Int. J. Electron. Commun. (AEÜ) 70 (2016) 341–353

Contents lists available at ScienceDirect

International Journal of Electronics and Communications (AEÜ)

journal homepage: www.elsevier.com/locate/aeue

Unified framework for human activity recognition: An approach using spatial edge distribution and ℜ-transform

D.K. Vishwakarma*, Rajiv Kapoor, Ashish Dhiman

Department of Electronics and Communication Engineering, Delhi Technological University, Bawana Road, Delhi 110042, India

ARTICLE INFO

Article history:
Received 23 March 2015
Accepted 17 December 2015

Keywords:
Human activity recognition
Spatial edge distribution
ℜ-transform
Fusion of translational and rotational features
Still images

ABSTRACT

In this paper, a unified approach for the recognition of human activity using the spatial edge distribution of gradients and orientation of the human silhouettes in a video sequence is presented. The spatial edge distribution is computed on still image at different levels of resolution of sub-images to extract out the shape of the activity posture. The fuzzy trapezoidal membership function is used to extract the key frames of the activity, and the single still key image is extracted according to the histogram distance. The temporal content of the activity is extracted by the computation of orientation of the silhouettes using ℜ-transform. The ℜ-transform is applied on the binary human silhouettes, and the extraction of human silhouettes from the video sequence is done using texture based segmentation techniques. The high dimensionality of the ℜ-transform features is handled by applying Local linear embedding (LLE) dimension reduction approach. A unified model is constructed by integrating the spatial edge distribution of gradients and temporal content of the activity. The performance of the developed model is demonstrated on publicly available datasets, and the highest classification accuracy achieved on each datasets is compared with the similar state-of-the-art techniques and shows the superior performance.

1. Introduction

In recent years, human activity recognition has been an active area of research in computer vision due to its potential applications in the field of surveillance, assistive healthcare, sports event analysis, robotics, terrorist activities, content-based video analysis and human-computer interactions [1–3]. However, human activity recognition is both challenging and multifaceted due to viewpoint variations, occlusion, cluttered background, intra-class motion variability and inter-class motion ambiguity. Faced with these challenges several researchers are trying to devise a general, competent and robust method for recognition of human activity.

Over the last few decades, numerous human activity recognition techniques have been proposed, which are mainly focused on the local and global feature based representations. Some of the popular existing approaches for the human activity recognition are as follows; motion flow or optical flow, point trajectories, space-time volume, Bag-of-words model, and spatio-temporal interest points (STIP's) [3–5]. The advantages of these approaches are that they do not require background subtraction and are efficient in handling partial occlusions but all of these techniques have their limitations like optical flow approach is less accurate when the video quality is poor and rough. Similarly, point trajectories based approach requires an efficient tracking of the human motion and if human is moving with variable speed then tracking trajectories may be inefficient. In case of STIPs the distribution of the interest points should be stable around the object. Even the bag of words approach is inadequate to capture the spatial and temporal information and only focus on the global saliency and ignores the structure of the body.

In these days, a trend in the human activity recognition (HAR) has been realized, where multiple features [6–8] are used to improve the recognition accuracy. These methods include global and local information and admit that an individual feature based methods are less effective as compared to the multiple features based methods. In this context more recently, researchers [9,10] supported that multiple features based fusion techniques can provide better performance than the individual features. Wu and Shao [12] presented the combination of local and holistic representation for human action recognition where they effectively worked on the Bag of correlated poses for the local representation and involved MHI/GEI images for the holistic representation. This combined approach is robust to viewpoint, scale and orientation but has plenty of scope for improvement using multiple features. Liu et al.

* Corresponding author. Tel.: +91 1127871044x1308/9971339840.
E-mail addresses: dvishwakarma@gmail.com, dkvishwakarma@dce.ac.in (D.K. Vishwakarma), rajivkapoor@dce.ac.in (R. Kapoor), ashish.dhiman1@gmail.com (A. Dhiman).

http://dx.doi.org/10.1016/j.aeue.2015.12.016
1434-8411/

[13] proposed the adaptive learning methodology where they used the genetic programming for the representation of spatio-temporal features. Simultaneously, they fused the color and motion information for high-level activity recognition. The main drawback was that they required a large number of generations for good results that further lead to heavy computation. Liu et al. [14] give a probabilistic model, which effectively combined the Gaussian process (GP) regression with sparse covariance matrix for realistic action recognition. However, GP is limited to large-scale computer vision tasks because modeling the large dataset with stochastic process remain a challenge. Hu et al. [15] introduced the spatial pose based exemplars to characterize the Human–Object Interaction (HOI) from still images but it did not work accurately for complex images. This approach does not provide the motion information in the short or long duration of time and hence, scarce in representing the human action from a video sequence. Shao et al. [16] presented the content based search algorithm for localization of human action in the video database. Their work mainly focused on the temporal and spatial localization to decrease the search time of the algorithm. Even though it has a limitation with the large online database, it opens the window for a robust and effective content-based searching algorithm with the existing human action recognition approaches.

Recently, the concept of still images [17–23] has emerged as a popular means for detecting a person's activity or behavior. In these approaches the visual appearance of the object is used to describe the information content. Wang et al. [19] introduce the concept of action recognition based on still images. The shape of human action is represented using a Canny edge detector, and similar body poses are clustered using the spectral clustering method. Li and Ma [20] gave "exemplarlet" based feature descriptor that contains enough visual information to identify in still images. Li and Fei-Fei [21] presented an integrated method that is based on the appearance information on still image and occurrence of action scenes. Thurau and Hlavac [24] proposed the method based on pose information of human action, where they determined the region of interest (ROI) images and further calculated the histogram of gradients with non-matrix factorization to represent the feature vectors. Lopes and Santos [22] proposed the transfer learning approach, where contextual information is extracted from the still images, and a similar approach is used by Zheng et al. [23] for the representation of human action by combining the poselet with contextual information. But in general it is observed that still image based human activity recognition is less effective due to missing temporal information. Hence, more holistic solution is that which utilizes the shape as well motion information together.

Motion temporal information at different orientation is extracted by using the $\Re$-transform [25–27] and extensively used for representing human activity. The $\Re$-transform is applied on the silhouettes of the human body and provides the orientation of the silhouettes. Also, the rate of change of orientation of the human body is different for dissimilar activity. Wang et al. [19] use $\Re$-transform to represent the low-level features due to its advantages of low computational complexity, geometrical invariance. Zhang et al. [25] used a simple approach to the representation of the human activity by using the shape information. They used the $\Re$-transform as a shape descriptor and reported that it works better for the activities that are being performed by the rotation of the human body. Similar work based on $\Re$-transform by Khan et al. [26] for representation of abnormal human activities were carried and it was observed that it is a good descriptor for the orientation based human activity. The properties of $\Re$-transform i.e. invariant to translation, scaling [28] and effectively depicts the change in rotation makes robust feature descriptor.

Nevertheless, most of the reviewed work reveals that an individual feature descriptor has advantages of their own as a single still image based feature representation does not require any background subtraction, morphological operation, or tracking trajectories, thus reducing the computation time and complexity of the system. Therefore, it can be said that these images become, occlusion free and robust to noise. But apart from these benefits, it is also observed that a single still image based technique requires effective positioning of the posture, and it alone does not always provide enough information for recognizing all kinds of activities because it does not contain the temporal information in short or long time duration. In earlier works, it has been observed that $\Re$-transform is effectively used to incorporate the spatio-temporal content of the human activity and found more effective for the abnormal activities representation as compared to the normal activities. In general, human activities are performed by the translation and rotation of human body, and the normal activities have more translation than the rotation while abnormal activities have more rotation than translation.

More recently, several researchers [6,7,10,11] have promoted that multiple feature based fusion techniques give better performance than the individual feature based technique. Hence, in this work a unified structure is proposed by considering the facts of action dynamics. The action dynamics of the human body state that an activity cannot be accomplished without translation and rotation characteristics of the human body. Therefore, for effective representation of human action, an effective descriptor may be formed by incorporating these two characteristics, which could be occlusion free, robust to noise, and computationally less complex.

In this paper, multiple features based integrated structure is proposed where the shape of human pose is recorded in single 2D posture and the motion of the human body is recorded in human silhouettes. The translation provides the change in shape, and sequence of orientation provides the nature of the activity. Due to the translation, the appearance of 2D human postures of different activities is different, and the single key pose is chosen, which have highest variations among the postures of the video sequences. The rotation of key binary human silhouette is computed by applying $\Re$-transform. The binary human silhouette is obtained using texture based foreground segmentation, which is a robust and reliable approach to the illumination change and noise [29]. The key contributions of the work are as follows:

- The appearance of 2D human pose is chosen from the video sequences using fuzzy logic based model. The edge of human body pose is extracted using canny edge detector.
- The edge spatial distribution of gradients at various orientation bins is computed at different sublevels of the single 2D posture for the representation shape of activity.
- The spatio-temporal motion content of human activity is extracted by applying the $\Re$-transform on the key poses of the binary human silhouette. The key poses of binary human silhouettes are chosen on the basis of high energy.
- An integrated model is constructed using the 2D shape information and spatio-temporal motion information of the human activity.
- The performance of the integrated model is measured on publicly available standard datasets using K-nearest neighbor (K-NN) and support vector machine (SVM).
- The comparative analysis of the result achieved through the proposed model is done with the earlier state-of-the-art methods, and an additional analysis of the speed of computation and robustness test of the proposed algorithm is done.

