ARTICLE IN PRESS

Robotics and Autonomous Systems ▮ (▮▮▮▮) ▮▮▮–▮▮▮

Contents lists available at ScienceDirect

Robotics and Autonomous Systems

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

A proposed unified framework for the recognition of human activity by exploiting the characteristics of action dynamics

Dinesh K. Vishwakarma *, Rajiv Kapoor, Ashish Dhiman

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

HIGHLIGHTS

- A combined algorithm based on shape and motion features of human activity.
- A single key pose is used for estimation of shape using edges.
- A single global key pose is extracted from video signal by exploiting local notion.
- The temporal motion feature is computed using $\Re$-transform.
- Robustness of the algorithm is demonstrated on the varied dataset.

ARTICLE INFO

Article history:
Received 29 June 2015
Received in revised form
7 November 2015
Accepted 16 November 2015
Available online xxxx

Keywords:
Human action recognition
Single still image
Fuzzy logic model
Edge based posture representation
Rotational feature

ABSTRACT

The aim of this paper is to present a novel integrated framework for the recognition of human actions using a spatial distribution of edge gradient (SDEG) of human pose and detailed geometric orientation of a human silhouette in a video sequence. The combined descriptor endows a wealthy feature vector dictionary having both the appearance and angular kinematics information that significantly wraps the local and global information and provides discriminative depiction for the action recognition. The SDEG is computed on a still image at different levels of resolution of sub-images, and still images of the human poses are extracted from the input video sequence using fuzzy trapezoidal membership function based on the normalized histogram distance between the contiguous segment frames. The change of geometric orientation of human silhouette with time is computed using normalized $\Re$-Transform. To validate the performance of the proposed approach, extensive experiments are conducted on five publicly available human action datasets i.e. Weizmann, KTH, Ballet Movements, Multi-view i3dPost, and IXMAS. The recognition accuracy achieved on these datasets demonstrates that the proposed approach has an abundant discriminating power of recognizing the variety of actions. Moreover, the proposed approach yields superior results when compared with similar state-of-the-art methods.

1. Introduction

Human activity recognition has been an active area of research in computer vision due to its potential applications in the field of surveillance, early diagnostic of stroke and rehabilitation of elder people, sports event analysis, robotics, terrorist activities, content-based video analysis, and human–computer interactions [1–3]. However, human action recognition is both challenging and multifaceted due to viewpoint variations, occlusion, cluttered background, intra-class motion variability and inter-class motion ambiguity. The answer to these problems is still a challenging task. Therefore, day by day researchers are trying to devise a general, competent and robust method for recognition of human action.

Over the last few decades, numerous action and activity recognition techniques have been proposed, which are mainly based on the local and global representations. Optical flow, point trajectories, space–time volume, Bag-of-words and sparse interest points (STIP's) [4] are common existing techniques used for human activity recognition. While these methods effectively handle partial occlusions and make background subtraction superfluous, almost all of them have their own set of limitations. Optical flow methodology results in the inaccurate analysis if the video quality is low and not smooth. Similarly, point trajectories desire an

* Corresponding author. Tel.: +919971339840 (Mob.), +91 11 27871044x1308 (Office).
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.robot.2015.11.013
0921-8890/

efficient tracking of the human motion, whereas the distribution of the interest points around the object should be stable in the STIP's approach.

Recently, the concept of still images [5–10] has emerged as a popular means of detecting human activity or behaviour as it focuses on the visual appearance of the object. These images do not require any morphological operation or tracking trajectories. Therefore, these methods provide the better handling under occlusion, less computation time, less complexity, and is effective in noise handling [11].

In earlier approaches [12–15] $\Re$-Transform is used to extract the temporal motion information in terms of orientation features, and extensively used for representing the human activity, where the activity is dominated by rotation kinematics like falling on the ground, vomiting, etc.

Nevertheless, most of the earlier works [16–21] admit that individual feature based methods are less effective as compared to the multiple features based methods. However, 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 given that it does not encompass the motion information. Furthermore, it is also observed that $\Re$-transform based motion temporal features are more effective for those activities that have dominant orientation changes rather than translation like bending, falling on the ground, and vomiting, etc.

Motivation and contributions of the work:

Over the last few years, the concept of still images [5–11] based recognition of object has gained popularity. As it focuses on the visual appearance of the object or human pose, which does not require the segmentation, morphological operation, and tracking trajectories. Therefore, by utilizing still images for the recognition of human activity, a better approach is framed, which can handle the existing problems of vision based activity recognition algorithms like occlusion, segmentation, noise, complexity and speed of computation. Still image has the richness of shape of an object while for the effective recognition of human activity, but both the shape and motion information are needed. Hence, to boost the recognition of human activity, motion information is incorporated to the still image based approach by computing the orientation of human silhouettes using $\Re$-Transform. The shape of human pose is extracted by computing SDEG. Another important principle behind this study is to analyse and compare the effect of Histogram oriented gradients (HOG), pyramid of histogram oriented gradients (PHOG), and $\Re$-Transform with the proposed model for the purpose of human activity recognition. The proposed model consists of a rich feature vector vocabulary having both the appearance and angular kinematics information, which overthrows the limitation of earlier approaches. The pose dictionary yields the human appearance representation, and sequence of orientation provides the nature of the activity. The main contributions of the work are as follows:

- 2-D human body poses are extracted from the video sequence using the fuzzy logic model based on normalized histogram distances between the segments of video.
- The structural appearance of human pose is represented by computing SDEG, which offers the shape information of a human pose at various orientation bins and decomposition levels.
- The temporal motion content of the human body is represented by the computation of geometric orientation using normalized $\Re$-Transform and it offers the geometry transformation and scale invariant.
- A novel integrated model is constructed by combining shape based appearance and orientation information of the human body action.

The rest of this paper is structured as follows: Section 2 gives the glimpses of prior work; Section 3 explains the details of the proposed methodology, which includes the abstraction of spatial edge distributions, and angular kinematics information using normalized $\Re$-transform. Section 4 demonstrates the experimental details and discussion of results.

2. Related work

Usually, human activity recognition methods can be divided into still images based on visual appearance, the global method based on the human silhouette, and local feature-based approaches. The details of prior work based on these approaches are explained in the subsequent sections, which includes their merits and demerits.

