Journal of ELECTRICAL ENGINEERING, VOL 70 (2019), NO6, 443–453

sciendo
PAPERS

Human action recognition using descriptor based on selective finite element analysis

Rajiv Kapoor[*], Om Mishra[*], Madan Mohan Tripathi[**]

This paper proposes a novel local descriptor evaluated from the Finite Element Analysis for human action recognition. This local descriptor represents the distinctive human poses in the form of the stiffness matrix. This stiffness matrix gives the information of motion as well as shape change of the human body while performing an action. Initially, the human body is represented in the silhouette form. Most prominent points of the silhouette are then selected. This silhouette is discretized into several finite small triangle faces (elements) where the prominent points of the boundaries are the vertices of the triangles. The stiffness matrix of each triangle is then calculated. The feature vector representing the action video frame is constructed by combining all stiffness matrices of all possible triangles. These feature vectors are given to the Radial Basis Function-Support Vector Machine (RBF-SVM) classifier. The proposed method shows its superiority over other existing state-of-the-art methods on the challenging datasets Weizmann, KTH, Ballet, and IXMAS.

Keywords: finite element analysis (FEA), stiffness matrix, discretization, support vector machine

1 Introduction

Over the last few decades, it has been observed that computers have transformed human life in almost every possible aspect. Along with the latest transformations, video data has become easily accessible and dominant in the present time. Every new reform has enabled hardware devices like mobile phones, tablets, digital cameras to create, store and share videos. The increasing number of accessible videos has also created the need to understand them. This idea has led to an extensive study of videos to recognize the action. Action recognition has its major application in the field of medical, sports and security.

Researchers modeled action features globally as well as locally. Global features rely on the localization of individual whose action is to be perceived. Localization is done through background subtraction or human tracking. The 2D template methods use 2D silhouettes for global representation [1, 2]. At this point Hu moments, Radon transform descriptors are also utilized to signify the activities. Global features can also be represented as space-time volumes. Spatial-temporal volume is made by stacking silhouettes over a given grouping [3, 4]. Global features may also represent motion information with the assistance of optical flow. It doesn't require background subtraction. An optical flow-based strategy where the movement of pixels have been contemplating is also used in global feature extraction [5, 6]. The disadvantage of this method is that it is very sensitive to noise because the motion descriptor can be corrupted due to noise that appeared in a dynamically changing background. Primary inconveniences of global features portrayal are that it unpleasantly depends on precise localization and background subtraction, so it is sensitive to a viewpoint and individual appearance. Likewise, it cannot give motion information of an activity which makes it sub-par compared to similar kinds of activities like jogging and running.

Local feature portrayal is utilized more frequently in recent times. It doesn't require precise restriction and background subtraction and furthermore demonstrates invariance in viewpoint and individual appearance. Local spatiotemporal descriptors are based on the bag-of-words model [7-9]. HOG/HOF, HOG3D, SURF, and MoSIFT are some major descriptors [10-13]. The principal burden of this portrayal is that they can't give basic/shape information. Its aftereffect is that it can give the same component descriptor for various activity classes. Furthermore, in recent years, a paradigm based on deep learning techniques is also very popular in the research community for human action recognition [14-16]. Unlike the handcrafted approaches discussed above the deep learning, the technique is fully automated. The efficiency of deep learning methods depends upon the design of the network. These methods need large datasets and parameters, on account of this, the complexity of the structure increases. Researchers are working on deep learning techniques to overcome this problem.

Silhouette analysis-based methodologies contributed significantly to human action recognition. They are used to find out both global and local features [17- 21]. In [17] and [18] silhouettes analysis is used to extract the global and local features to represent action video. [19] represented the local feature as pose correlogram and extended motion history image is used for global features.

[*]Department of Electronics & Communication, Delhi Technological University, Bawana Road, Delhi-110042, India, [**]Department of Electrical Engineering, Delhi Technological University, Bawana Road, Delhi-110042, India, mmtripathi@dce.ac.in

DOI: 10.2478/jee-2019–0077, Print (till 2015) ISSN 1335-3632, On-line ISSN 1339-309X

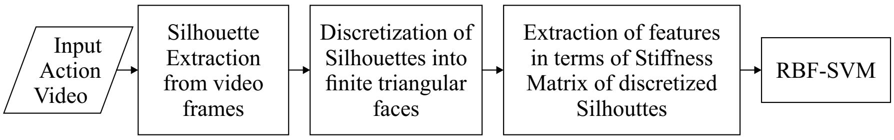

Input Action Video
Silhouette Extraction from video frames
Discretization of Silhouettes into finite triangular faces
Extraction of features in terms of Stiffness Matrix of discretized Silhouttes
RBF-SVM

Fig. 1. Workflow diagram of the proposed method

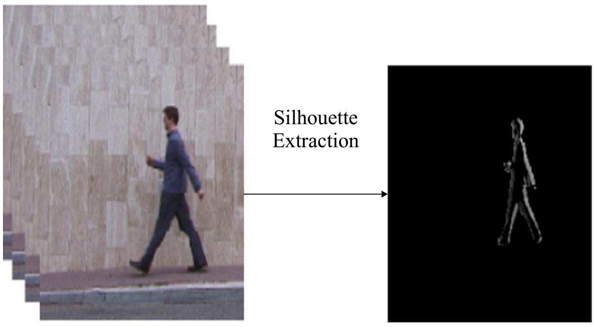


Fig. 2. Walk and the extracted silhouette

The three-dimensional Histogram of the oriented gradient is used to represent the action video in [20]. The method [21] proposed a new feature based on the negative space which is related to the region of the surrounding of the subjects.

The human body poses represent an action effectively. Multi-View key poses from the silhouette are extracted in [22]. They modeled the actions with spatiotemporal features. In [23] the binarized silhouette is used to find out the trace transform to represent the global feature of action The sequence of a silhouette is represented as the cube video to model the action in [24]. The multi-scale volumetric approach for action videos is used in [25, 26]. The action is modeled using sparse coding of image sequences in [26]. The silhouette-based analysis is also used in deep learning-based methodologies [27, 28, and 29]. The CNN and HMM are combined to represent long action video in [27]. The methodology used in [28] represented the actions through a neural mechanism. The two cortical areas, the primary cortex, and the middle cortex are used to extract the motion features. To capture complete motion information, [29] proposed a new descriptor that is capable of static, short term motion as well as long term motions.

Motivation: This paper introduces a new feature descriptor based on Finite Element Analysis (FEA) [30-31]. FEA has been used as a very powerful technique for the structural analysis of the system. In the FEA technique, the structure is converted into a finite number of elements. Wherever any deformation occurs in a body/structure these finite elements also get displaced from their previous position and the stiffness matrix of these elements shows how stiff the body/structure is against this deformation. This gives accurate and precise information about the structural deformation of the body. Similarly, when a person performs an action, his body gets deformed in different patterns. This motivates us to apply the concept of FEA on the silhouettes extracted from the action video. The proposed method offers a new local features descriptor that is solely capable of representing shape as well as motion features of the silhouette.

