Received: 28 February 2018 | Revised: 12 May 2018 | Accepted: 25 June 2018

DOI: 10.4218/etrij.2018-0102

SPECIAL ISSUE

WILEY ETRI Journal

Vector space based augmented structural kinematic feature descriptor for human activity recognition in videos

Sowmiya Dharmalingam[1] | **Anandhakumar Palanisamy**[2]

[1]Department of Information and Communication Engineering, Anna University, Chennai, Tamil Nadu, India.

[2]Department of Computer Technology, Madras Institute of Technology, Anna University, Chennai, Tamil Nadu, India.

Correspondence
Sowmiya Dharmalingam, Department of Information and Communication Engineering, Anna University, Chennai, Tamil Nadu, India.
Email: d.sowmiya86@gmail.com

A vector space based augmented structural kinematic (VSASK) feature descriptor is proposed for human activity recognition. An action descriptor is built by integrating the structural and kinematic properties of the actor using vector space based augmented matrix representation. Using the local or global information separately may not provide sufficient action characteristics. The proposed action descriptor combines both the local (pose) and global (position and velocity) features using augmented matrix schema and thereby increases the robustness of the descriptor. A multiclass support vector machine (SVM) is used to learn each action descriptor for the corresponding activity classification and understanding. The performance of the proposed descriptor is experimentally analyzed using the Weizmann and KTH datasets. The average recognition rate for the Weizmann and KTH datasets is 100% and 99.89%, respectively. The computational time for the proposed descriptor learning is 0.003 seconds, which is an improvement of approximately 1.4% over the existing methods.

KEYWORDS

human activity recognition, kinematic features, multiclass support vector machine classifier, structural features, vector space based augmented structural kinematic

1 | INTRODUCTION

Monitoring of human activities in public places is a significant research area in the field of computer vision. An activity is performed by each individual based on their unique style and personality. Activity recognition is very difficult to process because each activity may vary from person to person in terms of speed, postures, and dynamics in their movements. Each activity needs to be categorized and recognized with the help of the most essential features for recognizing each action class. Hence, a suitable representation methodology is required to build an action descriptor. A descriptor consists of the collected features represented in an organized form that is required for each action classification and understanding.

Distinctive feature descriptors are available in the literature for activity recognition. Human activity recognition requires optimal feature representation to achieve robust solutions. Ideally, the feature representation should be strong enough to handle varied poses, person localization in a cluttered environment, viewpoint variations, temporal variations, anthropometrics variations, and so on. Various state-of–the-art-descriptors have been developed based on local or global features. Although the two types of features provide the integrant capability of local or global characteristics, the individual representation of either the local or global features does not provide approbative action information. A vector space based augmented structural kinematic (VSASK) feature descriptor is proposed in this paper. The proposed structural kinematic descriptor is built by integrating the structural (pose) and kinematic (position and velocity) properties of the actor using an augmented matrix and represented in vector space for human activity learning.

Depending on the scenario, human activity takes different forms such as atomic actions, atomic actions in sequences, person to person interactions, and person to object interactions. In real scenarios, cameras are mounted everywhere for surveillance purposes. However, huge human resources are required to scrutinize the activity occurring, which is highly expensive. Automatic human activity recognition requires an equitable and impeccable technique for activity based understanding. Although researchers in the field of computer vision have conducted extensive research on automatic activity recognition, it is still an elusive goal to develop a robust descriptor to represent the features for activity classification [1–3].

Image feature representation is of two types: (i) global representation and (ii) local representation. In global representation, the whole region of interest is represented as silhouettes, edges, or optical flow. In local representation approaches, spatiotemporal interest points are first spotted and then the local patches around these points are calculated and combined to form the descriptor for feature representation. Various state-of–the-art-descriptors have been developed based on local or global features. Although the two types of features provide an integrant capability of local or global characteristics, individual representation of either the local or global features will not provide approbative action information [3,4]. In the literature, only very few studies have combined the local and global features. The harmonizing property achieved by combining these two features will certainly alleviate the drawbacks of the individual representation of local and global features.

The major contributions of the proposed VSASK descriptor are as follows:

- A descriptor named vector space based augmented structural kinematics (VSASK) is proposed for activity recognition.
- The structure (pose) and kinematics (position and velocity) features are integrated using augmented matrix schema and represented by means of vector space for each individual action class.
- The proposed structural kinematics descriptor integrates the local and global temporal information that complement the action characteristics for efficient activity recognition.

2 | RELATED WORK

In human activity recognition [5], the most significant process is to extract and represent the descriptive information required to develop a pattern base for learning different actions. The collective information (features) required for activity recognition of each individual activity form a pattern termed as a descriptor. The descriptors built are used for learning each activity through a suitable machine learning approach. The recent significant works related to the feature exaction and representation for human activity learning are discussed below.

A spatio temporal multi-feature based descriptor was developed by Jalal and others [6] for online human activity recognition achieving an accuracy of 94.1%. The features considered for the descriptor construction were skeleton joint features and the shape feature. Luo and others [7] proposed a new framework that uses both the RGB videos and depth maps for human activity recognition. The 3D joint features were represented using a sparse coding based temporal pyramid. The spatial and temporal gradient patterns were represented using center symmetric motion local ternary patterns. These features were fused for activity recognition and obtained a recognition accuracy of 86.2%.

Althloothi and others [8] presented a 3D shape and motion feature that was fused at the kernel level for activity recognition with an accuracy of 89%. The then existing 3D shape and joint features based on depth maps and RGB videos showed good recognition accuracy. However, the computational cost involved in the framework was high. The usage of depth map features and skeletal joint features or the combination of both affects the recognition accuracy in case of occlusion or ambiguous postures [9]. Hence, a novel methodology that provides the complete body kinematics and the whole body structure of each activity was required.

Yang and Tian [10] proposed a global feature vector, namely, super normal vector for human activity recognition. The differences between the spatio-temporal data of the skeletal joints were fused and represented using a supernormal vector. Liu and others [11] detailed a skeletal joint information based feature descriptor for activity recognition. The spatial and temporal information obtained through skeletal joints was learned using a convolution neural network to recognize different activities. In the case of skeleton based approaches, the accuracy rate is not appreciable when the human position is oblique to the camera. The skeletal joints are more suitable for applications such as hand gesture recognition rather than human activity recognition [12].