2.1. Still images based approaches

In recent, the still images based action recognition have grown tremendously due to its crucial advantages like no need of background subtraction, robust against occlusion, less complexity, etc. Initially, [6] introduced the concept of action recognition based on still images and in these images they have used the canny edge detector to represent the shape of the human action and features have clustered into similar body poses. Li and Ma [7] presented the concept of "exemplarlet" that had adequate visual information to identify the human action in still images. Shao et al. [22] used the idea of the random forest decision tree algorithm to search for the discriminating patches of the human action region. Li and Fei-Fei [8] represented the integrated method that is based on the appearance information of still image and occurrence of action scenes. Thurau and Hlavac [23] proposed the method based on pose information of the human action. They worked on the ROI images and computed the histogram of gradients with non-matrix factorization (NMF) to represent feature vectors. Lopes et al. [9] presented the transfer learning approach where contextual information from still images applied to the test video sequence. Zheng et al. [10] combined the poselet with contextual information for recognizing human action from still images. Hu et al. [24] 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. These approaches do not provide the motion information in a short or long duration of time and therefore, are scarce in representing the human action from a video sequence.

2.2. Global approaches

In global approaches, the representation of human action is done via appearance and motion of the actions based on the silhouettes and temporal models. Template models signify the actions in whole video sequence rather than the diminutive duration of the period. Bobick and Davis [25] were the first to use the notion of template representation and they formed MHI/MEI templates for action recognition. Efros et al. [26] perform recognition on low-resolution videos by correlating the optical flow measurements. Shao et al. [27] used the combination of motion and shape information for recognition of activities. They used the MHI images for the shape representation and Pyramid of Correlogram (PCOG) for the feature description. Shao et al. [28] introduced the novel Laplacian pyramid coding descriptor for the holistic representation of human action. This method is independent of tracking of features or localization of STIP's.

Please cite this article in press as: D.K. Vishwakarma, et al., A proposed unified framework for the recognition of human activity by exploiting the characteristics of action dynamics, Robotics and Autonomous Systems (2016), http://dx.doi.org/10.1016/j.robot.2015.11.013

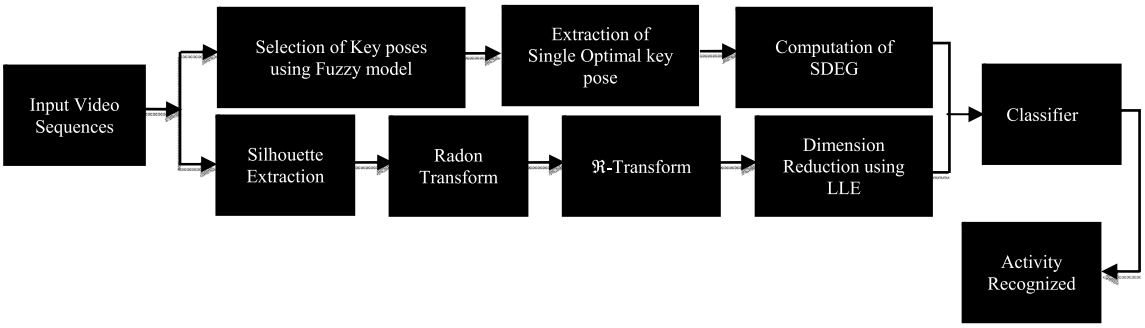

Fig. 1. Workflow diagram of human activity recognition.

Goudelis et al. [29] presented the method based on Trace transform for action recognition. It is a complex method but invariant to scaling, translation, and rotation. Khan and Sohn [13] introduced the abnormality activity recognition system based on ℜ-transform. Jalal et al. [14] represented the human action in-depth silhouette. For each depth silhouette, Radon Transform is employed, which provides invariance to the silhouette in terms of scaling and rotation.

2.3. Local approaches

Local methods depict the actions as a group of local descriptors or patches. Laptev et al. [30] introduced the concept of interest points and analysed the image in x–y–t dimensions, to detect the presence of interest points in space–time volumes. This method of generation of interest points is not stable and effective in complex actions. Laptev et al. [31] improved his results by proposing spatio-temporal histogram model. Nibeles et al. [32] presented the probabilistic latent semantic analysis (PLSA) and Latent Dirichlet Allocation (LDA) models over the space–time regions to perform unsupervised action recognition. These models did not provide the temporal and scale invariance. Later on, Zhao et al. [16] presented the combined approach representing both structure and appearance information based on STIP's. They used cuboids for representing the appearance information that are extracted from the STIP's. Shao et al. [28] presented the spatial–temporal steerable pyramid (STSP) approach for human action recognition. They used a Laplacian pyramid approach to decompose the video sequence and then multilevel steerable filters were used to extract the features in different directions and scale. Somasundaram et al. [33] introduced spatio-temporal feature vector based on the sparse representation for action recognition. Their methodology did not provide for large storage memory as it worked on the saliency approach, but the sparse representation made the system scale invariant. Tran et al. [34] proposed the part based model for the recognition of human action. This method was robust to partial occlusion and complex activities. Bregonzio et al. [17] presented the combination of appearance and distribution information of interest point for recognition of action. Although these methods showed improved results, they are not applicable to a dynamic background that has multiple people performing activities. It can also be observed that the space–time approaches are not suitable for recognizing multiple or more complex activities that are not periodic in nature, sensitive to partial occlusion, background variations.

3. Proposed methodology

As the earlier framework for action recognition used in [16–18,35,36], the proposed framework for action recognition is illustrated by assimilating the structural and rotational information of the action dynamics.

3.1. Overview of framework

The proposed framework is composed of computation of shape attributes using spatial distribution of edge gradients (SDEGs) of 2D postures and extraction of geometric points i.e. oriented points of silhouettes using normalized ℜ-transform. Incorporation of SDEG constructs the final feature vector with orientation features at the recognition stage. The SDEG feature gives the local region based level information of pose of action, which is computed on the still image, whereas the still image is the most optimal key frame of the activity sequence which is extracted using fuzzy approach in a video sequence. The temporal motion content of the activity is computed by applying ℜ-transform on binary silhouette images. The overview of the workflow diagram of the proposed framework is shown in Fig. 1 and the subsequent sections give the detailed explanation of every block.

3.1.1. Extraction of single key frame

For the accurate representation of body postures, the key poses of the activity from the video sequence are extracted and these key poses are compared on the basis of histogram distance for selecting the optimum single key frame for the structural representation of 2D pose of human activity. The extraction of single key pose is as illustrated in Fig. 2.

For selecting key poses, a stack of frames ($FFS^1, FFS^2, \ldots, FFS^{K+1}$) is formed by selecting the frames after a certain interval in the video sequence because the deviation in the video sequence does not vary instantaneously. These stacks of frames are converted into CIELab colour space because it closely conforms to human perception of colours and is device independent [37] as compared to the rest of colour spaces. The normalized histogram distances (*Norm HD*) are computed for all the three components L, a and b using Eq. (1).

$$HD_t = 1 - \frac{1}{3(MN)}\left[\sum_{j=1}^{n} \min\left(FFS_{L_j}^{t}, FFS_{L_j}^{t+1}\right) + \sum_{j=1}^{n} \min\left(FFS_{a_j}^{t}, FFS_{a_j}^{t+1}\right) + \sum_{j=1}^{n} \min\left(FFS_{b_j}^{t}, FFS_{b_j}^{t+1}\right)\right] \quad (1)$$

where FFS, t denote the frames of the stack, and frame number respectively. The size of frame is $M \times N$ and j stands for the histogram bin number varying from 1 to n, and L, a, b for the Luminance, and 'a' and 'b' components. For the selection of most optimized frame on the basis of normalized histogram distances a Fuzzy logic model is constructed.