2 Methodology of proposed framework

The proposed methodology can be described by the Fig. 1. The human silhouette is extracted from the frames of the action video. Then we discretized the silhouette of the human body into several finite elements (triangle faces). Then complete stiffness matrix of the silhouette is calculated by using FEA. The stiffness matrices are represented as feature vectors. The RBF-SVM classifier is used for the classification of actions.

2.1 Silhouette extraction

In the proposed method, features are acquired to distinguish different human actions which make the result more authentic. These features describe the deformation that occurs in the silhouette in terms of shape and motion information while performing an action. As silhouette moves, the finite elements also get deformed. The stiffness matrix of the silhouette narrates these features. The first step of the proposed method is silhouette extraction which is also a very challenging problem because it requires background subtraction. Background cluttering, illumination change, noises, etc. are some challenges for background subtraction. The GMM [32] is robust to problems discussed above and it also has the capability to deal with the critical issue like a shadow. We used GMM for background subtraction. Then the silhouette is extracted and normalized so that all the silhouettes become equal in size [19]. Fig. 2 shows the extracted silhouette from the video.

2.2 Discretization and shape function representation

The preliminary step for FEA is discretization i. e. modeling the silhouette structure into numbers of small elements as shown in Fig. 3(a). The number of elements in which geometry is divided is variable and can be determined by software like MATLAB, COMSOL, *etc.* which demand physics of the geometry. We selected the simple triangular element as the finite element. The reason behind using the triangular shape is that it is the simplest structure for numerical representation. We used MATLAB having an FEA toolbox in the proposed method.

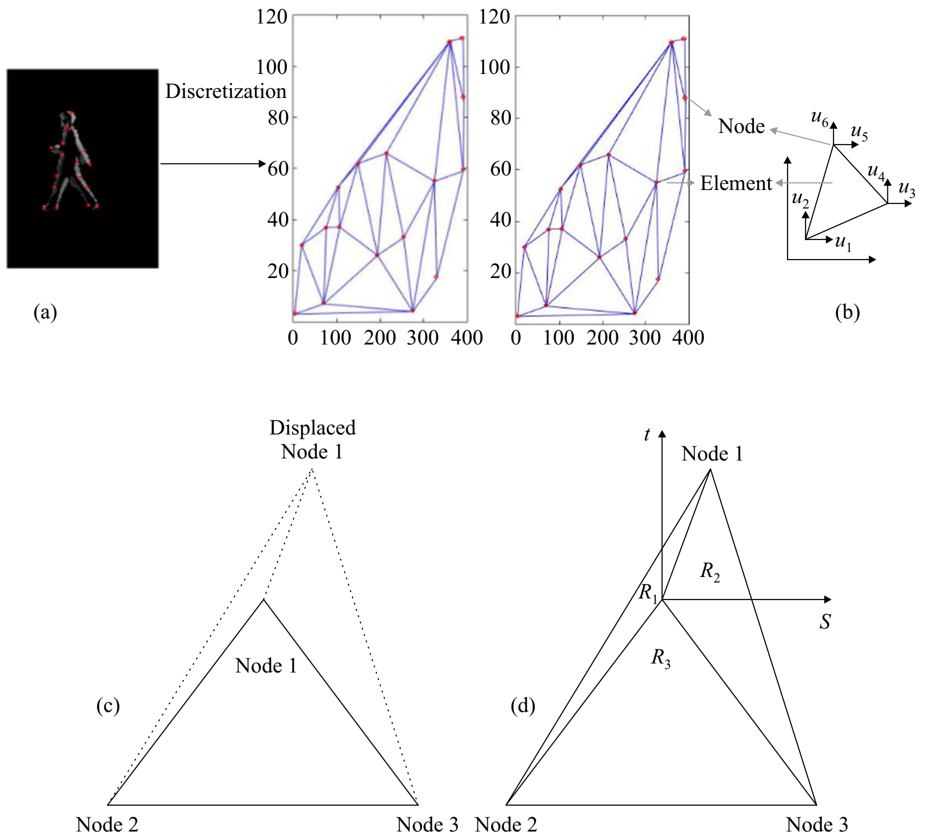


Fig. 3. (a) – segmented discretized silhouette, (b) – displacement of a mesh (triangle) of the human body and its mesh element, (c) – deformed triangle as node 1 is displaced, (d) – triangle divided into 3 parts

We referred Laptev *et al* [10] to find out the prominent points at the boundary of the silhouette. These prominent points are given as the nodes to the FEA toolbox. The silhouette is discretized into the finite triangular elements using these nodes as vertices of the triangle. The discretization of the silhouette is done in such a way that they do not overlap.

In Fig. 3(a) silhouette is discretized into finite triangular elements. The discretized structural representation of the silhouette is shown on the right side of Fig. 3(a) where X -axis and Y -axis are the spatial coordinates of the vertices of the triangle. The complete domain is divided into simpler parts; it provides precise and detailed representation for analysis. Each triangular element has three nodes. Every node has a displacement in the X direction and Y direction as shown in Fig. 3(b). Displacement vectors of the triangle can be represented as

$$U = \{u_1, u_2, u_3, u_4, u_5, u_6\}^\top \tag{1}$$

where, U is a displacement vector for each triangular element and u_1 and u_2 are the displacement of node 1 in X and Y direction respectively, u_3 and u_4 are the displacements of node 2 in X and Y direction and similarly u_5 and u_6 are the displacements of node 3.

When silhouette moves from one frame to another frame, nodes of the triangles are also displaced. To build the face correspondence between the frames we calculated the Euclidean distance among the nodes of the previous frame and next frame. The minimum distance will correspond to the same points. The displacement of the nodes of the triangle is found out from one frame to another frame. Shape function [30,31] of the triangle is used to represent the nodal displacement. Fig. 3(c) shows the displacement of node 1 as a dotted line from its previous point to the displaced point, which results in the deformation of the triangle if the other two nodes are fixed. Similarly, node 2 and node 3 are considered. An interior-point is taken inside the triangle to divide it into 3 regions as shown in Fig. 3(d). Let the total area of the triangle is R and R_1, R_2 and R_3 are the areas of three regions. This is represented by (2).

$$R = R_1 + R_2 + R_3 \tag{2}$$

From (2) the shape functions of all three regions are evaluated using (3)

$$J_1 = \frac{R_1}{R}, \; J_2 = \frac{R_2}{R}, \quad \text{and} \quad J_3 = \frac{R3}{R}, \tag{3}$$

where, J_1, J_2, and J_3 are the shape functions of regions R_1, R_2, and R_3. We assume the displacement of the interior point of the triangle in Fig. 3(d) is s and t in X and