The rest of this paper is structured as follows: Section 2 gives the proposed methodology, which includes the overview of proposed model, extraction key pose, fuzzy logic model, abstraction of spatial edge distributions and motion temporal information using $\Re$-transform, and silhouette extraction. Section 3 gives the details

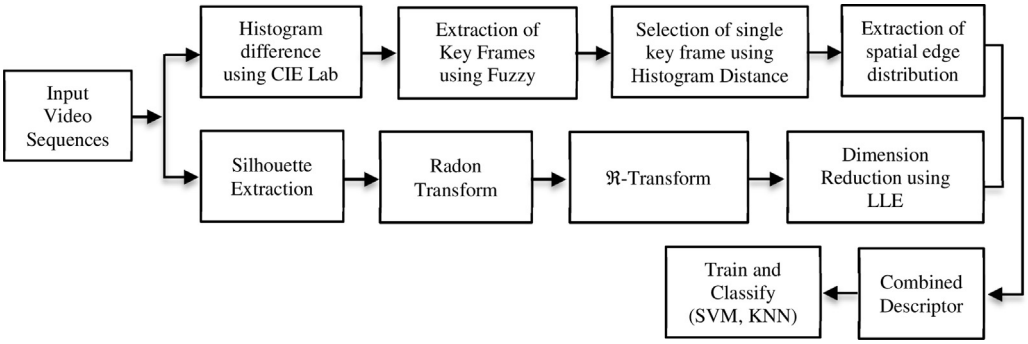

Fig. 1. Workflow diagram of human activity recognition.

of experimental works and discussion of results. The conclusion of the work is explained in Section 4.

2. Proposed methodology

Based on the earlier frameworks for activity recognition used in [6,7,11,12,30] the framework for the action recognition by integrating the translational-based appearance and rotational based motion information is illustrated, which includes overview of propose model, key pose selection, fuzzy logical model, computation of edge spatial distribution and spatio-temporal motion of the human silhouettes. The detailed explanation about the proposed framework is presented in the subsequent sections.

2.1. Overview of the framework

The proposed framework comprises abstraction of spatial edge distribution of 2D postures, extraction of the orientation of silhouettes by using ℜ-transform and integration of spatial edge distribution vector with orientation feature vector. The spatial edge distribution gives shape information of the action, which is computed on the still image, whereas still images are single key frame extracted from video sequence using fuzzy approach.

The workflow diagram of the proposed framework is shown in Fig. 1 and every block is explained in the upcoming sections.

2.2. Extraction of key poses

For the accurate representation of body posture, the key poses of the activity in video sequences are extracted, and these key poses are used to represent the human activity in the video. For selecting key poses, a stack of frames is formed by selecting the frames after a certain interval length in the video sequence because the deviation in the action does not vary instantaneously. These stacks of frames are converted into CIELab color space because it closely conforms to human perception of colors and device independent [31] as compared to the rest of color spaces. In this color space model, L stands for the luminance and a, b components define the color opponent dimensions. The histogram distances D between the frames are computed for all the three components L, a and b as per Eq. (1).

$$D = \left\| \sum_{i=1}^{M} \sum_{j=1}^{N} S_{ij}^{t} - S_{ij}^{t+1} \right\| \tag{1}$$

S, t denotes the frames of the stack, and frame number respectively. The size of frame is $M \times N$. The distance that we computed for the subsequent frames is used to make the Fuzzy logic model.

2.2.1. Fuzzy logic model

In some videos, the information content may be large, and hence, it is difficult to extract the key frames using normal distance metric because some of the information regarding the frames have the risk of being lost. Hence, the fuzzy approach is used to find the key frames, which as explained in Algorithm 1.

The fuzzy logic system is based on the probability model, where the knowledge about the system is partial rather than being accurate. The fuzzy set theory, given by Zadeh [32], was an extension of Boolean logic than the Crisp logic, and provides the measure of uncertainty. Crisp logic gives a hard decision whether it belongs to a group or not.

Consider an example of crisp function represented by set A. This function assigns a value $\mu_A(x)$ to every $x \in A$ such that:

$$\mu_A(x) = \begin{cases} 1 & \text{if and only if } x \in A \\ 0 & \text{if and only if } x \notin A \end{cases}. \tag{2}$$

Fuzzy logic deals with the membership function that defines the probability of a variable belonging to a particular group. The membership function assigns a variable value between 0 and 1. As the value increases from 0 to 1, the degree of membership value also starts increasing for the variable and the membership function is chosen on the basis of the set interval length. Fuzzy logic makes decisions by using simple and flexible 'IF...THEN' statements, which maps the set input to the set of output.

Algorithm 1. Key frames extraction.

Step 1: Input the video sequence and select the frames with the difference of certain frames.
Step 2: Convert these frames into CIELab color space.
Step 3: Compute the histogram distance for 'L', 'a', 'b' components between the frames using Eq. (1).
Step 4: Compute the mean, 'μ_d' for all consecutive frame differences as: $\mu_d = [count\,'L' + count\,'b' + count\,'a']/3$.
Step 5: Create the trapezoidal membership functions dynamically using the computed mean μ_d as shown in Fig. 2, where {small, medium, large} are the linguistic variables.
Step 6: Compute the value of parameters using: $A = (\mu_d - \mu_d * 0.4)$, $B = (\mu_d - \mu_d * 0.3)$, $C = (\mu_d - \mu_d * 0.2)$, $D = (\mu_d + \mu_d * 0.4)$, $E = (\mu_d + \mu_d * 0.5)$, $F = (\mu_d + \mu_d * 0.8)$.
Step 7: Define fuzzy rules, i.e. if the distance between segment frame and its neighboring frame is small or large then it is a key frame. The extracted key frames are as shown in Fig. 3.

2.2.2. Selection of single key frame

The extracted key frames are compared internally and ranked according to the histogram distance values. Higher distance values show that the corresponding frame has higher variations as

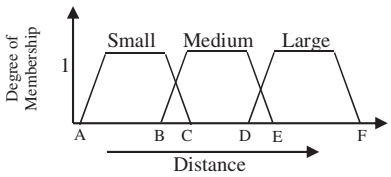

Fig. 2. Fuzzy trapezoidal membership function.

compared to the other frames. Histogram difference is normally considered as a global change in a video sequence, and it helps in detecting abrupt changes in the frame. If the extracted key frames are represented as ''$f_1, f_2, \ldots\ldots\ldots\ldots, f_n$'', then the histogram distance between the successive frames is computed and highest distance key frame is the single key frame that is highest pixel variation to others and denoted as a single still image as shown in Fig. 4(b) and 5.

$$\text{Single Key Frame} = arg\max(D_i), \quad i \in n \tag{3}$$

2.3. Computation of spatial edge distribution

The computation of spatial edge distribution feature is based upon the visual appearance of human body pose. Human body pose provides a significant amount of information for nonverbal communication and based on appearance, certain patterns of body movements are indicative of specific action. Human posture gives the information about the motion of the human body, and the 2-D representations of images give the spatial distribution of posture and characteristics of action or the attitude of the person. The shape based appearance is represented by dividing the ROI of key pose still image into sub-regions at multilevel, and orientations of the edges are counted on the finer scale and is further expressed in the form vector. Edges are more useful for the shape analysis and the Canny edge detector combines the derivative and smoothing properties efficiently to obtain the edges. To remove the edges which are insignificant, a threshold value is selected based on the pixel variation and only those edges are kept, which are having the pixel value greater than the threshold value. The detailed flow of step used for computation of spatial edge distribution is as explained in Algorithm 2.

Algorithm 2. Computation of spatial edge distribution.

Step 1 – Input video sequence $\mathcal{F}(x, y, t)$; t set of frames.
Step 2 – Select a single key frame as explained in Section 2.2.
Step 3 – Select ROI and normalize to the fixed dimension of 50×50, and denote as: $\mathfrak{B}(x, y, \varphi)$, where $0 \leq x, y \leq 50$.
Step 5 – Compute the edges of ROI using canny edge detector and denoted as: $\varepsilon(x, y, \varphi) = \text{Canny}\langle\mathfrak{B}(x, y, \varphi)\rangle$.
Step 6 – Compute the spatial edge distribution vector at different levels as follows:
a. At level-0, the magnitude $\mathcal{M}(x, y)$ and orientation $\varphi(x, y)$ at any point (x, y) of the entire image $\varepsilon(x, y, \varphi)$ is determined as:
$\mathcal{M}(x, y) = \sqrt{\mathcal{G}_x(x, y)^2 + \mathcal{G}_y(x, y)^2}$ and $\varphi(x, y) = \arctan\left[\frac{\mathcal{G}_y(x,y)}{\mathcal{G}_x(x,y)}\right]$.
Where $\mathcal{G}_x(x, y)$ and $\mathcal{G}_y(x, y)$ are image gradients along the x and y directions. Each sub-region is quantized into the 8-orientation bins, evenly spaced over 0°–360°. The obtained feature vector for entire image is of 8×1 dimension.
b. At level-1, the entire image $\varepsilon(x, y, \varphi)$ is divided into 4 sub-image regions, and denoted as: $\varepsilon(x, y, \theta) = \{\mathcal{S}_1(x, y, \varphi), \mathcal{S}_2(x, y, \varphi), \mathcal{S}_3(x, y, \varphi), \mathcal{S}_4(x, y, \varphi)\}$. The feature vector is computed as in **step 6-a**, and dimension is of $8 \times [1+4]$.
c. At level-2, each sub-image region is further divided into 4 sub-blocks, and feature vector is computed as in **step 6-a.** The obtained feature vector for 16 sub blocks is of $8 \times [1+4+16]$.
Step 7 – The final feature vector is formed by combining all the levels and obtained as: $\mathcal{F}_a = [8 \times 1] + [8 \times 1] \times 5 + [1 \times 8] \times 21 = 216$. The result of spatial edge distributions at different levels are as shown in Fig. 6.