Fuzzy logic model: In the video sequence frame to frame difference varies greatly so it is not possible to work with a fixed

Please cite this article in press as: D.K. Vishwakarma, et al., A proposed unified framework for the recognition of human activity by exploiting the characteristics of action dynamics, Robotics and Autonomous Systems (2016), http://dx.doi.org/10.1016/j.robot.2015.11.013

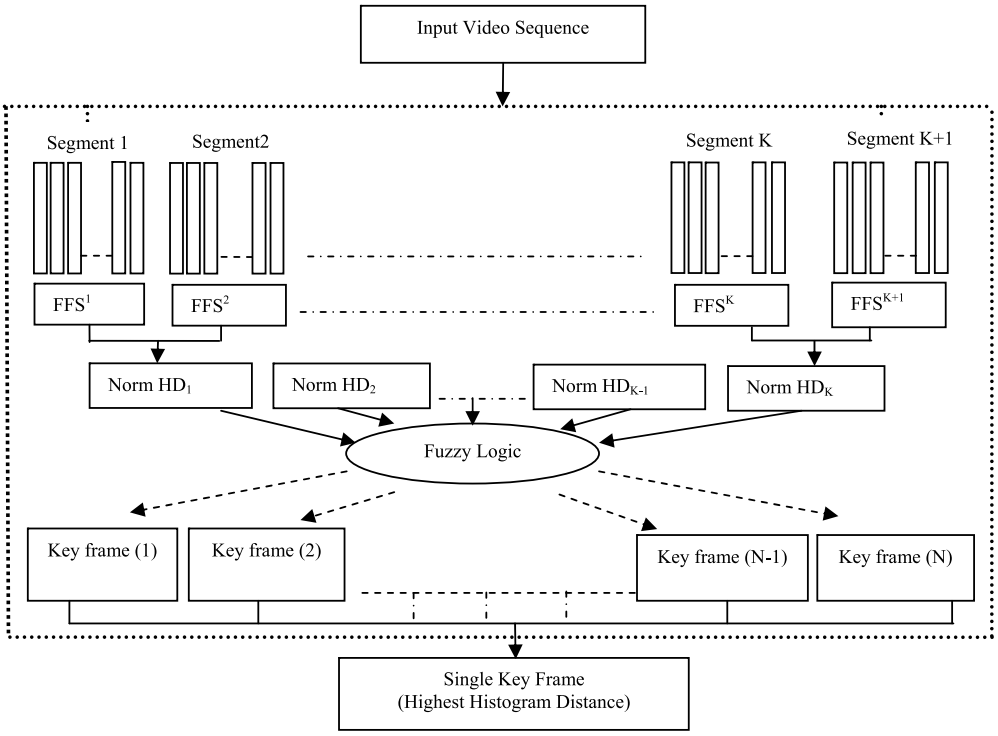

Fig. 2. Illustration of optimum single key frame extraction.

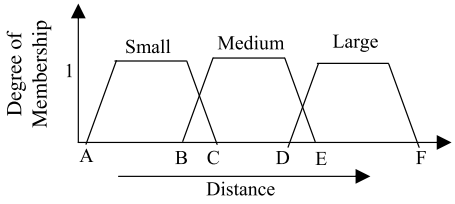

Fig. 3. Fuzzy trapezoidal membership function.

threshold for computing the key frames on the basis of normal histogram distance metric (see Fig. 3). Hence, the fuzzy approach is used to find the key frames, which is explained as Algorithm 1.

Selection of single key frame: The extracted key frames are compared internally and ranked according to the normalized histogram distance values. Histogram difference is typically considered as a global change in a video sequence and it helps in detecting abrupt changes in the frames. If the extracted key frames are represented as "Key frame (1), Key frame (2)…Key frame (N)" then the normalized histogram distance between the successive frames is computed and highest distance key frame is the optimum single key frame that has the highest pixel variation for the representation of 2D pose of the activity as shown in Fig. 4(a), (b).

3.2. Computation of SDEG

The computation of SDEG feature is based on 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 body posture hints both the enduring characteristics of a person (character, temperament, etc.) and the 2-D representations of the image give the spatial distribution of posture and characteristics of action or the attitude of the person.

For the representation of SDEG, the region of interest (ROI) of key pose is selected and further divided into sub-regions of still images at multilevel. The orientations of the edges are counted on the finer scale and expressed in the form vector. The pictorial representation of SDEG computation and feature representation is presented as Fig. 5 and stepwise flow is explained in Algorithm 2.

Fig. 6 shows that the SDEG feature vectors are almost discriminative because the representation of the patterns of the histograms of different activities is different with an increase in the degrees. At a lower degree (near zero) the peaks are higher and more variation in magnitude with increase in the degrees, although amplitude decreases but the peaks are more variant. Hence, these features are proficient in the representation of human activities but this representation may be less effective for the recognition of activities like "walking" and "jogging" because this does not have temporal information. Hence, to alleviate the limitations of structural information, it is further moved to incorporate the kinematics information, which is described in the subsequent section.

3.3. Computation of orientation of human silhouette

The orientation of silhouettes gives the directional as well as the temporal information of human body motion and these are computed using $\Re$-transform. The $\Re$-transform is computed by applying the Radon Transform (R_T) on the binary silhouettes of the human activity.

3.3.1. Extraction of human silhouette

The silhouette is the basic unit of human activity, which is formed by extracting the foreground object from the rest of the video sequence. A method for describing different textures as presented in [38] is used for silhouette extraction. The texture is always a reliable source of information for scene description and change detection. Entropy is one of the most important parameters

Please cite this article in press as: D.K. Vishwakarma, et al., A proposed unified framework for the recognition of human activity by exploiting the characteristics of action dynamics, Robotics and Autonomous Systems (2016), http://dx.doi.org/10.1016/j.robot.2015.11.013

Algorithm 1: Key Frame Extraction

Input: Video sequence of the activity.

Step 1: Divide the video sequence into a segment of group of frames. The first segment of frame is represented as FFS i.e. First Frame of the Segment group.

Step 2: Find a single key frame in each segments by comparing the histogram distance of each frames.

Step 3: Compute the normalized histogram distance in all 3 components 'L', 'a' and 'b' between two successive segments frames through Equation 1.