Y directions respectively. Displacements s and t can be represented with the help of the shape functions

$$s = J_1 u_1 + J_2 u_3 + J_3 u_5 \tag{4}$$

$$t = J_1 u_2 + J_2 u_4 + J_3 u_6 \tag{5}$$

In (4) and (5), $u_{1,}u_3$, and u_5 are the displacement in the X direction and $u_{2,}u_4$ and u_6 are the displacements in the Y direction of the vertices of the triangle. The shape functions are determined by the areas of the different regions of the triangle; thus, the regions of the triangle are dependent on each other and can be represented as

$$J_1 + J_2 + J_3 = 1 \tag{6}$$

From (6), we can say that if we know two shape functions, then the third function can be easily calculated

$$J_3 = 1 - J_1 - J_2 \tag{7}$$

Putting these values into (4) and (5) we get the displacement of the interior point of the triangle in terms of s and t in X and Y directions respectively are

$$s = (u_1 - u_5)J_1 + (u_3 - u_5)J_2 + u_5 \tag{8}$$

$$t = (u_2 - u_6)J_1 + (u_4 - u_6)J_2 + u_6 \tag{9}$$

2.3 Representation of feature vector

The displacement of the interior point represents the displacement of the triangle. As discussed above the interior point (x, y) has displacements s in X direction and t in the Y direction. Due to these displacements, a deformation is produced in the triangular element. This deformation is nothing, but the strain developed in the triangular element in X, Y, and shear direction. These are given follows

Strain in X, Y - direction and shear strain are

$$\phi_x = \frac{\partial s}{\partial x} \tag{10}$$

$$\phi_y = \frac{\partial t}{\partial y} \tag{11}$$

$$\phi_x y = \frac{\partial s}{\partial y} + \frac{\partial t}{\partial x} \tag{12}$$

These strains are written in the form of matrices

$$\phi = \begin{bmatrix} \frac{\partial s}{\partial x} \\ \frac{\partial t}{\partial y} \\ \frac{\partial s}{\partial y} + \frac{\partial t}{\partial x} \end{bmatrix} \tag{13}$$

Once we get the strain in the triangular element of the discretized silhouette, we found out the stiffness matrix, using FEA [30-31]

$$K_t = C^\top D C t_e R_e \tag{14}$$

where, k_t is the stiffness matrix for a triangle element, C is a displacement matrix subject to strains in X direction, Y direction, and shear strain, t_e is thickness of the body which is constant in case of silhouette, R_e is the area of the triangle and D is a constant matrix

$$D = \frac{\xi}{1-\tau} \begin{bmatrix} 1 & \tau & 0 \\ \tau & 1 & 0 \\ 0 & 0 & \frac{1-\tau}{2} \end{bmatrix} \tag{15}$$

where, ξ is Youngs modulus and τ is Poissons ratio and both are constants. We tuned these parameters and we discussed their values in the experimental result section. Further, we converted the stiffness matrix of the triangle k_t into a one-dimensional feature vector [19] by scanning the matrix from the top left to bottom right element by element. The stiffness matrix of the triangle having m rows and m columns will be converted into the one-dimensional feature vector having total $m \times m$ elements. Similarly, we calculated the feature vectors of all possible triangles of the silhouette. The complete stiffness matrix of the silhouette K_s is created by combining all feature vectors of triangles where rows of the matrix represent the triangles associated with the silhouette. To solve the issue that which feature vector of the triangle will be the first row of the Stiffness Matrix of the silhouette we adopted the following strategy:

We scanned the discretized silhouette from top left to bottom right (interior point of the triangle). The first triangle whose interior point is found first in scanning will represent the first row of the matrix K_s and the triangle whose interior point is found last will represent the last row of the K_s matrix. If a silhouette of a frame is discretized into n number of triangle face then the stiffness matrix of silhouette K_s will have n number of rows and $m \times m$ numbers of the column. Further, the complete stiffness matrix of a silhouette is converted into the feature vector with a similar procedure as discussed above. This feature vector represents the frame of the action video.

2.4 Dimension reduction and classification

A frame of an action video at time t is represented by the feature vector extracted from the proposed method. The length of the feature vector of a frame is

$$C = row \times column$$

of the stiffness matrix of the silhouette. Suppose an action sequence consists of S frames, then that action sequence has S feature vectors. This results in a very high dimensional feature space. To reduce the dimensional feature space, we applied Principal component analysis (PCA). Further, these reduced features are given to RBF-SVM classifier [33-34] to recognize the actions. The proposed methodology can be summarized in the form of the algorithm as follows

Algorithm:

Given an action video, feature vectors can be constructed as follows.

Step 1: Extraction of silhouettes from input video frames.

Step 2: Extraction of prominent points on the boundary of the silhouettes.

Step 3: Prominent points are given as nodes to the FEA toolbox, MATLAB.

Step 4: Silhouettes are discretized into finite triangular elements where nodes act as vertices of the triangle.

Step 5: Each triangular element is represented by three nodes displacement vector (U) of the triangle.

Step 6: Displacement matrix (C) of each triangle is calculated

Step 7: Stiffness matrix (k_t) is calculated for each triangle with the help of displacement matrix C.

Step 8: Complete the Stiffness matrix of the silhouettes is created by combining stiffness matrices (k_t) of all possible triangles of the silhouette.

Step 9: Stiffness matrix of the silhouette is represented as one-dimensional feature vectors

Step 10: Feature vector is calculated for all frames of the action video for all actions.

Step 11: The RBF-SVM Classifier is used for recognition.

3 Experimental result

We have developed our proposed method on MATLAB R2015a. The proposed algorithm has been tested on a system having hardware configuration processor Intel(R) Core (TM) i5-6200U CPU @2.30GHz 2.40 GHz with 8 Gb RAM and 64-bit operating system. To evaluate the performance of the proposed methodology, accuracy is used as the performance parameter in a leave-one-out cross-validation strategy. It can be represented by using a true positive rate (TPR) and a false-positive rate (FPR) represented using true positive (TP), true negative (TN), false posirtie (FP), and false negative (FN)

$$TPR = \frac{TP}{TP + FN} \tag{16}$$

$$FPR = \frac{FP}{FP + TN} \tag{17}$$

where, TPR represents positive cases that are correctly classified and FRP represents negative cases that are incorrectly classified as positive. Accuracy is calculated as

