


ITC 4/50 Information Technology and Control Vol. 50 / No. 4 / 2021 pp. 686-705 DOI 10.5755/j01.itc.50.4.27845	Novel Machine Learning for Human Actions Classification Using Histogram of Oriented Gradients and Sparse Representation	
	Received 2020/10/12	Accepted after revision 2021/08/19
	 http://dx.doi.org/10.5755/j01.itc.50.4.27845	

HOW TO CITE: Uma Maheswari, B., Sonia, R., Rajakumar, M. P., Ramya, J. (2021). Novel Machine Learning for Human Actions Classification Using Histogram of Oriented Gradients and Sparse Representation. *Information Technology and Control*, 50(4), 686-705. <https://doi.org/10.5755/j01.itc.50.4.27845>

Novel Machine Learning for Human Actions Classification Using Histogram of Oriented Gradients and Sparse Representation

B. Uma Maheswari

Department of Computer Science and Engineering; St. Joseph's College of Engineering, OMR, Chennai- 600 119, Tamilnadu, India; phone: 99400 91240; e-mail: mahespal2002@gmail.com

R. Sonia

B.S. Abdur Rahman Crescent Institute of Science and Technology Chennai- 600 048, Tamilnadu; India; e-mail: sonia.j25@gmail.com

M. P. Rajakumar

Department of Computer Science and Engineering; St. Joseph's College of Engineering; OMR, Chennai- 600 119, Tamilnadu, India; phone: 94440 65025; e-mail: rajranjhu@gmail.com;

J. Ramya

Department of Computer Science and Engineering; St. Joseph's College of Engineering, OMR, Chennai- 600 119, Tamilnadu, India; phone: 99400 91240, e-mail: ramsharsha@gmail.com

Corresponding author: mahespal2002@gmail.com

Recognition of human actions is a trending research topic as it can be used for crucial medical applications like life care and healthcare. In this research, we propose a novel machine learning algorithm for the classification of human actions based on sparse representation theory. In the proposed framework, the input videos are initially partitioned into several temporal segments of a predefined length. From these temporal segments, the key-cuboids are then obtained. These cuboids are obtained based on the locations having maximum variation in orientation. From these regions, key-cuboids are extracted. From the key-cuboids, Histogram of Oriented Gradient (HOG) features are extracted. This new descriptor has the capability to express the dynamic features in the action videos. Using these features, a single shared dictionary is created from the videos belonging to different classes using K-Singular Value Decomposition (K-SVD) algorithm. This dictionary has the combined features of all the action classes. This shared dictionary is generated during the training phase. During the testing phase, the features belonging to a test class is classified using a novel Sparse Representation Modeling based Action Recognition (SRMAR) Algorithm using Orthogonal Matching Pursuit (OMP) and the shared dictionary. The proposed framework was evaluated using popular benchmark action recognition datasets like KTH dataset, Olympic dataset and the Hollywood dataset. The results obtained using these datasets were represented in the form of a confusion matrix. Evaluation was performed using metrics like overall classification accuracy, specificity, precision, recall and F-score that were obtained from the confusion matrix. This system achieved a high specificity of about 99.52%, 99.16% and 96.15% for the KTH dataset, Olympic dataset and the Hollywood datasets, respectively. Similarly, the proposed framework attained very good precision of 97.64%, 90.46% and 73.39% for the KTH dataset, Olympic dataset and the Hollywood datasets, respectively. Also, the average value of recall achieved was 97.58%, 90.86% and 74.09% for the KTH dataset, Olympic dataset and the Hollywood datasets, respectively. It was also observed that the proposed machine learning algorithm achieved outstanding results compared to the existing state-of-the-art human action recognition frameworks in the literature.

KEYWORDS: Histogram of Oriented Gradients, Human action representation, OMP, K-SVD, Classification.

1. Introduction

Human action recognition is a recently trending and a challenging research area. It can be applied in a variety of real-time scenarios like video surveillance, elderly people monitoring, human-computer interaction, rehabilitation, telemedicine, robotics, assistive living, etc. Human action recognition can be broadly categorized into two main categories namely, the video-based [1]–[8] and the wearable sensor-based [9], [10] systems. In [1] multi-level discriminative patches were used for action representation. The classification was done using SVM algorithm.

Action recognition using Kinect depth images was presented in [2]. Here, deep convolutional neural networks were employed for classification. A new dataset based on movie clips was presented in [3]. Here, space time features and spatio-temporal bag of features were extracted from the movie data. Classification was done using non-linear support vector machine. A scheme for action recognition in video using factorized spatio-temporal neural network architecture was presented by Sun et al. in [4].

The global and local temporal structure of video sequences were exploited for action recognition in [5], [37]. In this paper, the 3D-convolution neural network structure was used for classification. A new scheme for pose-based action recognition was proposed by Wang et al. in [32–34]. Here, the skeletal features provided by Kinect depth sensor were used for feature extraction. Classification was done using kernel SVM. Another scheme for action recognition using dense trajectories was proposed in [7]. Here the commonly used SIFT descriptor was used for representing the action sequences. Here, classification was performed using bag-of-features approach. Ji et al. proposed a new CNN model architecture for action recognition in [8]. In this scheme, spatial and temporal features were extracted using convolution operations. Shoaib et al. [28] proposed a scheme for action recognition using two types of sensors namely the smart phone sensors and the wrist-worn sensors. Seven different window sizes were evaluated in this paper.

Brezmes et al. [4] proposed a scheme for action recognition using accelerometer data from mobile phone. Here, actions like walking, climbing down-stairs, climbing up-stairs, sitting, standing and falling were recognized. Sparse based algorithms are also employed in the literature. However, those systems used class-specific dictionaries. That is, one dictionary is created to represent a particular action. The major drawback of this system is that they increase the computational complexity of the system and also consume more space. However, in our work we have generated a single shared dictionary that has the features from all the action classes. This enabled to decrease the computational complexity and also the space requirements.

The wearable sensor-based systems are based on the usage of inertial sensors that is tied to the body of the user. Sensors like accelerometer, gyroscope, magnetometer, orientation sensor etc are used in these systems. Based on the readings from the inertial sensors, the actions performed by individuals are classified. However, the main drawbacks of this system are the possible breakage of the device, battery failure, inconvenience etc. To avoid these issues, video-based systems are popularly being used. In these systems, the video frames are processed and the features are extracted. Based on the extracted features, the human actions are classified. Sparse representation is employed recently as a valuable tool for classification [11], [29]. This makes use of the sparsity component of the real-time signals. Orthogonal basis like wavelet, Fourier basis elements can be used in sparse representation. However, the most commonly used basis is the over-complete basis. In these elements, the numbers of dictionary atoms are greater than the length of each atom.

In [23], Mei et al. proposed a scheme for vehicle classification using sparse representation. The sparsity was optimized using l_1 -regularized least square problem. Here, classification was done using outdoor infrared videos. In [11], [16], [31], the authors presented the usage of sparse representation theory in signal classification applications. The outcome of sparse based classification was compared with discriminative techniques like linear discriminant analysis. Many action recognitions schemes have been proposed in the literature using sparse representation [29]. These schemes make use of two types of dictionaries namely the class-specific or the shared dictionaries [13].

Due to the wide range applicability of human action representation (HAR), development of models that are capable of accurately classifying the actions is a vital task. The main issue faced by these systems is the delay in recognition of actions due to the implementation of complex algorithms. However, in this work, to avoid this issue, we have employed sparse representation (SR) theory for the classification of actions. This SR theory has very high speed of implementation in HAR applications due to the sparsity nature of the video frame data. In this work, a novel framework for human action recognition based on sparse representation is presented. Also, a new technique for the selection of key-cuboids is proposed. We also present a novel Sparse Representation Modeling based Action Recognition (SRMAR) algorithm.

1.1. Motivation and Justification

The main motivation behind this work is to classify human actions. Classification of human actions can be used for a wide range of applications like fall detection, abnormal action detection, rehabilitation, etc. Thus, the main motive behind this work is to develop a robust action classifier that can differentiate between actions performed by humans.

The existing action recognition frameworks employ traditional algorithms like k-nearest neighbor (k-NN) and support vector machine (SVM). These systems do not produce reliable classification results and are also time consuming. Also, the sensitivity of traditional algorithms is very low.

1.2. Outline of the Paper

Hence, in our research we propose a novel scheme for classification of human actions using video sequences and sparse representation theory.

The overall contributions of this paper are fourfold:

- a** A novel framework for human action recognition based on sparse representation is presented.
- b** A new technique for the selection of key-cuboids is proposed.
- c** A novel Sparse Representation Modelling based Action Recognition (SRMAR) algorithm is proposed.
- d** Evaluation is done using KTH dataset, Olympic dataset and the Hollywood dataset.

2. Literature Survey

Guha et al. [10] proposed a new methodology for classifying human actions based on sparse representation theory. In this work, spatio-temporal features were computed for the extraction of discriminative features. These features cuboid features and the local motion pattern features. Key point selection was performed using Harris detection.

These key points were identified for the selection of key-cuboids. To remove noise, Gaussian blurring was performed to the patches in the cuboids. From these de-noised patches, moment matrices were extracted. Using the moment matrices, the local motion pattern features were obtained. Two types of dictionaries were learned from these features namely the class-specific and the shared dictionaries.