Step 4: Compute the mean the mean of all pixels, 'μ_d' for all consecutive frame differences as: $\mu_D = [\mathbf{count}\ `\mathbf{L}` + \mathbf{count}\ `\mathbf{b}` + \mathbf{count}\ `\mathbf{a}`]/3$.
Comment: Merging of pixels with the histogram distance improves the coverage while avoiding the redundancy for the segment frames.

Step 5: Compute the values endpoint (A, B, C, D, E, and F) of the membership function, which is shown in Fig. 3. Where {small, medium, large} are the linguistic variables. $\mathbf{A} = (\mu_D - \mu_D * 0.4)$, $\mathbf{B} = (\mu_D - \mu_D * 0.3)$, $\mathbf{C} = (\mu_D - \mu_D * 0.2)$, $\mathbf{D} = (\mu_D + \mu_D * 0.4)$, $\mathbf{E} = (\mu_D + \mu_D * 0.5)$, $\mathbf{F} = (\mu_D + \mu_D * 0.8)$. These values are optimum for our results as it is obtained from the training data after multiple trials.

Step 6: Frame fuzzy rules, to select the single global key frame.

Rule 1: IF the distance between a segment frame and its neighbouring segment frame is "medium" THEN it is a key frame.

Rule 2: IF the distance between a segment frame and its neighbouring segment frame is "large" THEN it is a key frame.

Rule 3: IF the distance between a segment frame and its neighbouring segment frame is "small" THEN it is NOT a key frame.

Output: Key frames as shown in Fig. 4.

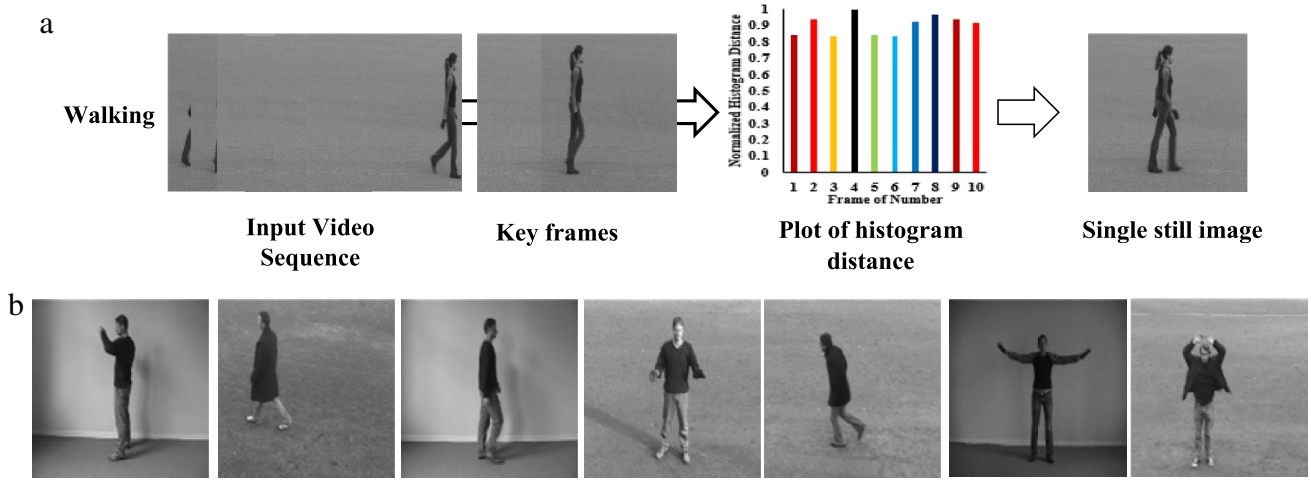

Fig. 4. (a) Illustration of single key pose extraction (b) depiction of single still images of different activities.

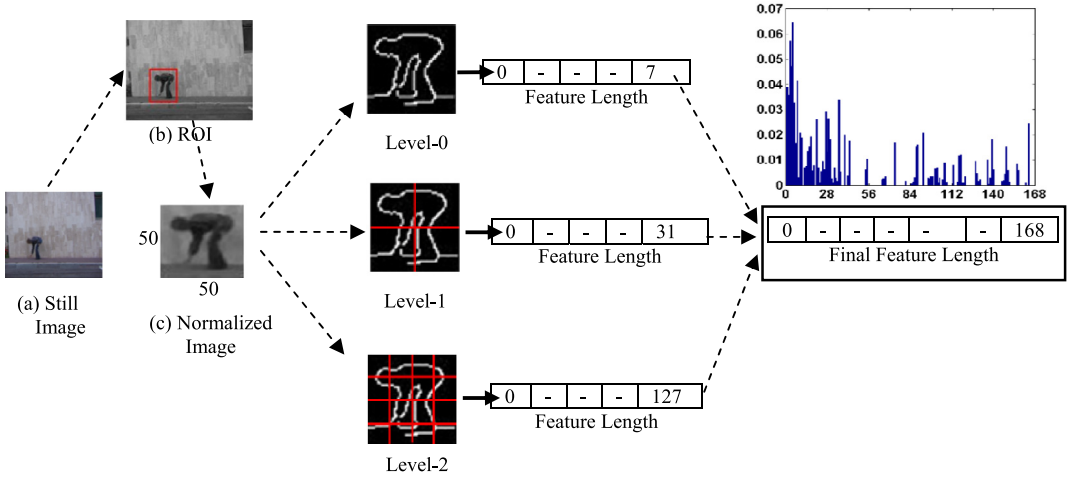

Fig. 5. Depicting the flow of computation of SDEG vector at 8-orientation bins.

Please cite this article in press as: D.K. Vishwakarma, et al., A proposed unified framework for the recognition of human activity by exploiting the characteristics of action dynamics, Robotics and Autonomous Systems (2016), http://dx.doi.org/10.1016/j.robot.2015.11.013

Algorithm 2: Computation of SDEG

Input: single still image

Step 1: Select ROI and normalize to the dimension of $\mathbf{50} \times \mathbf{50}$, which is denoted as: $\mathcal{N}(x, y)$, where $\mathbf{0} \leq x, y \leq \mathbf{50}$.

Step 2: Compute the edges of ROI using canny edge detector as: $\mathcal{E}(x, y) = \mathbf{Canny}(\mathcal{N}(x, y))$.