Fig. 7 shows that the spatial edge distribution of gradients of different activities postures and these distributions are distinct from each other because the representation of the patterns of the histograms of different activities is different with an increase in the degree. At a lower degree (near zero) the peaks are higher and much more variant in magnitude. Hence, these features are proficient in the representation of human activities. The spatial edge disribution of gradients are computed at the mixed levles of 0, 1, and 2. The compuation of these features are the magnitude of gradients at different orienation bins, which is easy and simple to compute. This representation may be less effective for the recognition of activities like "walking" and "jogging" because of reasonably similar histogram of these activities. Hence, to alleviate these shortcomings the temporal content of the activities are computed and integrated with the global information obtained through the kinematics of geometric orientation distribution points.

Fig. 3. Extracted key frames.

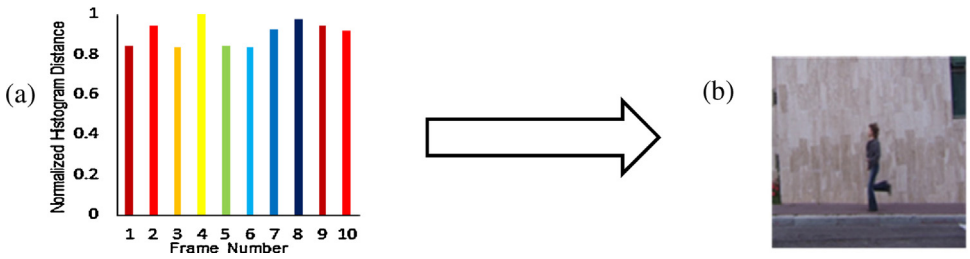

Fig. 4. (a) Plot of histogram distance of key frames (b) single frame i.e. third having maximum distance.

Fig. 5. Single key frames extracted from different activities of two different video datasets Row 1: Wiezmann, Row 2: KTH.

2.4. Computation of orientation features

The orientation features give the directional information of the object, and most of the human activities are perfomed by changing the orientation of the body. The directional information of human activity is computed using the ℜ-transform and further this information is integrated with the translated information, which is computed in the previous section. The ℜ-transform gives the orientation information of an object, and it is used for the abnormal human activity recognition [26]. The ℜ-transform is computed via Radon transform (RT), by applying the RT on the binary silhouettes of the human activity. Radon transform gives the directional features in the range of angle (0–179°) and is defined as the integral of a silhouette image $I(x, y)$ from $-\infty$ to ∞, and is denoted as:

$$R_f(\rho,\theta) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} I(x,y)\delta(\rho - x\cos\theta - \sin\theta)\partial x\partial y \tag{4}$$

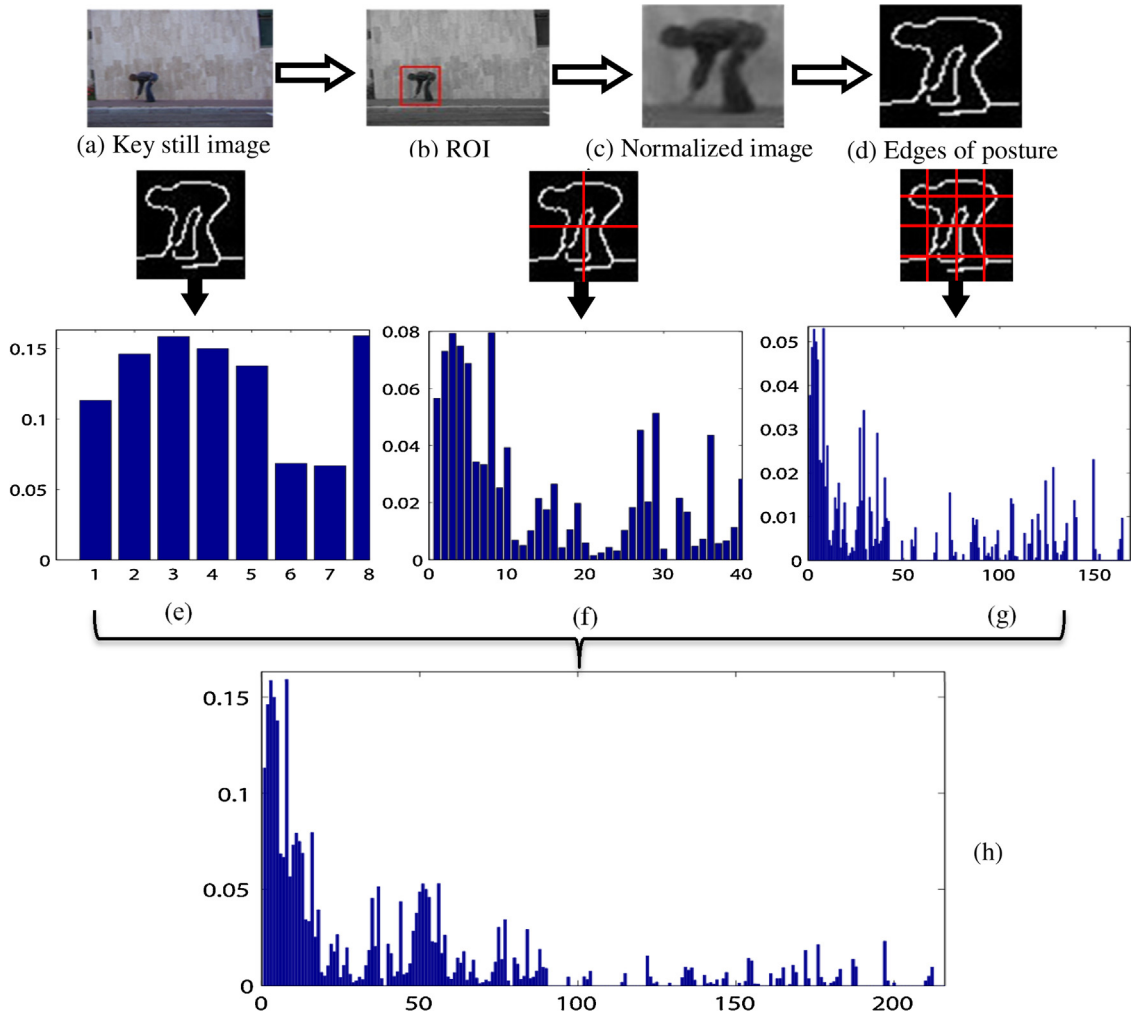

Fig. 6. Simulation results: (a)–(d) preprocessing steps, (e)–(h) shows the computation of spatial edge distribution vector at level 0, 1, 2 and final feature vector respectively.