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{18}$$

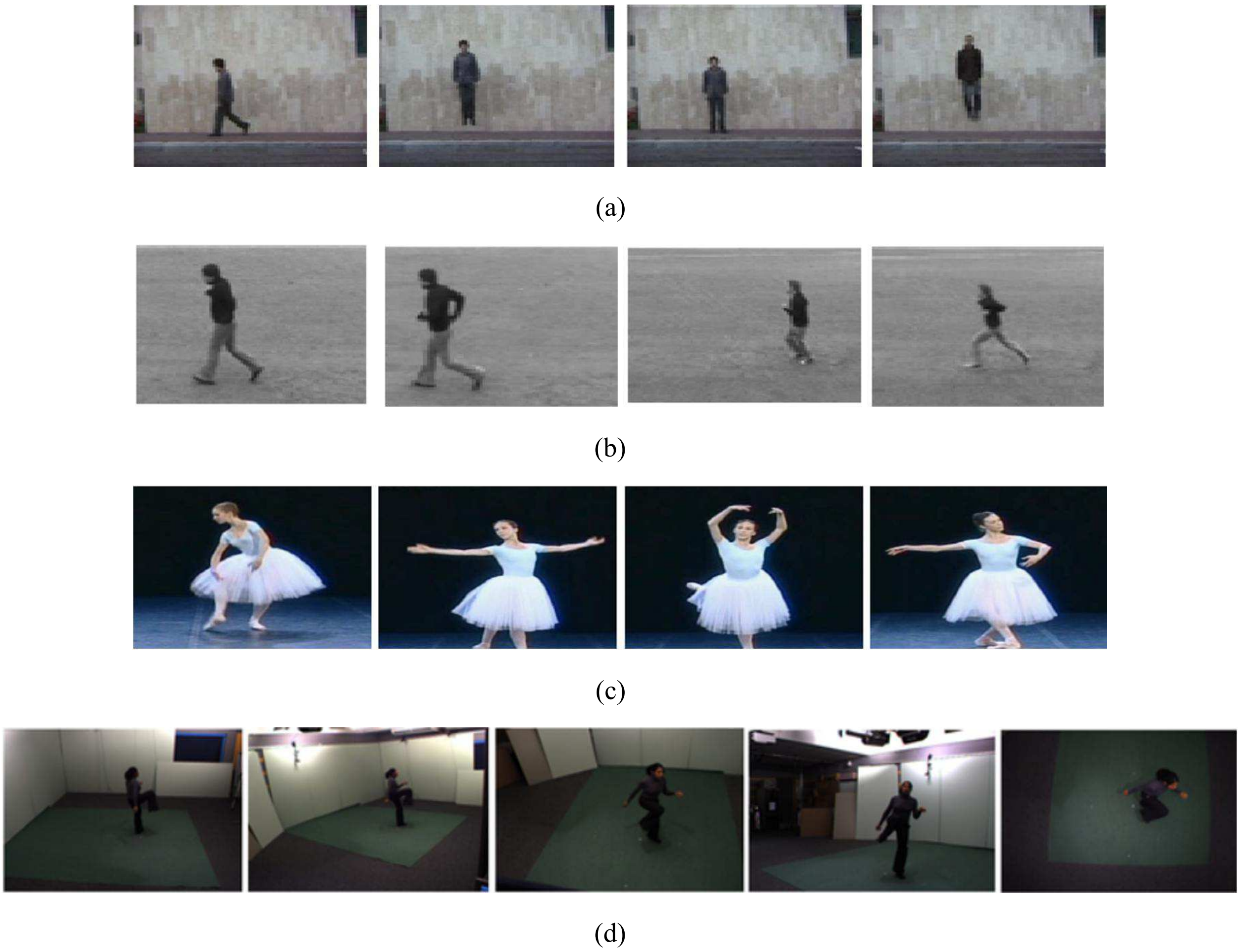
We have chosen four challenging datasets for action recognition namely Weizmann data set [4], KTH [35], Ballet [36] and IXMAS [37] to evaluate and compare our proposed method. In the Weizmann action dataset, there are ninety videos. The frame rate is 25 frames per second (fps) and resolution is 144180 pixels. It consists of nine different persons who performed a total of ten actions such as running, jumping, waving, bending, *etc.* The sample frame is shown in Fig. 4(a). The KTH dataset comprises six essential exercises, in particular: applauding, waving, boxing, walking, jogging and running. Activities in KTH have been recorded in four different lightings, indoor and outdoor situations and have 100 recordings. But the foundation for all recordings has been kept the same with a static camera with 25 fps and resolution of 160×120 pixels. The states of the recordings in the KTH informational index suffer from camera development and lighting impacts. The sample frame of this dataset is shown in Fig. 4(b).

Ballet is an expressive dance dataset, which comprises of profoundly complex artful dance poses of various onscreen characters. The specimen casings of the dataset are shown in Fig. 4(c). The dataset is acquired from a ballet dance DVD. The foundation in the dataset is basic. Every video grouping comprises of just a single performing artist. The dataset comprises of 44 videos. There are eight different unique activities performed in these videos. IXMAS is a very challenging dataset where 10 distinctive persons are performing every activity three times. These videos have been recorded from different view perspectives where seven different cameras used for recording. These activities incorporate scratching head, looking at the watch, strolling, taking a seat, etc. This dataset offers different challenges by introducing huge appearance change, intra-class varieties, and self-impediments, *etc.* XMAS results have been evaluated for five different camera views. The sample from the IXMAS dataset is shown in Fig. 4(d).

For the parameter settings, we tuned the important parameters on the KTH dataset and similar settings are applied to other datasets. These important parameters are the number of nodes, number of finite elements and feature dimension through PCA. The number of nodes is the prominent points extracted on the silhouette boundary. We have experimented on 5, 10, 15, 17, 20 and 25 prominent points which were considered as nodes. It is clear from Fig. 5.a that when a few prominent points are greater than 15, we achieve a good result. In the proposed method we have taken 17 numbers of nodes because accuracy is varying only 1-2% as we take several points greater than 17.

A next parameter is several finite elements. The number of finite elements has experimented as 10, 15, 20, 22 and 25. Fig. 5(b) shows that the number of elements greater than 20 is giving better accuracy.

