE (a)250, Λ_1=0 (b)250, Λ_1=0.001 (c)250, Λ_1=0.01(d)100, Λ_1=0 (e)100, Λ_1=0.001 (f)100, Λ_1=0.01 (g)50, Λ_1=0 (h)50, Λ_1=0.001 (i)50, Λ_1=0.01

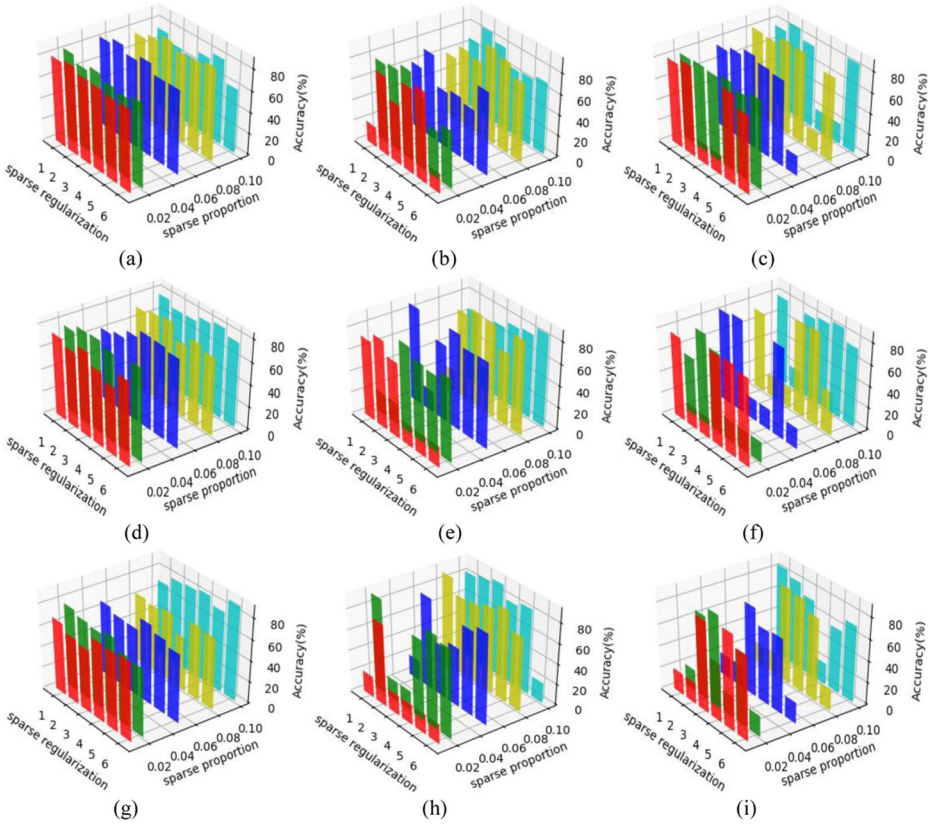


Fig. 9 Effect of sparse proportion, sparse regularization (Λ_2) and L_2 regularization (Λ_1) on accuracy for different combination of hidden nodes in a DLNN with two stacked AE's (a) 250–100, Λ_1=0 (b) 250–100, Λ_1=0.001 (c) 250–100, Λ_1=0.01 (d) 250–50, Λ_1=0 (e) 250–50, Λ_1=0.001 (f) 250–50, Λ_1=0.01 (g) 100–50, Λ_1=0 (h) 100–50, Λ_1=0.001 (i) 100–50, Λ_1=0.01

number of hidden nodes increases, the time complexity will also get increased exponentially. After several experiments are done based on accuracy and computational complexity, it is

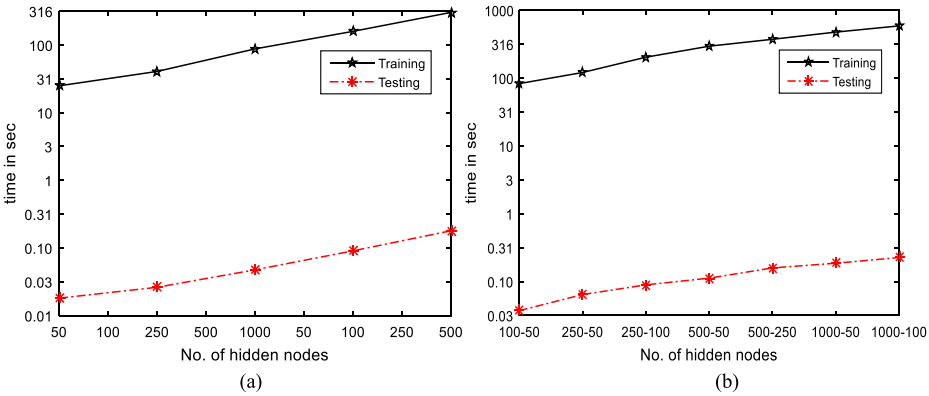


Fig. 10 Computational time Vs Number of hidden nodes in a DLNN with (a) single AE (b)Two stacked AE's

found that 100 and 50 neurons are sufficient at the first and second hidden layers respectively. Interestingly, the time taken for testing the new action sequence is only 0.037 s.

