Advanced Robotics

ISSN: 0169-1864 (Print) 1568-5535 (Online) Journal homepage: http://www.tandfonline.com/loi/tadr20

Integrated approach for human action recognition using edge spatial distribution, direction pixel and -transform

D.K. Vishwakarma & Rajiv Kapoor

To cite this article: D.K. Vishwakarma & Rajiv Kapoor (2015) Integrated approach for human action recognition using edge spatial distribution, direction pixel and -transform, Advanced Robotics, 29:23, 1553-1562, DOI: 10.1080/01691864.2015.1061701

To link to this article: http://dx.doi.org/10.1080/01691864.2015.1061701

Advanced Robotics, 2015
Vol. 29, No. 23, 1551–1561, http://dx.doi.org/10.1080/01691864.2015.1061701

Taylor & Francis
Taylor & Francis Group

FULL PAPER

Integrated approach for human action recognition using edge spatial distribution, direction pixel and ℜ-transform

D.K. Vishwakarma* and Rajiv Kapoor

Department of Electronics and Communication Engineering, Delhi Technological University, Delhi, India

(*Received 4 March 2015; revised 4 March 2015; accepted 1 June 2015*)

In this article, a simple yet proficient approach for the recognition of human action and Activity is presented. This method is based on the integration of translation and rotation of the human body. The proposed framework undergoes three major steps: (i) the shape of the human action/activity is represented through the computation of average energy images using edge spatial distribution of gradients along with the directional variation of the pixel values, (ii) the orientation-based rotational information of the human action is computed through ℜ-transform and (iii) a descriptor is developed by the fusion of translational features with rotational features. The fusion of features possesses the advantages exhibited by both local and global features of the silhouette and thus provides the discriminating feature representation for human activity recognition. The performance of descriptor is evaluated through a hybrid approach of support vector machine and the nearest neighbour classifiers on standard data set. The proposed method has shown superior results in terms of recognition accuracy in comparison with other state-of-the-art methods.

Keywords: human action recognition; texture segmentation; edge spatial distribution of gradients; ℜ-transform; hybrid SVM-NN; classifier

1. Introduction

With the innovation in the area of computer vision, the recognition of a human activity from a video or image sequence has become the epicentre of research, whose focus is the development of a rugged system which recognizes the human action efficiently and accurately. The human activity/action recognition (HAR) presents potential applications such as robotics, smart technologies, surveillance and human–computer interaction.

Recent surveys [1–3] in this field indicate that the different approaches are being used for the human activity recognition i.e. spatial temporal interest points (STIP), point trajectories, optical flow, bag of words and silhouette models. Still, recognition of human activity from a video (or sequence of images) remains a major challenge as it is dependent upon the complexity of background, motion of the camera, angular variations, illumination conditions, etc. The proposed framework is based on the combination of static posture information and rotational features which are extracted from the binary silhouettes. The posture representation is given by average energy silhouettes images which represents the global information about the human posture. To extract the static posture features, the edge distribution of gradients is used as a local descriptor which is computed on the region of interest (ROI) of the image, which helps us to remove the redundant information around the object. Another feature which is used to capture the structural information is the directional pixels which show the variations of sum of pixels in the x–y direction. This representation is integrated with rotation information because in every action or activity there is always some kind of rotation information like bending of arms or knees while running, walking, jumping, etc. Therefore, we used the ℜ-transform for the extraction of rotation features from the binary silhouettes which is invariant to translation and scaling except angular variations in the actions.

In more recent, it has been observed that multiple features [4–7]-based techniques have superior performance than individual features. Thus, the descriptor formed by the incorporation of more than one feature together the representation of the human action would be more efficient, robust, occlusion free and give us improved accuracy. Hence, we believe that our integrated model based on multiple features would be more efficient than the earlier one. The main contributions of the paper are as follows:

- Silhouette of the activity video sequences is obtained using a robust texture-based segmentation method, which utilizes the entropy of image is the basis of information content.
- Average energy images are formed by summing and averaging the silhouette images of the activities.

*Corresponding author. Email: dkvishwakarma@dce.ac.in

- Shape of the silhouette activity represented through average energy images is descripted using edge spatial distribution and variation of pixel values in x–y directions.
- The rotation of the silhouette frames is computed using ℜ-transform, which is variant to the rotation and invariant to the translation and scaling.
- Finally, a novel descriptor is proposed by integration of shape and rotation features of the activity.

The remainder of the paper is organized as follows: Section 2 describes the local and global approaches in the field of HAR. The technical details of the proposed methodology are explained in Section 3, which explains the binary silhouette extraction using texture-based segmentation, 2D shape feature vectors and rotational information using ℜ-transform. Section 4 highlights the detailed experimental results, comparison and analysis. Lastly, the conclusion is drawn, and the references are cited.

2. Related work

This section concisely describes the previous work in HAR related to global and local feature extraction models with their merits and demerits. Global approaches based on the silhouette images or contour of the body which gives the shape information, while the local representation works on the small patches which is used in object recognition.