These dictionaries were generated using K-means Singular Value Decomposition (K-SVD) algorithm. Classification was done using Orthogonal Matching pursuit (OMP) algorithm. This framework was analyzed using several benchmark video datasets. Alfaró et al. proposed a scheme for action recognition using sparse coding technique [2]. Here, instead of using all the frames for feature extraction, only the key frames were selected. This selection was done using sparse coding methodology. Alternate Direction Method was employed for solving the optimization problem to identify the key frames. Also, from the key-frames, the key sequences were then identified. Sequences that have high score values were selected as the key-sequences in this paper. Relative local temporal features were opted as the suitable features in this method. Inter-class relative descriptors were also employed for classification. The extraction of these features involved three main steps. In the first step, low-level feature representation was done. Here, HOG3D features were extracted. In the next step, local dictionaries were built from the extracted features. The dictionary learning was done by employing the K-SVD [26] algorithm. In the final step, a local similarity descriptor was generated using the previously created dictionaries. Sparse matrices were computed from the generated dictionaries and the feature descriptors using the OMP algorithm. Finally, the video classification was done using sum pooling technique.

Islam et al. presented a comparative study on various action recognition algorithms using skeletal features

[12]. In this work, the depth images and the skeletal data provided by Kinect Microsoft sensors for action recognition were employed for comparative study. This paper also presented a framework for action recognition. In this framework, the depth image sequences were first acquired from the depth cameras. Using these sequences, skeletal data was obtained. Transformation was performed using Euclidean group. Finally, classification was done using SVM. Evaluation was done using UT Kinect action dataset and Florence 3D action dataset. This system achieved an accuracy of 94.95% for the UT Kinect action dataset. The Lie Algebra Absolute pain attained accuracy of 95.96%. The Florence dataset achieved an accuracy of 81.82% for the absolute joint position feature.

Jalal et al. presented a new technique for the segmentation of humans and recognition of their actions using depth images [14]. Here, initially the depth maps were acquired. These maps were pre-processed using background subtraction technique. Using the foregrounds, the silhouettes were extracted. Using these silhouettes, two types of features were extracted. The first feature was the silhouette-oriented features and the second was the body joint based features. These features were grouped based on the k-means clustering algorithm. Using these clusters, each class was modeled using hidden Markov model. That is, code books were created representing each class. Using these models, the actions were recognized based on the maximum likelihood classifier. In particular, the code books that have minimum distance are selected. This technique was evaluated using MSR action recognition dataset. An accuracy of 88.9% was achieved in this paper. The proposed system was also analyzed using the IM daily depth activity dataset. Using this dataset, this methodology attained an accuracy of 66.60%. This system achieved excellent tracking accuracy, segmentation and recognition results.

Zhang et al. proposed a detailed survey on various action recognition techniques using video sequences [39]. In this paper, different analysis was done based on the types of data used, based on the types of features used for action representation and also based on the types of techniques used for classification. Recognition of human-object interaction is also evaluated in this paper. Features like motion history image, motion energy image, SIFT features, histogram of oriented gradient features and spatio-temporal features were

discussed. Papers involving trajectory-based features like improved dense trajectories were analyzed.

Multiple deep learning techniques like 3D convolutional neural networks and long short-term memory techniques used in action recognition were discussed. Types of performance evaluation metrics were also investigated. Different types of datasets available for research in action recognition domain were also listed in this work. It included datasets like Hollywood, HMDB51, UCF50, kinetic and Olympic datasets

A scheme for the recognition of human action using R-transform and Zernike moments was proposed by Dhiman et al. in [8]. Here, initially the depth action sequences are obtained. Using these sequences, binary silhouettes are generated. Two types of features are extracted from these binary images namely the R-transform features and the Zernike moment features. The Zernike moments are obtained for various angles like 0° , 30° , 45° , 60° , 90° and 180° . The scale and rotation invariant properties of the R-transform are also presented in this paper using various illustrations with depth maps. These features are combined to form a shape descriptor. Analysis was performed using UR fall detection dataset. Evaluation was performed using two different classifiers namely the k-NN and SVM classifier. R-transform alone and the combination of R-transform and Zernike features were evaluated. It was seen that; the combination of R-transform with the Zernike features achieved the highest performance. High accuracy of 96.5% was achieved using the combination of these features with the k-NN classifier. SVM classifier attained an accuracy of 95.5%.

A scheme for action recognition based on the combination of video and wearable data was presented by Wei et al. in [35]. Here fusion of the data from two different modalities was performed using feature-level and decision level fusion. Convolutional neural networks were employed for classification. The neural network layer comprised of input layer, convolutional layer, normalization layer, and max pooling layer. Analysis was performed using UTD-MHAD dataset. It was inferred that, the accuracy achieved using video only was 76%. Similarly, the accuracy achieved using inertial data alone was 90.3%. The feature-level fusion using the data from the two modalities attained an accuracy of 94.1%. However, the decision level fusion attained the highest accuracy of 95.6%. It was concluded that usage of combination of two

modalities produces robust results for action recognition compared to a single modality framework.

Khan et al. proposed a scheme for action recognition using multilayer neural networks. Here, initially the problem of poor level of contrast in the foreground of the video frames are enhanced using contrast stretching technique. The CIE images are then transformed to LAB format. Then the luminance channel is alone selected for further processing.

The motion of the video sequence is then estimated. From the motion estimated data, two types of data are extracted. They are the foreground and the saliency map. Then frame segmentation is done using velocity estimation technique. Then threshold-based technique is used for the extraction of foreground. These data are then fused using morphological operations. From the fused data, the shape and texture features are then obtained. HOG and SFTA features are extracted. Suitable features are selected based on the Euclidean distance in between the fused data. Top 500 features with best entropy values are selected in the feature selection process. Finally, classification is done using multi-level perceptron neural network architecture.

Minhas et al. employed extreme learning machines (ELM) for the classification of action data [24]. Here, initially, spatio-temporal features were extracted from the input training videos. In addition, local static features were also extracted. The dimensionality reduction of these features was performed using 2D bi-directional PCA technique. This technique was used for the reduction of 2D matrix data unlike the 1D data reduction using traditional PCA technique. Comparison of PCA and 2D PCA was also done in this paper. The region-of-interest was extracted using motion estimation technique.

Then feature mining was performed using Page Rank methodology. From the data, similarity graph was then constructed. Finally using the constructed graph, vocabulary data was formed. This data was subjected to ELM training using which the final classification was performed. Zemgulys et al. [38] has proposed a scheme for the classification of basketball signals using HOG and SVM classifier. In this work, the sign language of the basket-ball referee was used as the input data. Initially, the data was preprocessed by converting the RGB signal to the black and white domain. Then edges were detected using Sobel opera-

tor. Features were extracted from the edge data. Here, HOG features were used. Analysis was done using different cell sizes for the HOG extraction. This included cell size of [2 2], [4 4] and [8 8]. Finally, classification was done using SVM classifier. Dataset was collected from the YouTube medium. This scheme achieved an accuracy of 97.50%.

Li et al. [20] has proposed a new scheme for the recognition of hand gesture using convolutional neural networks. In this paper, both spatial and temporal features were employed for the hand gesture recognition. Classification was done using a three-dimensional neural network architecture. This system achieved an accuracy of 65.35%. Liu et al. [21] has designed a new system for dynamic hand gesture recognition. Here two-dimensional convolutional neural network structure was employed. Here, the input hand gesture image was initially encoded in the form of a feature vector. This generated feature vector is used for creating a new image, using which classification was performed. This new image possessed the spatio-temporal information of the gesture.

Zheng et al. [41] has proposed a system for the optimization of neural networks. This was done based on stochastic gradient descent algorithm. The functions involved in the network structure were optimized using this algorithm. The learning schedule was designed in a layer-wise manner to increase the speed of optimization. The authors of [42] have presented a paper based on convolutional network pruning. According to this paper, a new pruning technique called Drop-path was introduced. In this technique, for every neural network layer, the influence of neurons was ordered. This helped to achieve a reduced model size. This system was evaluated using two different benchmark classification datasets.

The authors of [43] presented a new system for image classification using data augmentation. Here, the accuracy of neural networks was improved using network optimization. The generalization ability of the entire neural network system was improved using this augmentation technique. This scheme achieved an accuracy of about 93.41% for the coarse-grained dataset.

In [44], the authors have presented a system for increasing the generalization capability of neural network structure using two-stage technique. In this technique, the feature boundary of the neural net-

work structure was optimized. Also, the network was retrained to regularize the generated feature boundary. In this way a two-level training was done to improve the overall performance of the system.

3. Proposed Methodology

The activity diagram of the proposed methodology is shown in Figure 1. In the proposed methodology, the input video sequences belonging to various action classes are segmented in the temporal domain with a temporal length of n . From the temporal segments m key cuboids are obtained. These regions are obtained based on the locations having maximum variation in orientation.

A novel descriptor called Modified Histogram of Oriented Gradient (MHOG) features are then obtained from the key cuboids. These features are then subjected to dimensionality reduction using principle component analysis (PCA). Finally, the features obtained from all the action classes are used for the generation of a single shared dictionary. Using the shared dictionary, classification is performed.