Lillo and others [13] introduced a hierarchical method for activity recognition. The geometric features, in addition to the motion features extracted, were represented as a sparse composition for human activity learning. However, the model performance was not good for joint features with noise, and accuracy was degraded by 7.2%. The feature and motion based approach is robust in case of occlusion. However, this approach is not suitable for bulky data sets. Tran and Torresani [14] proposed the space-time volume descriptors that are learned using exemplar-movement SVMs and showed an improvement of 1% in accuracy.

The space time trajectory information was extracted and represented in a vector form for human activity learning. However, the authors were not certain about the movement clues for each action class required for good activity recognition.

Zhang and Parker [15] proposed a new feature descriptor, multichannel orientation histogram (MCOH), to represent the 4D color depth local spatio temporal features. Ho and others [16] proposed 3D postures for different activities that were learned by the max-margin classifier. The joint kinematics, color, and depth information were used for posture learning. Vishwakarma and Kapoor [17] proposed a feature descriptor obtained by dividing the human silhouette poses into grids and cells. This approach yielded an accuracy rate of 96%. A hybrid classification model SVM-NN was proposed for feature learning. Chaaraoui and others [18] represented human postures using silhouettes. The key postures were recognized using the Euclidian distance based *K*-means clustering technique. A less expensive feature descriptor has been constructed using the contour points of human silhouettes for posture learning that obtains a recognition rate of 92%.

Bayat and others [19] used the acceleration information obtained through the tri-axial accelerometer in Android phones for activity learning using the available classifiers with a recognition rate of 88.15%. Kwon and others [20] extracted the sensor based features such as acceleration and angular velocity in time, frequency domain using mobile phones with an average recognition accuracy of 83.33%. These features were learnt using the hierarchical clustering approach. However, a major drawback of the sensor based data is that it varies from person to person, and data cannot be obtained in case the sensor does not work or the sensor is dropped or lost by the person without their knowledge. Hence, the model performance may not achieve appreciable results [21].

Cuntoor and others [22] proposed a probabilistic approach representing various events using the hidden Markov model (HMM). This approach is robust to variations in the viewing direction. Human behaviors were analyzed at the semantic level using this blob-based approach. However, the blob-based approaches are sensitive to illumination variation. Duque and others [23] proposed a graph based method known as dynamic oriented graph (DOG). It detects abnormal behaviors using unsupervised learning with an accuracy rate of 82.5%. However, the DOG classifier was not tested with real data in various environments.

Zhang and others [24] developed a grammar system by transforming the motion trajectory information into primitives achieving a recognition rate of 88.3%. The concealed temporal relationship in the primitives was discovered with the help of a rule stimulation algorithm using the minimum description length (MDL). However, the activities that did not follow the rule set were not recognized. Hsieh and others [25] proposed a Delaunay triangulation technique that divides a discrete human posture into different triangular meshes. The string matching algorithm was adopted to recognize various actions with a recognition rate of 94.83%.

Lee and others [26] presented a stratified method for activity recognition to reduce the computational load and to make the system robust to noise. Kalman filtering has also been applied to increase the accuracy of the motion estimation. Ben and others [27] used a voting scheme to update the parameters of the stick model. The model is constructed from video sequences to analyze different actions with a recognition rate of 84.68%. However, this approach is view dependent.

The literature reveals that the most robust features for human activity recognition are the motion and shape information of each action class. Computer vision researchers have used different techniques to extract and represent these features. They have used these features either individually (shape or motion) or in combination (shape and motion). However, a good activity recognition rate is achieved by combining the shape and motion features. Even with these combinations, extracting the joint motion cues was not very efficient. Hence, the whole-body kinematics and the typical structural information can be more useful for achieving a good activity recognition rate. In addition to the extraction of the features, a compact representation is important for efficient activity learning. Hence, there is a significant need to construct a robust descriptor for efficient activity recognition. Table 1 consolidates the literature review pertaining to human activity recognition

In this work, the structural (shape) information and kinematics (position and velocity) information are fused. These features are augmented and represented in vector space for learning activities. The benefit of the proposed method is that it considers the whole body structural and kinematic information that improves the activity recognition rate as demonstrated by the experimental results. The feature representation of the proposed descriptor significantly reduces the computational time. The proposed descriptor handles circumstances such as variations in view, scale, time, color, object size, localization of an actor in a cluttered environment, and ambiguous postures. An overview of the proposed descriptor is given in Figure 1.

3 | DESCRIPTOR CONSTRUCTION

The proposed augmented structural kinematic descriptor has been developed on the basis of the Fourier and Kalman theory. The foreground binary silhouettes, $Z(t) = B_1, B_2, \ldots, B_m$ in each frame of a video say, v_1, v_2, ... , v_m (sequential frames in each video) are segmented using the background subtraction algorithm. The structural

TABLE 1 Brief introduction to some similar experiments

Feature descriptor for human activity recognition			
Multi features	**Author**	**Year**	**Remarks**
Motion	Ben et al. [27]	2002	View dependent
Human postures	Hsieh et al. [25]	2008	Human postures represented as triangular meshes
Grammar rules using trajectory information	Zhang et al. [24]	2011	The activities that don't follow the rule set were not recognized
3D-Joint features and spatial and temporal gradient patterns	Luo et al. [7]	2014	High computational cost and lower recognition accuracy
3D-Shape and motion features	Althloothi et al. [8]	2014	
Silhouettes Poses into grids and cells	Vishwakarma et al. [17]	2015	
4D color depth local Spatio temporal features	Zhang et al. [15]	2016	
Skeletal joint features and shape features	Jalal et al. [6]	2017	
3D postures	Ho et al. [16]	2016	
Super normal vector	Yang et al. [10]	2017	The recognition accuracy is not appreciable in case human is oblique to the camera
Spatial and temporal features of skeletal joints.	Liu et al. [11]	2017	
Space time trajectory features	Tran et al. [14]	2016	Difficulty in predicting the movement clues
Geometric and motion features	Lillo et al. [13]	2017	Not suitable for bulky dataset
Sensor data			
Acceleration	Bayat et al. [19]	2014	Sensor data used varies from person to person leads to inaccurate results
Acceleration and angular velocity	Kwon et al. [20]	2014	
Dynamic oriented graph	Duque et al. [23]	2014	Real time data are not tested