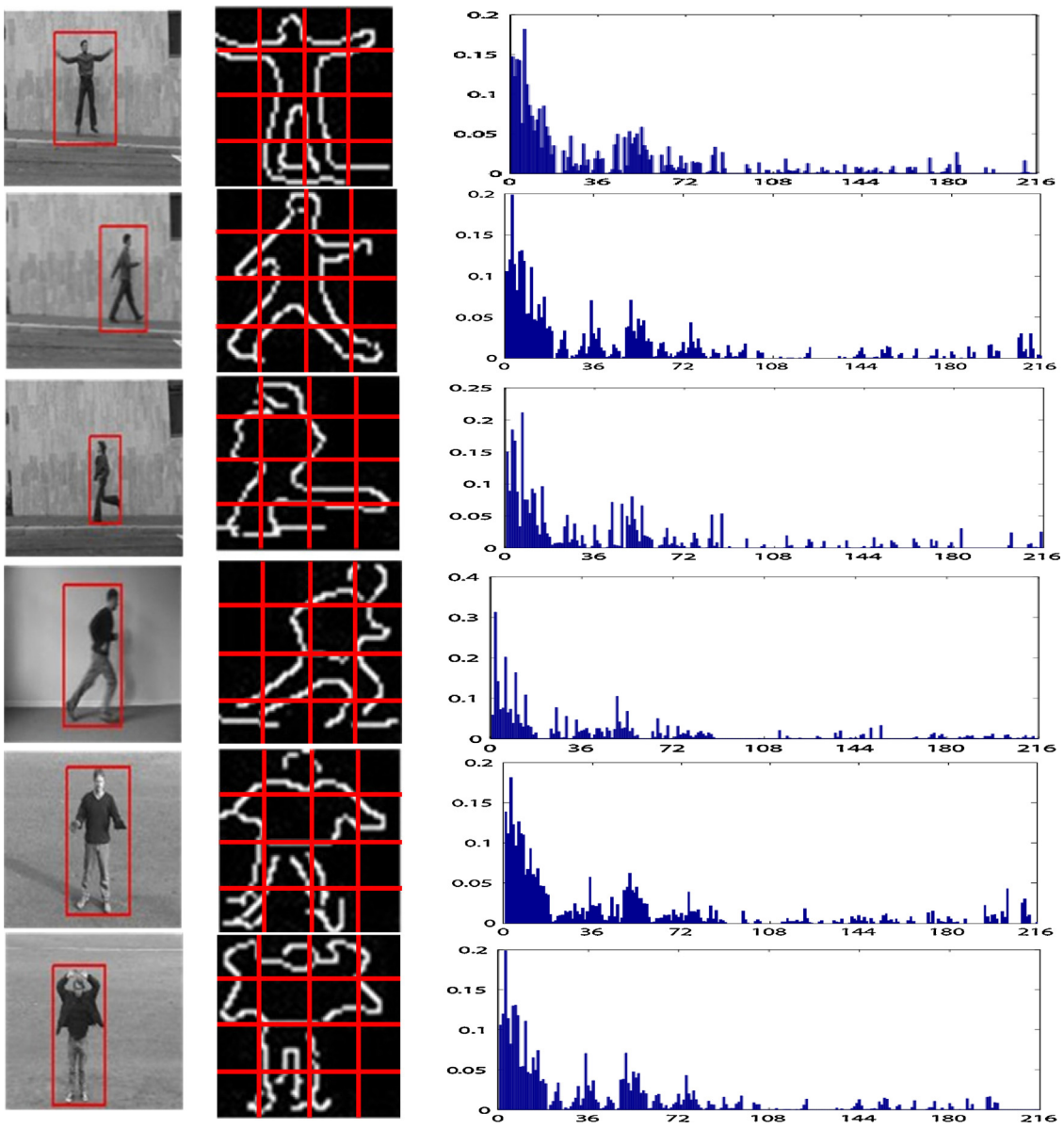

Fig. 7. Result of spatial edge distribution vector of different activities: Column 1: ROI of still images, Column 2: edges of postures, Column 3: spatial edge distribution at final level.

where $\delta(.)$ is defined as the Dirac delta function that is zero everywhere except at the origin. The perpendicular distance from the origin to the radon line is defined as:

$$\rho = x\cos\theta + y\sin\theta \quad \text{for}(0 \le \theta \le \pi), (-\infty \le \rho \le \infty) \tag{5}$$

and shown in Fig. 8(b), where 'θ' is the angle between the horizontal axis and the projection line.

RT cannot restore all the parameters of the original geometric transformation when translated, rotated, or scaled the image. Tabbone et al. [28] introduce the $\Re$-transform that is defined as integral transform of the squared values of RT and expressed as:

$$\Re(\theta) = \int_{-\infty}^{\infty} R_f^2(\rho,\theta)\partial\rho \tag{6}$$

RT gives the feature representation in 2-D while $\Re$-transform gives the compact representation in 1-D. Normalization of $\Re$-transform will further improve the similarity measure, and thus,

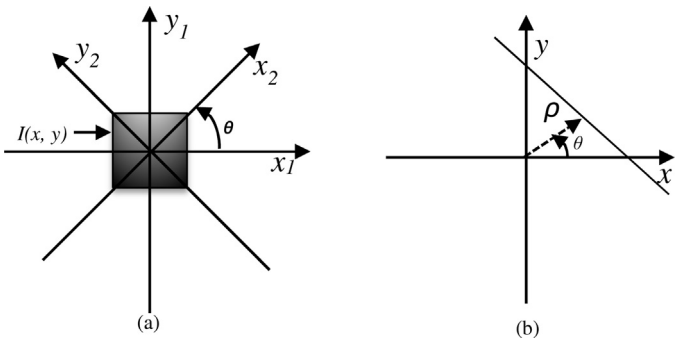

Fig. 8. (a) Illustrates the projection of lines over 2-D function $I(x, y)$. (b) The position of the projection line.

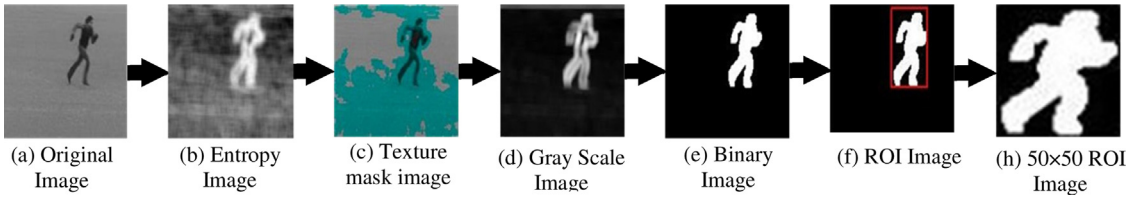

Fig. 9. Illustration of finding a normalized ROI binary silhouette image.

representation of features in a very compact manner can be achieved. The normalized form of the ℜ-transform is defined by Khan et al. [26].

$$\Re\text{norm}(\theta) = \frac{\int_{-\infty}^{\infty} \Re(\theta)\mathrm{d}\theta}{\max(\Re(\theta))} \tag{7}$$

The RT is applied on the ROI of binary silhouette images, which are obtained from the key poses of the frames. Binary silhouette images are obtained using texture based segmentation method (Haralick et al. [33]) and texture is one of the important features to identify the objects.

2.4.1. Silhouette extraction

The silhouette is the basic unit of human activity, which is formed by extracting the foreground object from the rest of the video sequences. A method for describing different textures as presented in [33] is used for silhouette extraction. Entropy is one of the most important parameters that describe the texture information in an image and can be expressed as:

$$\zeta = \sum_i \sum_j \rho(i,j)\log(\rho(i,j)) \tag{8}$$

where $\rho(i,j) = \frac{M(i,j)}{\sum_{i,j} M(i,j)}$ is the probability density function; where i and j are indices to the co-occurrence matrix M. The entropy of the image is used to describe the complexity of the background and a higher value indicates greater complexity in the image background.

The filter matrix is generated for a pixel and its entropy is calculated in a 9×9 neighborhood mask. Converting this filter matrix in binary form gives an image with white spots in different areas. Applying this mask to the raw image provides a silhouette image which is as shown in Fig. 9.

The segmented image may contain different white blocks, but not all of them are of human silhouettes. By comparing the sizes of these blocks, the image with the largest area is selected, which is a human silhouette.

The key poses of the frames are the most discriminating frames that are extracted using the concept of the size of the silhouettes, which is explained as Algorithm 3.

Algorithm 3. Extraction of key poses of the human silhouette.

Step 1: Input the Video Sequence.
Step 2: Extract the silhouette using texture based segmentation.
Step 3: Apply Weiner filter for smoothing.
Step 4: Calculate the size of each silhouette images using

$$U_t = \sum_i^M \sum_j^N ||I(i,j,t)||^2.$$

Step 5: Find the mean value (μ) as: $\mu = \frac{[U_1, U_2 \ldots\ldots U_n]}{n}$, where $U_1, U_2, \ldots . U_n$ denotes the size of silhouette images and 'n' is the *total* number.
Step 6: If the size of the silhouette is greater than the mean value, then it is a key frame.
Step 7: Otherwise EXIT.

2.4.2. Properties of ℜ-transform and Radon transform

The important properties of RT and ℜ-transform described in Table 1 shows that it has the robustness to the scale, translation, and rotation.

These properties are experimentally verified using MATLAB 2012b and are shown in Fig. 10.

In Fig. 10, the rotated image shows more variation in the brighter portion of RT than the other images because there is more deviation in the pixel values corresponding to projection lines. The magnitude of the translated image varies as compared to the scaled image, but the signal representation of ℜ-transform remains the same. Therefore, from Table 1 and Fig. 10 it can be said that ℜ-transform is invariant to translation and scaling in the plane, but orientation in the plane will change due to the phase shift. Hence, ℜ-transform is an efficient method for describing the motion temporal of the actions that have more variation in the orientation such as bending, walking, and running. The representation of ℜ-transform of different activities is as shown in Fig. 11, and it can be visualized that ℜ-transform representation of different activities is different. Hence, it can be said that ℜ-transform is a discriminating feature,

Table 1
Comparison of Radon and ℜ-transform properties.