Laptev [8] introduced the notion of interest points for action recognition, but it is not stable and effective for complex actions. Dollar et al. [9] worked on these limitations and proposed the interest points based on the separable linear filters. Gorelick et al. [10] extended the 2D motion template to 3D space time volume. Niebles et al. [11] presented the probabilistic latent semantic analysis and latent Dirichlet allocation models over the space time regions, but they do not provide temporal and scale invariance. Onofri et al. [12] used the combination of histograms of oriented gradients with histogram of optical flow for the classification of actions based on the subsequence. This method is comparatively stable and robust in representation compared to part-based models. Zhen et al. [13] classified the human actions based on the spatial–temporal steerable pyramid which is compact in representation and acts as an effective global descriptor. Somasundaram et al. [14] introduced spatio-temporal features based on sparse representation. This method performs well for large and scale invariant data set. Shao et al. [15] introduced the novel Laplacian pyramid coding technique which is independent of tracking of features or localization of STIP's. Bobick and Davis [5] used the concept of template representation where they formed MHI/MEI templates for action recognition. The representation is simple, but they are dependent upon time parameter, variation of action speed and view dependent. Efros et al. [16] perform the recognition on low-resolution videos by correlating the optical flow measurements. Hung et al. [17] presented an activity-based human identification approach using average energy image and adaptive discriminant analysis. Ahmad et al. [2] formed the silhouette energy images, to differentiate the shape and motion properties for human actions recognition. Whytock et al. [18] presented a novel action recognition method based on gait-energy image (GEI) and used the HOG as local descriptors. Shao et al. [19] proposed the recognition of activities on the combination of motion and shape analysis. Zhang et al. [20] used the ℜ-transform as a shape descriptor for the representation of HAR and reported that its performance is better for the activities having the rotation motion. Khan et al. [21] use ℜ-transform for representation of abnormal human activities and find the improved and accurate results for the rotation-based human activity.

The prior work reveals that there is a need of an extensive technique, which consists of the translation and orientation characteristics of the action dynamics. So that the efficient representation of the human action/activity can be made possible, and in this work, we have proposed indiscriminate method, which can represent the human action efficiently by utilizing both the characteristics of action dynamics.

3. Proposed methodology

The proposed method is based on the integration of local and global information which enables the recognition of human activity in an effectual mode. The overview of the framework is as depicted in Figure 1.

The edge distribution of gradients and directional pixels is used as local descriptors for the structural information obtained through average silhouette energy image. On the other hand, the ℜ-transform gives the global information obtained through the binary silhouettes which represents the rotational information of the activity. The major steps of the proposed method are explained in subsequent sections.

3.1. *Texture-based segmentation*

The performance of silhouette-based models is dependent upon the accurate extraction of the silhouettes from video/image sequences. In our method, we have proposed the texture-based segmentation for the silhouette extraction. The flow diagram of silhouette extraction is as shown in Figure 2. Entropy is one of the important parameter to describe the texture information which is explained by Haralick et al. [22], to describe the complexity of the background and expressed by the Equation (1).

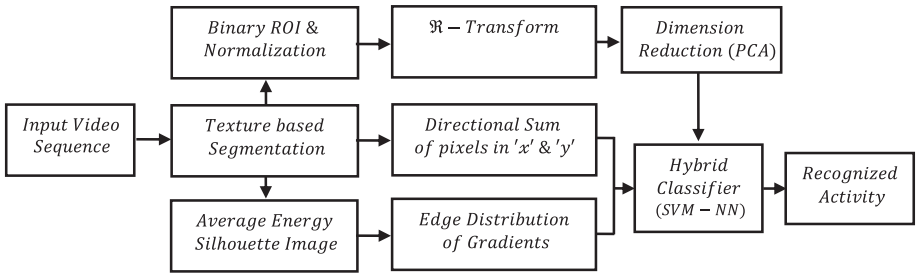

Figure 1. Work flow diagram of proposed methodology.

$$\zeta = \sum_i \sum_j \rho(i,j) \log(\rho(i,j)) \tag{1}$$

where $\rho(i,j) = \frac{M(i,j)}{\sum_{i,j} M(i,j)}$, the probability density function, and here, i and j are the indices to the co-occurrence matrix M. The neighbourhood filter mask of $[9 \times 9]$ is formed for each pixel which applied over the texture images generated from the entropy. After the filtering process, wiener filter of $[15 \times 15]$ window is used for smoothening of the images. These images are filled with grey pixels and further converted to binary silhouette images and these images are as shown in Figure 2.

3.2. *Average energy silhouette image*

Han and Bhanu [23] introduced the GEI concept which is similar to average silhouette energy image. In our method, we align the binary silhouettes to a reference point and then compress the sequence into a single-image based on incremental procedure that takes into account the changes of subsequent silhouettes. The calculation of the average energy silhouette image is done as defined by the equation:

$$A_E(x,y) = \frac{1}{\zeta} \sum_{\tau=1}^{\zeta} |\psi_\tau(x,y)|^2, \tag{2}$$

where 'ζ' is the number of frames in a complete cycle, '$\psi_\tau(x,y)$' is defined as binary silhouette frame at time instant 'τ' and x, y are the pixel values in 2D image coordinate. The advantage of using these images is that it removes the time-dependency limitations [5] and easily differentiates the similar representation actions of different categories on the basis of the intensity of pixel values, and these images are as shown in Figure 3.