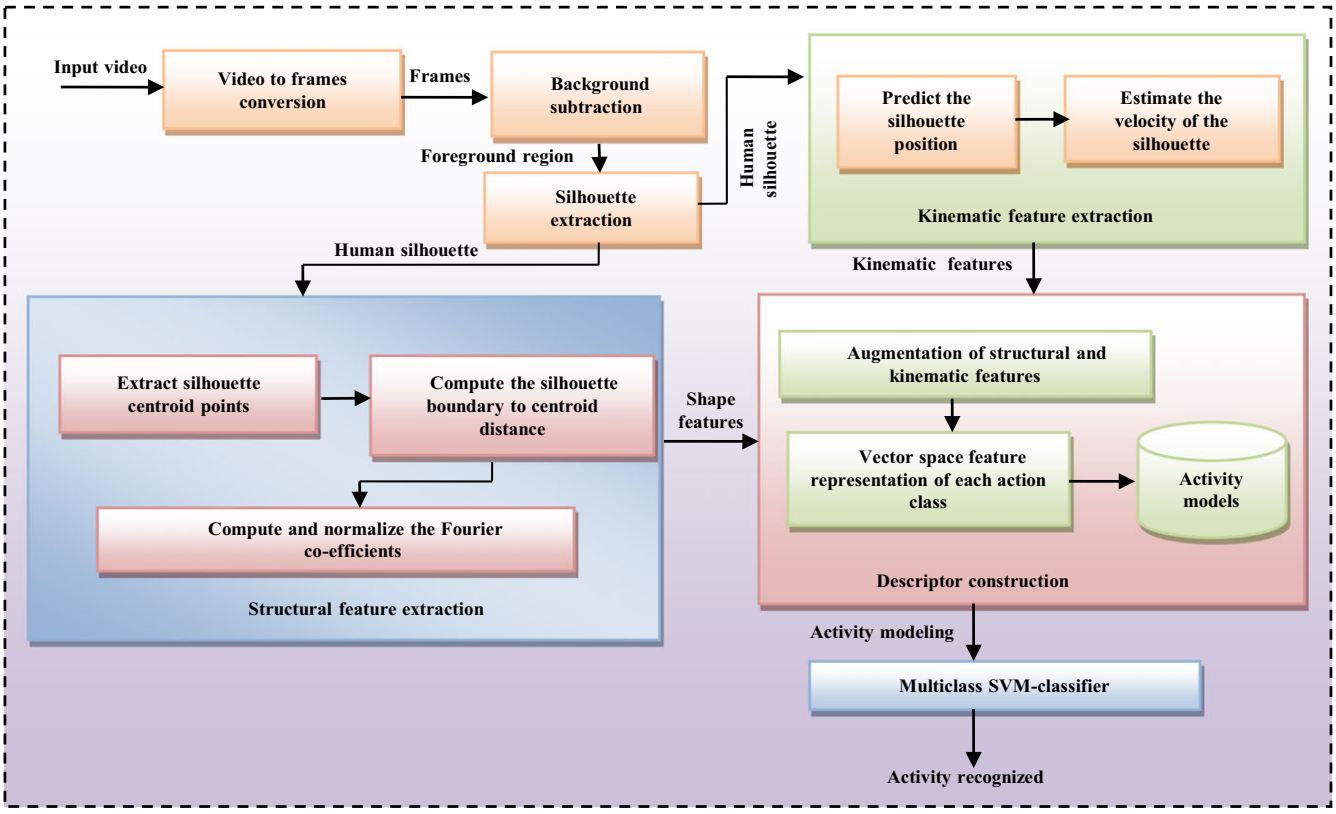

FIGURE 1 Overview of the proposed vector space based augmented structural kinematic (VSASK) feature descriptor for activity recognition

and kinematic information of the various actions in each video frame needs to be integrated and represented for ensuring efficient activity recognition. $Z(t)$ denotes the human silhouettes at time t. Let (x_b, y_b) represent the boundary coordinates of the binary silhouettes, where $b = 0, 1, \ldots, N-1$.

Let X denote the state of an actor; $\bar{X}$-predict the actor state; $\hat{X}$-update the actor state; here, the state denotes the position and velocity of an actor. Let us consider an actor at time t where the actor's state and shape are to be predicted and represented by the augmented structural kinematic descriptor. The centroid coordinates x_c, y_c of the actor $Z(t)$ at time t are calculated as:

$$x_c = \frac{1}{N}\sum_{b=0}^{N-1} x_b; y_c = \frac{1}{N}\sum_{b=0}^{N-1} y_b \tag{1}$$

where N is the number of boundary coordinates.

The distance r_b from the centroid to the boundary coordinates is calculated as follows:

$$r_b = \sqrt{(x_b - x_c)^2 + (y_b - y_c)^2}. \tag{2}$$

The discrete Fourier transform coefficient a_n for the distance r_b is calculated as:

$$a_n = \frac{1}{N}\sum_{b=0}^{N-1} r_b \exp\left(\frac{-j2\pi nb}{N}\right) \tag{3}$$

where $n = 0, 1, \ldots, N-1$ The Fourier coefficient a_n is normalized to represent an actor shape that is invariant to translation, rotation, and scaling by the following equation. s_l is denoted as:

$$s_l = \frac{|a_n|}{|a_0|} \tag{4}$$

- a_0: The first Fourier transform coefficient
- s_l: The normalized Fourier transform coefficient, where $l = 1, 2, \ldots, \frac{N}{2}$.

The Fourier transform coefficient of an actor $Z(t)$ holding the shape information is represented as:

$$V = [s_1\, s_2 \ldots s_l] \tag{5}$$

- V: A matrix holding the shape information.

The actor state is predicted based on the following state prediction equation:

$$\bar{X}_t = A.\hat{X}_{t-1} + BU \tag{6}$$

- A, B, U: state transition matrix

Let us initialize

$\hat{X}_{t-1} = \left[x_{t-1} y_{t-1} v_{x(t-1)} v_{y(t-1)}\right]$

where $(x_{t-1}, y_{t-1}) = (x_c, y_c)$ represent the centroid coordinates of the actor at time $t-1$. $v_{x(t-1)}, v_{y(t-1)}$, are the corresponding velocities in the x, y directions.