Properties	Radon transform	ℜ-transform
Scaling by 'α'	$[(R_f(\alpha\rho,\theta))] = \frac{1}{\alpha^2}(R_f(\alpha\rho,\theta))$	$\frac{1}{\alpha^3}\int_{-\infty}^{\infty} R_f^2(x,\theta+\theta_O)\partial x = \frac{1}{\alpha^3}\Re(\theta)$
Translation by (x_o, y_o) vector	$R_f(\rho,\theta) = R_f(\rho - x_o\cos\theta - y_o\sin\theta, \theta)$	$\Re(\theta) = \int_{-\infty}^{\infty} R_f^2((\rho - x_o\cos\theta - y_o\sin\theta),\theta)\partial\rho$
Periodicity	$R_f(\rho,\theta) = R_f(\rho,\theta+2k\pi)$, k is any integer value	$\Re(\theta\pm\pi) = \Re(\theta)$, period is π
Rotation by 'θ_o'	$R_f(\rho,\theta) = R_f(\rho,\theta+\theta_1)$	$\Re(\theta) = \int_{-\infty}^{\infty} R_f^2(\rho,\theta+\theta_O)\partial\rho$

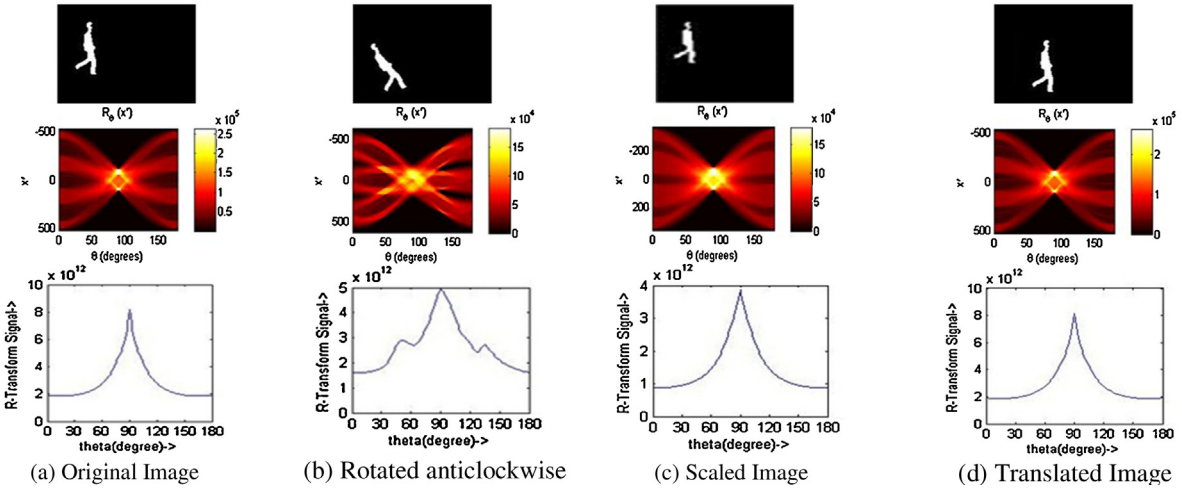

Fig. 10. Shows $\Re$-transform is variant to the rotation and invariant to scaling and translation Row 1: silhouette images, Row 2: Radon transform (RT), Row 3: $\Re$-transform signal.

which gives the directional information by calculating the pixel variations at different angles. But in some activity where the change in orientation is quite similar, as in the case of walking and running, it may not be able to discriminate exactly. It can also be said that the $\Re$-transform alone cannot represent all kinds of human actions efficiently. The actions that have similar orientation characteristics like walking and running can be distinguished by the spatial change in the video sequence.

The $\Re$-transform is computed on the binary silhouette of size 50×50 and after concatenation, it can be represented as a 1×2500 feature vector. The $\Re$-transform gives the reduced dimension of a silhouette image to 1×180 and for single action representation the dimension of the feature vector is 7×180. Further, to represent features in a reduced set, the LLE is applied.

2.4.3. Dimension reduction using local linear embedding

LLE is defined as the unsupervised manifold learning based dimension reduction method. Unlike PCA, LDA, KPCA which is based on the maximization of the variance, LLE searches the nearest neighbors around the data point in high-dimensional space. For each nearest neighbor, weights are assigned, which describe whether the data points are linear to neighbors or not. The weights are assumed to be invariant against translational, rotational, and scaling parameters, which possesses the local properties as in the original space. These weights reconstruct the data points into another dimensional space $\simeq\mathcal{D}\simeq$ and furthermore in the $\simeq\mathcal{D}\simeq$ dimensional space, data points are mapped into lower dimension 'd' ($d \leq \mathcal{D}$). While transforming the features data points from 1-D space to another dimensional space, it maintains

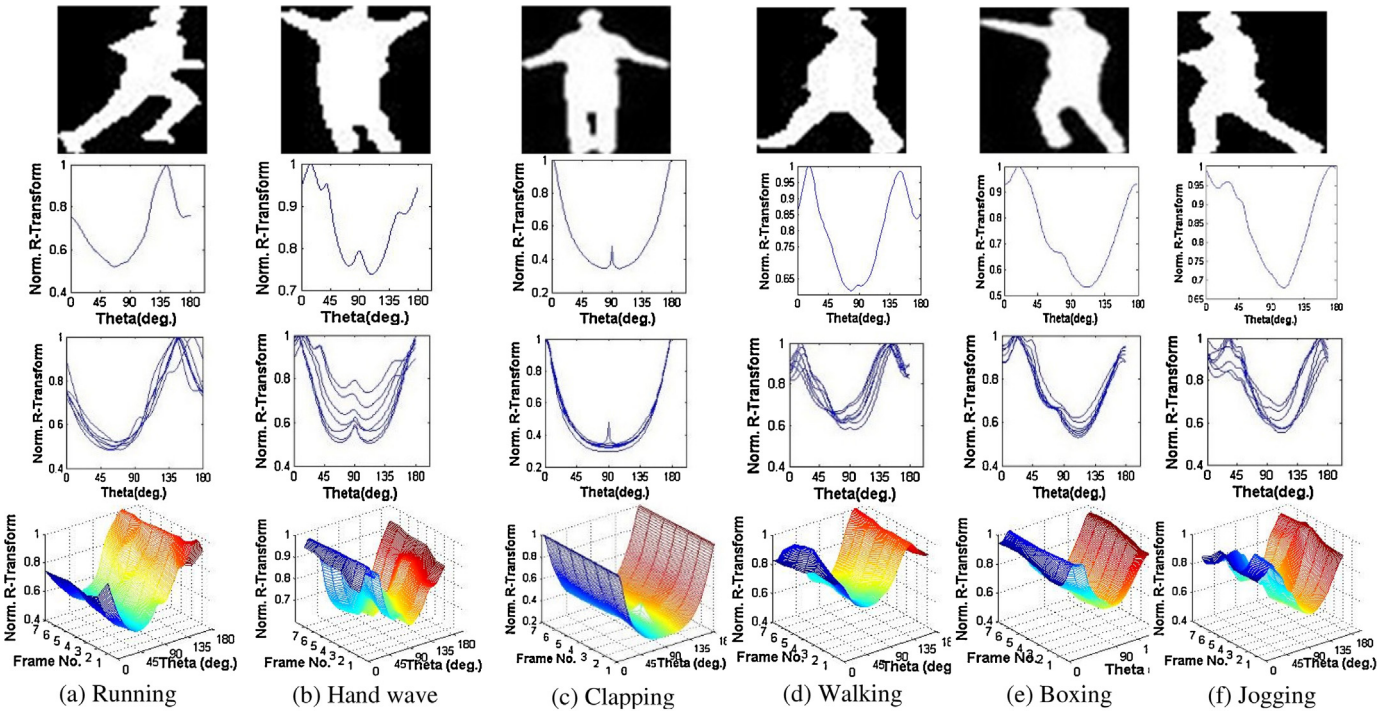

Fig. 11. Representation of $\Re$-transform for different activities: Row 1: 50×50 silhoutte image, Row 2: $\Re$-transform, Row 3: $\Re$-transforms of key frames, Row 4: 3D $\Re$-transform of sequence.

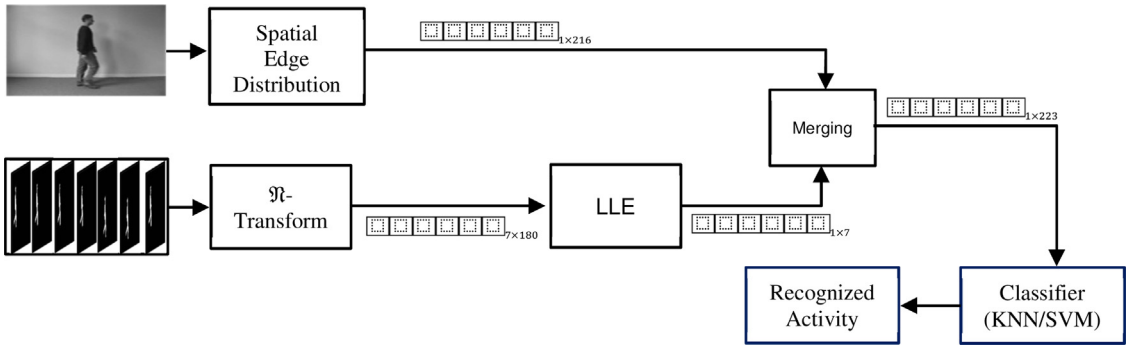

Fig. 12. Flow diagram of combing features.

the local characteristics of the features, which is explained as Algorithm 4.