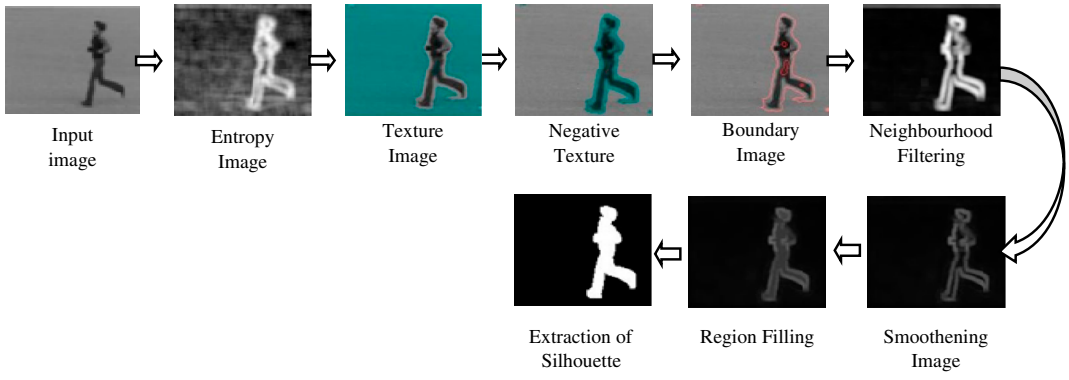

Figure 2. Flow diagram of silhouette extraction using texture-based segmentation of jogging activity.

Figure 3. Showing the variation of average silhouette energy images intensities of different types of human activities, where higher intensity values represent the activity performed by specific part of human body and lower intensity values represent the whole body.

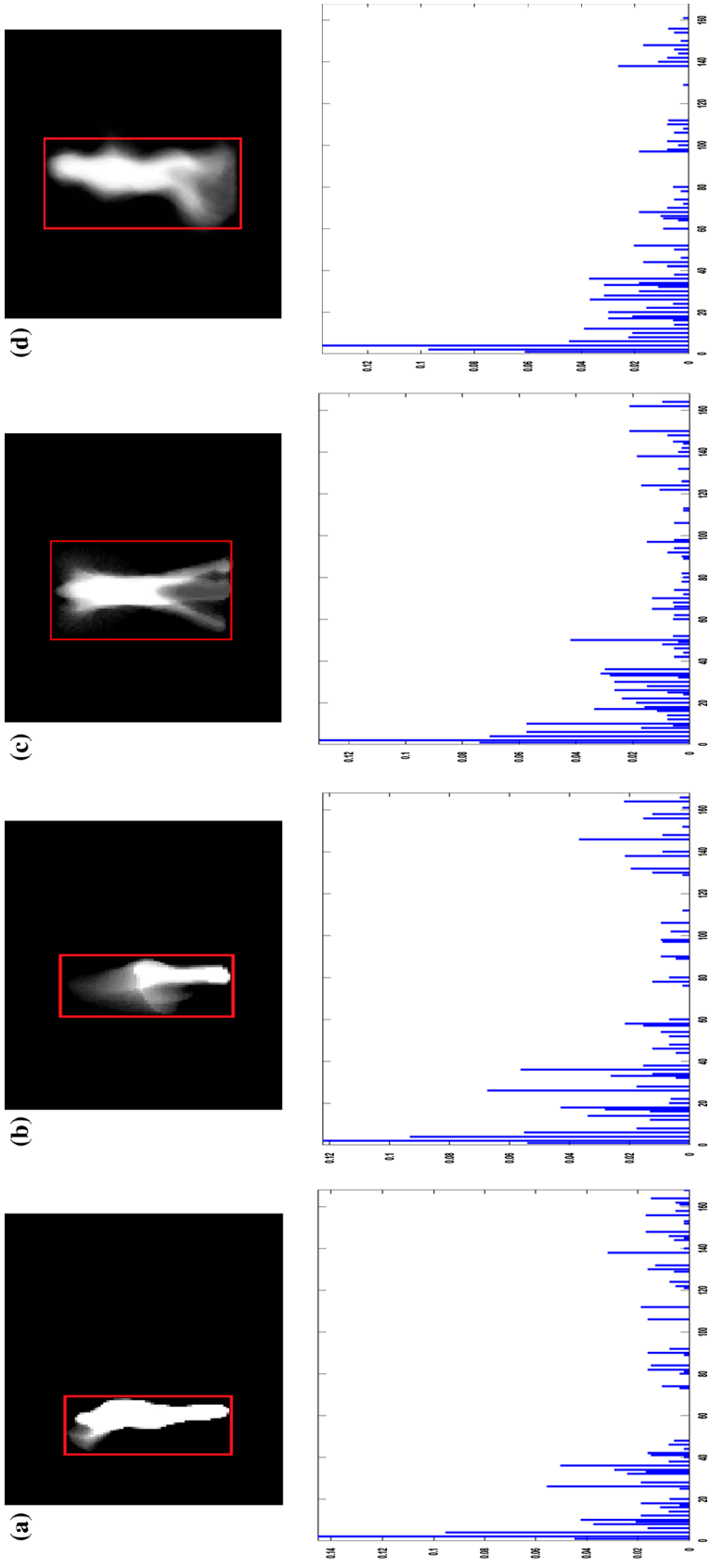

Figure 4. Depiction of edge spatial distribution of gradients for different activities at level-2 (a) one hand wave, (b) bending, (c) jumping jack, (d) running.

3.3. Edge spatial distribution of gradient

The concept of edge spatial distribution of gradients was introduced by the Bosch et al. [24] in object classification. It acts as a local descriptor which is applied on ROI images that will give more localized, noise-free information and structural information about the object. ROI also helps in making these images invariant to scale and translational variations and further divided into sub-levels, and for each level, we compute the orientation of the edges at finer scale. It reduces the computational time and complexity of the system.

3.3.1. Algorithm for computation of edge spatial distribution of gradient

Step 1: Let $f(x, y)$ represent the average silhouette energy image computed using Equation (2).

Step 2: Select the ROI and normalized to the dimension of 50×50 and represented as: $f_r(x,y)$, where $0 \le x, y \le 50$.