3.1. Identification of Key Cuboids

Using the input video sequences, temporal segmentation is first done. Let n represent the length of each temporal segment. Each of these segments are then segmented into non-overlapping cuboids. Instead of extracting features from all the cuboids, in our work we have used only the key cuboids of action for the feature extraction. This is done to eliminate the regions having minimal changes in order to create discriminative features. Figure 2 shows how the temporal segmentation of action videos is performed.

From the segmented non-overlapping cuboids, the key cuboid is obtained using the following technique. The feature used for the identification of key cuboid is variance. For every cuboid, the variance along the temporal direction for every pixel in the first frame in the cuboid is computed. The sum of the variances for all the pixel locations in first frame is then computed. This gives the value of the total variance across each cuboid. The cuboids that have the highest variance values are then selected as the key cuboids for the feature extraction.

Figure 1
Swimlane Activity diagram for classification of human actions

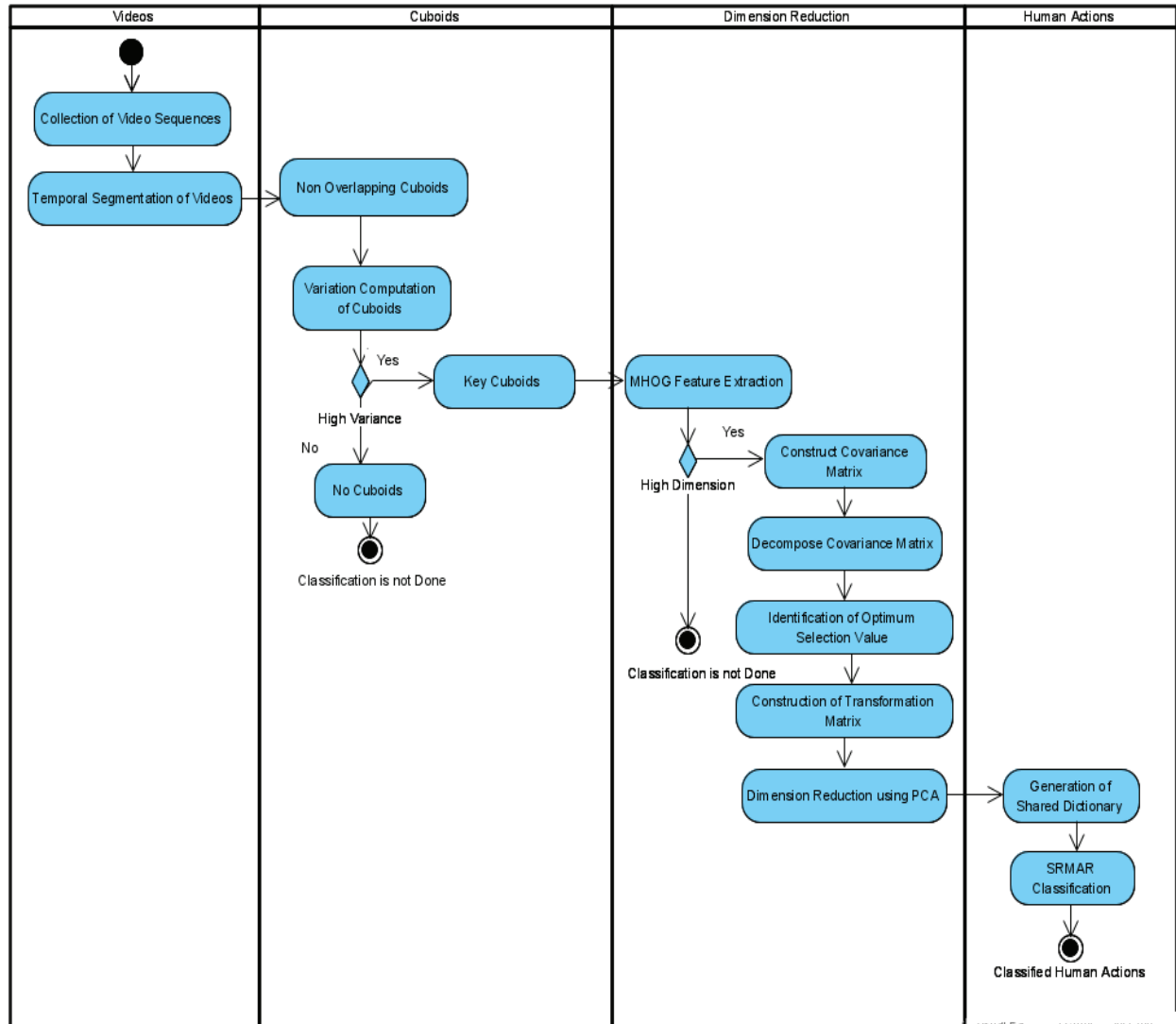


Figure 2
Temporal segmentation of action sequences

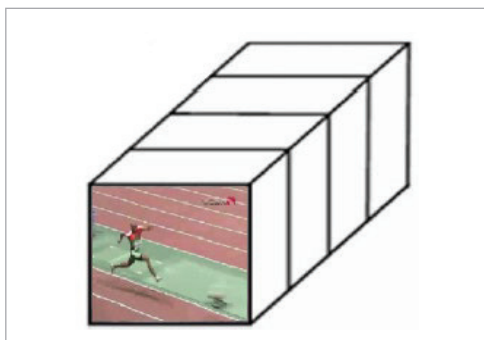
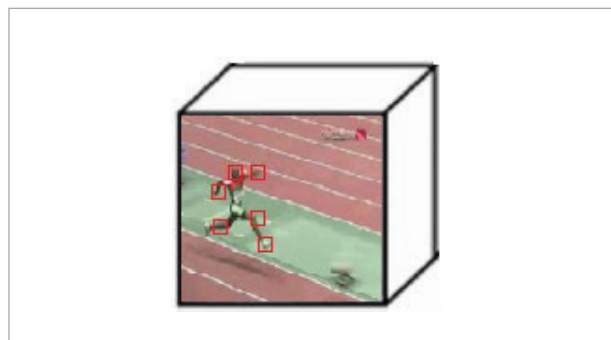


Figure 3
Selection of key cuboids



3.2. Extraction of Modified Histogram of Oriented Gradient (MHOG) Features

In this work, we employed Histogram of Oriented Gradient (HOG) features [7]. In our work we have selected the temporal length =16. Thus, the size of each cuboid is 24×24×16. To obtain the HOG features, these cuboids are stacked together in a horizontal manner to get a rectangular patch of size 24×384. HOG is formed using a cell of size 4×4. Every block has 4 such cells. These blocks are arranged in a 50% overlapping manner.

Thus, the total number of blocks obtained for a patch size of 24×384 turns out to be 5×95=475. In HOG, each block is represented using 4 independent cells and each cell is represented with its 9 orientation bins. Each orientation bin represents a particular angle from 0 to 180. Hence, the total number of HOG features extracted from one key-cuboid is 17100 (475 blocks × 4 cells × 9 bins). We have selected =32 cuboids, hence the size of the feature vector for each video is 32×17100.

3.3. Dimensionality Reduction

Since the dimension of the extracted feature is too high, they must be reduced using dimensionality reduction technique. In this work, we used PCA for dimensionality reduction.

The steps for feature selection using PCA is given below.

Step 1:

Using the feature matrix $f_m \in \mathbb{R}^{p \times q}$, construct the covariance matrix c_v using

$$c_v = \frac{1}{p} ([f_m - \bar{f}_m]^T [f_m - \bar{f}_m]). \tag{1}$$

Here, f_m represents the mean feature matrix and p is the number of features, q is the number of segments and \bar{f}_m is the mean.

Step 2:

Decompose the covariance matrix using eigen value decomposition. This is represented as

$$c_v = \omega \Lambda \omega^T \tag{2}$$

Here $\omega = [v_1, v_2, \dots, v_q] \in \mathbb{R}^{q \times q}$ is a matrix comprising of eigen vectors and the diagonal elements of Λ represents the eigen values ϕ_i where $i \in 1, 2, \dots, q$.

Step 3:

To identify the optimum selection value s out of q eigen values, compute the following

$$\tau = \frac{\sum_{i=1}^s \phi_i}{\sum_{i=1}^q \phi_i} \times 100 \tag{3}$$

Here τ represents the threshold.

Step 4:

The transformation matrix is constructed as

$$T = [v_1, v_2, \dots, v_s], s < q \tag{4}$$

Step 5:

Dimensionality reduction is achieved using

$$x = (f_m - \bar{f}_m) * T. \tag{5}$$

3.4. Classification Using Sparse Representation Modeling

Let x_j represent the input data matrix (dimension reduced feature matrix) of class j . The training data from all the classes are combined to form a single concatenated data matrix using $X = \{x_j, j = 1, 2, \dots, C$, where C refers to the total number of action classes. The input data matrix is trained to form a single shared dictionary using k-singular value decomposition (K-SVD) algorithm [2]. This algorithm is a generalized version of k-means clustering algorithm. It alternated between the dictionary update step and the sparse coding step. K-SVD algorithm establishes the representation of a single in the form if an over-complete dictionary ψ with a dictionary length K . The sparse coding step in K-SVD is based on the optimization of the following problem:

$$\arg \min_{D, \alpha_i} \|\alpha_i\|_0 \text{ s.t. } \|x_i - \psi \alpha_i\|_2 \leq \gamma. \tag{6}$$

In the above equation, y refers to the reconstruction error. Find the solution for the above equation is an NP hard problem [3]. Hence, orthogonal Matching Pursuit (OMP) algorithm [4] which is a greedy optimization algorithm is used for solving the above equation. This algorithm chooses an optimal parameter α_i using the dictionary ψ that produces a minimal residual error γ .