The more discretized element we have, the more will be the accuracy of the representation of the body structure. But the tradeoff is that more discretized elements will increase the complexity in terms of time. Thus, we



(a)

(b)

(c)

(d)

Fig. 4. Datasets:(a) Weizmann,(b) KTH,(c) Ballet,(d)– IXMAS(5 cameras)

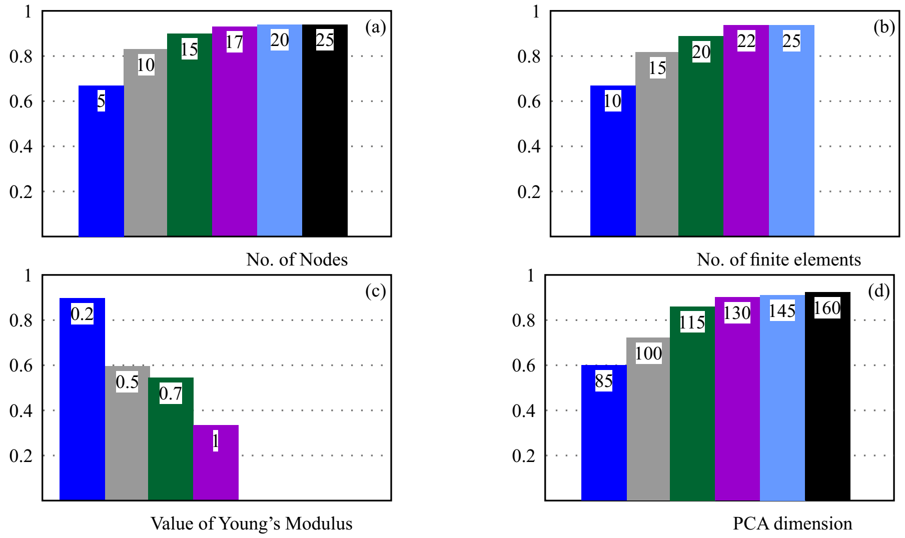


Fig. 5. Parameter setting for: (a) a number of nodes,(b) number of finite elements,(c) Young's Modulus,(d)– feature dimension through PCA

have taken 22 numbers of triangular faces in the proposed method. These 22 numbers of finite elements are taken in such a manner that these triangles do not overlap. This makes the structure simple and attains better accuracy. As far as Young's modulus mentioned in equation (15) is concerned, we have experimented on its normalized values 0. 2, 0. 5, 0. 7 and 1. 0 as shown in Fig. 5(c). We got the highest accuracy when the value of Youngs Modulus was 0. 2. The possible reason behind it could be the value of Youngs modulus is higher for the rigid body and lower for the flexible body. As the human body is very flexible while performing an action, the lower value 0. 2 gives a better result. The Poissons ratio mentioned in equation (15) is used for the material property and it is a constant

Table 1. Confusion matrix for Weizmann Dataset (R-Running, W-Walking, J-Jumping, JJ-Jumping Jack, S-Skipping, JP-Jumping at a place, SJ-Side Jumping, B-Bending, W-Waving with one hand, WB-Waving with both hands)

	R	W	J	JJ	S	JP	SJ	B	W	WB
R	0.95	0.05	0	0	0	0	0	0	0	0
W	0	1	0	0	0	0	0	0	0	0
J	0	0	1	0	0	0	0	0	0	0
JJ	0	0	0	1	0	0	0	0	0	0
S	0	0	0	0	0.96	0.04	0	0	0	0
JP	0	0	0	0	0.02	0.98	0	0	0	0
SJ	0	0	0	0	0	0	1	0	0	0
B	0	0	0	0	0	0	0	1	0	0
W	0	0	0	0	0	0	0	0	1	0
WB	0	0	0	0	0	0	0	0	0	1

Table 2. Confusion matrix for KTH Dataset (A- Applauding, W-Waving, B- Boxing, WK- Walking, J-Jogging, R-Running)

	A_1	A_2	A_3	A_4	A_5	A_6
A	1	0	0	0	0	0
W	0	1	0	0	0	0
B	0	0.02	0.98	0	0	0
WK	0	0	0	1	0	0
J	0	0	0	0	0.96	0.04
R	0	0	0	0	0.02	0.98

Table 3. Confusion matrix for Ballet LR- Left to right-Hand Opening, RL- Right to left-Hand Opening, J-Jumping, H-Hopping, S-Swinging leg, ST-Standing, T-Turning

	A_1	A_2	A_3	A_4	A_5	A_6	A_7
LR	1	0	0	0	0	0	0
RL	0	1	0	0	0	0	0
J	0	0	0.97	0.03	0	0	0
H	0	0	0.10	0.91	0	0	0
S	0	0	0	0	1	0	0
ST	0	0	0	0	0	1	0
T	0	0	0	0	0	0	1

Table 4. Confusion matrix for IXMAS Dataset: W- Walking, WA- Waving, P- Punching, K- Kicking, T- Throwing, P- Pointing, PU-Picking Up, G- Getting Up, S- Sitting Down, TA-Turning Around, F-Folding arms, C-Checking Watch, SH-Scratching Head

	W	WA	P	K	T	P	PU	G	S	TA	F	C	SH
W	1	0	0	0	0	0	0	0	0	0	0	0	0
WA	0	0.92	0.08	0	0	0	0	0	0	0	0	0	0
P	0	0	0.97	0	0.03	0	0	0	0	0	0	0	0
K	0	0	0	1	0	0	0	0	0	0	0	0	0
T	0	0	0.03	0	0.94	0.03	0	0	0	0	0	0	0
P	0	0	0	0	0.03	0.97	0	0	0	0	0	0	0
PU	0	0	0	0	0	0	0.94	0.06	0	0	0	0	0
G	0	0	0	0	0	0	0.8	0.92	0	0	0	0	0
S	0	0	0	0	0	0	0	0	1	0	0	0	0
TA	0	0	0	0	0	0	0	0	0	1	0	0	0
F	0	0	0	0	0	0	0	0	0	0	0.97	0.03	0
C	0	0	0	0	0	0	0	0	0	0	0.04	0.94	0
SH	0	0	0	0	0	0	0	0	0	0	0	0	1

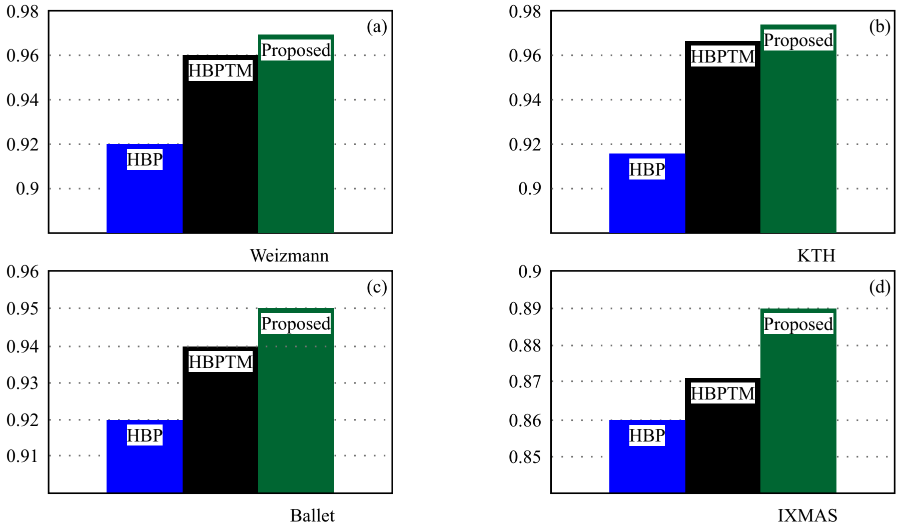


Fig. 6. Comparison of the proposed method with similar methods for four datasets Weizmann, KTH, Ballet, and IXMAS

value that lies between 0-0. 5. In the proposed method we got the optimized result when the value of was 0. 5. The thickness of the silhouette discussed in equation (14) remains constant for all frames in a video and in the proposed methodology; we have taken the value of thickness as 1. The last parameter is the feature dimension through PCA. The result of PCA for different dimensions is represented in Fig. 5(d). We have experimented on different values of dimension such as 85, 100, 115, 130, 145 and 160. Here dimension 130 is showing better results in terms of accuracy and complexity.