Step 3: Computation of edge spatial distribution vector at various levels as follows:

I. At level-0, the magnitude $\mathcal{M}(x,y)$ and orientation $\mathcal{O}(x,y)$ at point (x, y) is computed on entire image $f_r(x, y)$ as:

$$\mathcal{M}(x,y) = \sqrt{\mathcal{G}_x(x,y)^2 + \mathcal{G}_y(x,y)^2} \text{ and } \mathcal{O}(x,y) = \arctan\left[\frac{\mathcal{G}_x(x,y)}{\mathcal{G}_y(x,y)}\right] \quad (3)$$

where $\mathcal{G}_x(x,y)$and $\mathcal{G}_y(x,y)$ denote the gradient of image along x and y directions. Each sub-region is quantized into eight orientation bins $(0^\circ - 180^\circ)$.

II. At level-1, the entire image $f_r(x, y)$ is divided into four sub-image regions and denoted as:

$$f_r(x,y) = \{\mathcal{B}_1(x,y), \mathcal{B}_2(x,y), \mathcal{B}_3(x,y), \mathcal{B}_4(x,y)\} \quad (4)$$

Compute the features vectors through step 3-I.

III. At level-2, each sub-image region $\mathcal{B}(x,y)$ is further divided into four sub-blocks, and feature vector is computed as step 3-I.

Step 4: The feature vector is formed by joining all these levels together, which is as expressed by Equation (5) and depicted as in Figure 4 for different activities of average energy images and represented as:

$$\mathcal{F}e = [1 \times 8] + 4 \times [1 \times 8] + 16 \times [1 \times 8] \quad (5)$$

3.4. Directional pixels computation

Directional pixels calculate the sum of the pixel values of the shape representing the action along x and y directions, and the variations of these values store the information about the action. The calculation of pixel values in $x - y$ direction is defined by:

$$H_X(k) = \sum_{l=0}^{n-1} \frac{A_E(k,l)}{\max(A_E(k))}, 0 < k < m-1, \; V_Y(l) = \sum_{k=0}^{m-1} \frac{A_E(k,l)}{\max(A_E(l))}, 0 < l < n-1 \quad (6)$$

where m and n are the rows and columns of the image, respectively. The mean of pixel values is computed to remove the effects of noise in x and y direction as:

$$\mu_x = \frac{1}{m}\sum_{k=1}^{m} H_X(k), \mu_y = \frac{1}{n}\sum_{l=1}^{n} V_Y(l) \quad (7)$$

hence, the feature vector represented as: $\mathcal{F}_d = [\mu_x, \mu_y]$.

In Figure 5, hand wave activity, peak at the centre shows that more pixels are variation as we move column wise (y-pixel) and as we move in x-axis, it represents the presence of more pixels in upper portion (x-pixel).

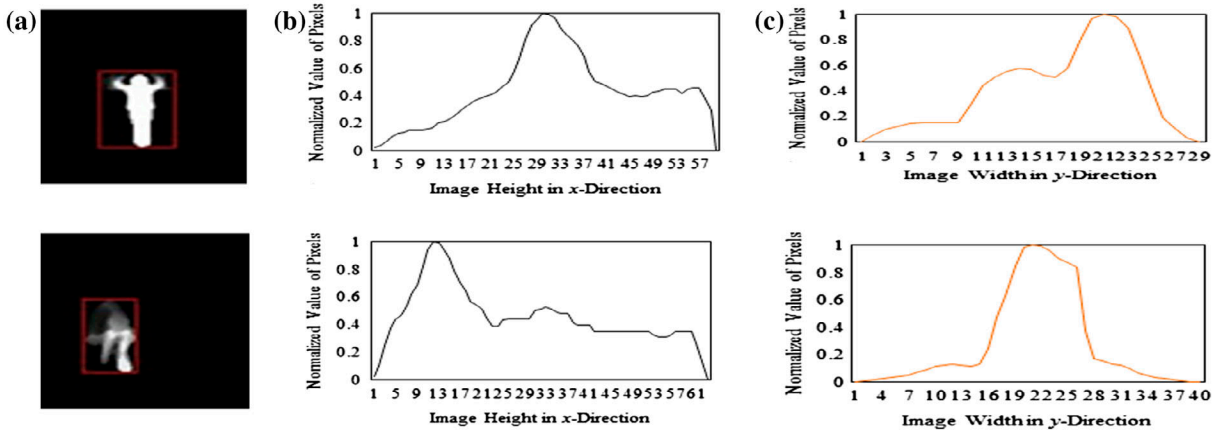

Figure 5. (a) Shows the average energy images of hand-waving and bending activity, (b) pixel values variation in the x-direction, (c) pixel values variation in the y direction.

In bending activity, the major part of the information is present in the lower half of the image and thus the graph shows peaks in the later part of the graph.

The shape-based feature computed in Sections 3.3 and 3.4 are mixed with the rotational information of the human action, which is computed on the binary silhouettes of the human body poses through ℜ-transform.

3.5. Rotational features computation

The action dynamics of human body states that any activity or action cannot be possible to perform without translation and rotation. The rotational features are computed using ℜ-transform. ℜ-transform is having robustness to the translation and scaling due to its invariant properties as it is shown in Figure 6. The sensitiveness of ℜ-transform can be observed through Figure 6(b), where the change in rotation reflects the higher variation in ℜ-transform signal. The discriminative property of ℜ-transform can also be seen from Figure 7, where different types of activities are presented and their ℜ-transform signal is different. Hence, we have used ℜ-transform, to compute the rotational-based orientation information of the silhouettes.