Using training data of each class, calculate the training sparse coefficient matrix using OMP algorithm.

Then, the training sparse histogram vector of each class is formed. Then, the test sparse coefficient matrix is calculated using test data using OMP. The test sparse histogram vector of each class is then formed. The Manhattan distance between the training sparse histogram vector and test sparse histogram vector is then calculated. The Pearson's Correlation between the training sparse histogram vector and test sparse histogram vector is calculated using which the action label is identified

Algorithm: Proposed Sparse Representation Modeling based Action Recognition (SRMAR) Algorithm

Input:

Data matrix of training video $x_j \in \mathbb{R}^{s \times n_j}$

Data matrix of test video x_t

Output:

Action label \mathcal{Y}

Steps:

- Concatenate the data matrix of training video using $X = \{x_j, j = 1, 2, \dots, C\}$
- From the concatenated training data matrix X , generate the over complete dictionary $\psi \in \mathbb{R}^{s \times K}$ using K-SVD.
- Find the training sparse coefficient matrix $\alpha_j \in \mathbb{R}^{K \times n_j}$ using training data of each class x_j using OMP. It can be represented as $\alpha_j = [a_1^j a_2^j \dots a_{n_j}^j]$.
- Form the training sparse histogram vector $H_j \in \mathbb{R}^{K \times 1}$ of each class using $H_j = \sum_{i=1}^{n_j} \alpha_i^j$. It can be represented as $H_j^T = [h_1^j h_2^j \dots h_K^j]$.
- Compute the test sparse coefficient matrix $\alpha_t = [a_1 a_2 \dots a_{n_t}]$ using test data x_t using OMP.
- Form the test sparse histogram vector $H_t \in \mathbb{R}^{K \times 1}$ of each class using $H_t = \sum_{i=1}^{n_t} \alpha_i$. It can be represented as $H_t^T = [h_1 h_2 \dots h_K]$.
- Compute the Manhattan distance between the training sparse histogram vector and test sparse histogram vector using $m(H_j, H_t) = \sum_{i=1}^K |h_i^j - h_i|$.
- Compute the Pearson's Correlation between the training sparse histogram vector and test sparse histogram vector using

$$p(H_j, H_t) = \frac{\sum_{i=1}^K (h_i^j - \bar{h}_i^j)(h_i - \bar{h}_i)}{\sqrt{\sum_{i=1}^K (h_i^j - \bar{h}_i^j)^2} \sqrt{\sum_{i=1}^K (h_i - \bar{h}_i)^2}}, \quad (7)$$

where \bar{h}_i^j and \bar{h}_i represents the mean of h_i^j and h_i , respectively.

Finally, compute the action label y using

$$y = \arg \min_{j=1,2,\dots,C} \frac{m(H_j, H_t)}{p(H_j, H_t)}. \quad (8)$$

The human actions such as walking, jogging, running, boxing and waving are identified using the Jaccard distance between probability density functions for the various dataset. Let X represent the input data matrix (dimension reduced feature matrix) of class j , where $X = \{x_j, j = 1, 2, \dots, C\}$, where C refers to the total number of action classes.

The probability density function is

$$f_h(x) = \frac{1}{Nh} K\left(\frac{x - x_j}{h}\right), \quad (9)$$

K is the kernel used in the probability density function and h is the flattening parameter which is a Gaussian function for the feature set.

4. Results and Discussions

We have evaluated the proposed framework using three publicly available datasets namely the KTH dataset [25], Olympic dataset [27] and the Hollywood dataset [5]. Simulations were performed using MATLAB 12b in Intel Core i3 processor with 4GB RAM.

The value of reconstruction error and dictionary size K was chosen to be 0.01 and 200, respectively in our work. Three datasets were used for analysis namely, KTH, Olympic and Hollywood datasets. The number of classes was 6, 16 and 32 for the KTH, Olympic and Hollywood dataset, respectively.

4.1. Dataset Description

4.1.1 KTH Dataset:

This dataset comprises of 2391 video sequences. It includes 6 different human actions. The actions comprise of walking, jogging, running, boxing, waving and clapping. Sample frame from each action in KTH dataset is shown in Appendix A. All the videos were captured with homogeneous background. Static camera was used for capturing all the sequences. The frame rate of the camera was 25fps. The special resolution of the videos was 160×120.

4.1.2. Olympic Dataset:

The Olympic dataset has 783 video sequences involving 16 different sport-based activities. It contains the

video samples of different athletes performing sporting actions. Sample of each action is shown in Appendix B. All these sequences were obtained from YouTube. It included actions like high jump, long jump, triple jump, pole vault, discus, hammer, javelin, shot put, basket-ball lay-up, bowling, tennis serve, platform, springboard, snatch, clean jerk and vault.

4.1.3. Hollywood Dataset:

The Hollywood dataset involves movie clips from 32 different movies. It has a total of 8 different actions. A sample frame showing each action is shown in Appendix C. The actions include answer phone, get out of car, hand shake, hug person, kiss, sit down, sit up and stand up. These actions were selected from unconstrained videos such as in feature films, sitcoms, or news segments.

These actions were automatically selected from movies based on the movie scripts. The movie scripts contain text description like the scene, dialogs, characters, etc.

4.2. Performance Evaluation

To evaluate the classification performance, metrics like overall accuracy, recall, precision, specificity and F-score were employed [18].

Table 1 shows the comparison of accuracy for the KTH dataset with other state-of-the-art techniques proposed in the literature. From Table 1 we clearly infer that the proposed machine learning classification algorithm based on sparse representation produces the highest classification accuracy. It reaches about 97.61%. This shows the reliability of the proposed classifier.

Table 2 shows the comparison of accuracy for the Olympic dataset with other state-of-the-art techniques proposed in the literature. From Table 2 we clearly infer that the proposed machine learning classification algorithm i.e., SRMAR produces the highest classification accuracy. It reaches about 90.76%. This shows the credibility of the proposed classifier.

Table 1

Comparison of overall accuracy for KTH dataset

References	Methods	Classifier used	Dataset used	Overall accuracy (%)
Laptev et al. [19]	Bag-of-features	SVM	KTH	91.80
Niables et al. [25]	Temporal structure modeling	Latent SVM	KTH	91.30
Sun et al. [30]	Slow feature analysis	Deep learning	KTH	93.10
Castrodad et al. [6]	Deep layer model learning	Sparse representation	KTH	96.30
Alfaro et al. [3]	Spatio-temporal dictionaries	Sparse representation	KTH	95.70
Jaouedi et al. [15]	Gaussian Mixture Model	Deep Learning	KTH	96.30
Proposed SRMAR	SRMAR	Sparse representation	KTH	97.61

Table 2

Comparison of overall accuracy for Olympic dataset

References	Methods	Classifier used	Dataset used	Overall accuracy (%)
Niables et al. [25]	Temporal structure modelling	Latent SVM	Olympic	72.10
Liu et al. [22]	Data-driven attributes	Latent SVM	Olympic	74.40
Jiang et al. [17]	Trajectory motion modelling	SVM	Olympic	80.60
Alfaro et al. [3]	Spatio-temporal dictionaries	Sparse Representation	Olympic	81.30
Gaidon et al. [9]	Motion hierarchies	Cluster tree	Olympic	85.00
Proposed SRMAR	SRMAR	Sparse representation	Olympic	90.76

Table 3

Comparison of overall accuracy for Hollywood dataset

References	Methods	Classifier used	Dataset used	Overall accuracy (%)
Laptev et al. [19]	Bag-of-features	SVM	Hollywood	38.40
Wang et al. [34]	Spatio-temporal modeling	SVM	Hollywood	47.40
Wu et al. [36]	Lagrangian particle advection	SVM	Hollywood	47.60
Sun et al. [30]	Slow feature analysis	Deep learning	Hollywood	48.10
Zhou et al. [45]	Split-and-merge algorithm	SVM	Hollywood	50.50
Proposed SRMAR	SRMAR	Sparse representation	Hollywood	73.05

Table 3 shows the comparison of accuracy for the Hollywood dataset with other state-of-the-art techniques proposed in the literature. From Table 3 we clearly infer that the proposed

SRMAR classification algorithm produces the highest classification accuracy. It reaches about 73.05%. This shows the excellence of the proposed classifier.