Algorithm 4. LLE computation.

Step 1: Define training set vector $Z = \{x^{(n)} \in \mathbb{R}^d\}_{n=1}^{\mathcal{N}}$, of $d \times \mathcal{N}$ dimension.
Step 2: Construct the weight matrix $'\mathcal{W}$ to record the linear reconstruction coefficients, where the ith row $[\mathcal{W}_{ij} = 0]_{1 \le j \le \mathcal{N}}$ contains the initial coefficients of the sample 'i' from its neighbors.
Step 3: The set containing all the neighbors of $x^{(i)}$ is denoted as $\Psi(i)$
$\Psi(i) = |\Psi(i)|$ = neighbors of $x^{(i)}$.
Step 4: Define linear coefficients $\mathcal{W}^{(i)}$ for each sample i with the least reconstruction error using $\varepsilon(\mathcal{W}) = argmin_{\mathcal{W}} \left\| x^{(i)} - \sum_j \mathcal{W}_{ij} x_j \right\|^2$.
Step 5: Two constraints are set on $\mathcal{W}^{(i)}$
i. $\mathcal{W}_j^{(i)} = 0$, if $j \neq \mathcal{N}(i)$ (Neighborhood constraint). It means that $x^{(i)}$ is only reconstructed from its neighbor.
ii. $\sum_{j=1}^{\mathcal{N}} \mathcal{W}_j^{(i)} = 1$ (For translation invariant).
Step 6: Re-embedding in the reduced feature space, $\mathcal{Z} = \{z^{(n)} \in \mathbb{R}^m\}_{n=1}^{\mathcal{N}}$ of $m \times \mathcal{N}$ dimension and to maintain the local structures in $\mathcal{Z}$, it is represented as: $\phi(\mathcal{Z}) = \sum_{i=1}^{\mathcal{N}} \left\| \mathcal{Z}^{(i)} - \sum_{j=1}^{\mathcal{N}} \mathcal{W}_{ij} \mathcal{Z}^{(j)} \right\|^2$.

To get the coefficient in reduced feature set, the function $\phi(\mathcal{Z})$ is minimized. This method is better for representation of a non-linear distribution of data points and also maintains the local structures [34]. LLE is dependent upon the representation of data points in higher dimensional space; if they are different in representation, then LLE cannot ensure their reconstruction in lower dimensional space. In this case, the dimension of ℜ-transform feature set of 7×180 is reduced to 1×7 by applying the LLE algorithm.

2.5. Final feature vector formation

The final feature vector is formed by integrating the feature set obtained in Sections 2.3 and 2.4 which is shown in Fig. 12. The formed feature vectors are used for training and testing of the video sequence.

The spatial edge distribution feature vectors are computed as explained in the previous section and have a dimension of 1×216. Further, these feature vectors are combined with ℜ-transform features computed on the binary silhouettes. The ℜ-transform computed feature vector had the dimension of 7×180. Here, the number of key frames used are 7. The dimension of ℜ-transform feature is reduced using LLE, which gives the reduced discriminative feature vector of 1×7. Finally, the spatial edge feature vectors and ℜ-transform feature vectors are combined and give the resultant feature vector of $[1 \times 216 + 1 \times 7] = 1 \times 223$ dimension. The resultant feature vector is used to represent the activities of the dataset and these are classified using K-NN and SVM.

2.5.1. K-nearest neighbors

The K-nearest neighbor (K-NN) is a simple machine learning algorithm for classifying objects based on a similarity measure (distance function). It is called the non-parametric method as it does not learn an explicit mapping from the training data. There are various distance functions which are used to measure the similarity, but they depend on the type of features of the data. If the distance is small, then there is more similarity between the two samples. Euclidean distance, Mahalanobis distance and Hamming distance are some of the common distance functions. If the features are real valued then, Euclidean distance is used and if the features are binary valued, then Hamming distance is used. The Euclidean distance between any two points $w = (x_1, x_2, \ldots\ldots\ldots, x_n)$ and $z = (y_1, y_2, \ldots\ldots\ldots, y_n)$ is as given in [35];

$$\text{Distance}(w, z) = \sqrt{\sum_{i=1}^{n=m} (x_i - y_i)^2} \tag{9}$$

K-NN classifier is simple and easy to implement. It gives better results in the case of large training data with the high value of 'K' and it uses the local information of data which makes it highly adaptable.

2.5.2. Support vector machine

SVM algorithm is based on the supervised learning [36] i.e. it has prior knowledge about the features that are to be used to predict the state of the class labels. It transforms the set of vectors into higher dimensional space where it continuously finds the optimal hyperplane that separates the class labels. The hyperplane is called a decision boundary that distinguishes the two classes and maximizes the separation, margin between itself and those lying nearest to it. The nearest set of points are called support vectors. Support vectors are features of the samples which are closest to the hyperplane of an SVM. Therefore, the determination of the location of most important data is near the hyperplane and which is formed by the set of lines drawn between the class samples. Initially, it is used for classification of two classes, but later on it is extended to multiclass problems. In two class problems, it is one of the most efficient and robust classifier. Hence, by utilizing the advantages of the two classes of SVM, it is extended to multiclass SVM. The frequently used approaches of multiclass SVMs are: (I) one-against-rest (II) one-against-one (III) directed acyclic graph SVM. In this paper, the multiclass SVM, the one-against-rest approach has been used. To perform the classification of the two classes, a nonlinear kernel based SVM can be applied by mapping the input data into the higher dimensional feature space. Consider N training samples of feature set $\mathcal{T}_k$. Whereas:

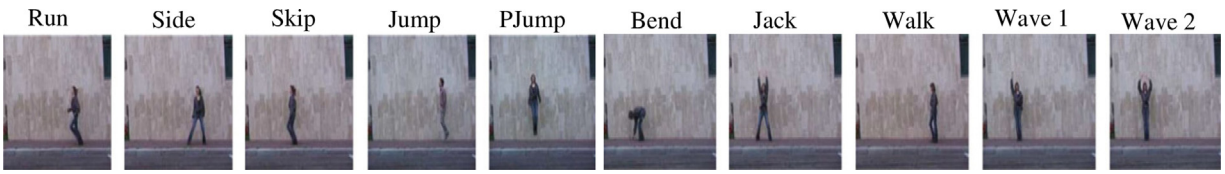

Fig. 13. Sample frames of Weizmann human action dataset.

$$\mathcal{T}_k = \{(x_i, y_i) \| x_i \in \mathbb{R}^m, y_i \in (-1, 1)\} \tag{10}$$

where m is the dimensional feature vector signifying the ith training sample and $y_i \in (-1, 1)$ is the class label of x_i. An optimal hyper plane can be expressed as:

$$\psi(x) = sgnf\left(\sum_i^N \alpha_i y_i \mathcal{K}(x_i, x_j) + \mathfrak{b}\right) \tag{11}$$

where *sgnf*(.) represents the signum function, and $\mathcal{K}(x_i, x_j)$ is a nonlinear predefined kernel function that satisfies the condition of Mercer's [25]. In this work the radial basis function used is as follows:

$$\mathcal{K}(x_i, x_j) = \exp(-p||x_i - x_j||^2), \quad p > 0 \tag{12}$$

The coefficient α_i and b can be determined using the concept of maximization of hyperplane value and written as:

$$argmax\left[\sum_i^N \alpha_i - \frac{1}{2}\sum_{i,j} \alpha_i \alpha_j y_i y_j \mathcal{K}(x_i, x_j)\right] \tag{13}$$

where $\underbrace{\min}_{\alpha} \sum_i \alpha_i y_i = 0, \quad 0 \le \alpha_i \le C, \quad \forall i \quad \text{and} \quad y_i = \mp 1.$

where C is the penalty parameter that signifies the tradeoff between the maximizing the margin and minimizing the training set error. Although, the SVM is a binary classifier, it can still be extended for M-class classification in human activity recognition. One of the widely used multiclass techniques is one-against-rest [37], and the same approach is used in this work.

3. Experiment and results

In order to test the performance of proposed fusion method, an experiment is conducted on three publicly available datasets i.e. Weizmann [38], KTH [39] and Ballet movement [40] datasets. The classification accuracy is computed in terms of recognition rate using SVM and K-NN classifier with a testing scheme of leave-one-out manner. In leave-one-out test scheme, one sequence is used as a test video while others are used as training sets. This procedure continues until all the sequences are tested for single activity, and it further proceeds to other activities. The advantage of this method is that whole information of the dataset is used to calculate the precision. The effectiveness of the approach is measured in terms of average recognition rate (ARR), which is defined as:

$$\text{ARR} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \times 100 \quad \text{(In percentage)} \tag{14}$$

where TP, TN, FP, and FN are the true positive, true negative, false positive, and false negative. The highest average recognition rate achieved on each dataset is compared with the similar state-of-the-art techniques.

Weizmann dataset: This dataset is introduced by Gorelick et al. [38]. This dataset contains 90 videos with a frame rate of 15 fps and each frame having a size of 144×180. In a video sequence, nine people are performing ten different actions, categorized as walk, run, jump-jack, bend, jumping forward on one leg, jumping on two legs in the forward direction, jumping in place, sideways jump, one hand wave, two hand wave is shown in Fig. 13.