Tabbone et al. [25] introduced the ℜ-transform, which is an integral transform of the squared values of radon transform (RT), and it represents the features in reduced 1-D compared to the 2-D representation of RT. RT is defined as the integral of a silhouette image $\mathcal{I}(x,y)$ from $-\infty$ to ∞ and gives the directional features in the range (0°–179°), denoted as:

$$R_T(\rho,\theta) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \mathcal{I}(x,y).\delta(\rho - x\cos\theta - \sin\theta)\partial x\partial y \tag{8}$$

where $\delta(.)$ is the Dirac delta function. ℜ-transform is defined as:

$$\Re(\theta) = \int_{-\infty}^{\infty} R_T^2(\rho,\theta)\partial\rho \tag{9}$$

The normalization of ℜ-transform gives improvement of the resemblance and represents the features in a compact manner and expressed as:

$$\Re_{\text{norm}}(\theta) = \int_{-\infty}^{\infty} \frac{\Re(\theta)d\theta}{\max(\Re(\theta))} \tag{10}$$

In computation of orientation features, the key poses of the human silhouettes are extracted from a videos sequence using the concept of energy. The higher energy is computed as the sum of total number white pixels counts in the silhouettes. These extracted key poses are highly discriminative among the whole frames of the video sequence. In this work, for our convenience, only seven key frames are selected for the computation of orientation feature. The ℜ-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 ℜ-transform gives the dimension of a silhouette image to 1 × 180 and for single action representation the dimension of feature vector is 7 × 180. Further, to reduce discriminative set of features, the principal component analysis (PCA) is used. The operation of PCA [26] reveals the internal structure of the silhouette in a manner that efficiently explains the variance in the data. The dominant features of data set are obtained by solving the eigenvalue problem of covariance matrix (C) of the data set, represented as:

$$k = U^T CU \tag{11}$$

The dimension of [7 × 180] feature vectors reduced to discriminative feature vector of size [1 × 7] with the help of PCA.

3.6. Final feature vector computation

The final feature descriptor is formed by integrating the feature set obtained in Sections 3.2 and 3.3. The method of integration of feature vectors is shown in Figure 8.

The shape-based edge spatial distribution feature vectors, and directional pixels feature vectors are combined with the rotational features vector computed using the ℜ-transform, and it forms a novel descriptor for the representation of human activity. The edge spatial distribution feature vectors and directional pixel feature vectors are computed as explained in Sections 3.3 and 3.4 and have a dimension of 1 × 168 and 1 × 2, respectively. Further, these feature vectors are combined with ℜ-transform features computed on the binary silhouettes. The ℜ-transform computed feature vector has the dimension of 7 × 180. Here, the number of key frames used is 7. The dimension of ℜ-transform feature is reduced using PCA, which gives the reduced discriminative feature vector of 1 × 7. Finally, the edge spatial feature vectors and ℜ-transform feature vectors are combined together and give the resultant feature vector of $[1 \times 168 + 1 \times 2 + 1 \times 7] = [1 \times 177]$ dimension. The performance of final descriptor is evaluated by conducting an experiment on the standard data set.

4. Experimental results, comparison and analysis

The accuracy of the proposed method is computed by conducting the experiment on publically available

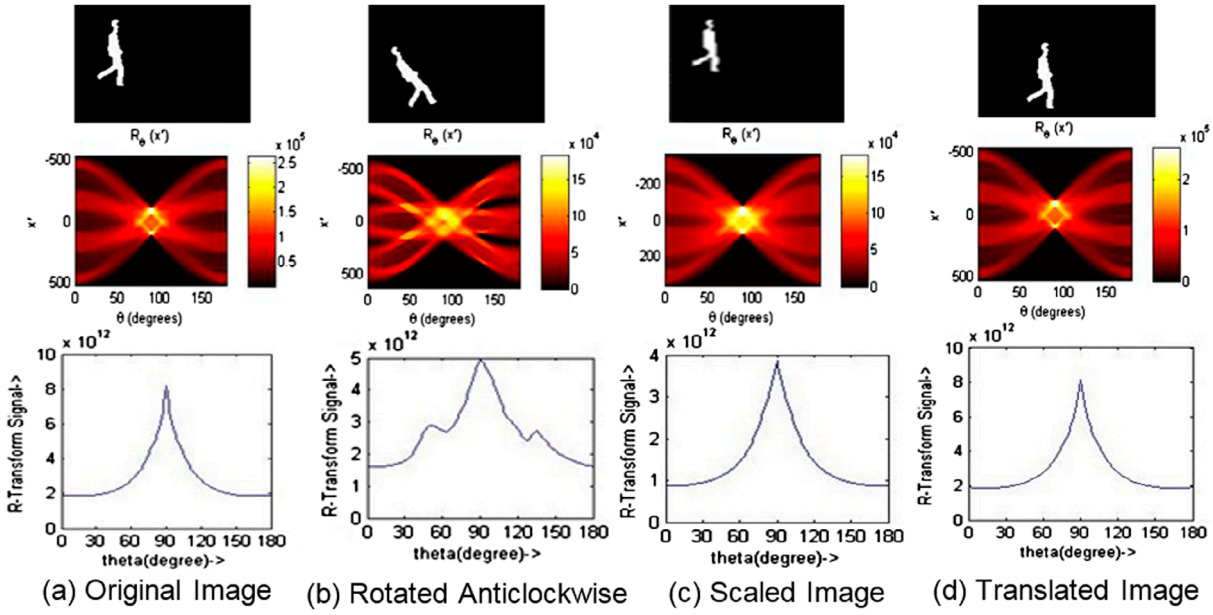

Figure 6. Depiction of ℜ-transform properties Row 1: silhouette images, Row 2: Radon transform (RT), Row 3: ℜ-transform signal of (a) original image, (b) rotated image anticlockwise, (c) scaled image and (d) translated image.