4.4 Performance analysis on UT interaction dataset

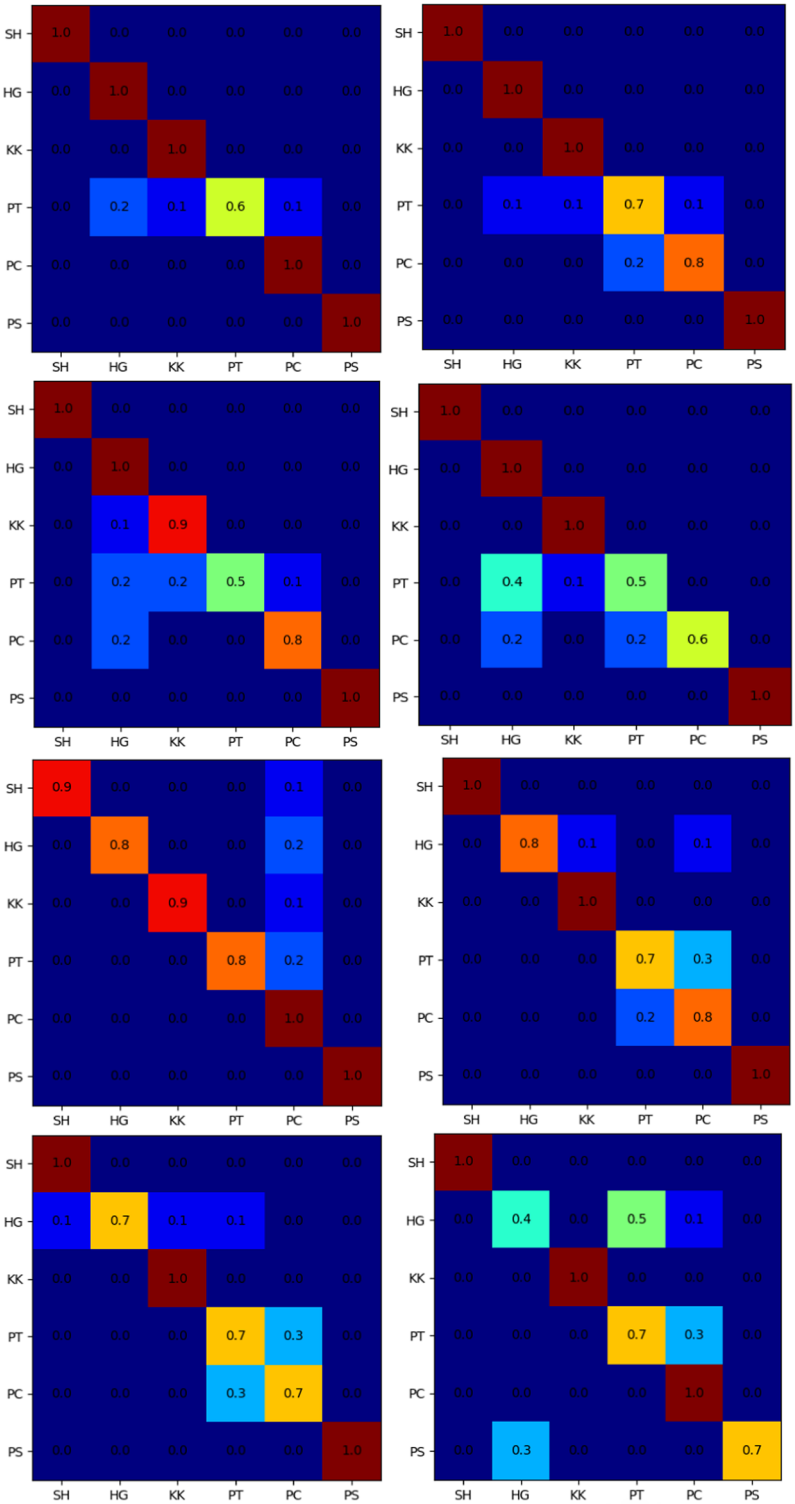
As this dataset contains two sets, the one with static background and the other with dynamic background, the background subtraction and GMM is applied over set1 and set2 respectively. There are total of 10 video sequences for each action. For each iteration, one sequence is used for testing and the remaining sequences are used for training. Figure 11 presents the confusion matrix on UT interaction dataset. From the figure, it is noticed that the recognition accuracy for some actions such as shake hand, hug and push are excellent on UT interaction set1. The actions such as point and punch are considered as the most confusing one. Because, both of the actions involve only the hand movement with little variation. Moreover, the foreground image obtained for UT interaction set2 is not much clear compared to UT interaction set1 because of the cluttered background and occlusion. The movement patterns corresponding to interactions such as kick, punch and push of UT interaction dataset have similarity at some instants, hence they are misclassified.

The performance analysis on UT interaction dataset is depicted in Fig. 12. The best classification accuracy of 93% is attained on set1 and 90% on set2 for $a = 0$. It is 91% and 88% on set1 and set2 for $a = 0.2$ with 40% and 50% of feature reduction respectively. However, the accuracy is much more reduced for $a = 0.5$ because equal importance is given to both feature reduction and classification accuracy. But, it reduces the features for about 52% and 65% on set1 and set2 respectively. Table 1 compares the recognition rate of the proposed method with the results reported in the literature. The results obtained are comparable to the state-of-the-art performances. In set 2, there are few challenges that include cluttered background with moving trees, bystander and camera jitter. The method in [32], requires computation of spatiotemporal interest points and histogram of gradient to generate the feature descriptor followed by K-means clustering to extract the visual words. Similarly in [40], all the interaction sequences are pre-processed in such a way that the main actor who involved in the interaction always stands on the left side. Also, the length of the HOG and HOF descriptors used in [55] are huge i.e. 204 features from single spatio temporal interest point. Despite of these methods, the proposed method significantly gives good performance only with 19 features extracted from single spatio-temporal volume.

4.5 Performance analysis on SBU interaction dataset

This dataset contains both human-human interaction and human-object interaction. Here, the depth images of the sequence is explicitly given, hence thresholding is done to obtain the foreground image. Figure 13 illustrates the confusion matrix on SBU interaction dataset. The diagonal elements are the one that are correctly classified and the rest of the values are the misclassified values. It is shown that the interactions apart, depart, push and shake hand work excellently with 100% classification rate.