Step 3: Compute the spatial edge distribution at different levels as follows:

a. At level-0, the magnitude $\mathbf{M}(x, y)$ and orientation $\varphi(x, y)$ at any point (x, y) of the entire image $\mathcal{E}(x, y)$ is computed as: $\mathbf{M}(\mathbf{x}, \mathbf{y}) = \left[\mathcal{G}_{\mathbf{X}}(\mathbf{x}, \mathbf{y})^2 + \mathcal{G}_{\mathbf{y}}(\mathbf{x}, \mathbf{y})^2\right]^{\mathbf{0.5}}$ and $\varphi(x, y) = \mathbf{arctan}\,(\mathcal{G}_y(x, y)/\mathcal{G}_{\mathbf{x}}(x, y)$. Where $\mathcal{G}_{\mathbf{x}}(x, y)$ and $\mathcal{G}_{\mathbf{y}}(x, y)$ are image gradients along x and y directions. Each sub-region is quantized into the 8-orientation bins and evenly spaced over the range of $(\mathbf{0^0 - 360^0})$ orientations. The obtained feature length for the entire image is of $\mathbf{8 \times 1}$ size.

b. At level-1, the entire image $\mathcal{E}(x, y)$ is divided into 4 sub-image regions, and denoted as: $\mathcal{E}(x, y)$=$\{\mathcal{S}_1(x, y), \mathcal{S}_2(x, y), \mathcal{S}_3(x, y), \mathcal{S}_4(x, y)\}$. The feature length is obtained as in *step 3-a*, and dimension is of $\mathbf{8} \times [\mathbf{1} + \mathbf{4}]$.

c. At level-2, each sub-image region is further divided into 4 sub-blocks, and feature vector is computed as in *step 3-a*.The obtained feature vector for 16 sub-blocks is of $\mathbf{8} \times [\mathbf{1} + \mathbf{4} + \mathbf{16}]$.

Output: histogram of spatial distribution as shown in Fig.5, which has dimension of $K \sum_{l=0}^{2} \mathbf{4}^l$, where K is the number of orientation bins.

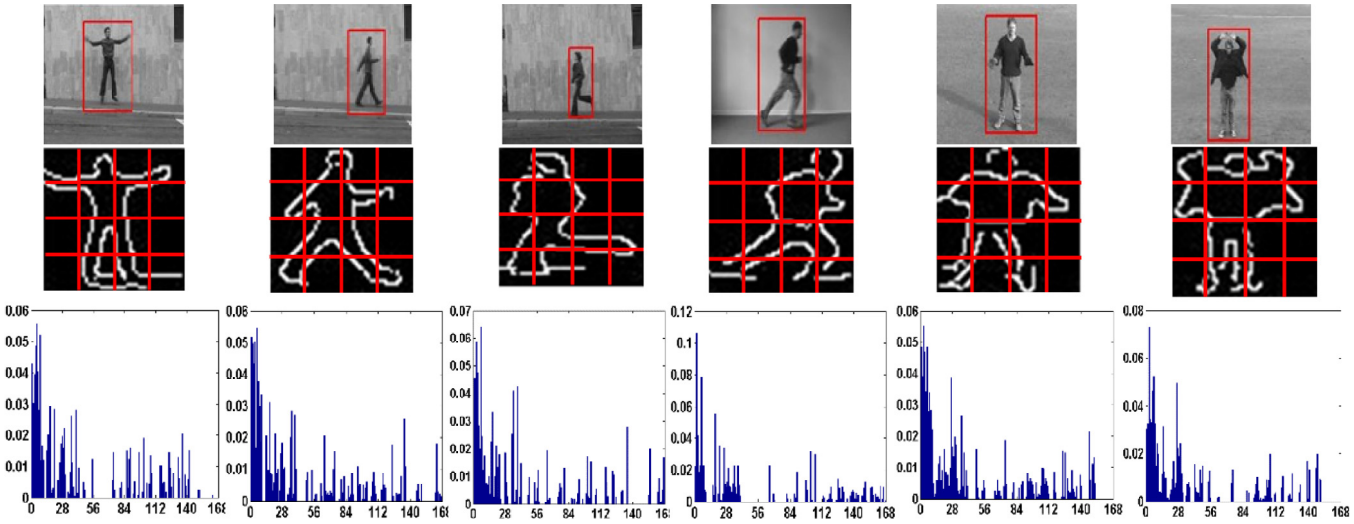

Fig. 6. Result of SDEG vector of different activities: Row 1: ROI of still Images, Row 2: Edges of postures, Row 3: Spatial Edge Distribution at level '2'.

that describe the texture information in an image and can be expressed as:

$$\zeta = \sum_i \sum_j \rho\,(i, j) \log\,(\rho\,(i, j)) \tag{2}$$

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 neighbourhood mask. Converting this filter matrix to a binary form gives an image with white spots at different areas. Applying this mask over the raw image provides a silhouette image from this raw image as shown in Fig. 7. The segmented image may contain different white blocks, but not all of them are of human silhouettes. By comparing the size of these blocks, the image with the largest area is selected, which is a human silhouette. To compute the motion temporal information, 25 key frames are extracted from the video sequence, which have the significant amount of energy.

3.3.2. Computation of rotation feature

Consider a sequence of silhouette image $I_t\,(x, y)$ of the human activity, where 't' is the frame number and subsequently $\Re$-Transform is defined via $\Re(\theta)$ as follows:

$$\Re\,(\theta) = \int_{-\infty}^{\infty} R_T^2(\rho, \theta) \partial \rho. \tag{3}$$

Radon Transform (R_T) 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 ∞, denoted as:

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

Please cite this article in press as: D.K. Vishwakarma, et al., A proposed unified framework for the recognition of human activity by exploiting the characteristics of action dynamics, Robotics and Autonomous Systems (2016), http://dx.doi.org/10.1016/j.robot.2015.11.013

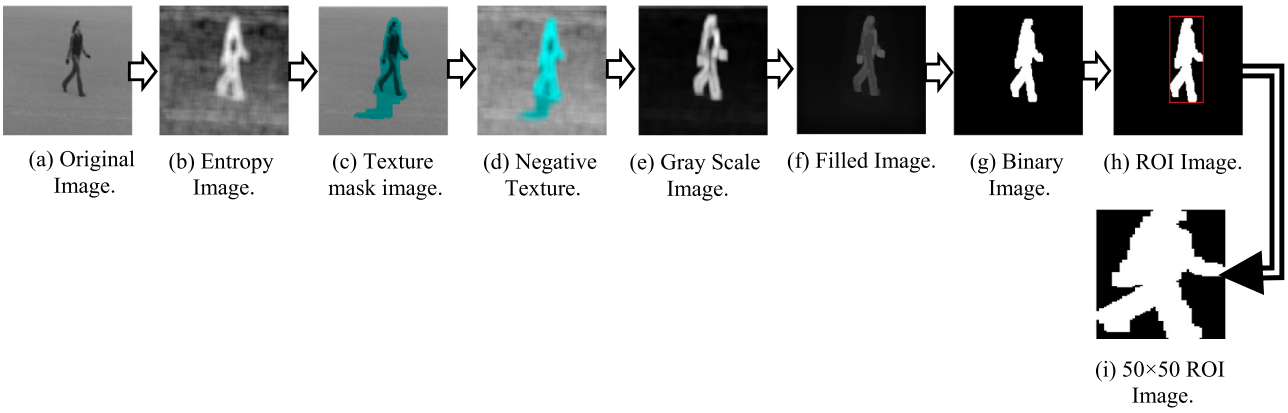

Fig. 7. Flow of steps for extraction of binary silhouette image.