The current state $\hat{X}_t$ is predicted and updated using the following state update equation:

$$\hat{X}_t = \bar{X}_t + K(Z_t - \bar{Z}_t) \tag{7}$$

- Z_t: Actual position of an actor
- $\bar{Z}_t$: Estimated position of an actor
- K: Kalman Gain
  $K = \bar{P}_t C^T S^{-1}$
- $\bar{P}_t$: Covariance matrix of prediction state
- S, C^T: constants, T – transpose of a matrix

The updated state of an actor is

$$\hat{X}_t = \left[x_t\ y_t\ v_{xt}\ v_{yt}\right] \tag{8}$$

where $(x_t, y_t) = (x_c, y_c)$, the centroid co-ordinates of the object at time 't' (the current state). (v_{xt}, v_{yt}) denote the corresponding velocities in the x, y directions. The matrix V holds the shape information and the matrix $\hat{X}_t$ holds the position and velocity of an actor $Z(t)$. By the notion of augmented matrix, V and $\hat{X}_t$ are combined to form an efficient structural kinematic descriptor.

$$d_t = \left(\hat{X}_t | V\right) = \left[x_t\ y_t\ v_{xt}\ v_{yt}\ s_1\, s_2 \ldots, s_l\right]. \tag{9}$$

The above augmented feature vector d_t represents the position, velocity, and shape information of the moving actor $Z(t)$, at time t in a video sequence. The shape, position, and velocity information are integrated through the descriptor for activity recognition. Let the feature descriptor for each activity class be denoted by a vector space D_i where i denotes the number of action classes such as walking, jogging, and boxing. The descriptor for each activity class i is constructed as a vector space, $D_i = \{d_1, d_2, \ldots, d_m\}$, $d_t \varepsilon R^{l+4}$; here, m is the number of frames in a video sequence. R denotes the dimension of augmented feature vector d_t. The algorithm for the proposed descriptor construction is given below.

Algorithm 1. Computation of the proposed vector space based augmented structural kinematic feature descriptor

Input: Foreground human binary silhouettes $Z(t) = B_1, B_2, \ldots, B_m$ at time t. m denotes the number of video frames. (x_b, y_b) represents the boundary coordinates of the binary silhouettes, where $b = 0, 1, \ldots, N-1$. (x_t, y_t) represents the position of an actor at time t. (v_{xt}, v_{yt})represents the velocity of an actor in x and y directions at time t

Output: Augmented structural kinematic vector space descriptor D_i.

$v_{xt} = v_{yt} = 0$ //intial velocity in x & y direction

1: for $k \leftarrow 1$ to m do
2: x_c, $y_c \leftarrow$ extract_centroid (B_k, t).
3: for $b \leftarrow 0$ to $N-1$ do
4: $r_b \leftarrow$ calculate the centroid distance(x_b, y_b, x_c, y_c, t).
5: $a_n \leftarrow$ compute discrete Fourier transform (r_b).
6: $s_l \leftarrow \frac{|a_n|}{|a_0|}$ //s_1-normalized Fourier coefficients
7: $V \leftarrow [s_1 s_2, \ldots, s_l]$//*wherel* $= \frac{N}{2}$
8: end for
9: $\vec{X}_t \leftarrow$ predict actor state (x_{t-1} y_{t-1} $v_{x(t-1)}$ $v_{y(t-1)}$)
10: $\hat{X}_t \leftarrow$ update actor state$\left(x_t\ y_t\ v_{xt}\ v_{yt}\right)$
11: $d_k \leftarrow$ Feature Augmentation (x_t, y_t, v_x, v_{yt}, s_1, s_2, ... , s_l)
12: $D_i \leftarrow [d_1, d_2, \ldots, d_m]$ //D_i-vector space for each action class i
13: end for

4 | HUMAN ACTIVITY RECOGNITION

The proposed descriptor is different from the previous works wherein each action video is generally represented by a D-dimensional feature vector or D-dimensional matrix. In the proposed descriptor, augmented feature vectors of each action class are collected in vector space $D_i = \{d_1, d_2, \ldots, d_m\}, d_t \varepsilon R^{l+4}$, where i denotes the number of action classes.

In the proposed vector space representation model, a set of videos of the same action class can be represented as a vector in a common vector space, whereas in a D-dimensional feature vector representation, each video is represented as a separate feature descriptor. Hence, in the proposed work, a single individual descriptor is built for each action class with reduced computational time for descriptor learning. The activity recognition rate of the proposed action defined vector space is more effective than that of the video-based feature vector representation or the matrix representation.

A multiclass SVM classifier is used to recognize various human activities. An SVM is a distinct machine learning technique that classifies data using an optimal hyperplane drawn to separate data that belong to any one of the class labels. A multiclass SVM is an extension of binary SVM that assigns class labels (supervised learning) to the input feature spaces to classify them into their corresponding classes by support vectors.

The training data are $(D_1, Y_1), \ldots, (D_N Y_N)$, where $D_i \varepsilon R^{m(l+4)}$ is the input vector space and $Y_i \varepsilon \{1, \ldots, N\}$ are the activity labels of the feature space. The multiclass SVMs are formed as a solution to the optimization problem. The decision parameter of SVM uses the maximum margin to classify the ith class, from the remaining classes,

$$Y_i(D_i) = W_i^t D_i + b_i, \tag{10}$$

where $W_i^t \varepsilon R^{m(l+4)}$ is a weight vector and b_i is a scalar.

The function $Y_i(D_i) = 0$ denotes the best partitioning hyperplane with the maximum margin. For any data that are linearly independent, the support vectors that belong to a particular class i are given by the condition $Y_i(D_i) = 1$ and the data that belong to the other classes are given by the condition $Y_i(D_i) = -1$. A traditional SVM will classify each input vector space D_i to its corresponding activity class $Y_i(D_i)$ only under the following condition:

$$Y_i(D_i) > 0. \tag{11}$$

For any feature vector in the vector space, D_i is grouped into class$i$. The decision for any data D_i that belongs to some class i is based on the sign of the decision function, hence the outcome is distinct. A vector space D_i is unclassifiable if $Y_i(D_i)$ fails under above condition (11). To solve this problem, the feature space is classified into a class based on the following decision function that has the largest value:

$$\arg\max_i Y_i(D_i). \tag{12}$$

5 | EXPERIMENTAL RESULTS

The proposed VSASK descriptor was experimentally analyzed using the Weizmann dataset and KTH datasets. The Weizmann dataset contains 10 varied activities performed by nine people with 5,687 frames and a total of 93 video sequences. The KTH dataset contains a total of 599 videos of 25 people performing six varied activities. MATLAB 2014a was used for conducting the experiments. To classify different activities, multiclass SVM was used. Throughout the experiments, all the frames were used to obtain the structural and kinematics features for robust feature representation.

A total of 8,171 VSASK descriptors were constructed for activity recognition. The structural and kinematic features are fused using augmented matrix schema. The vector space is a collection of vectors of the same size. The size of a single feature vector in the vector space is 1×68. For each action class, if there are m frames there will be m feature vectors in the vector space. For each individual action class, a learning model has been developed in vector space that is used by the multiclass SVM for classifying the activity.