Fig. 11 Confusion matrix on UT interaction dataset (a) set1:With PSO ($a = 0$) (b) set1:With PSO ($a = 0.2$) (c) set1: With PSO ($a = 0.5$) (d) set1: Without PSO (e) set2:With PSO ($a = 0$) (f) set2:With PSO ($a = 0.2$) (g) set2:With PSO ($a = 0.5$) (h) set2: Without PSO


SH HG KK PT PC PS
1.0 0.0 0.0 0.0 0.0 0.0
0.0 1.0 0.0 0.0 0.0 0.0
0.0 0.0 1.0 0.0 0.0 0.0
0.0 0.2 0.1 0.6 0.1 0.0
0.0 0.0 0.0 0.0 1.0 0.0
0.0 0.0 0.0 0.0 0.0 1.0
1.0 0.0 0.0 0.0 0.0 0.0
0.0 1.0 0.0 0.0 0.0 0.0
0.0 0.0 1.0 0.0 0.0 0.0
0.0 0.1 0.1 0.7 0.1 0.0
0.0 0.0 0.0 0.2 0.8 0.0
0.0 0.0 0.0 0.0 0.0 1.0
1.0 0.0 0.0 0.0 0.0 0.0
0.0 1.0 0.0 0.0 0.0 0.0
0.0 0.1 0.9 0.0 0.0 0.0
0.0 0.2 0.2 0.5 0.1 0.0
0.0 0.2 0.0 0.0 0.8 0.0
0.0 0.0 0.0 0.0 0.0 1.0
1.0 0.0 0.0 0.0 0.0 0.0
0.0 1.0 0.0 0.0 0.0 0.0
0.0 0.0 1.0 0.0 0.0 0.0
0.0 0.4 0.1 0.5 0.0 0.0
0.0 0.2 0.0 0.2 0.6 0.0
0.0 0.0 0.0 0.0 0.0 1.0
0.9 0.0 0.0 0.0 0.1 0.0
0.0 0.8 0.0 0.0 0.2 0.0
0.0 0.0 0.9 0.0 0.1 0.0
0.0 0.0 0.0 0.8 0.2 0.0
0.0 0.0 0.0 0.0 1.0 0.0
0.0 0.0 0.0 0.0 0.0 1.0
1.0 0.0 0.0 0.0 0.0 0.0
0.0 0.8 0.1 0.0 0.1 0.0
0.0 0.0 1.0 0.0 0.0 0.0
0.0 0.0 0.0 0.7 0.3 0.0
0.0 0.0 0.0 0.2 0.8 0.0
0.0 0.0 0.0 0.0 0.0 1.0
1.0 0.0 0.0 0.0 0.0 0.0
0.1 0.7 0.1 0.1 0.0 0.0
0.0 0.0 1.0 0.0 0.0 0.0
0.0 0.0 0.0 0.7 0.3 0.0
0.0 0.0 0.0 0.3 0.7 0.0
0.0 0.0 0.0 0.0 0.0 1.0
1.0 0.0 0.0 0.0 0.0 0.0
0.0 0.4 0.0 0.5 0.1 0.0
0.0 0.0 1.0 0.0 0.0 0.0
0.0 0.0 0.0 0.7 0.3 0.0
0.0 0.0 0.0 0.0 1.0 0.0
0.0 0.3 0.0 0.0 0.0 0.7

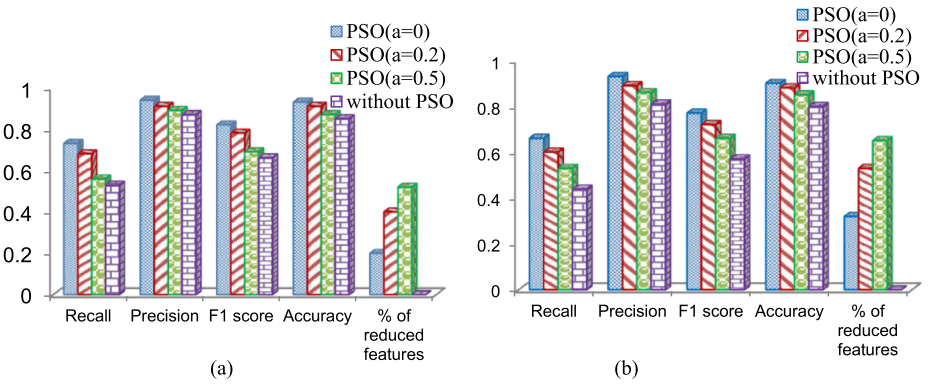


Fig. 12 Performance metrics obtained using DLNN (a) UT interaction set1 (b) UT interaction set2

From the figure, it is clear that almost in all the cases the interaction kick performs worse, because it includes whole body action instead of leg alone. The action punch is sometimes misclassified as push because both of the interactions have similar pose. The different performance metrics obtained on SBU interaction dataset is given in Fig. 14. The figure shows that the multi-objective PSO with a = 0, 0.2, 0.5 gives the classification accuracy of 93%, 91% and 90% respectively. Moreover, the percentage of selected feature that is fed to the classifier is reduced from 70% to 60% and 55% for a = 0.2 and 0.5 respectively.

The results of the recently reported techniques in the literature are also presented for comparison. Compared to Harris corner features, it is noticed that a significant improvement in performance is obtained when combined with DWT. The result of the proposed method is compared to the recently reported techniques as demonstrated through Table 2. Overall, the best performance is achieved for the proposed DWT feature with multi-objective PSO. In the literature, the best reported result is 90% whereas the proposed method achieves an accuracy of 91%.