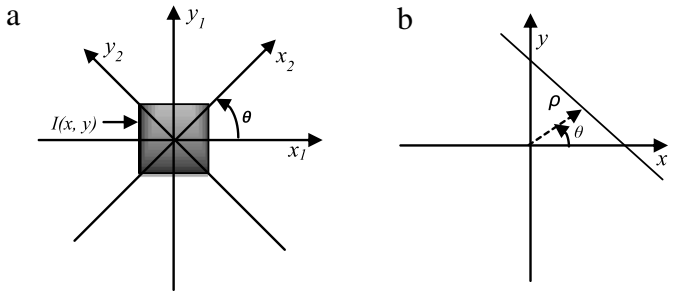

Fig. 8. (a) Illustrates the projection of lines over 2-D function $I(x, y)$, (b) shows the position of the projection line.

where $\delta(.)$ is defined as the Dirac delta function which is zero everywhere except at the origin, where it approaches infinity. The perpendicular distance ρ from the origin to the radon line is defined as Eq. (5) and shown in Fig. 8(a), (b).

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

where 'θ' is the angle between horizontal axis and the projection line.

R_T cannot restore all the parameters of the original geometric transformation when the image is translated, rotated, or scaled. Hence, [39] introduced the $\Re$-transform which is invariant to translation and scaling parameters but with orientations it provides sufficient discriminative change.

The normalized $\Re$-transform ($\Re_{\text{norm}}(\theta)$) improves the similarity measure and compactness of feature representation, which is defined as:

$$\Re_{\text{norm}}(\theta) = \frac{\int_{-\infty}^{\infty} \Re(\theta)\, d\theta}{\max(\Re(\theta))}. \tag{6}$$

Properties of $\Re$-Transform: The fundamental properties of $\Re$-transform are described by Tabbone et al. [39] which shows that it is invariant against the scaling and translation but sensitive against the rotational characteristics. These properties are proved in work by taking bending activity silhouette as shown in Fig. 9 for the human activity sequence under scaling, translation and rotation.

In Fig. 9, rotation in the image shows more variation in the brighter portion of R_T when compared to other images because in the rotation, 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 $\Re$-transform remains the same. The rotational sensitivity of $\Re$-transform is used for the representation of motion temporal information of different activities. The $\Re$-transform representation of different actions is as shown in Fig. 10.

Fig. 10 shows the normalized $\Re$-transform signal representation of different activities and it is observed that the representation is significantly different for various actions. The geometrical profiles of normalized $\Re$-transform for jogging and walking actions look similar due to the similar postures of the actions. Hence, it can be clinched that $\Re$-transform representation is capable of describing the motion characteristics of the human action but alone it cannot sufficiently provide the information for distinguishing the actions. The $\Re$-transform of a single frame of the human silhouette is not much effective in representing the temporal motion information as compared to the set of frames of silhouette, but it results in a high dimension. For effective and compact representation of $\Re$-transform, the local linear embedding (LLE) [40] unsupervised manifold learning non-linear dimension reduction method is used. Non-linear dimensionality reduction in comparison to the linear dimension reduction technique discards the correlated information and works on the maximization of information. Linear dimension reduction technique principal component analysis (PCA) [41] when applied to the dataset gives resulting values which are not well organized. Therefore, to reduce the dimension of feature set the non-linear LLE approach is used.

3.4. Final feature vector computation

The final feature descriptor is constructed by assimilating the SDEG and normalized $\Re$-transform, which gives a novel integrated structure for the representation of human activity. This assimilation gives a rich descriptor having both spatial and temporal information. The flow of assimilation is as shown in Fig. 11, which is based on the experiment data used in the proposed approach.

The dimension of feature vector obtained by $\Re$-Transform on one frame of size 50×50 is 2500×180 and in the proposed scheme 25 key frames are used to compute the rotation feature. Hence, the final dimension of the $\Re$-Transform is very high. Therefore, a

Please cite this article in press as: D.K. Vishwakarma, et al., A proposed unified framework for the recognition of human activity by exploiting the characteristics of action dynamics, Robotics and Autonomous Systems (2016), http://dx.doi.org/10.1016/j.robot.2015.11.013