5.1 | Performance metrics used for the evaluation

Activity recognition is a multiclass problem with i different action classes. The performance of the proposed method is evaluated based on the metrics, namely, precision, recall, F-score, and accuracy. The metrics values are estimated

based on the true positives, true negatives, false positives, and false negatives.

Let each action class be denoted as i, where TP_i (True Positives) is the total count of the videos precisely recognized as action class i. FP_i (False Positives) is the total count of videos recognized as action class i, but that actually do not belong to action class i. FN_i (False Negatives) is the total count of videos belonging to action class i, but not recognized as action class i. TN (True Negatives) is the total count of videos accurately categorized as not belonging to action class i.

The performance of the proposed VSASK descriptor is evaluated by means of the F-score, which is equal to the harmonic mean of recall (ρ) and precision (π), which are defined as follows:

$$\pi = \frac{TP_i}{TP_i + FP_i}, \tag{13}$$

$$\rho = \frac{TP_i}{TP_i + FN_i} \tag{14}$$

where i denotes the corresponding action class.

The F-score values range between (0, 1). A larger F-score value indicates good performance. The F-score is calculated as follows:

$$F = \frac{2\pi\rho}{\pi + \rho}. \tag{15}$$

The most common and simplest performance evaluation measure is the accuracy (A). The accuracy is calculated as follows:

$$A = \frac{TP_i + TN_i}{TP_i + FP_i + FN_i + TN_i}. \tag{16}$$

5.2 | Description of the benchmark datasets

The proposed approach of this study focuses on static camera human activity recognition. Static camera human activity recognition is required for elderly people and patient monitoring, where the videos are captured from a single viewpoint. In reality, environments such as rooms in a house or a garden outside the house usually have a single surveillance camera (static view). In home environments, usually people do not have multi-view (multiple) cameras inside a single room or outside the house (car parking or garden). It is also more expensive to have multi-view cameras in these scenarios. The patient resting inside the room or the elderly walking in the garden are commonly captured by a static camera in the home environment. In all these scenarios, the elderly person or the patient tends to be alone (single). Hence, there is a need to focus on a single actor performing actions. Therefore, the KTH and Weizmann datasets, which are the standard benchmark datasets for a static camera view with a single actor performing actions, are used for experimental analysis. The KTH and Weizmann datasets are used worldwide by many researchers in the field of human activity recognition.

The benchmark datasets, namely Weizmann & KTH, are used for evaluating the proposed descriptor. The datasets contain versatile actions such as jumping, walking, and bending, that are reasonably helpful in recognizing various activities. Sample frames from the Weizmann and KTH datasets where the actors perform various activities such as walking, running, jumping, skipping, bending, and hand waving, are shown in Figure 2.

5.2.1 | Weizmann action dataset

The Weizmann dataset was developed mainly to recognize human activities in a simple environment. The Weizmann dataset contains 10 human actions performed by nine people. The dataset was developed from a static viewpoint. Ground truth data of foreground silhouettes obtained through background subtraction are available. Various activities available in this dataset are walking, running, jumping, skipping, bending, jumping in place of two legs (Pjump), jumping-jack (jack), galloping sideways (side), skipping, waving two hands (wave2), and waving one hand (wave1). The video sequences are captured in simple background environments with a frame rate of 50 fps. The resolution of the video sequence is 180×144 pixels. The dataset provides a collection of a large number of actions suitable for evaluating the accuracy of the proposed approach for activity recognition. The dataset contains approximately 5,687 frames, and a total of 93 video sequences.

5.2.2 | KTH action dataset

The KTH Institute of Technology developed a dataset for activity recognition. This is one of the largest video datasets of different human actions performed under different environmental scenarios. The videos were captured with a static background and a single viewpoint. The KTH dataset contains six human actions, namely, walking, jogging, running, boxing, hand waving, and hand clapping, performed multiple times by 25 persons in four different scenarios, namely, outdoors, outdoors with scale variation, outdoors with different clothes, and indoors. The KTH dataset consists of 599 video sequences where each video contains any one of the above listed actions. The videos were taken in static view of a homogenous background with a frame rate of 25 fps. The video sequence is downsampled to a resolution of 160×128 pixels.

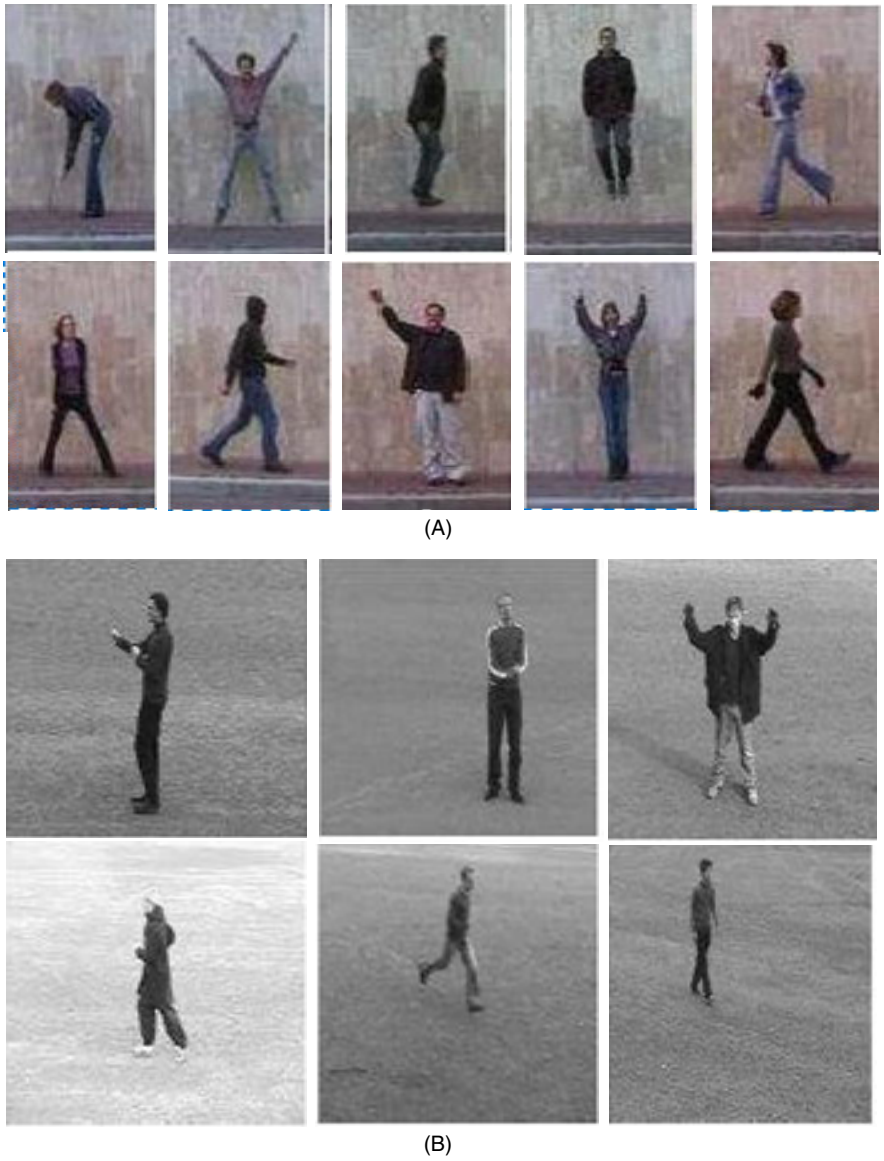

FIGURE 2 (A) Screenshots from the Weizmann dataset. (B) Screenshots from the KTH dataset