4.6 Performance analysis on Weizmann dataset

Different from UT and SBU interaction datasets, the Weizmann dataset is meant for single person human action. The foreground mask for the action sequence is explicitly given and is taken for further processing. The confusion matrix on Weizmann dataset is illustrated in

Table 1 Performance on UT interaction dataset in comparison with the state-of-the-art methods

Method	Accuracy (%)	
	Set 1	Set 2
Hong et al. [32]	95	86
Zhang et.al [55]	98	100
Sener et.al [40]	95	91
Nijun et.al [28]	91	85
Proposed-DWT (a = 0.2)	81	75
Proposed-DWT + Harris (a = 0.2)	91	88

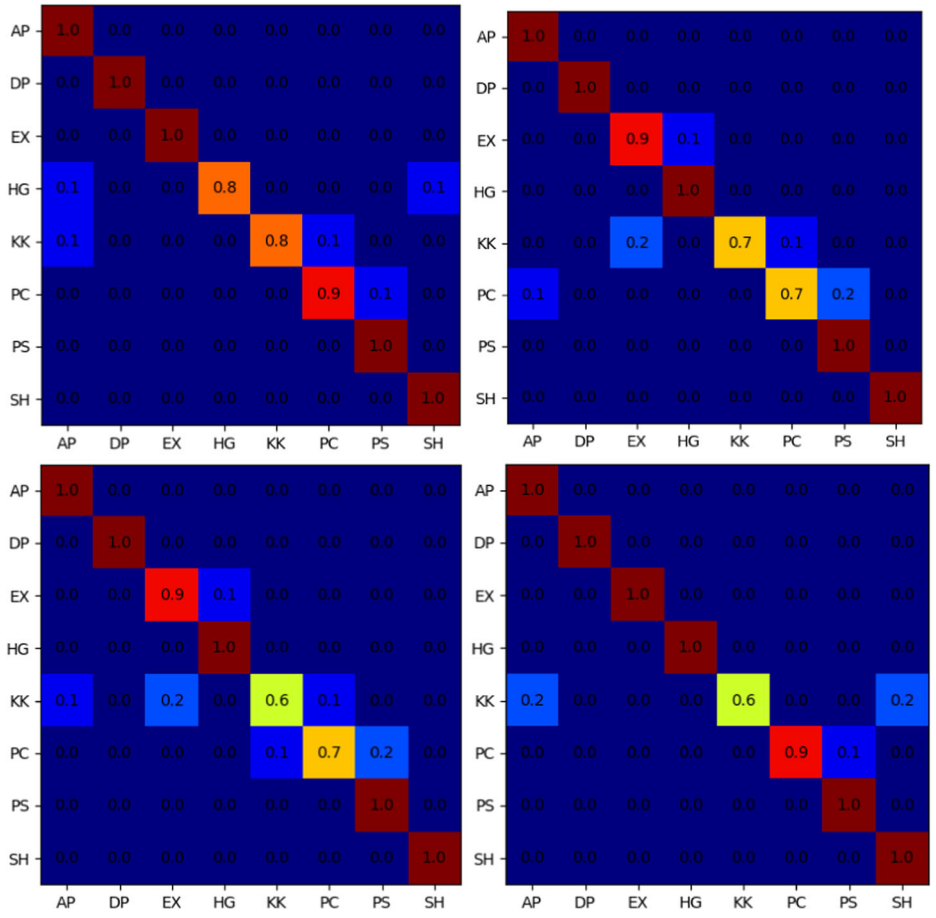


Fig. 13 Confusion matrix on SBU interaction dataset (a) With PSO (a = 0) (b) With PSO (a = 0.2) (c) With PSO (a = 0.5) (d) Without PSO.

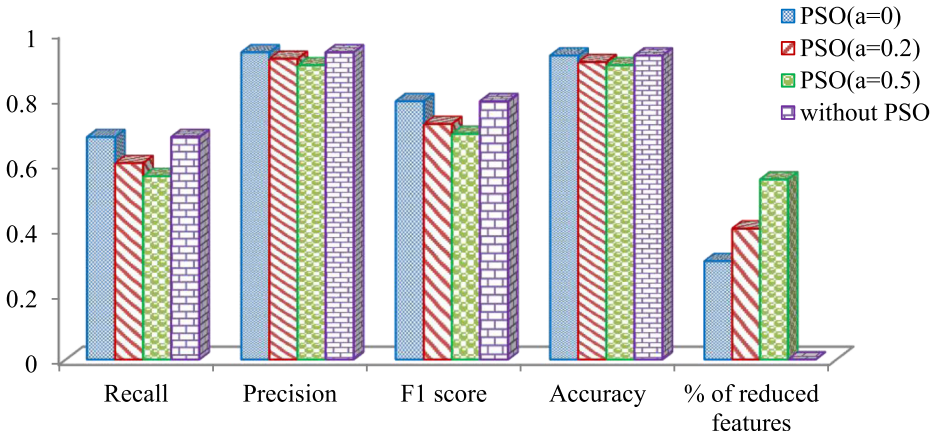


Fig. 14 Performance metrics obtained using deep learning network on SBU interaction dataset

Table 2 Performance on SBU interaction dataset in comparison with the state-of-the-art methods

Method	Accuracy (%)
Thien et al. [16]	90
Wenbo et.al [29]	84
Yanli et.al [19]	87
Yanli et.al[20]	87
Proposed-DWT (a = 0.2)	80
Proposed-DWT + Harris (a = 0.2)	91

Fig. 15. From the figure, it shows that the classification accuracy for some of the actions such as bend, jack, side, skip and wave are high up to 100%. It can be seen that the misclassification occur between the actions jump, run, walk and wave. All of these actions have similar structural and motion pattern, hence it is difficult to discriminate those actions. Therefore, some of the action samples have higher possibility to be recognized among others.