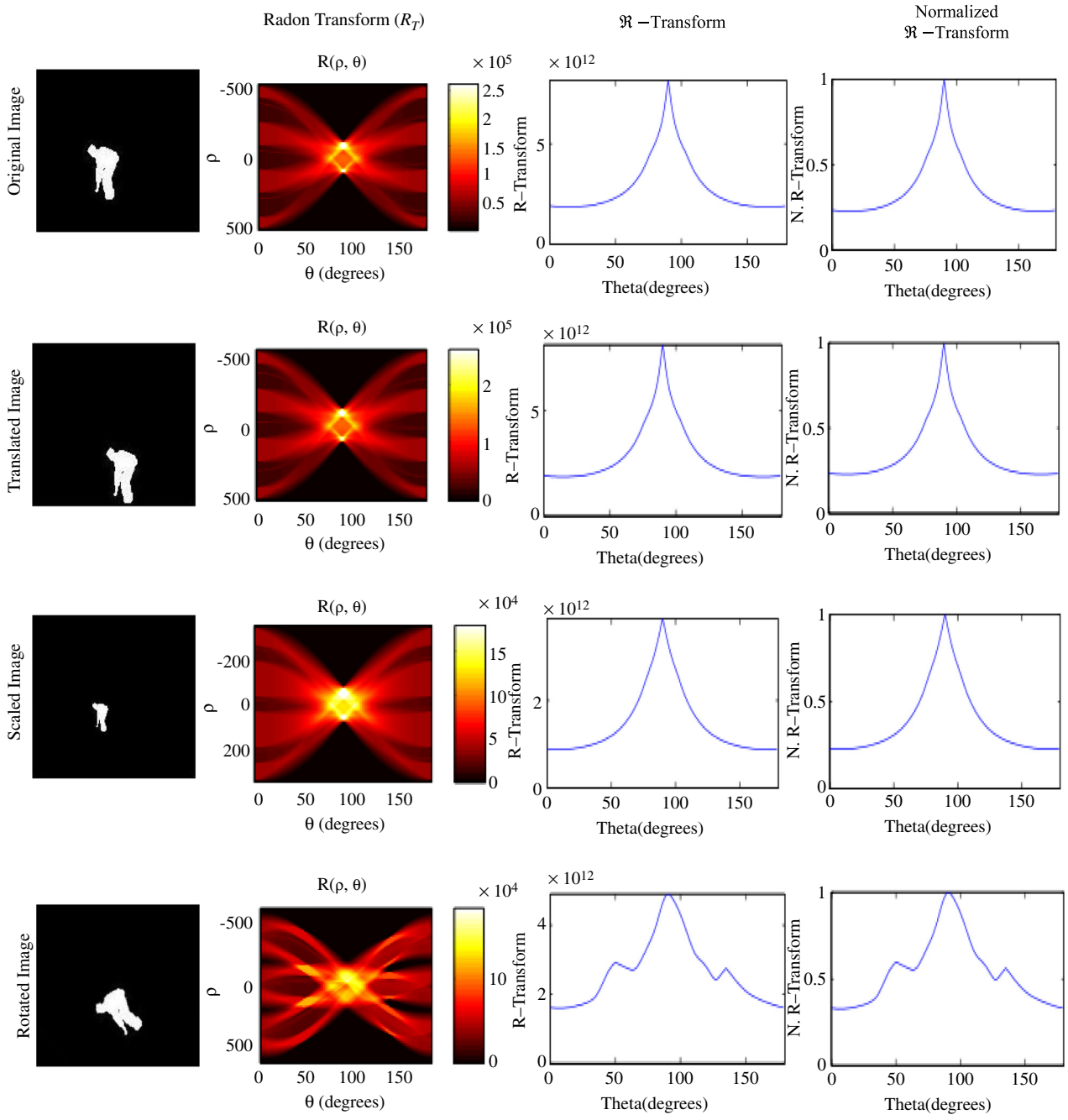

Fig. 9. Shows $\Re$-Transform is variant to the rotation and invariant to scaling and translation Column 1: silhouette images at various conditions, Column 2: Radon Transform (R_T), Column 3: $\Re$-Transform signal, Column 4: Normalized $\Re$-Transform signal.

dimension reduction technique LLE is used and which gives the size of $\Re$-Transform feature as 1×25. Similarly, for the SDEG computation, there are 8-orientation bins and decomposition level-2 is used in the proposed approach, which gives the size of 1×168 as explained in the computation of SDEG section. The final feature vector is obtained by concatenation of these two feature vectors together. Hence, the size of the feature vector is $[1 \times 168 + 1 \times 25 = 1 \times 193]$.

4. Experimental result and discussion

In order to demonstrate the effectiveness of the proposed approach, an extensive experiment is conducted on five publicly available and widely used human action datasets—the Weizmann [42], KTH [43], Ballet movement [44], Multi-view id3Post [45] and IXMAS [46] datasets. The sole reason behind the use of these datasets is to show the robustness of the proposed approach against the variation of illumination, viewing angle, inter-class similarity, and intra-class dissimilarity of the activities. The performance is measured in terms of classification accuracy using Multi-class SVM classifier in leave-one-out cross validation (LOOCV) routine.

Weizmann dataset: The dataset was introduced by Gorelick et al. [42] and comprises of 90 videos with a frame rate of 25 fps and each frame having a size of 144×180. There are 9 people (male, female), who have performed 10 different actions and categorized as 'Run', 'sideways jump', 'Skip', 'jumping', 'jump in place', 'bending', 'jumping jack', 'walk', 'one hand wave (Wave 1)', and 'two hand (Wave 2)'. The sample image postures of these activities are as shown in Fig. 12.

Please cite this article in press as: D.K. Vishwakarma, et al., A proposed unified framework for the recognition of human activity by exploiting the characteristics of action dynamics, Robotics and Autonomous Systems (2016), http://dx.doi.org/10.1016/j.robot.2015.11.013

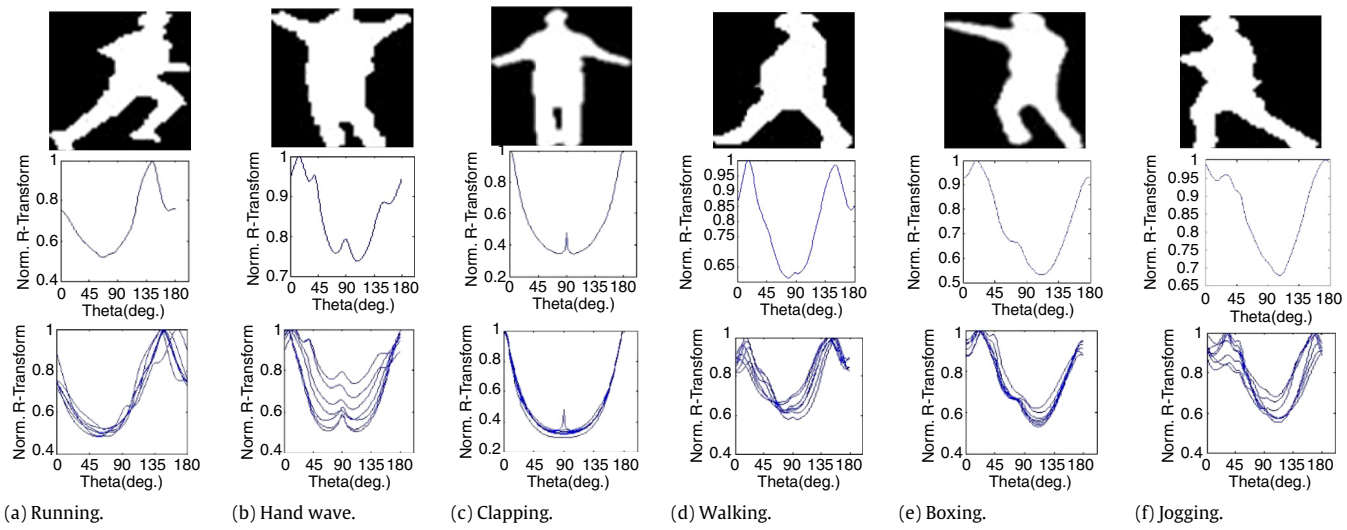

Fig. 10. Representation of $\mathfrak{R}$-transform for different activities: Row 1: 50 × 50 Silhouette Image, Row 2: $\mathfrak{R}$-transform, Row 3: $\mathfrak{R}$-transforms of few key frames.