KTH dataset: This dataset is introduced by Schuldt et al. [39] and it is a more challenging dataset as compared to the Weizmann dataset. The dataset consists of six basic activities, namely; 'hand-clapping' 'hand-waving', 'jogging', 'jumping', 'running', and 'walking'. Each activity has 100 videos for four different scenarios in different lighting conditions, indoor and outdoor conditions. All these video sequences are recorded in a uniform background with a static camera of frame rate 25 fps and further down-sampled to the spatial resolution of 160×120 pixels. Sample images of the dataset are shown as in Fig. 14.

Ballet dataset: Ballet dataset [40] is the one of the most challenging dataset, which is consist of eight ballet movements. The sample postures of this dataset are as shown in Fig. 15.

The name of various activity postures areas: "Hopping (HP)," "Jump (JP)," "Left-to-Right Hand Opening (LRHO)," "Leg Swinging (LS)," "Right-to-Left Hand Opening (RLHO)," "Standing with Hand Opening (SHO)," "Stand Still (SS)" and "Turning (TR)". The key challenges of the dataset are the considerable amount of intra-class dissimilarities in terms of spatial and temporal scale, speed, and clothing.

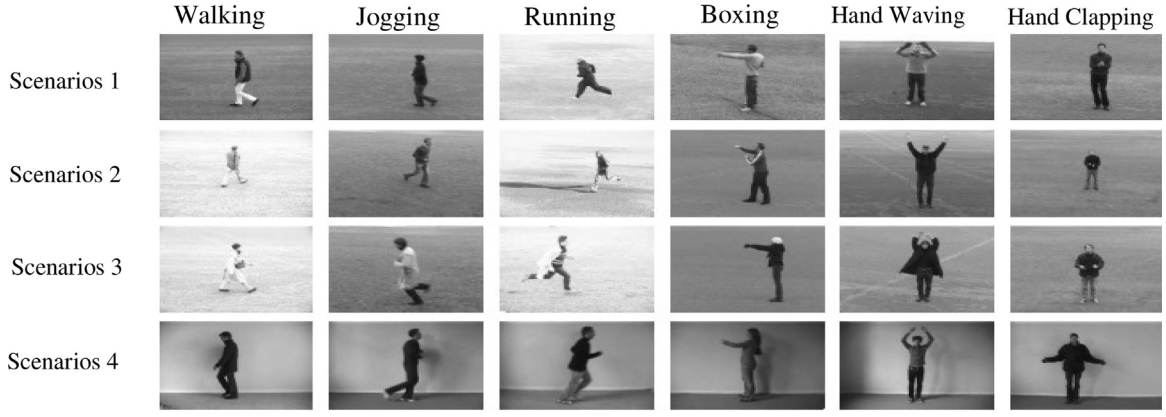

Fig. 14. Sample frames of KTH dataset.

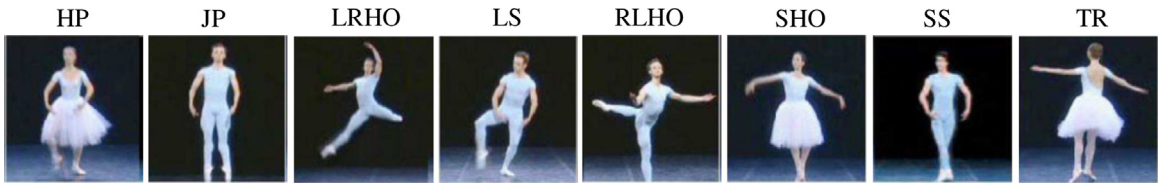

Fig. 15. Images of the Ballet data set depicting eight movement of actions.

Table 2
The classification result on Weizmann dataset.

Activities	Bend	Jack	Jump	Pjump	Run	Side	Skip	Walk	Wave 1	Wave 2	ARR (%)
KNN (%)	100	100	100	100	89	89	89	100	89	100	95.6
SVM (%)	100	100	100	100	100	100	100	100	100	100	100

Table 3
The classification result on KTH dataset.

Activities	Boxing	Hand waving	Hand clapping	Jogging	Running	Walking	ARR (%)
KNN (%)	100	92	92	92	78	100	92.3
SVM (%)	100	100	92	92	89	100	95.5

Table 4
The classification result on Ballet dataset.

Activities	HP	JP	LRHO	LS	RLHO	SHO	SS	TR	ARR (%)
KNN (%)	83	82	100	91	85	100	86	87	89.25
SVM (%)	88	87	100	100	100	93	92	94	94.25

3.1. Performance evaluations

Tables 2–4 show the performance of proposed methodology on Weizmann, KTH and Ballet movement datasets. From these tables, it can be clearly perceived that our proposed method gives a comparable result with the state-of-the-art methods.

From Tables 2–4 the highest recognition accuracy achieved on each dataset is through the SVM in comparison with K-NN. The ARR achieved on Weizmann dataset is 100% due to the less variation in intra-class activity, and the clean background. Due to clean background, the accurate extraction of the silhouette is accomplished which favors to achieve high accuracy. The KTH dataset is more challenging as compared to Weizmann dataset due to variations in lighting conditions, camera angle and human size within the sequence. Therefore, the silhouette extraction is quite difficult as compared to the previous data set. Hence, we achieved less ARA (95.50%) as compared to Weizmann dataset.

The Ballet movement data set is highly complex data set in terms of intra-class dissimilarities like execution, speed, clothing, etc. The ARA achieved on this data set is 94.25% that is less than the Weizmann and KTH dataset due to the complex actions. The maximum error is caused due to high similarity between the hopping and the jumping action. In some actions, it was observed that there is occlusion in the extracted silhouette frames, which poses difficulties in the computations of orientation features from the $\Re$-transform but the multiple fusion of the orientation and spatial distribution gives significant improvement in the recognition accuracy. The spatial edge distribution of gradients vector greatly improves the accuracy of the ballet dataset as it gives distinguishable features, which are more evident as compared to the $\Re$-transform characteristics. For $\Re$-transform, the human silhouette is extracted but as given that the dataset is complex as in some cases, hands and body parts get occluded which creates the confusion in recognizing the action. But with fusion techniques it involves all the parameters and efficiently improves the recognition in this dataset.

The performance of the proposed approach is also assessed in terms of processing speed. The algorithm is implemented using MATLAB 2012a on a computer, which has a specification of corei5, 1.60 GHz Intel processor with 2 GB RAM. For a single video in each dataset i.e. Weizmann, KTH and Ballet, the feature extraction time is 34 s, 45 s and 43 s respectively. For testing a video sequence, the classification time for each dataset i.e. Weizmann, KTH and Ballet are 0.0012 s, 0.0018 s and 0.0015 s respectively. The computation time can be further reduced if all algorithms are implemented in C++/Open CV on high configuration machine.

3.2. Comparison of results

The highest recognition rate achieved on these datasets are compared with similar state-of-the-art techniques, which is as presented in Tables 5 and 6.

Table 6 shows the comparison of our result with the similar state-of-the-art methods, which use the Ballet data set. In this experiment the similar setup is used as Fathi and Mori [40], Wang and Mori [45], Ming et al. [46], Iosifidis et al. [47], and Vishwakarma and Kapoor [44] thus this comparison is fair. From Table 6, it can be seen that the ARR achieved through proposed approach is highest among the compared earlier techniques. Hence, it can be said that the proposed approach gives better recognition result.

Table 5
Comparison of ARR on KTH and Weizmann datasets.

Methods	Weizmann (%)	KTH (%)
Dollar et al. [41]	85.20	81.17
Niebles et al. [42]	90.00	83.33
Ikizler et al. [43]	100.00	89.40
Bregonzio et al. [7]	96.66	94.33
Dou and Li [11]	94.20	92.70
Vishwakarma and Kapoor [44]	97.7	92.4
Proposed method	100.00	95.50

Table 6
Comparison of ARR on Ballet movement dataset.

Methods	Fathi and Mori [40]	Wang and Mori [45]	Ming et al. [46]	Iosifidis et al. [47]	Vishwakarma and Kapoor [44]	Proposed method
ARR (%)	51	91.3	90.8	91.1	92.75	**94.25**

Table 7
Independent test result.

Input sequences (Weizmann-testing)	Predicted sequences (KTH-training) (%)			
	Running	Walking	Jogging	Waving
Running	94	2	4	0
Walking	0	98	2	0
Waving	0	0	0	100

3.3. Independent test analysis

In order to evaluate the robustness of the algorithm under the independent test sample of a video sequence, a training and testing experiment is performed using support vector machine as a classifier. In this process, two different environmental settings of the video sequences (KTH and Weizmann) of the similar kind of activities are chosen for training and testing. The running, walking and 2-hand waving activities of Weizmann dataset is similar to the running, walking, jogging, and waving activities of KTH dataset. These four activities of the KTH dataset are labeled to train the classifier because KTH dataset is having the adequate number of video samples compared to Weizmann dataset, which is considered to be a good training environment. The three activities of the Weizmann dataset are used for testing and the result of testing is as shown in Table 7.