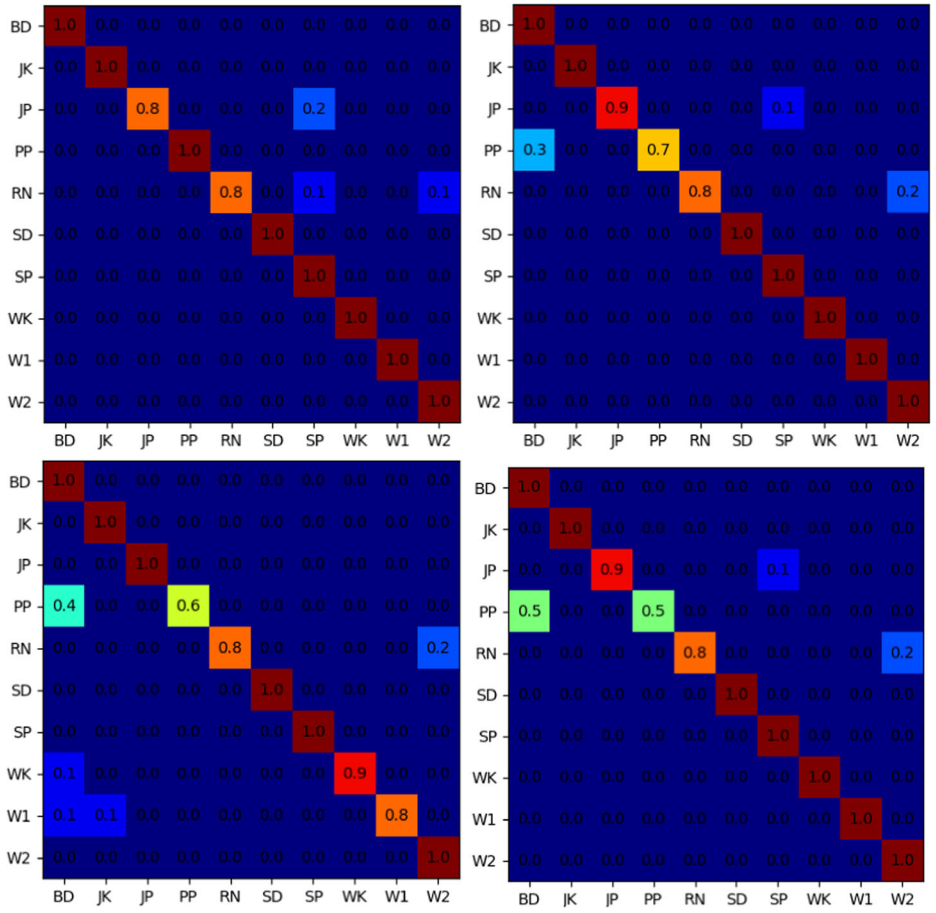


Fig. 15 Confusion matrix on Weizmann dataset (a) With PSO (a = 0) (b) With PSO (a = 0.2) (c) With PSO (a = 0.5) (d) Without PSO.

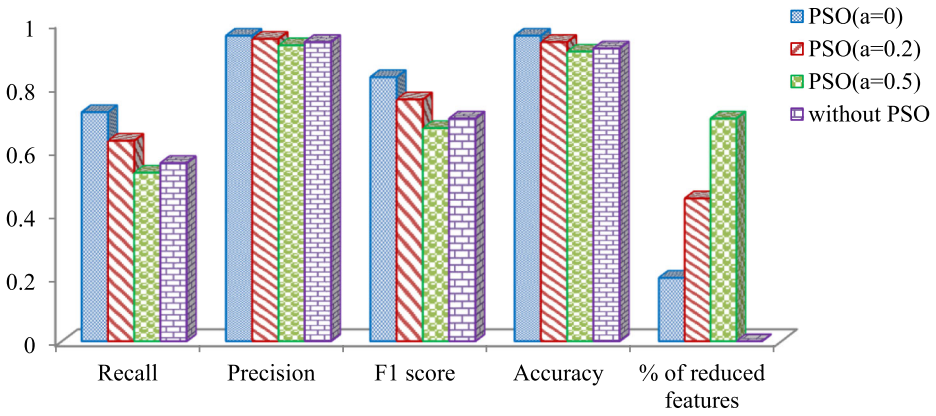


Fig. 16 Performance metrics obtained using DLNN on Weizmann dataset

The different performance metrics obtained on Weizmann dataset is given in Fig. 16. The results show that maximum recognition rate of 96% and 94% are achieved when multi-objective PSO is used for feature reduction with a = 0 and 0.2 respectively. But, the amount of feature selected for classification is reduced from 80% to 55% when multi objective function of a = 0.2 is used rather than the one with single objective function. The multi-objective PSO with a = 0.5 gives the classification accuracy of 91% with 80% reduction in feature set. The comparisons of the proposed work with the previously reported works are illustrated in Table 3.

The proposed work performs favourably against the other methods reported in [2, 55]. But, the proposed work gives comparable performance as compared with the method used in [4]. However, the methods reported in [27,41] produces higher accuracy compared to the proposed one. In [27], the action sequences are decomposed into several saliency action units and each unit is subjected to classification. Moreover, the action sequence is tested with the network trained over other dataset. On the other hand, the method in [41] employs 3D local multi-scale implementation and space-time saliency detection which adds complexity compared to the proposed work.