To further depict the performance of the proposed machine learning framework, we have compared the results produced by the proposed algorithm with the traditional classification algorithms like K-NN and SVM. Evaluation was performed using the commonly used KTH dataset. Figure 4 shows the comparison of

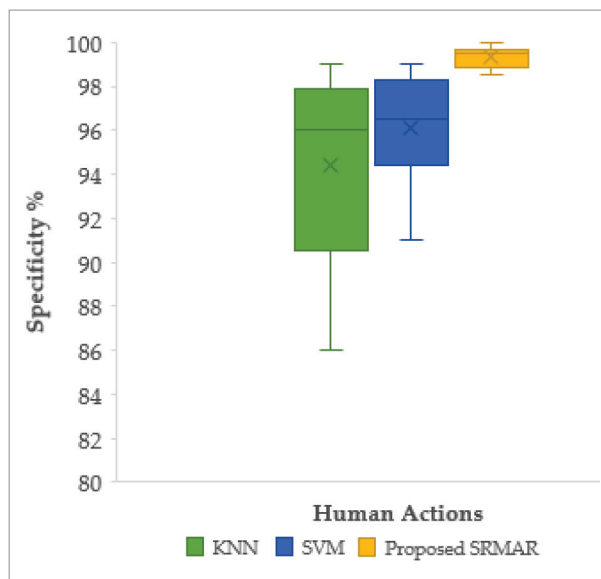
specificity. From Figure 4, it is obviously seen that the proposed algorithm produces highest values of specificity for all the individual actions. The highest specificity achieved is 99.99% for the action Boxing.

Figure 5 shows the comparison of precision. From Figure 8, it is obviously seen that the proposed algorithm produces highest values of precision for all the individual actions. The highest precision achieved is 99.99% again for the action Boxing.

Figure 6 shows the comparison of recall. From the Figure 6 it is obviously seen that the proposed algorithm produces highest values of recall for all the individual actions. The highest recall achieved is 98.87% for the action Jogging.

Figure 4

Specificity Measure of Proposed SRMAR for Human Actions

**Figure 5**

Precision Measure of Proposed SRMAR for Human Actions

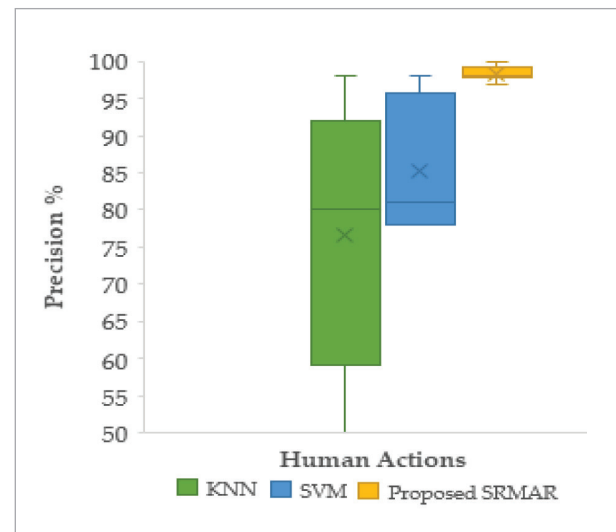


Figure 6
Recall Measure of Proposed SRMAR for Human Actions

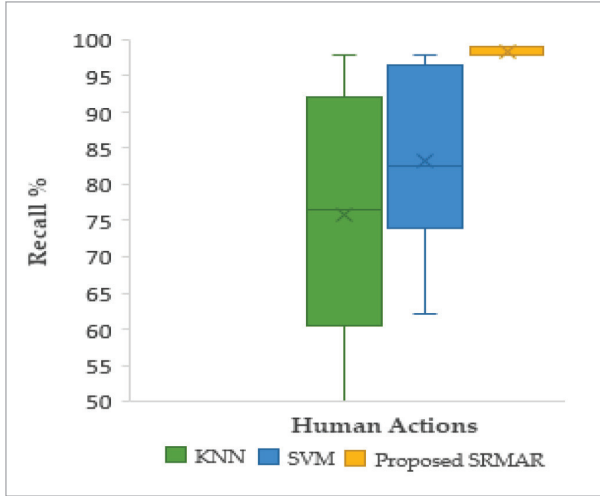
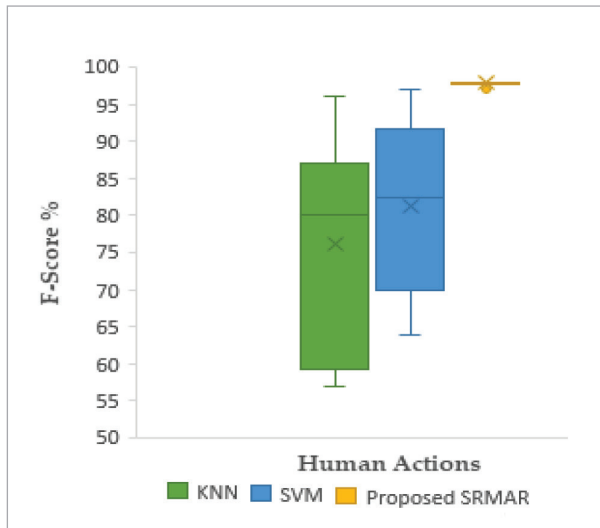


Figure 7 shows the comparison of F-score. Obviously shows that the proposed algorithm produces highest values of F-score for all the individual actions. The highest F-score achieved is 98.29% for the action Boxing.

Figure 7
F-Score Measure of Proposed SRMAR for Human Actions



The performance metric values achieved for each individual action in the KTH dataset is shown in Table 4. We see that our classifier has attained very good performance in terms of almost all the actions.

The performance metric values achieved for each individual action in the Olympic dataset is shown in

Table 4
Performance metrics for each action in KTH dataset

Actions	Specificity (%)	Precision (%)	Recall (%)	F-score (%)
Walking	99.55652	97.87224	97.87224	97.86724
Jogging	99.12279	95.65207	98.87629	97.23246
Running	99.5575	97.84936	97.84936	97.84436
Boxing	99.99998	99.99989	96.66656	98.29998
Waving	99.56894	97.49988	96.29618	96.88929

Table 5. We see that our classifier has attained very good performance in terms of almost all the actions. The highest value of specificity attained is 99.65% for the action Pole Vault. The highest value of precision attained is 95.74% for the action Pole Vault. The highest value of recall attained is 97.97% for the action Discus. And, the highest value of F-score attained is 95.71% for the action Javelin.

Table 5
Performance metrics for each action in Olympic dataset

Actions	Specificity (%)	Precision (%)	Recall (%)	F-score (%)
High jump	99.56634	94.68075	86.40768	90.35025
Long jump	99.2119	91.5887	85.96484	88.68271
Triple jump	99.04097	90.3508	94.49533	92.3716
Pole vault	99.65546	95.74458	94.73674	95.23299
Discus	99.13569	90.65412	97.9797	94.16967
Hammer	99.28379	93.93932	89.20857	91.50785
Javelin	99.29203	93.89306	97.61897	95.71477
Shot put	98.9547	86.66657	72.22216	78.78284
Basketball lay up	98.80749	84.26957	91.4633	87.7142
Snatch	98.70129	86.23845	93.06921	89.51873
Clean jerk	98.89923	84.14624	91.99988	87.89299
Vault	99.39182	93.45786	95.238	94.33453

The performance metric values achieved for each individual action in the Hollywood dataset is shown in

Table 6. We see that our classifier has attained very good performance in terms of almost all the actions.

Table 6

Performance metrics for each action in Hollywood dataset

Actions	Specificity (%)	Precision (%)	Recall (%)	F-score (%)
Answer phone	96.18207	77.39124	87.25482	82.02259
Get out of car	96.57141	74.99992	86.74688	80.44186
Handshake	92.65535	50.94335	71.9999	59.66359
Hug person	96.64334	68.42096	76.47048	72.21714
Kiss	93.22032	63.33328	56.71638	59.83749
Sit down	97.12642	74.35888	66.66659	70.29796
Stand up	97.12989	79.78715	61.98342	69.76246

Figure 8 shows the bar graph that depicts the performance metric values for each action in the KTH dataset. The excellence of the proposed framework is clearly seen from the bar graph.

Figure 8

Evaluation of Human Action Classifier for KTH dataset

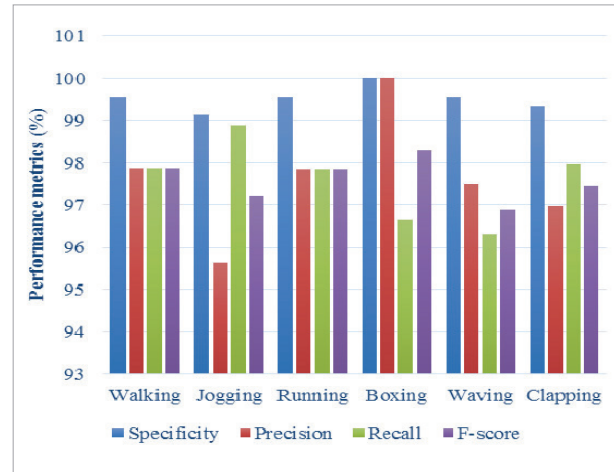


Figure 9 shows the bar graph that depicts the performance metric values for each action in the Olympic dataset. The performance of classification of the proposed framework is clearly seen from the bar graph