5.3 | Performance analysis of the proposed descriptor on benchmark datasets

The confusion matrix for the Weizmann and KTH human action datasets is given in Tables 2 and 3. The performance analysis of multiclass activity recognition is conducted by considering the prediction of the individual action class of a video sequence tabulated in the confusion matrix. The confusion matrix is a metric that matches the actual classes with the predicted classes. In this matrix, the horizontal rows denote the actual action classes and the vertical columns denote the predicted action classes. The proposed model is evaluated based on each video recognition. The diagonal values represent the recognition accuracy rate of the corresponding activity that is recognized and classified correctly. The off-diagonal values represent the rates of misclassified activities.

Tables 4 and 5 provide the performance measures such as recall, precision, *F*-score, and accuracy for various activities in the Weizmann and KTH datasets. The tabulated results prove that the proposed descriptor is robust in terms of activity recognition. The proposed VSASK descriptor correctly recognizes all activities, walk, run, jump, skip, bend, Pjump, jack, wave1, wave2, gallop, with 100% accuracy and an *F*-score value of 1 for the Weizmann dataset. The performance of the proposed VSASK descriptor is notably good in terms of various performance measures owing to the robust structural and kinematic information fused through the augmented representation.

The proposed descriptor enhances the classification results for ambiguous activity classes in the KTH dataset. Because certain actions appear similar, such as boxing and handclapping, there is a greater chance of misclassification. The activities such as walking, running, jogging, and hand

TABLE 2 Confusion matrix for a video sequence of various actions in Weizmann human action dataset (average recognition accuracy: 100%)

		Predicted action classes									
		Walk	**Run**	**Jump**	**skip**	**Bend**	**Pjump**	**Jack**	**wave1**	**wave2**	**Gallop**
Actual action classes	Walk	100									
	Run		100								
	Jump			100							
	Skip				100						
	Bend					100					
	Pjump						100				
	Jack							100			
	wave1								100		
	wave2									100	
	Gallop										100

TABLE 3 Confusion matrix for the video sequence of various actions in KTH human action dataset (average recognition accuracy: 99.89%)

		Predicted action classes					
		Walking	**Running**	**Jogging**	**Boxing**	**Hand waving**	**Hand clapping**
Actual action classes	Walking	100					
	Running		100				
	Jogging			100			
	Boxing				99.67		0.33
	Hand waving					100	
	Hand clapping				0.33		99.67

TABLE 4 Performance measure of the proposed VSASK descriptor for Weizmann dataset

Activities	Recall	Precision	*F*-score	Accuracy (%)
Walk	1	1	1	100
Run	1	1	1	100
Jump	1	1	1	100
Skip	1	1	1	100
Bend	1	1	1	100
Pjump	1	1	1	100
Jack	1	1	1	100
wave1	1	1	1	100
wave2	1	1	1	100
Gallop	1	1	1	100

waving are recognized with a remarkable accuracy rate of 100% and an *F*-score of 1. The analogous activities that share sizeable similarities in motion and shape, such as jogging-running-walking, are very well distinguished using the proposed descriptor. Some videos under the boxing categories are misclassified as handclapping because of the similarity in the motion and shape. The accuracy for boxing and handclapping is 99.67%. This misclassification of boxing as hand clapping is considered as a false negative in the boxing category and a false positive in the hand clapping category. Hence, the *F*-score value for the activities, boxing (0.9999) and hand clapping (0.99), are considerably reduced. The experimental results validate the commendable performance of the proposed VSASK descriptor.

TABLE 5 Performance measure of the proposed VSASK descriptor for KTH dataset

Activities	Recall	Precision	*F*-score	Accuracy (%)
Walking	1	1	1	100
Running	1	1	1	100
Jogging	1	1	1	100
Boxing	0.98	1	0.99	99.67
Hand waving	1	1	1	100
Hand clapping	1	0.9802	0.99	99.67

5.4 | Validation protocol

Table 6 shows the comparison of the proposed VSASK descriptor performance with that of many state-of-the-art-descriptors [28–45]. The proposed descriptor shows a remarkable performance for the Weizmann (100%) and KTH (99.89%) datasets. The proposed approach yields a significantly high accuracy rate compared with some of the approaches in the literature. The evaluation protocol used for evaluating the proposed method is the leave one out cross validation (LOOCV) process. In LOOCV, for each class, one video is used for testing and the remaining videos are used for training. The process may be repeated by changing the test video for each iteration.

Figure 3 depicts the comparison of computational time per descriptor learning of the proposed VSASK descriptor with that of state-of-the-art descriptors. Owing to the integrated feature schema and reduced dimensions of the feature vectors, the proposed VSASK descriptor achieves a reduced computational time of 0.003 s when compared with the existing descriptors [44], [46–48].