5 Conclusion

In this work, human action is modelled as a sequence of transient phenomena and characterized using orthogonal wavelet coefficients extracted from the video volume. It is shown that the wavelet feature combined with Harris corner points have attained better discriminative

Table 3 Performance on Weizmann dataset in comparison with the state-of-the-art methods

Method	Accuracy (%)
Saad et.al [4]	95
Maryam et.al [2]	92
Zhang et.al [55]	93
Chang et.al [27]	99
Hae et.al [41]	97
Proposed-DWT (a = 0.2)	88
Proposed-DWT + Harris (a = 0.2)	94

capability. Further, the effect of feature reduction upon classification rate has been studied. The number of feature values is varied based on its significant value through multi-objective PSO and its recognition performance has been investigated. It is found that, the proposed work achieves dimensionality reduction while optimizing the performance of the classifier. This method attains significantly better performance with 50% reduction in the feature set. The experimental evaluations show that this method is computationally simple and is a promising candidate for human action classification.

In future, this method would be further optimized to increase the accuracy of classification. It also include the exploration of different spatial and temporal features which further aid in better recognition of human actions in real time activities. Another interesting direction would be modelling video hashing based human action recognition system which has attracted substantial attentions because of the transformation of higher dimensional feature vectors into compact binary codes with better classification rate.

Acknowledgements The first author is a recipient of DST INSPIRE Fellowship and wishes to thank DST, India for the same.

References

1. Abdelgawad H, Shalaby A, Abdulhai B, Gutub AAA (2014) Microscopic modeling of large scale pedestrian–vehicle conflicts in the city of Madinah, Saudi Arabia. J Adv Transp 48(6):507–525. https://doi.org/10.1002/atr.1201
2. Al-Berry MN, Ebied HM, Hussein AS, Tolba MF (2014) Human action recognition via multi-scale 3D stationary wavelet analysis. In 14th Int Conf on hybrid intelligent systems IEEE Kuwait pp.254–259. https://doi.org/10.1109/HIS.2014.7086208
3. Al-Berry MN, Salem MAM, Ebeid HM, Hussein AS, Tolba MF (2016) Fusing directional wavelet local binary pattern and moments for human action recognition. IET Comput Vis 10(2):153–162. https://doi.org/10.1049/iet-cvi.2015.0087
4. Ali S, Shah M (2008) Action recognition in videos using kinematic features and multiple instance learning. IEEE Trans Pattern Anal 32(2):288–303. https://doi.org/10.1109/TPAMI.2008.284
5. Berlin SJ, John M (2016) Human interaction recognition through deep learning network. In IEEE Int Carnahan Conf on Security Technology USA pp 1–4. https://doi.org/10.1109/CCST.2016.7815695
6. Cai J, Yu J, Imai F, Tian Q (2016) Towards temporal adaptive representation for video action recognition. In IEEE Conf on Image Processing USA pp4155–4159. https://doi.org/10.1109/ICIP.2016.7533142
7. Cheng J, Liu H, Wang F, Li H, Zhu C (2015) Silhouette analysis for human action recognition based on supervised temporal T-SNE and incremental learning. IEEE Trans Image Process 24(10):3203–3217. https://doi.org/10.1109/TIP.2015.2441634
8. Chuang LY, Tsai SW, Yang CH (2011) Improved binary particle swarm optimization using cat fish effect for feature selection. Expert Syst Appl 38(10):12699–12707. https://doi.org/10.1016/j.eswa.2011.04.057
9. Curtis S, Zafar B, Gutub A, Manocha D (2013) Right of way. Vis Comput 29(12):1277–1292. https://doi.org/10.1007/s00371-012-0769-x
10. Gong M, Liu J, Li H, Cai Q, Su L (2015) A multiobjective sparse feature learning model for deep neural networks. IEEE Trans Neural Netw Learn Syst 26(12):3263–3277. https://doi.org/10.1109/TNNLS.2015.2469673
11. Gong M, Zhao J, Liu J, Miao Q, Jiao L (2015) Change detection in synthetic aperture radar images based on deep neural networks. IEEE Trans Neural Netw Learn Syst 27(1):125–138. https://doi.org/10.1109/TNNLS.2015.2435783
12. Gorelick L, Blank M, Shechtman E, Irani M, Basri R (2007) Actions as space-time shapes. IEEE Trans Pattern Anal 29(12):2247–2253. https://doi.org/10.1109/TPAMI.2007.70711
13. Han Y, Zhang P, Zhuo T, Huang W, Zhang Y (2018) Going deeper with two-stream ConvNets for action recognition in video surveillance. Pattern Recogn Lett 107:83–90. https://doi.org/10.1016/j.patrec.2017.08.015