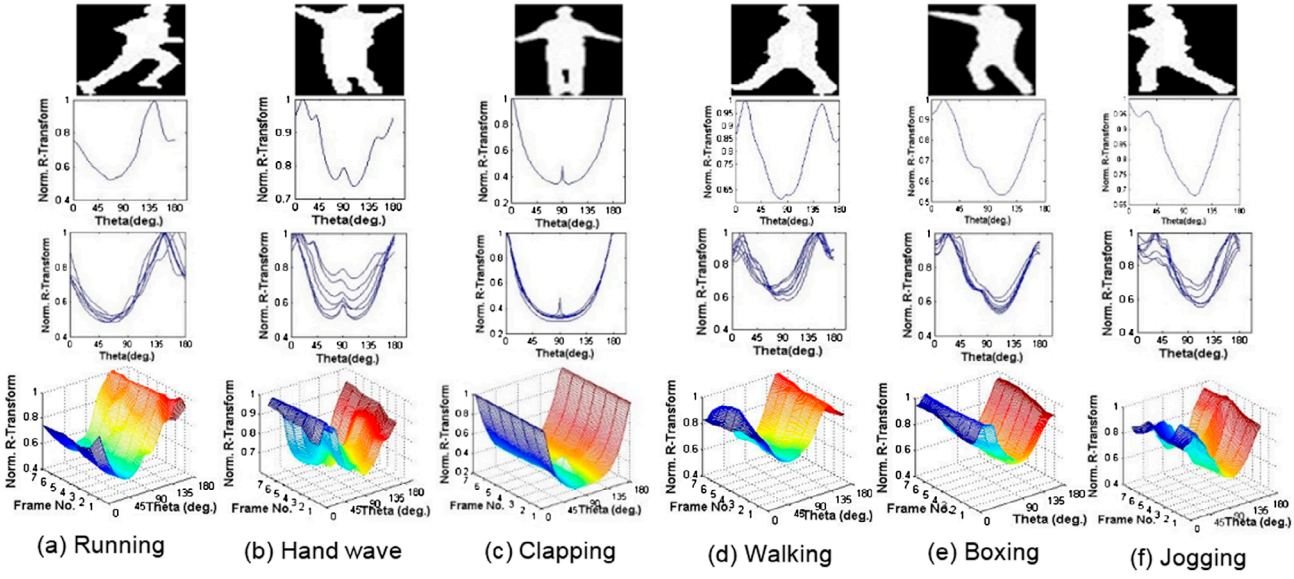

Figure 7. Representation of ℜ-transform for different activities: Row 1: 50 × 50 Silhoutte Image, Row 2: ℜ-transform, Row 3: ℜ-transforms of key frames, Row 4: 3D ℜ-transform of sequence.

standard data sets of Weizmann and KTH. To test the feasibility of proposed method, the classification is done using a cascaded model of support vector machine (SVM) and nearest neighbour classifier [27] as shown in Figure 9. To utilize the benefit of individual performances of the nearest neighbour and SVM, a hybrid classification approach is proposed. In this work, initially, the SVM is used to classify the input feature set, and in these feature sets, few are correctly classified and few are wrongly classified. The wrongly classified

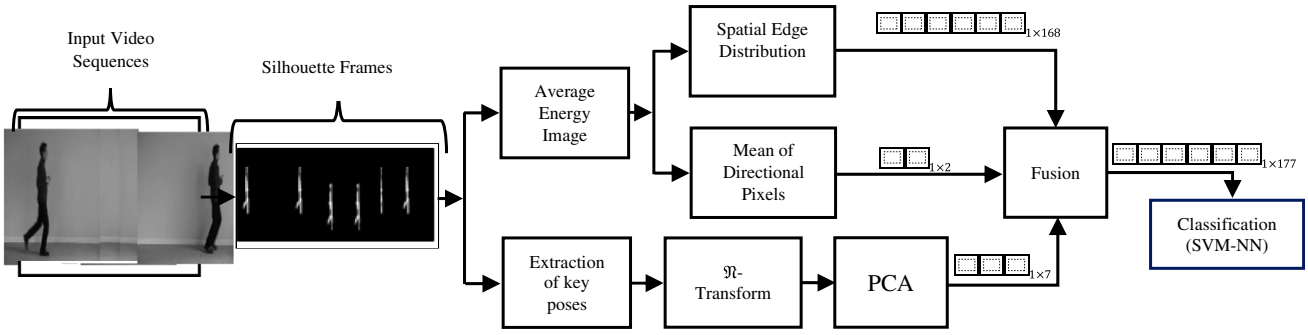

Figure 8. Flow diagram of integration of different features.

feature sets lie near the separating hyper-plane, and these are the support vectors. Further, these support vectors are classified using NN and are considered as representative points. The advantage of this method is that whole information of data set is used to calculate the precision. The algorithms are performed in MATLAB 2012a, running on a corei5, 1.60 GHz Intel processor with 4 GB RAM.

4.1. *Weizmann data set*

This data set [10] captures the human at a frame rate of 15 fps and size of the frame is 144 × 180 with a total of 90 videos. The data set has 9 people, each performing 10 different actions which are categorized as 'walk', 'run', 'jump_jack', 'bend', 'jump forward on one leg', 'jump', ' jumping in place', 'sideways jump', 'one hand wave', 'two hand wave' and as shown in Figure 10.