Figure 9

Evaluation of Human Action Classifier for Olympic dataset

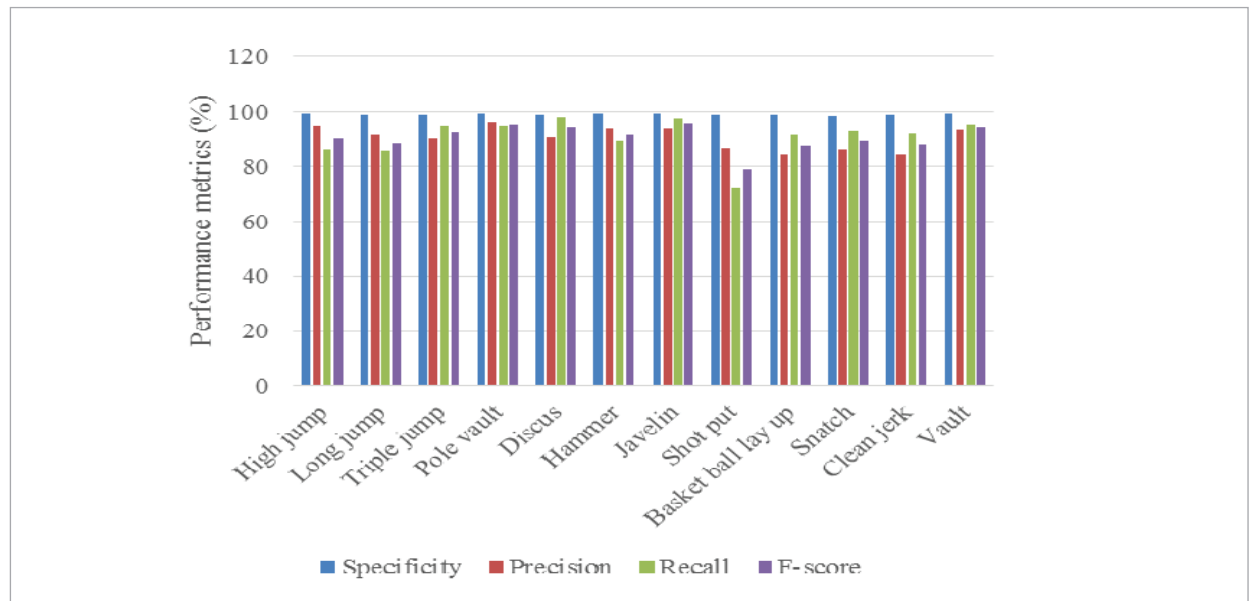


Figure 10 shows the bar graph that depicts the performance metric values for each action in the Hollywood dataset. We clearly infer that our classifier has attained very good performance in terms of almost all

Figure 10

Evaluation of Human Action Classifier for Hollywood dataset

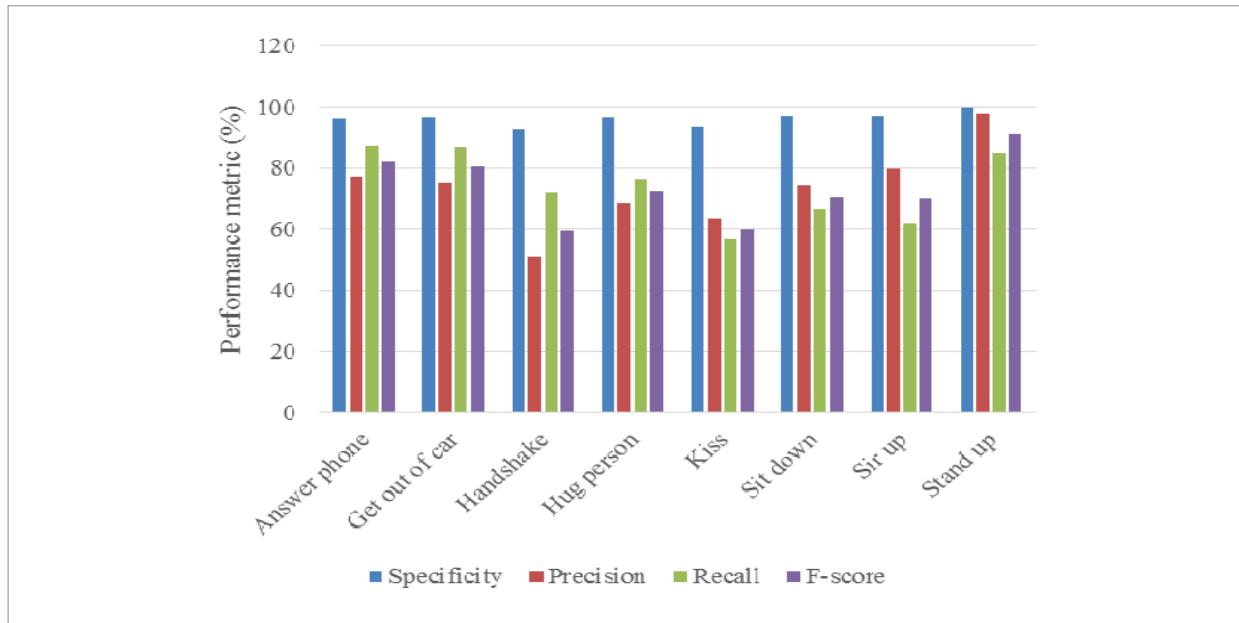


Table 7

Average Performance Measures of Human Action Classifier

Action Dataset	Specificity (%)	Precision (%)	Recall (%)	F-Score (%)
KTH	99.56115	97.77469	97.51213	97.62667
Olympic	99.16173	90.46917	90.86703	90.52276
Hollywood	95.64697	69.89068	72.54835	70.60616
Average %	98.12328	86.04485	86.97584	86.25186

the actions. The highest value of specificity attained is 99.70% for the action Stand up. The highest value of precision attained is 97.95% for the action Stand up. The highest value of recall attained is 87.25% for the action Answer phone. And, the highest value of F-score attained is 90.99% for the action Stand up.

The average performance of the human action classification are listed in Table 7. The data set used in the classification model is imbalanced. The measure G-mean evaluates the proposed SRMAR model on imbalanced data sets and indicates the classification performance. The geometric mean of the human actions classification is shown in Table 8.

$$G\text{-mean}1 = \text{SQRT}(\text{Precision} \times \text{Recall})$$

$$G\text{-mean}2 = \text{SQRT}(\text{Recall} \times \text{TN})$$

The statistical hypothesis test Wilcoxon signed-rank test is used to compare two related samples. Calculate the test statistic

$$W = \sum_{i=1}^n (\text{sgn}(\text{GMean} 2 - \text{GMean} 1) \cdot R_i)$$

sgn is the sign function and R_i denote the rank.

The hypothesis is “There is no significant difference between two means”. The difference between the geometric means are listed in Table 9

The hypothesis is “There is no significant difference between two means”. The difference between the geometric means are listed in Table 8.

Table 8

Geometric mean of SRMAR model using Sparse Representation

S.No	Actions	Specificity (%)	Precision (%)	Recall (%)	F-Score (%)	TN	G-Mean 1	G-Mean 2
1	Walking	100	98	98	98	97	98	98
2	Jogging	99	96	99	97	98	97	99
3	Running	100	98	98	98	92	98	95
4	Boxing	100	100	97	98	95	98	96
5	Waving	100	97	96	97	88	97	92
6	High jump	100	95	86	90	90	90	88
7	Long jump	99	92	86	89	80	89	83
8	Triple jump	99	90	94	92	99	92	97
9	Pole vault	100	96	95	95	96	95	95
10	Discus	99	91	98	94	100	94	99
11	Hammer	99	94	89	92	95	92	92
12	Javelin	99	94	98	96	93	96	95
13	Shot put	99	87	72	79	95	79	83
14	Basketball lay up	99	84	91	88	93	88	92
15	Snatch	99	86	93	90	91	90	92
16	Clean jerk	99	84	92	88	94	88	93
17	Vault	99	93	95	94	99	94	97
18	Answer phone	96	77	87	82	94	82	91
19	Get out of car	97	75	87	80	92	81	89
20	Handshake	93	51	72	60	96	61	83
21	Hug person	97	68	76	72	99	72	87
22	Kiss	93	63	57	60	99	60	75
23	Sit down	97	74	67	70	95	70	80
24	Stand up	97	80	62	70	87	70	74
	Average %	98	86	87	86	94	86	90

Table 9

Difference between Geometric means of SRMAR model using Sparse Representation

i	Actions	G-Mean 1	G-Mean 2	G-Mean 2 - G-Mean 1	
				sgn	abs
1	Walking	98	98		0
2	Jogging	97	99	-1	2
3	Running	98	95	1	3
4	Boxing	98	96	1	2
5	Waving	97	92	1	5
6	High jump	90	88	-1	2
7	Long jump	89	83	1	6
8	Triple jump	92	97	-1	5
9	Pole vault	95	95		0
10	Discus	94	99	-1	5
11	Hammer	92	92		0
12	Javelin	96	95	1	1
13	Shot put	79	83	-1	4
14	Basketball lay up	88	92	-1	4
15	Snatch	90	92	-1	2
16	Clean jerk	88	93	-1	5
17	Vault	94	97	-1	3
18	Answer phone	82	91	-1	3
19	Get out of car	81	89	-1	9
20	Handshake	61	83	-1	22
21	Hug person	72	87	-1	15
22	Kiss	60	75	-1	15
23	Sit down	70	80	-1	10
24	Sit up	70	74	-1	4

Table 10

Ranking of Geometric means of SRMAR model using Sparse Representation

i	Actions	sgn	abs	Ri	sgn.Ri
1	Walking	-	0	-	-
9	Pole vault	-	0	-	-
11	Hammer	-	0	-	-
12	Javelin	1	1	4	4
2	Jogging	1	2	6.5	6.5
4	Boxing	1	2	6.5	6.5
6	High jump	1	2	6.5	6.5
15	Snatch	1	2	6.5	6.5
3	Running	1	3	10	10
17	Vault	-1	3	10	-10
18	Answer phone	-1	3	10	-30
13	Shot put	-1	4	13	-13
14	Basketball lay up	1	4	13	13
21	Sir up	-1	4	13	13
5	Waving	1	5	16.5	16.5
8	Triple jump	-1	5	16.5	-5
10	Discus	-1	5	16.5	-16.5
16	Clean jerk	1	5	16.5	16.5
7	Long jump	1	6	19	19
19	Get out of car	-1	9	20	-20
23	Sit down	-1	10	21	-21
21	Hug person	1	15	22.5	22.5
22	Kiss	1	15	22.5	22.5
20	Handshake	-1	22	24	-22

The above action sequences are arranged in order by absolute difference is shown in Table 10.