14. Hasan M, Roy-Chowdhury AK (2015) A continuous learning framework for activity recognition using deep hybrid feature models. IEEE Trans Multimedia 17(11):1909–1922. https://doi.org/10.1109/TMM.2015.2477242
15. Huang CL, Dun JF (2008) A distributed PSO-SVM hybrid system with feature selection and parameter optimization. Appl Soft Comput 8(4):1381–1391. https://doi.org/10.1016/j.asoc.2007.10.007
16. Huynh-The T, Banos O, Le BV, Bui DM, Lee S, Yoon Y, Le-Tien T (2015) PAM-based flexible generative topic model for 3D interactive activity recognition. In Proc Int Conf on advanced Technologies for Communications Vietnam pp.117–122. https://doi.org/10.1109/ATC.2015.7388302
17. Imtiaz H, Mahbub U, Schaefer G, Ahad MAR (2013) A multi-resolution action recognition algorithm using wavelet domain features. In 2nd IAPR Asian Conf on pattern recognition IEEE pp.537-541. DOI https://doi.org/10.1109/ACPR.2013.143
18. Ji S, Xu W, Yang M, Yu K (2012) 3D convolutional neural networks for human action recognition. IEEE Trans Pattern Anal 35(1):221–231. https://doi.org/10.1109/TPAMI.2012.59
19. Ji Y, Ye G, Cheng H (2014) Interactive body part contrast Mining for Human Interaction Recognition. In Proc Int Conf on Multimedia and Expo Workshops China pp 1–6. https://doi.org/10.1109/ICMEW.2014.6890714
20. Ji Y, Cheng H, Zheng Y, Li H (2015) Learning contrastive feature distribution model for interaction recognition. J Vis Commun Image Represent 33:340–349. https://doi.org/10.1016/j.jvcir.2015.10.001
21. Ji X, Cheng J, Ta D, Wu X, Feng W (2017) The spatial Laplacian and temporal energy pyramid representation for human action recognition using depth sequences. Knowl-Based Syst 122:64–74. https://doi.org/10.1016/j.knosys.2017.01.035
22. Ji X, Cheng J, Feng W, Tao D (2018) Skeleton embedded motion body partition for human action recognition using depth sequences. Signal Process 143:56–68. https://doi.org/10.1016/j.sigpro.2017.08.016
23. KaewTraKulPong P, Bowden R (2002) An improved adaptive background mixture model for real-time tracking with shadow detection. In: Remagnino P, Jones GA, Paragios N, Regazzoni CS (eds) Video-Based Surveillance Systems. Springer, Boston, pp 135–144. https://doi.org/10.1007/978-1-4615-0913-4_11
24. Kim S, Guy SJ, Hillesland K, Zafar B, Gutub AAA, Manocha D (2015) Velocity-based modeling of physical interactions in dense crowds. Vis Comput 31(5):541–555. https://doi.org/10.1007/s00371-014-0946-1
25. Kong Y, Fu Y (2016) Human interaction recognition using patch-aware models. IEEE Trans Image Process 25(1):167–178. https://doi.org/10.1109/TIP.2015.2498410
26. Kumar SU, Inbarani HH (2017) PSO-based feature selection and neighborhood rough set-based classification for BCI multiclass motor imagery task. Neural Comput & Applic 28(11):3239–3258. https://doi.org/10.1007/s00521-016-2236-5
27. Li C, Yuen PC (2011) A boosted co-training algorithm for human action recognition. IEEE Trans Circ Syst Vid 21(9):1203–1213. https://doi.org/10.1109/TCSVT.2011.2130270
28. Li N, Cheng X, Guo H, Wu Z (2014) A hybrid method for human interaction recognition using Spatio-temporal interest points. In 22nd Int Conf on pattern recognition Washington USA pp.2513–2518 https://doi.org/10.1109/ICPR.2014.434
29. Li W, Wen L, Choo Chuah M, Lyu S (2015) Category-blind human action recognition: a practical recognition system. In Proc IEEE Int Conf on Computer Vision Chile pp 4444–4452. https://doi.org/10.1109/ICCV.2015.505
30. Li J, Wu Y, Zhao J, Lu K (2016) Low-rank discriminant embedding for multiview learning. IEEE Trans Cybernetics 47(11):3516–3529. https://doi.org/10.1109/TCYB.2016.2565898
31. Li J, Jing M, Lu K, Zhu L, Shen HT (2019) Locality preserving joint transfer for domain adaptation. IEEE Trans Image Process 28(12):6103–6115. https://doi.org/10.1109/TIP.2019.2924174
32. Liu H, Liu M, Sun Q (2014) Learning directional co-occurrence for human action classification. In Proc IEEE Int Conf on Acoustic, Speech and Signal Processing Italy pp 1235–1239. https://doi.org/10.1109/ICASSP.2014.6853794
33. Liu M, Liu H, Sun Q (2014) Action classification by exploring directional co-occurrence of weighted Stips. In Proc Int Conf on image processing pp.1460-1464. https://doi.org/10.1109/ICIP.2014.7025292
34. Ma M, Fan H, Kitani KM (2016) going deeper into first-person activity recognition. In Proc IEEE Conf on Computer Vision and Pattern Recognition USA pp 1894-1903. https://doi.org/10.1109/CVPR.2016.209
35. Nikouei SY, Chen Y, Song S, Xu R, Choi BY, Faughnan TR (2018) Real-time human detection as an edge service enabled by a lightweight cnn. In 2018 IEEE Int Conf on EDGE computing (EDGE) San Francisco USA pp.125-129. https://doi.org/10.1109/EDGE.2018.00025
36. Poppe R (2010) A survey on vision-based human action recognition. Image Vis Comput 28(6):976–990. https://doi.org/10.1016/j.imavis.2009.11.014