6 | CONCLUSION

The vector space based augmented structural kinematic (VSASK) feature descriptor has been proposed for human activity recognition. The features of shape, location, and velocity of an actor were extracted and augmented to

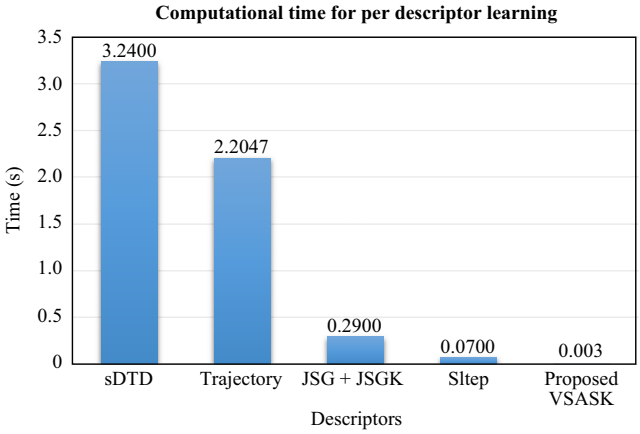

FIGURE 3 Computational time for per descriptor learning of the proposed descriptor with state of art descriptors

TABLE 6 Performance measure of the proposed VSASK descriptor for KTH dataset

				Accuracy (%)	
Author	**Year**	**Descriptor**	**Classifier**	**Weizmann**	**KTH**
Niebles et al. [28]	2008	STW	PLSA	90.00	83.33
Ikizler et al. [29]	2009	HOR	SVM	97.53	77.31
Yu et al. [30]	2011	STIP	Random forest	N/A	91.80
Wang et al. [31]	2012	HOF + HOG	SVM	95.60	93.33
Zhao et al. [32]	2013	3DSC + HOG3D	BoW	N/A	92.12
Reddy et al. [33]	2013	3D-SIFT	SVM	N/A	89.79
Roshtkhari et al. [34]	2013	STV	Codebook	91.90	81.20
Ballan et al. [35]	2013	H3DGrad + HOF	SVM	92.10	92.41
Rahman et al. [36]	2014	NSAD	Fuzzy	95.56	94.46
Chua et al. [37]	2014	HOOG + HOOF	SVM	96.67	83.94
Iosifidis et al. [38]	2014	DBoWs	SVM	N/A	92.13
Eweiwi et al. [39]	2015	HOG	SVM	51.8 (for both data sets)	
Yao et al. [40]	2015	Silhouette	Fuzzy	94.03	N/A
Yao et al. [41]	2016	PooLST	SVM	N/A	95.37
Yao et al. [41]	2016	CPST	SVM	N/A	94.91
Zhao et al. [42]	2017	LEMT + SIPS	RMM	98.92	N/A
Qian et al. [43]	2017	DCI	NN	97.78	95.17
Qian et al. [43]	2017	DCI	SRC	97.78	96.66
Shi et al. [44]	2017	sDTD	CNN-RNN	N/A	96.80
Xu et al. [45]	2017	IPs + FS	SVM	99.10	95.80
Proposed descriptor		**VSASK**	**SVM**	**100.00**	**99.89**

form a new descriptor. The augmented features were represented in vector space for each action class. The proposed descriptor was experimentally analyzed on the benchmark datasets, namely, Weizmann and KTH. The experimental results reveal that the proposed VSASK descriptor significantly outperforms the state-of-the-art methods. The computational time per descriptor construction is comparatively very less. Hence, activity learning is also achieved within a lower computational time (0.003 s). The proposed descriptor shows a remarkable performance for the Weizmann (100%) and KTH (99.89%) datasets. Thus, the proposed approach has the potential to be applied for activity recognition on a large dataset of unimpeded videos.

ORCID

Sowmiya Dharmalingam http://orcid.org/0000-0001-9998-1467

REFERENCES

1. O. D. Lara and M. A. Labrador, *A survey on human activity recognition using wearable sensors*, IEEE Commun. Surveys Tuts. **15** (2013), no. 3, 1192–1209.
2. L. Liu et al., *Learning spatio-temporal representations for action recognition: A genetic programming approach*, IEEE Trans. Cybern. **46** (2016), no. 1, 158–170.
3. Y. Gao et al., *Violence detection using oriented violent flows*, Image Vis. Comput. **48** (2016), 37–41.
4. X. Fang et al., *Action recognition using edge trajectories and motion acceleration descriptor*, Mach. Vis. Appl. **27** (2016), no. 6, 861–875.
5. F. Han et al., *Space-time representation of people based on 3D skeletal data: A review*, Comput. Vis. Image Underst. **158** (2017), 85–105.
6. A. Jalal et al., *Robust human activity recognition from depth video using spatiotemporal multi-fused features*, Pattern Recogn. **61** (2017), 295–308.
7. J. Luo, W. Wang, and H. Qi, *Spatio-temporal feature extraction and representation for RGB-D human action recognition*, Pattern Recogn. Lett. **50** (2014), 139–148.
8. S. Althloothi et al., *Human activity recognition using multi-features and multiple kernel learning*, Pattern Recogn. **47** (2014), no. 5, 1800–1812.
9. Y. Song et al., Combining rgb and depth features for action recognition based on sparse representation, *Proceedings of the 7th International Conference on Internet Multimedia Computing and Service*, ACM, Aug 2015, pp. 49.
10. X. Yang and Y. L. Tian, *Super normal vector for human activity recognition with depth cameras*, IEEE Trans. Pattern Anal. Mach. Intell. **39** (2017), no. 5, 1028–1039.
11. M. Liu, H. Liu, and C. Chen, *Enhanced skeleton visualization for view invariant human action recognition*, Pattern Recogn. **68** (2017), 346–362.
12. M. Liu, H. Liu, and C. Chen, *3D action recognition using multi-scale energy-based global ternary image*, IEEE Trans. Circuits Syst. Video Technol. (2017).
13. I. Lillo, J. Carlos Niebles, and A. Soto, *Sparse composition of body poses and atomic actions for human activity recognition in RGB-D videos*, Image Vis. Comput. **59** (2017), 63–75.
14. D. Tran and L. Torresani, *EXMOVES: Mid-level features for efficient action recognition and video analysis*, Int. J. Comput. Vision **119** (2016), no. 3, 239–253.
15. H. Zhang and L. E. Parker, *Code4d: color-depth local spatio-temporal features for human activity recognition from rgb-d videos*, IEEE Trans. Circuits Syst. Video Technol. **26** (2016), no. 3, 541–555.
16. E. S. L. Ho et al., *Improving posture classification accuracy for depth sensor-based human activity monitoring in smart environments*, Comput. Vis. Image Underst. **148** (2016), 97–110.
17. D. K. Vishwakarma and R. Kapoor, *Hybrid classifier based human activity recognition using the silhouette and cells*, Expert Syst. Appl. **42** (2015), no. 20, 6957–6965.
18. A. Andre Chaaraoui, P. Climent-Pérez, and F. Flórez-Revuelta, *Silhouette-based human action recognition using sequences of key poses*, Pattern Recogn. Lett. **34** (2013), no. 15, 1799–1807.
19. A. Bayat, M. Pomplun, and D. A. Tran, *A study on human activity recognition using accelerometer data from smartphones*, Procedia Comp. Sci. **34** (2014), 450–457.
20. Y. Kwon, K. Kang, and C. Bae, *Unsupervised learning for human activity recognition using smartphone sensors*, Expert Syst. Appl. **41** (2014), no. 14, 6067–6074.
21. W.-Y. Deng, Q.-H. Zheng, and Z.-M. Wang, *Cross-person activity recognition using reduced kernel extreme learning machine*, Neural Netw. **53** (2014), 1–7.
22. N. P. Cuntoor, B. Yegnanarayana, and R. Chellappa, *Activity modeling using event probability sequences*, IEEE Trans. Image Process. **17** (2008), no. 4, 594–607.
23. D. Duque, H. Santos, and P. Cortez, Prediction of abnormal behaviors for intelligent video surveillance systems, *Symposium on Computational Intelligence and Data Mining*, IEEE, Mar 2007, pp. 362–367.
24. Z. Zhang, T. Tan, and K. Huang, *An extended grammar system for learning and recognizing complex visual events*, IEEE Trans. Pattern Anal. Mach. Intell. **33** (2011), no. 2, 240–255.
25. J.-W. Hsieh et al., *Video-based human movement analysis and its application to surveillance systems*, IEEE Trans. Multimedia **10** (2008), no. 3, 372–384.
26. S.-W. Lee et al., *Hierarchical active shape model with motion prediction for real-time tracking of non-rigid objects*, IET Comput. Vision **1** (2007), no. 1, 17–24.
27. J. Ben-Arie et al., *Human activity recognition using multidimensional indexing*, IEEE Trans. Pattern Anal. Mach. Intell. **24** (2002), no. 8, 1091–1104.
28. J. Carlos Niebles, H. C. Wang, and L. Fei-Fei, *Unsupervised learning of human action categories using spatial-temporal words*, Int. J. Comput. Vision **79** (2008), no. 3, 299–318.
29. N. Ikizler and P. Duygulu, *Histogram of oriented rectangles: A new pose descriptor for human action recognition*, Image Vis. Comput. **27** (2009), no. 10, 1515–1526.
30. G. Yu et al., *Fast action detection via discriminative random forest voting and top-k sub volume search*, IEEE Trans. Multimedia **13** (2011), no. 3, 507–517.