We applied the leave-one-out strategy for cross- validation. Tab. 1, Tab. 2, Tab. 3 and Tab. 4 shows the confusion matrices resulted from applying the proposed method on the datasets Weizmann, KTH, Ballet, and IXMAS respectively. These confusion matrices show that most of the actions are 100% classified except some similar types of actions. Thus, we got an accuracy of 97. 8%, 96. 4%, 95. 2% and 90. 3% for the Weizmann, KTH, Ballet, and IXMAS respectively. We compared the proposed method with the other Silhouette analysis based human action recognition methods such as Human Body Pose Model (HBPM) [17, 18, 22] and Human Body Pose Temporal Model (HBPTM) [19, 38] for these datasets. Chaaraoui *et al* . [22] extracted the features of the silhouette as contour points and action is learned from the multi-views of cameras. The multi-view learning makes the method capable of differentiating different persons performing the same action. Wu *et al* . [19] proposed the Human Body Pose Temporal Model where a 2-D silhouette mask is converted into a 1-D feature vector. They represented the action as the correlogram of poses extracted from the silhouette. H. Han *et al* . [25] represented the human body shapes with sparse geometrical features using the Bandlets transformation. They used the AdaBoost to select the features. As the proposed method gives the precise change in the human body shape due to the change in the small elements of the silhouette, it makes it better as compared to other methods. Figure 6 shows that the proposed method shows a better result as compared to other silhouette analysis-based methods using Weizmann, KTH, Ballet, and IXMAS datasets.

We also compared other state-of-the-art methodologies with the proposed method on all four datasets in Tab. 5-Tab. 8. Different testing strategies are used in these methodologies. We mentioned these testing strategies in Tab. s along with the classifier that they have used. Tab. 5 shows a comparison of the proposed method with other methodologies on the Weizmann dataset. Goudelis *et al* . [23] proposed a new feature extraction technique based on the Trace Transform. They represented the spatiotemporal feature in terms of the trace transform from the binarized silhouette. They used the SVM classifier and leave-one-person-out cross-validation testing strategy. They achieved 94.6% accuracy.

The methods [19] and [28] achieved a higher accuracy of 96. 3% and 97. 3% respectively. Liu *et al* . [28] modeled human action with the neural mechanism. They used new feature vectors using two cortical areas one is the primary cortex and the second is the middle temporal cortex for motion. Later, they used the SVM classifier to recognize the action. The proposed method achieved an accuracy of 97. 8%. Since we have discretized the silhouette into a smaller triangle, similar actions such as walking and running are recognizable better than the other methods.

Unlike the Weizmann dataset, the KTH dataset offers more challenging environments. It has different setups having different lighting conditions for different actions. Moreover, the shadow is also a very big challenge in this dataset. So, to deal with these problems, GMM is a better strategy for background subtraction and silhouette extraction. Rahman *et al* . [38] used the negative space-based feature of human pose and motion features to model the actions. To classify the actions, they used Nearest Neighbor Classifier. The leave-one-out strategy is used for cross-validation. They showed an accuracy of

Table 5. Comparison of the proposed method with similar methods on Weizmann dataset

Method	Year	Classifier and Test Scheme	Accuracy
[25]	2015	SVM	81.5
[22]	2013	KNN, LOSO	91.7
[19]	2013	SVM, LOSO	96.3
[23]	2013	SVM, LOPO	94.6
[24]	2014	KNN, LOO	91.4
[26]	2017	NNC	95.3
[27]	2016	CNN-HMM	90.1
[28]	2017	SVM	97.3
Proposed Method		SVM, LOO	$\boxed{97.8}$

Table 6. Comparison of the proposed method with similar methods on KTH dataset

Method	Year	Classifier and Test Scheme	Accuracy
[25]	2015	Adaboost, SVM, LOO	94.2
[23]	2013	SVM, LOPO	92.7
[26]	2017	NNC, LOO	93.6
[27]	2016	CNN-HMM	94.4
[28]	2017	SVM	91.3
[38]	2014	KNN, LOO	95.1
[39]	2015	KNN	90.8
[29]	2017	CNN-RNN	95.8
Proposed Method		SVM, LOO	$\boxed{96.4}$

Table 7. Comparison of the proposed method with similar methods on Ballet dataset

Method	Year	Classifier and Test Scheme	Accuracy
[17]	2017	SVM-NN, LOOCV	94.2
[18]	2015	SVM-NN, LOOCV	93.8
[41]	2009	S-CTM, LOO	89.8
[42]	2014	RVM, LOO	90.4
[43]	2014	SVM, LOO	90.3
Proposed Method		SVM, LOOCV	95.2

Table 8. Comparison of the proposed method with similar methods on IXMAS dataset

Method	Year	C1	C2	C3	C4	C5	Overall Accuracy
[44]	2011	89.1	83.4	89.3	87.2	89.2	87.8
[45]	2010	84.2	85.2	84.1	81.5	82.6	82.7
[46]	2013	86.5	83.8	86.1	84.5	87.4	87.2
[47]	2016	91.3	85.7	89.3	90.2	86.5	87.5
Proposed Method		90.8	90 .6	92.4	91.2	90.6	90.2

95. 1% as shown in Tab. 6. In another method, Shi *et al* . [29] proposed new motion descriptor sequential deep trajectory descriptors for long term motion video. The CNN-RNN network is used to learn the motion. They achieved a comparable accuracy of 95. 8% as compared to [38]. We have used a leave-one-out strategy and the proposed method achieved 96. 4% accuracy which is better than other methods.

Table 7 shows a comparison of the proposed method with the other state-of-the-methods for the Ballet dataset. Vishwakarma *et al* . [18] used silhouette-based analysis and extracted the feature vectors based on human poses. They have used SVM, LDA and Neural Network-based hybrid classifiers to recognize the action. They achieved an accuracy of 93. 8%. Vishwakarma *et al* . [17] used a new silhouette analysis where they first found out the average energy image of a silhouette. The spatial distribution of gradient is applied on average energy image to make it a global feature and the temporal feature is found out by Radon transform of the silhouettes. These features are given to the hybrid classifier and they achieve a higher accuracy of 94. 2%. In both methods [17] and [18] they used leave-one-out cross-validation. The proposed method shows better accuracy of 95. 3%. As discussed above discretized silhouette into small triangles helps to recognize the actions in an expressive dataset like Ballet dance. In this dataset, the closely the performer's expressions are observed the better results could be achieved.