37. Rapantzikos K, Avrithis Y, Kollias S (2007) Spatiotemporal saliency for event detection and representation in the 3D wavelet domain: potential in human action recognition. In Proc Sixth ACM Int Conf Image and Video Retrieval USA pp 294–301. https://doi.org/10.1145/1282280.1282326
38. Ryoo MS, Aggarwal JK (2010) UT-interaction dataset. ICPR contest on semantic description of human activities (SDHA). IEEE Int Conf on pattern recognition workshops pp.1-6
39. Sargano AB, Wang X, Angelov P, Habib Z (2017) Human action recognition using transfer learning with deep representations. In 2017 Int joint Conf on neural networks (IJCNN) IEEE Anchorage USA pp.463-469. https://doi.org/10.1109/IJCNN.2017.7965890
40. Sener F, Ikizler-Cinbis N (2015) Two-person interaction recognition via spatial multiple instance embedding. J Vis Commun Image Represent 32:63–73. https://doi.org/10.1016/j.jvcir.2015.07.016
41. Seo HJ, Milanfar P (2010) Action recognition from one example. IEEE Trans Pattern Anal 33(5):867–882. https://doi.org/10.1109/TPAMI.2010.156
42. Shin HC, Orton MR, Collins DJ, Doran SJ, Leach MO (2012) Stacked autoencoders for unsupervised feature learning and multiple organ detection in a pilot study using 4D patient data. IEEE Trans Pattern Recognit Mach Intell 35(8):1930–1943. https://doi.org/10.1109/TPAMI.2012.277
43. Siddiqi M, Ali R, Rana M, Hong EK, Kim E, Lee S (2014) Video-based human activity recognition using multilevel wavelet decomposition and stepwise linear discriminant analysis. Sensors 14(4):6370–6392. https://doi.org/10.3390/s140406370
44. Sun L, Jia K, Yeung DY, Shi BE (2015) Human action recognition using factorized Spatio-temporal convolutional networks. In Proc IEEE Int Conf on computer vision Chile pp.4597-4605. https://doi.org/10.1109/ICCV.2015.522
45. Szegedy C, Liu W, Jia Y et al (2015) Going deeper with convolutions. In Proc IEEE Int Conf on Computer Vision and Pattern Recognition (CVPR) Boston USA pp 1–9. https://doi.org/10.1109/CVPR.2015.7298594
46. Tong M, Li M, Bai H, Ma L, Zhao M (2019) DKD–DAD: a novel framework with discriminative kinematic descriptor and deep attention-pooled descriptor for action recognition. Neural Comput & Applic 1-18. https://doi.org/10.1007/s00521-019-04030-1
47. Wang H, Schmid C (2013) Action recognition with improved trajectories. In Proc IEEE Int Conf on computer vision pp.3551-3558. https://doi.org/10.1109/ICCV.2013.441
48. Wang L, Qiao Y, Tang X (2015) Action recognition with trajectory-pooled deep-convolutional descriptors. In Proc IEEE Conf on computer vision and pattern recognition USA pp.4305-4314. https://doi.org/10.1109/CVPR.2015.7299059
49. Wang L, Xu Y, Cheng J, Xia H, Yin J, Wu J (2018) Human action recognition by learning spatio-temporal features with deep neural networks. IEEE Access 6:17913–17922. https://doi.org/10.1109/ACCESS.2018.2817253
50. Weinland D, Ronfard R, Boyer E (2006) Free viewpoint action recognition using motion history volumes. Comput Vis Image Underst 104(2–3):249–257. https://doi.org/10.1016/j.cviu.2006.07.013
51. Xue B, Zhang M, Browne WN (2012) Particle swarm optimization for feature selection in classification: a multi-objective approach. IEEE Trans Cybernetics 43(6):1656–1671. https://doi.org/10.1109/TSMCB.2012.2227469
52. Xue B, Zhang M, Browne WN (2014) Particle swarm optimization for feature selection in classification: novel initialization and updating mechanisms. Appl Soft Comput 18:261–276. https://doi.org/10.1016/j.asoc.2013.09.018
53. Xue B, Zhang M, Browne WN, Yao X (2016) A survey on evolutionary computation approaches to feature selection. IEEE Trans Evol Comput 20(4):606–626. https://doi.org/10.1109/TEVC.2015.2504420
54. Yun K, Honorio J, Chattopadhyay D, Berg TL, Samaras D (2012) Two-person interaction detection using body-pose features and multiple instance learning. In IEEE Computer Society Conf on Computer Vision and Pattern Recognition Workshops Providence RI pp 28–35. https://doi.org/10.1109/CVPRW.2012.6239234
55. Zhang Z, Tao D (2012) Slow feature analysis for human action recognition. IEEE Trans Pattern Anal 34(3):436–450. https://doi.org/10.1109/TPAMI.2011.157
56. Zhang Y, Gong DW, Cheng J (2017) Multi-objective particle swarm optimization approach for cost-based feature selection in classification. IEEE ACM Trans Comput Biol Bioinf (TCBB) 14(1):64–75. https://doi.org/10.1109/TCBB.2015.2476796
57. Zhu L, Shen J, Xie L, Cheng Z (2016) Unsupervised visual hashing with semantic assistant for content-based image retrieval. IEEE Trans Knowl Data Eng 29(2):472–486. https://doi.org/10.1109/TKDE.2016.2562624
58. Zhu L, Huang Z, Liu X, He X, Sun J, Zhou X (2017) Discrete multimodal hashing with canonical views for robust mobile landmark search. IEEE Trans Multimedia 19(9):2066–2079. https://doi.org/10.1109/TMM.2017.2729025

59. Zhu L, Huang Z, Li Z, Xie L, Shen HT (2018) Exploring auxiliary context: discrete semantic transfer hashing for scalable image retrieval. IEEE T Neural Netw Learn Syst 29(11):5264–5276. https://doi.org/10.1109/TNNLS.2018.2797248
60. Ling Shao, Ruoyun Gao, Yan Liu, Hui Zhang, (2011) Transform based spatio-temporal descriptors for human action recognition. Neurocomputing 74 (6):962-973

S. Jeba Berlin received B.E and M.E degree in the Department of Electronics and Communication Engineering from Anna University. She is currently pursuing Ph.D. degree at Anna University, Chennai, India. Her current area of research includes video analytics algorithms for video surveillance.

Mala John did her M.Sc and M.Tech from Indian Institute of Technology (IIT) Madras and IIT Delhi respectively. She did her PhD in the area of wavelets based image denoising in Anna University. She is a Professor in the Department of Electronics Engineering, Madras Institute of Technology Campus of Anna University, India. Her research interests include Communication, Signal Processing & Video Analytics.