4.1.1. *Results of Weizmann data set*

In this data set, the extraction of binary silhouettes does not pose a great challenge and ℜ-transform signals representations are smooth and variant. The edge distribution of gradients representation of activities is also finer and distributed. Therefore, the recognition rate achieved on this data set is 100%.

4.2. *KTH data set*

This is the challenging data set,[28] which has six different activities that are 'clapping', 'boxing', 'walking', 'hand waving', 'jogging', and 'running', and in four different scenarios (s1–s4). The videos have frame rate of 25 fps, and the size of frame is 160 × 120 pixels. The sample images of the data set are as shown in Figure 11.

4.2.1. *Results of KTH data set*

This data set has two similar classes of activities i.e. jogging and running which effectively decrease the accuracy compared to other activities which are easily distinguishable. Moreover, different scenarios in the activity lead to difficulties in the extraction of the silhouettes. The average recognition rate (ARR) achieved on this data set is 95.50% by the SVM–NN classifier, and the details of the result are presented in Table 1.

The overall accuracy achieved in this method is compared with the techniques of others on similar state of the art and gives superior result as depicted in Table 2. It is also noticed that the edge distribution plays an important role in the recognition. If we have better distribution around ROI, then our results further improve. The effect of directional pixels mean value is not so high as compared to edge distribution and ℜ-transform gives the rotational variations of the postures.

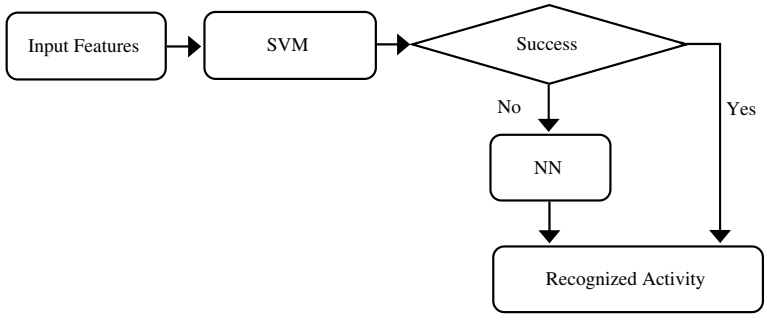

Figure 9. Hybrid 'SVM-NN' classifier.

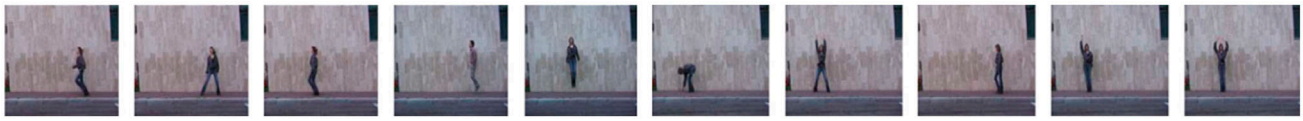

Figure 10. Sample frames of Weizmann human action data set.

Figure 11. Sample images of KTH data set.

Table 1. Classification result of KTH data set.

Activities	Boxing	Hand-waving	Hand-clapping	Jogging	Running	Walking	ARR (%)
SVM-NN (%)	100	100	92	92	89	100	95.5

Table 2. State-of-the-art comparison with the techniques of others.

Techniques	Weizmann Data set (%)	KTH Data set (%)
Dollar et al. [29]	85.20	81.17
Niebles et al. [11]	90.00	83.33
Klaser et al. [30]	84.30	91.40
Hung et al. [17]	92.50	35.04
Benmokhtar [31]	89.0	92.5
Proposed method	100.00	95.50

5. Conclusions

This article proposed a new method for the activity recognition based on the combined information obtained from the ℜ-transform and averaging of the energy silhouette. A feature vector is generated from the averaged energy silhouettes using an edge distribution of gradients and directional pixels. As the amount of levels increase in the edge distribution of gradients vector, we get better quality of edge distribution and accurate results, but the complexity of the system increases because of the increase in the vector size. The characteristics of edge distribution are also dependent on the ROI. The enhanced ROI, i.e. the region with minimal noise, when estimated in the image can bestow improve edge distribution of gradients. The ℜ-transform is applied to the extracted normalized silhouette, and then, dimension reduction is done to reduce the size of feature vector. This integrated technique supplies numerous distinctive feature vectors, which

escort us to a robust and noise-free action recognition model. In the future, one may proceed with complex and varied data sets to further check the robustness of the system and to acquire better and more accurate results.

Disclosure statement

No potential conflict of interest was reported by the authors.

Notes on contributors

D.K. Vishwakarma received the Bachelor of Technology (BTech) from Dr RML Avadh University, Faizabad, Uttar Pradesh, India, in 2002 and the Master of Technology (MTech) from Motilal Nehru National Institute of Technology, Allahabad, Uttar Pradesh, India in the year 2005. Currently, he is working as an assistant professor, in the Department of Electronics and Communication Engineering, at Delhi Technological University, Delhi, India-110042. His research interests include human computer interaction, pattern recognition, human pose estimation and recognition, and hand gesture recognition. He is also a reviewer of various Elsevier and IET journals.

Rajiv Kapoor received his BE, ME degree from Delhi University, India, and PhD from Punjab University, India, in the field of Electronics & Communication. He is currently a professor in the Department of Electronics & Communication, Delhi Technological University, Delhi, India. He is also an editor of ST Micro Electronics International Journal in the field of Electronics Design. He has authored more than 50 research papers. He is also a reviewer of various IEEE, IET, Elsevier and Springer Journals. His research interest includes Image Processing, Object Tracking, Pattern Recognition and Character Recognition.