Abbreviations used in Tab. 5-Tab. 8 are SVM: support vector machine, KNN: k-nearest neighbor, LOSO: leave-one-sequence -out, LOPO: leave-one-person-out, LOO: leave-one-out, NNC: nearest neighbor classifier, CNN: convolutional neural network, HMM: hidden Markov model, RNN: recurrent neural network, SVM-NN: support vector machine-neural network, LOOCV: leave-one-out cross-validation, S-CTM, RVM: relevance vector machine.

IXMAS dataset has five different camera views. Methods [42, 46-47] show almost similar accuracies which are around 87%. Wang *et al* . [47] used the Bag-of-visual-word method based on local features. Then they used the cross-view approach to deal with the problem of view change due to different cameras. The proposed method has achieved higher Accuracy for all the views. We got an average of 90. 2% accuracy. To deal with the problem of variation in viewpoint we used view-invariant interest point on the boundary of the silhouette which acted as the vertices of the triangles during the discretization step. Tab. 8 shows a comparison of the proposed method the other state-of-the-methods for IXMAS dataset

For run-time analysis, we have used NVIDIA GPU and MATLAB 2015a with a Parallel computing toolbox. Time taken for several modules in the proposed methodology is calculated. We have analyzed the run-time of the proposed method on all the datasets discussed above. The average run-time of the proposed dataset and is given below stepwise:

Extraction of silhouette from action video (Sec): 0. 41
Silhouette discretization into triangular faces (Sec): 0. 54
Calculation of the silhouette stiffness matrix (Sec): 1. 45
Classification (Sec): 0. 52
Total time for action recognition (Sec): 2.92

Thus, the run-time of the proposed method is fairly good.

6 CONCLUSION

This is a new method to recognize human action through Finite Element Analysis (FEA). A new feature descriptor where the feature vectors of the video frames are expressed in terms of the stiffness matrix of the silhouette extracted from the frames of the video is applied. This offers uniqueness to this method, as it can extract both shapes as well as motion information. The feature vectors extracted from the proposed method are given to the RBF-SVM classifier. Validation of the proposed method has been performed in different challenging environments. The limitation of the methodology is that it requires accurate silhouette extraction. The proposed method shows its superiority as compared to other existing methods of applying them on challenging standard datasets such as Weizmann, KTH, Ballet, and IXMAS.

References

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

[2] R. Souvenir and J. Babbs, "Learning the viewpoint manifold for action recognition, *IEEE International Conference on Computer Vision Pattern Recognition (CVPR'08)*, pp. 1-7, 2008.

[3] M. Blank, L. Gorelick, E. Shechtman, M. Irani, and R. Basri, "Action as space-time shapes, *IEEE International Conference on Computer Vision (ICCV'05)*, vol. 2, pp. 1395-1402, 2005.

[4] M. Blank, L. Gorelick, E. Shechtman, M. Irani, and R. Basri, "Action as space-time shapes, *IEEE Transaction on Pattern Analysis Machine Intelligence*, vol. 29, no. 12, pp. 2247-2253, 2007.

[5] K. Guo, P. Ishwa, and J. Konrad, "Action recognition from video using feature covariance matrices, *IEEE Transaction on Image Processing*, vol. 22, no. 6, pp. 2479-2494, 2013.

[6] Y. Chen, Z. Li, X. Guo, Y. Zhao, and A. Cai, "A spatio-temporal interest point detector based on vorticity for action recognition, *IEEE International Conference on Multimedia Expo Workshop*, pp. 1-6, 2013.

[7] M. Laptev, C. Marszalek, and B. Schmid, "Learning realistic human actions from movies, *IEEE Conference on Computer Vision Pattern Recognition*, pp. 1-8, 2008.

[8] S. Savarese, A. Delpozo, J. C. Niebles, and L. Fei-fei, "Spatial-temporal correlations for unsupervised action classification, *Proceedings*, of the IEEE Workshop on Motion Video Computing, pp. 1-8, 2008.

[9] M. S. Ryoo and J. K. Aggarwal, "Spatio-temporal relationship match: Video structure comparison for recognition of complex human activities, *IEEE 12th International Conference on Computer Vision*, pp. 1593-1600, 2009.

[10] I. Laptev and T. Lindeberg "Space-time interest points, *Proceedings Ninth IEEE International Conference on Computer Vision*, pp. 432-439, 2003.

[11] A. Klaser, M. Marszalek and C. Schmid, "A spatio-temporal descriptor based on 3D-gradients, *Proceedings of British Machine Vision Conference*, pp. 995-1004, 2008.

[12] G. Willems, T. Tuytelaars, and L. Van Gool, "An efficient dense scale-invariant spatio-temporal interest point detector, *ECCV 5303*, pp. 650-663, 2008.

[13] M. Chen and A. Hauptmann, "MoSIFT: Recognizing human actions in surveillance videos, *CMU-CS-09-161 2009*,.

[14] N. Ballas, L. Yao, C. Pal, and A. Courville, "Delving deeper into convolutional networks for learning video representations, *International Conference on Learning Representations 2016*,.

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

[16] L. Sun, K. Jia, D. Yeung, and B. E. Shi, "Human action recognition using factorized spatio-temporal convolutional networks, *IEEE International Conference on Computer Vision (ICCV)*, pp. 4597-4605, 2015.

[17] D. K. Vishwakarma and K. Singh, "Human activity recognition based on the spatial distribution of gradients at sub-levels of average energy silhouette images, *IEEE Transactions on Cognitive Development Systems*, vol. 9, no. 4, pp. 316-327, 2017.

[18] D. K. Vishwakarma and R. Kapoor, "Hybrid classifier based human activity recognition using the silhouettes ands cells, *Expert Systems with Applications*, vol. 42, no. 20, pp. 6957-6965, 2015.

[19] D. Wu and L. Shao, "Silhouette analysis-based action recognition via exploiting human poses, *IEEE Transactions on Circuits Systems for Video Technology*, vol. 23, no. 2, pp. 236-243, 2013.

[20] D. Weinland, M. Ozuysal, and P. Fua, "Making action recognition robust to occlusions viewpoint changes," *European Conference on Computer Vision* (ECCV), pp. 635-648, 2010.

[21] B. Saghafi and D. Rajan, "Human action recognition using Pose-based discriminant embedding, *Signal Processing: Image Communication*, vol. 27, no. 1, pp. 96-111, 2012.