From the test result shown in Table 6, it is observed that the proposed algorithm is robust due to significant prediction rate. There are cases, where prediction rate is less due to inter-similarity of actions, but results are optimum and satisfactory considering the variations of datasets.

4. Conclusion and future work

In this paper, two important characteristics of human activity i.e. appearance and orientation are used to recognize the human activity. For representation of appearance and rotation the still images based edge spatial distribution and human silhouette based $\Re$-transform are used. A single still image is extracted from a video sequence using histogram distances between the key poses of the frames, and further spatial edge distribution is determined. The spatial distributions of edge gradients are computed at different levels of various orientation bins. For the computation spatial distribution feature, as the number of levels for the computation of spatial edge distribution increases, the dimension of feature vector increases but the recognition accuracy does not increase significantly.

The $\Re$-transform gives the orientation feature of human activity, which is computed on the silhouettes of the activities and silhouettes are extracted using texture based segmentation method. The orientation provides the knowledge about the flow of action relating to time and the global change in the object. As the increase in the number of silhouette frames, the recognition accuracy is incremental but the computation time increases significantly, which may lead to the high complexity of the system.

A final feature descriptor is formed by incorporating the appearance and orientation features of the activity. The final descriptor offers numerous distinctive feature vectors, which escort us to robust and noise free action modeling. This approach is an example of the integration of global rotation characteristics with localized structural information, and it is believed that this integration methodology can be further explored on the varied datasets making an improvement and combining the wide variety of existing human action techniques.

In future, this work can be extended to the more complex as well as varied data sets, which may be close to the real life unconstraint environment.

References

[1] Turaga P, Chellappa R, Subrahmanian VS, Udrea O. Machine recognition of human activities: a survey. IEEE Trans Circuits Syst Video Technol 2008;18(11):1473–88.
[2] Holte M, Moeslund T, Nikolaidis N, Pitas I. 3D human action recognition for multi-view cameras systems. In: 3DIMPVT; 2011.
[3] Agrawal J, Rayoo M. Human activity analysis: a review. ACM Comput Surv 2011:16–43.
[4] Lim HC, Vats E, Chan CS. Fuzzy human motion analysis: a review. Pattern Recognit 2015;48(5):1773–96.
[5] Ziaeefard M, Bergevin R. Semantic human activity recognition: a literature review. Pattern Recognit 2015;48(8):2329–45.
[6] Shao L, Gao R, Liu Y, Zhang H. Transform based spatio-temporal descriptors for human action recognition. Neurocomputing 2011;74(6):962–73.
[7] Bregonzio M, Xiang T, Gong S. Fusing appearance and distribution information of interst points for action recognition. Pattern Recognit 2012;45(3):1220–34.
[8] Shao L, Ji L, Liu Y, Zhang J. Human action segmentation and recognition via motion and shape analysis. Pattern Recognit Lett 2012;33:438–45.
[9] Zhang Y, Lu H, Zhang L, Ruan X. Combining motion and appearance cues for anomaly detection. Pattern Recognit 2016;51(March):443–52.
[10] Zhao D, Shao L, Zhen X, Liu Y. Combining appearence and structural features for human action recogntion. Neurocomputing 2013;113(3):88–96.
[11] Dou J, Li J. Robust human action recognition based on spatio-temporal descriptors and motion temopral templates. Optik 2014;125(7):1891–6.
[12] Wu D, Shao L. Silhouette analysis-based action recognition via exploiting human poses. IEEE Trans Circuits Syst Video Technol 2013;23(2):236–43.
[13] Liu L, Shao L, Li X, Lu K. Learning spatio-temporal representations for action recognition: a genetic programming approach. IEEE Trans Cybern 2016;46(1):158–70.
[14] Liu L, Shao L, Zheng F, Li X. Realistic action recognition via sparsely-constructed Gaussian processes. Pattern Recognit 2014;47(12):3819–27.
[15] Hu J-F, Zheng W-S, Lai J, Gong S, Xiang T. Recognising human-object interaction via exemplar based modelling. In: IEEE International Conference on Computer Vision (ICCV). 2013.
[16] Shao L, Jones S, Xuelong L. Efficient search and localization of human actions in video databases. IEEE Trans Circuits Syst Video Technol 2014;24(3):504–12.
[17] Guo G, Lai A. A survey on still image based human action recogntion. Pattern Recognit 2014;47:3343–61.
[18] Bosch A, Zisserman A, Munoz X. Representing shape with a spatial pyramid kernel. In: ACM International conference on Image and Video Retrieval. 2012.
[19] Wang Y, Jiang H, Drew M, Li Z, Mori G. Unsupervised discovery of action classes. In: IEEE Conference on Computer Vision and Pattern Recognition. 2006.
[20] Li P, Ma J. What happening in a still picture? IEEE Asian Conference on Pattern Recognition. 2011.
[21] Li JL, Fei-Fei L. What, where and who? Classifying events by scene and object recognition. In: IEEE Conference on Computer Vision. 2007.
[22] Lopes A, Santos E, Valle E, Almeida J, Araujo A. Transfer learning for human action recognition. In: International Conference on Graphics Patterns and Images. 2011.
[23] Zheng Y, Zhang Y, Li X, Liu B. Action recognition in still images using a combination of human pose and context information. In: 19th International Conference on Image Processing (ICIP). 2012.
[24] Thurau C, Hlavac V. Pose primitive based human action recognition in videos or still images. In: IEEE Conference on Computer Vision and Pattern Recognition. 2008.
[25] Zhang H, Liu Z, Zhao H. Recognizing human activities by key frame in video sequence. J Softw 2010;5(8):818–25.
[26] Khan Z, Sohn W. Abnormal human activity recogntion system based on R-transform and Kernel Discriminant Technique for Eldely Home Care. IEEE Trans Consum Electron 2011;57(4):1843–50.
[27] Jalal A, Uddin M, Kim T. Depth video based human activity recognition system using translation and scaling invariant features for life logging at smart home. IEEE Trans Consum Electron 2012;58(3):863–71.
[28] Tabbone S, Wendling L, Salmon JP. A new shape descriptors defined on the Randon transform. Comput Vis Image Underst 2006;102(1):42–51.

[29] Li L, Leung MKH. Integrating intensity and texture differences for robust change detection. IEEE Trans Image Process 2002;11(2):105–12.
[30] Vishwakarma DK, Kapoor R. Integrated approach for human action recognition using edge spatial distribution, direction pixel, and R-transform. Adv Robot 2015;29(23):1551–61.
[31] Zeng P, Chen Z. Perceptual quality measure using JND model of the human visual system. In: in EEE International Conference on Electric Information and Control Engineering. 2011.
[32] Zadeh LA. Fuzzy sets. Inf Control 1965;8(3):338–53.
[33] Haralick R, Shanmugam K, Dinstein I. Textural features for image classification. IEEE Trans Syst Man Cybern 1973;6:610–21.
[34] Roweis ST, Saul LK. Nonlinear dimensionality reduction by locally linear embedding. Science 2000;260:2323–6.
[35] Cover TM, Hart PE. Nearest neighbor pattern classification. IEEE Trans Inf Theory 1967;13(1):21–7.
[36] Vapnik VN. An overview of statistical learning theory. IEEE Trans Neural Netw 1999;10(5):989–99.
[37] Qian H, Mao Y, Xiang W, Wang Z. Recognition of human activities using SVM multi-class classifier. Pattern Recognit 2010;31(2):100–11.
[38] Gorelick L, Blank M, Shechtman E, Irani M, Basri R. Actions as space-time shapes. IEEE Trans Pattern Anal Mach Intell 2007;29(12):2247–53.
[39] Schuldt C, Laptev I, Caputo B. Recognizing human actions: a local SVM approach. In: Proceedings of the International conference on Pattern Recognition. 2004.
[40] Fathi A, Mori G. Action recognition by learning mid-level motion features. In: Proceedings of IEEE conference on Computer Vision and Pattern Recognition. 2008.
[41] Dollar P, Rabaud V, Cottrell G, Belongie S. Behavior recognition via sparse spatio-temporal features. In: 2nd Joint IEEE International Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance. 2005.
[42] Niebles JC, Wang H, Fei-Fei L. Unsupervised learning of human action categories using spatial-temporal words. Int J Comput Vis 2008;79(3):299–318.
[43] Ikizler N, Duygulu P. Histogram of oriented rectangles: a new pose descriptor for human action recognition. Image Vis Comput 2009;27(10):1515–26.
[44] Vishwakarma DK, Kapoor R. Hybrid classifier based human activity recognition using the silhouette and cells. Expert Syst Appl 2015;42(20):6957–65.
[45] Wang Y, Mori G. Human action recognition by semi-latent topic models. IEEE Trans Pattern Anal Mach Intell 2009;31(10):1762–4.
[46] Ming XL, Xia HJ, Zheng TL. Human action recognition based on chaotic invariants. J South Central Univ 2013;20:3171–9.
[47] Iosifidis A, Tefas A, Pitas I. Discriminant bag of words based representation for human action recognition. Pattern Recognit Lett 2014;49:185–92.