ORCID

D.K. Vishwakarma http://orcid.org/0000-0002-1026-0047
Rajiv Kapoor http://orcid.org/0000-0001-9522-8670

References

[1] Aggarwal JK, Cai Q. Human motion analysis: a review. Comput. Vision Image Understand. 1999;73:428–440.
[2] Ahmad M, Lee SW. Variable silhouette energy image representations for recognizing human actions. Image Vision Comput. 2010;28:814–824.
[3] Poppe R. Vision-based human motion analysis: an overview. Comput. Vision Image Understand. 2007;108:4–18.
[4] Shao L, Gao R, Liu Y, Zhang H. Transform based spatio-temporal descriptors for human action recognition. Neurocomputing. 2011;74:962–973.
[5] Bobick A, Davis J. The recognition of human movement using temporal templates. IEEE Trans. Pattern Anal. Mach. Intell. 2001;23:257–267.
[6] Bregonzio M, Xiang T, Gong S. Fusing appearance and distribution information of interest points for action recognition. Pattern Recogn. 2012;45:1220–1234.
[7] Zhao D, Shao L, Zhen X, Liu Y. Combining appearance and structural features for human action recognition. Neurocomputing. 2013;113:88–96.
[8] Laptev I. On space-time interest points. Int. J. Comput. Vision. 2005;64:107–123.
[9] Dollar P, Rabaud V, Cottrell G, Belongie S. Behavior recognition via sparse spatio-temporal features. In: Proceedings of the 2nd IEEE International Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance; Beijing, China; 2005. p. 65–72.
[10] Gorelick L, Blank M, Shechtman E, Irani M, Basri R. Actions as space-time shapes. IEEE Trans. Pattern Anal. Mach. Intell. 2007;29:2247–2253.
[11] Niebles J, Wang H, Fei-Fei L. Unsupervised learning of human action categories using spatial-temporal words. Int. J. Comput. Vision. 2008;79:299–318.
[12] Onofri L, Soda P, Iannello G. Multiple subsequence combination in human action recognition. IET Comput. Vision. 2014;8:26–34.
[13] Zhen X, Shao L, Li X. Action recognition by spatio-temporal oriented energies. J. Inf. Sci. 2014;281:295–309.
[14] Somasundaram G, Cherian A, Morellas V, Papanikolopoulos N. Action recognition using global spatio-temporal features derived from sparse representations. Comput. Vision Image Understand. 2014;123:1–13.
[15] Shao L, Zhen X. Spatio-temporal Laplacian pyramid coding for action recognition. IEEE Trans. Cybern. 2014;44:817–827.
[16] Efros A, Berg A, Mori G, Malik J. Recognizing action at a distance. In: Proceedings of 9th IEEE International Conference Computer Vision (ICCV); Nice, France; 2003. p. 726–733.
[17] Hung TY, Lu, J, Hu J, Tan YP, YGe. Activity-based human identification. In: Proceedings of IEEE international conference on Acoustics, Speech and Signal Processing (ICASSP); Vancouver, BC; 2013. p. 2362–2366.
[18] Whytock T, Belyaev A, Robertson N. GEI + HOG for action recognition. In: Proceedings of BMVC; Surrey, UK; 2012. p. 1–12.
[19] Shao L, Ji L, Liu Y, Zhang J. Human action segmentation and recognition via motion and shape analysis. Pattern Recogn. Lett. 2013;33:438–445.
[20] Zhang H, Liu Z, Zhao H. Recognition of human activities by key frame in video sequence. J. Softw. 2010;5:818–825.
[21] Khan ZA, Sohn W. Abnormal human activity recognition system based on R-transform and kernel discriminant technique for elderly home care. IEEE Trans. Consum. Electron. 2011;57:1843–1850.
[22] Haralick M, Shanmugam K, Dinstein I. Textural features for image classification. IEEE Trans. Syst. Man Cybern. Part A Syst. Humans. 1973;3:610–621.
[23] Han J, Bhanu B. Individual recognition using gait energy image. IEEE Trans. Pattern Anal. Mach. Intell. 2006;28:316–322.
[24] Bosch A, Zisserman A, Munoz X. Representing shape with a spatial pyramid kernel. In: Proceedings of the ACM International Conference on Image and Video Retrieval; Amsterdam, Netherlands; 2007. p. 401–408.
[25] Tabbone S, Wendling L, Salmon JP. A new shape descriptor defined on the Radon transform. Comput. Vision Image Understand. 2006;102:42–51.

[26] Jolliffe IT. Principal component analysis. 2nd ed. New York (NY): Springer; 2003.
[27] Vishwakarma DK, Kapoor R. Hybrid classifier based human activity recognition using the silhouette and cells. Expert Syst. Appl. 2015;42:6957–6965.
[28] Schuldt C, Laptev I, Caputo B. Recognizing human actions: a local SVM approach. In: Proceedings of the IEEE International Conference on Pattern Recognition; Cambridge; 2004. p. 32–36.
[29] Dollar P, Rabaud V, Cottrell G, Belongie S. Behavior recognition via sparse spatio-temporal features. In: Proceedings of the 2nd IEEE International Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance; Beijing, China; 2005. p. 65–72.
[30] Klaser A, Marszalek M, Schmid C. A spatio-temporal descriptor based on 3D-gradients. In: Proceedings of BMVC; Leeds, UK; 2008.
[31] Benmokhtar R. Robust human action recognition scheme based on high-level feature fusion. Multimedia Tools Appl. 2014;69:253–275.