31. H. Wang et al., *Supervised class-specific dictionary learning for sparse modeling in action recognition*, Pattern Recogn. **45** (2012), no. 11, 3902–3911.
32. D. Zhao et al., *Combining appearance and structural features for human action recognition*, Neurocomputing **113** (2013), 88–96.
33. K. K. Reddy and M. Shah, *Recognizing 50 human action categories of web videos*, Mach. Vis. Appl. **24** (2013), no. 5, 971–981.
34. M. Javan Roshtkhari and M. D. Levine, *Human activity recognition in videos using a single example*, Image Vis. Comput. **31** (2013), no. 11, 864–876.
35. L. Ballan et al., Recognizing human actions by using effective codebooks and tracking, *Advanced Topics in Computer Vision*, Springer, London, 2013, pp. 65–93.
36. S. Atiqur Rahman et al., *Fast action recognition using negative space features*, Expert Syst. Appl. **41** (2014), no. 2, pp. 574–587.
37. T. Wee Chua and K. Leman, A novel human action representation via convolution of shape-motion histograms, *International Conference on Multimedia Modeling*, Springer, Jan 2014, pp. 98–108.
38. A. Iosifidis, A. Tefas, and I. Pitas, *Discriminant bag of words based representation for human action recognition*, Pattern Recogn. Lett. **49** (2014), 185–192.
39. A. Eweiwi, M. Shahzad Cheema, and C. Bauckhage, *Action recognition in still images by learning spatial interest regions from videos*, Pattern Recogn. Lett. **51** (2015), 8–15.
40. B. Yao et al., *A fuzzy logic-based system for the automation of human behavior recognition using machine vision in intelligent environments*, Soft. Comput. **19** (2015), no. 2, 499–506.
41. L. Yao, Y. Liu, and S. Huang, *Spatio-temporal information for human action recognition*, EURASIP J. Image Video Process. **39** (2016), 1–9.
42. Y. Zhao et al., *Region-based mixture models for human action recognition in low-resolution videos*, Neurocomputing **247** (2017), 1–5.
43. H. Qian et al., *Recognizing human actions from silhouettes described with weighted distance metric and kinematics*, Multimed. Tools Appl. **76** (2017), no. 21, 21889–21910.
44. Y. Shi et al., *Sequential deep trajectory descriptor for action recognition with three-stream CNN*, IEEE Trans. Multimed. **19** (2017), no. 7, 1510–1520.
45. K. Xu, X. Jiang, and T. Sun, *Two-stream dictionary learning architecture for action recognition*, IEEE Trans. Circuits Syst. Video Technol. **27** (2017), no. 3, 567–576.
46. S. Singh, C. Arora, and C. V. Jawahar, *Trajectory aligned features for first person action recognition*, Pattern Recogn. **62** (2017), 45–55.
47. M. Li and H. Leung, *Graph-based approach for 3D human skeletal action recognition*, Pattern Recogn. Lett. **87** (2017), 195–202.
48. X. Ji et al., *The spatial laplacian and temporal energy pyramid representation for human action recognition using depth sequences*, Knowl.-Based Syst. **122** (2017), 64–74.

AUTHOR BIOGRAPHIES

Sowmiya Dharmalingam received her B.E. degree in computer science and engineering from the Sri Krishna College of Engineering and Technology, Anna University, Chennai, Tamil Nadu, India in 2008. She received her M.E. degree in computer science with specialization in knowledge engineering and computational linguistics from the College of Engineering, Anna University, Chennai, Tamil Nadu, India in 2010. She is currently pursuing a full time PhD in information and communication engineering, Madras Institute of Technology, Anna University, Chennai, Tamil Nadu, India since 2012. Her research areas include moving object detection, human activity recognition, and crowd analysis.

Anandhakumar Palanisamy received his PhD degree in information and communication engineering, Anna University, Chennai, Tamil Nadu, India in 2006. He received his M.E. degree in computer science and engineering from the Government College of Technology, Bharathiar University, Coimbatore, Tamil Nadu, India in 1997. He received his B.E. degree in electronics and communication engineering from the Government College of Engineering, Salem, Tamil Nadu, India in 1994. He is currently working as professor in the Department of Computer Technology, Madras Institute of Technology, Anna University, Chennai, Tamil Nadu, India. He has been serving in Anna University for more than 20 years. His research interests include computer vision, artificial neural networks, multimedia applications, and cloud computing.