From the above table, Wilcoxon signed-rank test = 25.5. $W_{crit} = 89$. Hence $|W| < W_{crit}$; the hypothesis is rejected. So there is two geometric means are same for human actions pair. From the experimental analysis we infer that, this system achieved a high specificity of about 99.52%, 99.16% and 96.15% for the KTH dataset, Olympic dataset and the Hollywood datasets, respectively.

Similarly, the proposed framework attained very good precision of 97.64%, 90.46% and 73.39% for the KTH dataset, Olympic dataset and the Hollywood datasets, respectively. Also, the average value of recall achieved was 97.58%, 90.86% and 74.09% for the KTH dataset, Olympic dataset and the Hollywood datasets, respectively. Moreover, the average value of F-score achieved was 97.59%, 90.52% and 73.15% for the KTH dataset, Olympic dataset and the Hollywood datasets, respectively.

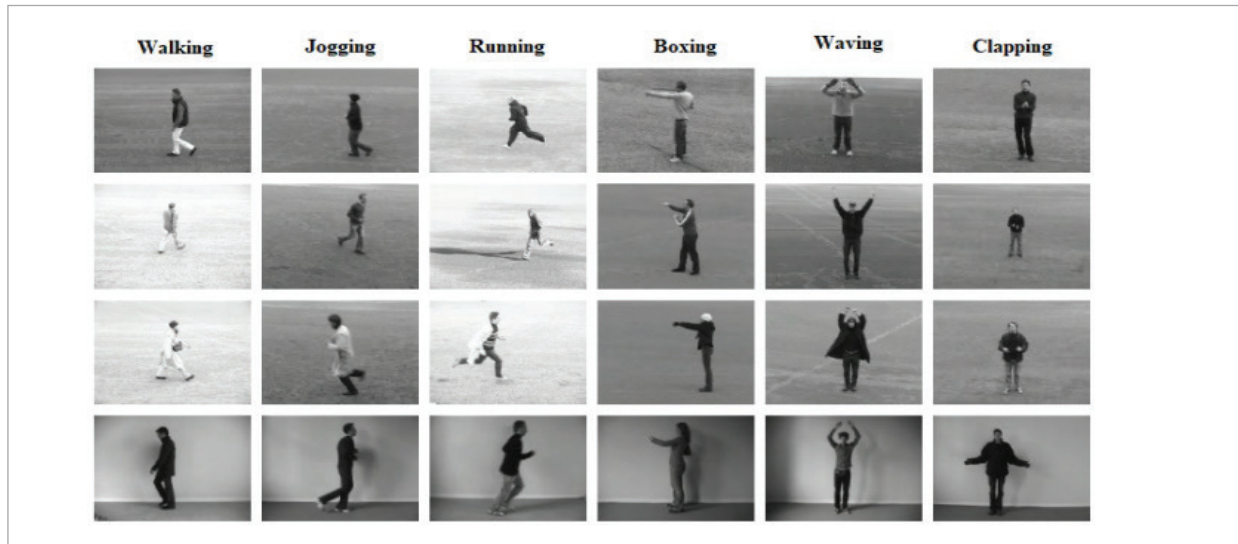
5. Conclusion and Future Work

In this work we presented a novel scheme for human action recognition using video sequences based on sparse representation theory. Here, initially the videos were partitioned into several temporal segments. From the temporal segments, the key-cuboids of interest were then obtained. From the key-cuboids, Histogram of Oriented Gradient (HOG) features were extracted. The dimension of these features was reduced using PCA technique.

Finally, classification was performed using the proposed Sparse Representation Modeling based Action Recognition (SRMAR) Algorithm. This system achieved a high accuracy of about 97.61%, 90.76% and 73.05% for the KTH dataset, Olympic dataset and the Hollywood datasets, respectively. The proposed system was compared with the state-of-the-art action recognition works in the literature. In addition, we also compared our work with the existing traditional classifiers like k-NN and SVM. It was shown that the proposed classifier produces incredible results.

Appendix A

Sample frames from KTH dataset



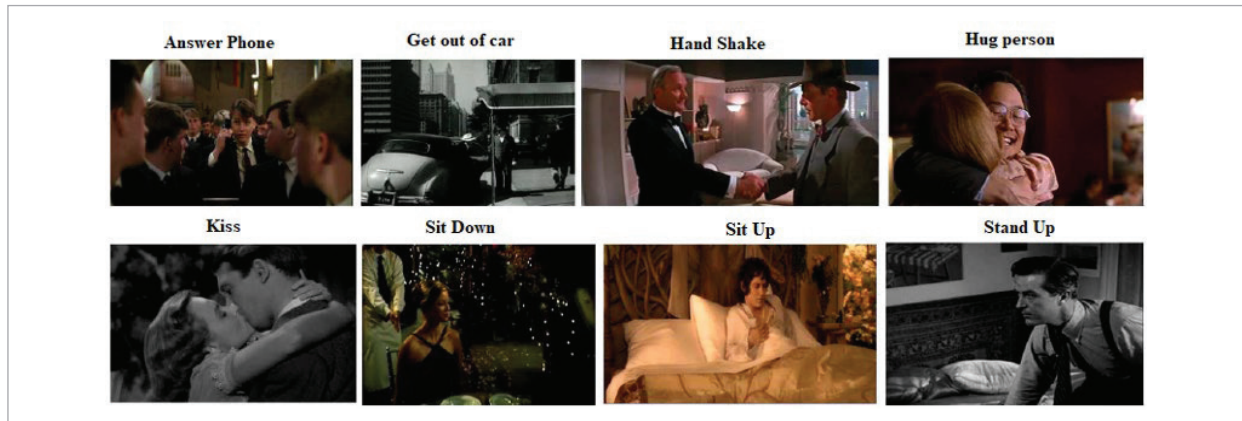
Appendix B

Sample frames from Olympic dataset



Appendix C

Sample frames from Hollywood dataset



References

- Aharon, M., Elad, M., Bruckstein, A. K-SVD: An Algorithm for Designing Over Complete Dictionaries for Sparse Representation. *IEEE Transactions on Signal Processing*, 2006, 54(11), 4311-4322. <https://doi.org/10.1109/TSP.2006.881199>
- Alfaro, A., Mery, D., Soto, A. Action Recognition in Video using Sparse Coding and Relative Features. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2016, 2688-2697. <https://doi.org/10.1109/CVPR.2016.294>
- Alfaro, A., Mery, D., Soto, A. Human Action Recognition from Inter-Temporal Dictionaries of Key-Sequences. *Lecture Notes in Computer Science*, 2014, 419-430. https://doi.org/10.1007/978-3-642-53842-1_36
- Brezmes, T., Gorricho, J. L., Cotrina, J. Activity Recognition from Accelerometer Data on a Mobile Phone. *Lecture Notes in Computer Science*, 2009, 5518(2), 796-799. https://doi.org/10.1007/978-3-642-02481-8_120
- Carreira, J., Zisserman, A. Quo Vadis. Action Recognition? A New Model and the Kinetics Dataset. *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition*, 2017, 6299-6308. <https://doi.org/10.1109/CVPR.2017.502>
- Castrodad, A., Sapiro, G. Sparse Modeling of Human Actions from Motion Imagery. *International Journal of Computer Vision*, 2012, 100(1), 1-15. <https://doi.org/10.1007/s11263-012-0534-7>
- Dalal, N., Triggs, B. Histograms of Oriented Gradients for Human Detection. *Proceedings IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2005, 1, 886-893. <https://doi.org/10.1109/CVPR.2005.177>
- Dhiman, C., Vishwakarma, D. K. A Robust Framework for Abnormal Human Action Recognition using R-Transform and Zernike Moments in Depth Videos. *IEEE Sensors Journal*, 2019, 19(13), 2019, 5195-5203. <https://doi.org/10.1109/JSEN.2019.2903645>
- Gaidon, A., Harchaoui, Z., Schmid, C. Activity Representation with Motion Hierarchies. *International Journal of Computer Vision*, 2014, 107(3), 219-238. <https://doi.org/10.1007/s11263-013-0677-1>
- Guha, T., Ward, R. K. Learning Sparse Representations for Human Action Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(8), 2012, 1576-1588. <https://doi.org/10.1109/TPAMI.2011.253>
- Huang, K., Aviyente, S. Sparse Representation for Signal Classification. *Advances in Neural Information Processing Systems*, 2007, 609-616.