[22] A. A. Chaaraoui, P. C. Pérez, and F. Florez-Revuelta, "Silhouette-based human action recognition using sequences of key poses, *Pattern Recognition Letters*, vol. 34, no. 15, pp. 1799 -1807, 2013.

[23] G. Goudelis, K. Karpouzis, and S. Kollias, "Exploring trace transform for robust human action recognition, *Pattern Recognition*, vol. 46, no. 12, pp. 3238-3248, 2013.

[24] R. Touati and M. Mignotte, "MDS-based multi-axial dimensionality reduction model for human action recognition, *Canadian Conference on Computer Robot Vision*, pp. 262-267, 2014.

[25] H. Han and X. J. Li, "Human action recognition with sparse geometric features, *The Imaging Science Journal*, vol. 63, no. 1, pp. 45-53, 2015.

[26] Y. Fu, T. Zhang, and W. Wang, "Sparse coding-based space-time video representation for action recognition, *Multimedia Tools Applications*, vol. 76, no. 10, pp. 12645-12658, 2017.

[27] J. Lei, G. Li, J. Zhang, Q. Guo, and D. Tu, "Continuous action segmentation recognition using hybrid convolutional neural network-hidden Markov model, *IET Computer Vision*, vol. 10, no. 6, pp. 537-544, 2016.

[28] H. Liu, N. Shu, Q. Tang, and W. Zhang, "Computational model based on the neural network of visual cortex for human action recognition, *IEEE Transactions on Neural Networks Learning Systems*, vol. 29, no. 5, pp. 1427-1440, 2017.

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

[30] 2D Triangular Elements, The University of New Mexico, http://www.unm. edu/ bgreen/ME360/2D%20Triangular %20 Elements.pdf. Accessed 24 February 2010,.

[31] D. K. Jha, T. Kant, and R. K. Singh, "An accurate two dimensional theory for deformation stress analysis of functionally graded thick plates, *International Journal of Advanced Structural Engineering*, pp. 6-7, 2014.

[32] J. Dou and J. Li, "Robust human action recognition based on spatiotemporal descriptors motion temporal templates, *Optik*, vol. 125, no. 7, pp. 1891-1896, 2014.

[33] Q. Song, W. Hu, and X. Wenfang, "Robust support vector machine for bullet hole image classification, *IEEE Transaction on Systems Man Cybernetics*,, vol. 32no. pp. 440-448, 2002.

[34] S. S. Keerthi C.-J. Lin, "Asymptotic Behaviors of Support Vector Machines with Gaussian Kernel, *Neural Computation vol*, 15, no, 7,, pp. 1667-1689, 2003.

[35] C. Schuldt, I. Laptev, and B. Caputo, "R, *ognizing human actions: a local SVM approach*, Proceedings of the 17th International Conference on Pattern Recognition Cambridge, UK, 2004,.

[36] T. Guha and R. K. Ward, "Learning sparse representations for human action recognition, *IEEE Transaction on Pattern Analysis Machine Intelligence*, vol. 34, no. 8, pp. 1576-1588, 2012.

[37] D. Weinland, R. Ronfard, and E. Boyer, "Free viewpoint action recognition using motion history vol. s, *Computer Vision Image Understanding*, vol. 104, no. 2-3, pp. 249-257, 2006.

[38] S. A. Rahman, I. Song, M. K. H. Leung, I. Lee, and K. Lee, "Fast action recognition using negative space features, *Expert Systems Applications*, vol. 41, no. 2, pp. 574-587, 2014.

[39] I. Gomez-Conde and D. N. Olivieri, "A KPCA spatio-temporal differential geometric trajectory cloud classifier for recognizing human actions in a CBVR system, *Expert Systems Applications*, vol. 42, no. 13, pp. 5472-5490, 2015.

[40] L. Juan and O. Gwun, "A comparison of SIFT, PCA-SIFT and SURF, *International Journal of Image Processing*, vol. 3, no. 4, pp. 143-152, 2009.

[41] Y. Wang and G. Mori, "Human action recognition using semilatent topic models, *IEEE Transactions on Pattern Analysis Machine Intelligence*, vol. 31, no. 10, pp. 1762-1764, 2009.

[42] L.-M. Xia J.-X. Huang, and L.-Z. Tan, "Human action recognition based on chaotic invariants, *Journal of Central University*, vol. 20, no. 11, pp. 3171-3179, 2014.

[43] A. Iosifidis A Tefas and I. Pitas, *Discriminant bag of words based representation for human action recognition*, Pattern Recognition Letters, vol. 49, no. 1, pp. 185-192, 2014.

[44] X. Wu, D. Xu, L. Duan, and J. Luo, "Action recognition using context appearance distribution features, *IEEE Conference on Computer Vision Pattern Recognition (CVPR)*, pp. 489-496, 2011.

[45] D. Weinland, M. Özuysal, and P. Fu, "Making action recognition robust to occlusions viewpoint changes", *European Conference on Computer Vision* (ECCV), pp. 635-648, 2010.

[46] E.-A, Mosabbeb, K. Raahemifar, and M. Fathy, "Multi-view human activity recognition in distributed camera sensor networks, *Sensors*, vol. 13, no. 7, pp. 8750-8770, 2013.

[47] J. Wang, H. Zheng, J. Gao, and J. Cen, "Cross-view action recognition based on a statistical translation framework, *IEEE Transactions on Circuits Systems for Video Technology*, vol. 26, no. 8, pp. 1461-1475, 2016.

Received 18 October 2019

Rajiv Kapoor (Dr) is a Professor in Delhi Technological University, Delhi, India. He worked as Principal in AIACTR, Delhi, India in diverted capacity. He is PhD in Electronics and Communication Engineering from Punjab Engineering College, India. His Research interest include vision/speech-based tracking, activity recognition vision/speech based, signal processing, pattern recognition. He has published more than 100 research articles in leading journals, conference proceedings and books including IEEE, Springer, Elsevier, *etc*

Om Mishra is a research scholar in Department of Electronics & Communication, Delhi Technological University, Delhi, India. He has worked as an Assistant Professor in GB Pant Government Engineering College, New Delhi, India. He received Master of Engineering in Electronics & Communication Engineering from Delhi College of Engineering (Presently DTU), Delhi, India. His research interest includes vision-based activity recognition, signal processing, pattern recognition.

Madan Mohan Tripathi (Dr) is a Professor in Electrical Engineering Department of Delhi Technological University, Delhi, India. He has also worked as Scientist with the Institute for Plasma Research, India and National Institute of Electronics & Information Technology, India. He is PhD in Electrical Engineering from GB Technical University, India. His research interests include Artificial Intelligence applications, renewable energy and power system restructuring. He has published approximately 100 research articles in leading journals and conference proceedings including IEEE, Springer, Elsevier, *etc*