12. Islam, S., Farhad Bulbul, M., Islam, M. S. A Comparative Study on Human Action Recognition Using Multiple Skeletal Features and Multiclass Support Vector Machine. *Machine Learning and Applications: An International Journal*, 2018, 5(1/2), 01-15. <https://doi.org/10.5121/mlaij.2018.5201>
13. Jain, A., Gupta, A., Rodriguez, M., Davis, L. S. Representing Videos using Mid-Level Discriminative Patches. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2013, 2571-2578. <https://doi.org/10.1109/CVPR.2013.332>
14. Jalal, A., Kamal, S., Azurdia-Meza, C. A. Depth Maps-Based Human Segmentation and Action Recognition Using Full-Body Plus Body Color Cues Via Recognizer Engine. *Journal of Electrical Engineering & Technology*, 2019, 14(1), 455-461. <https://doi.org/10.1007/s42835-018-00012-w>
15. Jaouedi, N., Boujnah, N., Bouhleb, M. S. A New Hybrid Deep Learning Model for Human Action Recognition. *Journal of King Saud University-Computer and Information Sciences*, 2020, 32(4), 447-53. <https://doi.org/10.1016/j.jksuci.2019.09.004>
16. Ji, S., Yang, M., Yu, K. 3D Convolutional Neural Networks for Human Action Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2013, 35(1), 221-231. <https://doi.org/10.1109/TPAMI.2012.59>
17. Jiang, Y. G., Dai, Q., Xue, X., Liu, W., Ngo, C. W. Trajectory-Based Modeling of Human Actions with Motion Reference Points. *Lecture Notes in Computer Science*, 2012, 7576(5), 425-438. https://doi.org/10.1007/978-3-642-33715-4_31
18. Khare, M., Gwak, J., Jeon, M. Complex Wavelet Transform-Based Approach for Human Action Recognition in Video. *IEEE International Conference on Control, Automation and Information Sciences 2017*, 157-162. <https://doi.org/10.1109/ICCAIS.2017.8217568>
19. Laptev, I., Marszałek, M., Schmid, C., Rozenfeld, B. Learning Realistic Human Actions from Movies. *IEEE Conference on Computer Vision and Pattern Recognition*, 2008. <https://doi.org/10.1109/CVPR.2008.4587756>
20. Li, J., Yang, M., Liu, Y., Wang, Y., Zheng, Q., Wang, D. Dynamic Hand Gesture Recognition using Multi-direction 3D Convolutional Neural Networks. *Engineering Letters*. 2019, 27(3).
21. Liu, Y., Yang, M., Li, J., Zheng, Q., Wang, D. Dynamic Hand Gesture Recognition using 2D Convolutional Neural Network. *Engineering Letters*. 2020, 28(1).
22. Liu, J., Kuipers, B., Savarese, S. Recognizing Human Actions by Attributes. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2011, 3337-3344. <https://doi.org/10.1109/CVPR.2011.5995353>
23. Mei, X., Ling, H. Robust Visual Tracking and Vehicle Classification Via Sparse Representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2011, 33(11), 2259-2272. <https://doi.org/10.1109/TPAMI.2011.66>
24. Minhas, R., Baradarani, A., Seifzadeh, S., Jonathan Wu, Q. M. Human Action Recognition Using Extreme Learning Machine Based on Visual Vocabularies. *Neurocomputing*, 2010, 73(10-12), 1906-1917. <https://doi.org/10.1016/j.neucom.2010.01.020>
25. Niebles, J. C., Chen, C. W., Fei-Fei, L. Modeling Temporal Structure of Decomposable Motion Segments for Activity Classification. *Lecture Notes in Computer Science*, 2010, 6312(2), 392-405. https://doi.org/10.1007/978-3-642-15552-9_29
26. Rubinstein, R., Zibulevsky, M., Elad, M. Efficient Implementation of the K-SVD Algorithm using Batch Orthogonal Matching Pursuit. *CS Technion*, 2008, 1-15.
27. Schüldt, C., Laptev, I., Caputo, B. Recognizing Human Actions: A Local SVM Approach. *International Conference on Pattern Recognition*, 2004, 3, 32-36. <https://doi.org/10.1109/ICPR.2004.1334462>
28. Shoaib, M., Bosch, S., Incel, O. D., Scholten, H., Haviga, P. J. M. Complex Human Activity Recognition Using Smartphone and Wrist-Worn Motion Sensors. *Sensors*, 2016, 16(4). <https://doi.org/10.3390/s16040426>
29. Sivalingam, R., Somasundaram, G., Bhatawadekar, V., Morellas, V., Papanikolopoulos, N. Sparse Representation of Point Trajectories for Action Classification. *IEEE International Conference on Robotics and Automation*, 2012, 3601-3606. <https://doi.org/10.1109/ICRA.2012.6224777>
30. Sun, L., Jia, K., Chan, T. H., Fang, Y., Wang, G., Yan, S. DL-SFA: Deeply-Learned Slow Feature Analysis for Action Recognition. *IEEE Conference on Computer Vision and Pattern Recognition 2014*, 2625-2632. <https://doi.org/10.1109/CVPR.2014.336>
31. Sun, L., Jia, K., Yeung, D. Y., Shi, B. E. Human Action Recognition using Factorized Spatio-Temporal Convolutional Networks. *IEEE International Conference on Computer Vision*, 2015, 4597-4605. <https://doi.org/10.1109/ICCV.2015.522>
32. Wang, C., Wang, Y., Yuille, A. L. An Approach to Pose-Based Action Recognition. *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2013, 915-922. <https://doi.org/10.1109/CVPR.2013.123>

33. Wang, H., Kläser, A., Schmid, C., Liu, C. L. Action Recognition by Dense Trajectories. *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2011, 3169-3176. <https://doi.org/10.1109/CVPR.2011.5995407>
34. Wang, H., Ullah, M. M., Kläser, A., Laptev, I., Schmid, C. Evaluation of Local Spatio-Temporal Features for Action Recognition. *British Machine Vision Conference*, 2009. <https://doi.org/10.5244/C.23.124>
35. Wei, H., Jafari, R., Kehtarnavaz, N. Fusion of Video and Inertial Sensing for Deep Learning - Based Human Action Recognition. *Sensors*, 2019, 19(17), 1-13. <https://doi.org/10.3390/s19173680>
36. Wu, S., Oreifej, O., Shah, M. Action Recognition in Videos Acquired by a Moving Camera using Motion Decomposition ff Lagrangian Particle Trajectories. *IEEE International Conference on Computer Vision*, 2011, 1419-1426. <https://doi.org/10.1109/ICCV.2011.6126397>
37. Yao, L., Torabi, A., Cho, K., Ballas, N. Describing Videos by Exploiting Temporal Structure. *IEEE International Conference on Computer Vision*, 2015, 4507-4515. <https://doi.org/10.1109/ICCV.2015.512>
38. Žemgulys, J., Raudonis, V., Maskeliūnas, R., Damaševičius, R. Recognition of Basketball Referee Signals from Videos using Histogram of Oriented Gradients (HOG) and Support Vector Machine (SVM). *Procedia Computer Science*, 2018, 130, 953-960. <https://doi.org/10.1016/j.procs.2018.04.095>
39. Zhang, H. B., Zhang, Y. A Comprehensive Survey of Vision-Based Human Action Recognition Methods. *Sensors*, 19(5), 2019, 1-20. DOI: 10.3390/s19051005. <https://doi.org/10.3390/s19051005>
40. Zhang, H., Nasrabadi, N. M., Yanning Zhang, Huang, T. S. Multi-Observation Visual Recognition via Joint Dynamic Sparse Representation, *IEEE International Conference on Computer Vision*, 2011, 595-602. <https://doi.org/10.1109/ICIP.2011.6116301>
41. Zheng, Q., Tian X., Jiang N., Yang, M. Layer-Wise Learning based Stochastic Gradient Descent Method for the Optimization of Deep Convolutional Neural Network. *Journal of Intelligent & Fuzzy Systems*, 2019, 37(4), 5641-54. <https://doi.org/10.3233/JIFS-190861>
42. Zheng, Q., Tian, X., Yang, M., Wu, Y., Su, H. PAC-Bayesian Framework Based Drop-Path Method for 2D Discriminative Convolutional Network Pruning. *Multidimensional Systems and Signal Processing*, 2019, 1-35. <https://doi.org/10.1007/s11045-019-00686-z>
43. Zheng, Q., Yang, M., Tian, X., Jiang, N., Wang D. A Full Stage Data Augmentation Method in Deep Convolutional Neural Network for Natural Image Classification. *Discrete Dynamics in Nature and Society*, 2020, 1-11. <https://doi.org/10.1155/2020/4706576>
44. Zheng, Q., Yang, M., Yang, J., Zhang, Q., Zhang, X. Improvement of Generalization Ability of Deep CNN via Implicit Regularization in Two-Stage Training Process. *IEEE Access*, 2018, 15844-15869. <https://doi.org/10.1109/ACCESS.2018.2810849>
45. Zhou, Z., Shi, F., Wu, W. Learning Spatial Temporal Extents of Human Actions for Action Detection. *IEEE Transactions on Multimedia*, 2015, 17(4), 512-525. <https://doi.org/10.1109/TMM.2015.2404779>