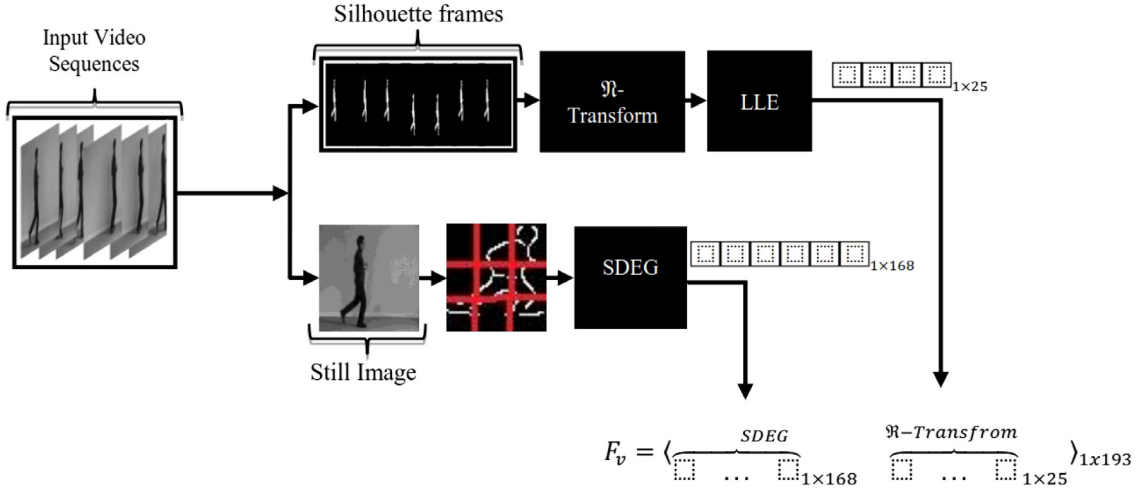

Fig. 11. Illustration of formation of final feature vector.

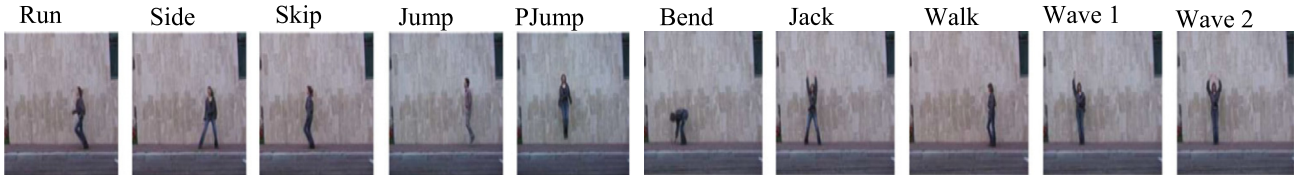

Fig. 12. Sample frames of Weizmann human action dataset.

KTH dataset: The dataset was introduced by Schuldt et al. [43] and it is a more challenging dataset as compared to the Weizmann dataset. The dataset consists of six basic activities that are named as 'hand-clapping', 'hand-waving', 'jogging', 'jumping', 'running', and 'walking'. There are 100 videos of each activity in four different scenarios, which comprise the variations of recording conditions like illumination, indoor, outdoor, zoom-in and zoom-out. All these video sequences are recorded in a background with a static camera of frame rate 25 fps and which is further down-sampled to the spatial resolution of 160 × 120 pixels. The sample images of the datasets are shown as in Fig. 13.

Ballet dataset: Ballet dataset [44] is one of the highly complex human action dataset due to the intra-class difference and inter-class similarity. The dataset comprises of eight ballet human movements i.e. 'Left-to-right hand (B1)', 'Right-to-left hand opening (B2)', 'Standing hand opening (B3)', 'Leg Swinging (B4)', 'Jumping (B5)', 'Turning (B6)', 'Hopping(B7)', and 'Standing still (B8)'. The sample image of different actions of the dataset are as shown in Fig. 14.

i3DPost dataset: The i3DPost Multi-View database [45] consists of 12 human actions i.e. 'Walk (P1)', 'Run (P2)', 'Jump(P3)', 'Bend (P4)', 'Hand-Wave(P5)', 'Jump in Place(P6)', 'Sit–stand-up (P7)', 'Run–Fall (P8)', 'Walk–Sit (P9)', 'Run–Jump–Walk (P10)', 'Handshake (P11)', and 'Pull (P12)' with high resolutions of 1920 × 1080, frame rate of 25 fps, where 6 actions (walk, run, jump, bend, hand-wave and jump-in-place) belong to the single action category, 4 actions (sit–stand-up, run–fall, walk–sit, and run–jump–walk) fit into the combined action and 2 actions (handshake, pull) belong to the interaction category. For each activity of the dataset, 8 persons performed the action, and it is recorded from 8 different view-

Please cite this article in press as: D.K. Vishwakarma, et al., A proposed unified framework for the recognition of human activity by exploiting the characteristics of action dynamics, Robotics and Autonomous Systems (2016), http://dx.doi.org/10.1016/j.robot.2015.11.013

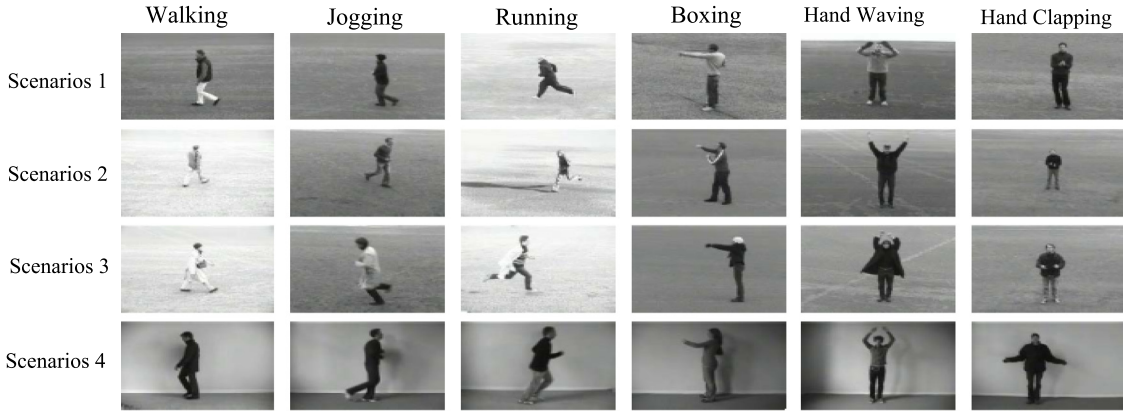

Fig. 13. Sample frames of KTH dataset.

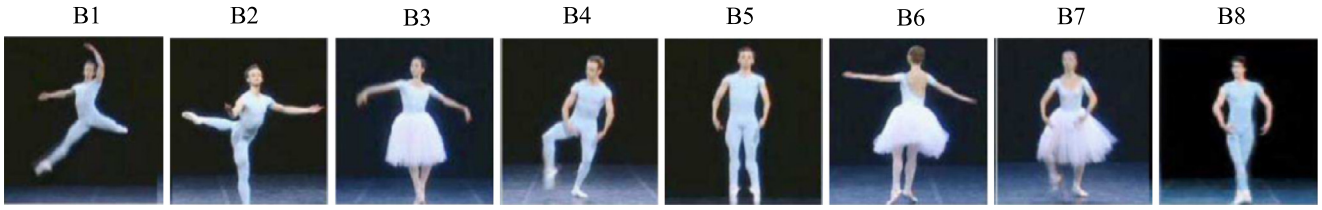

Fig. 14. Images of the Ballet dataset depicting eight movement of actions.

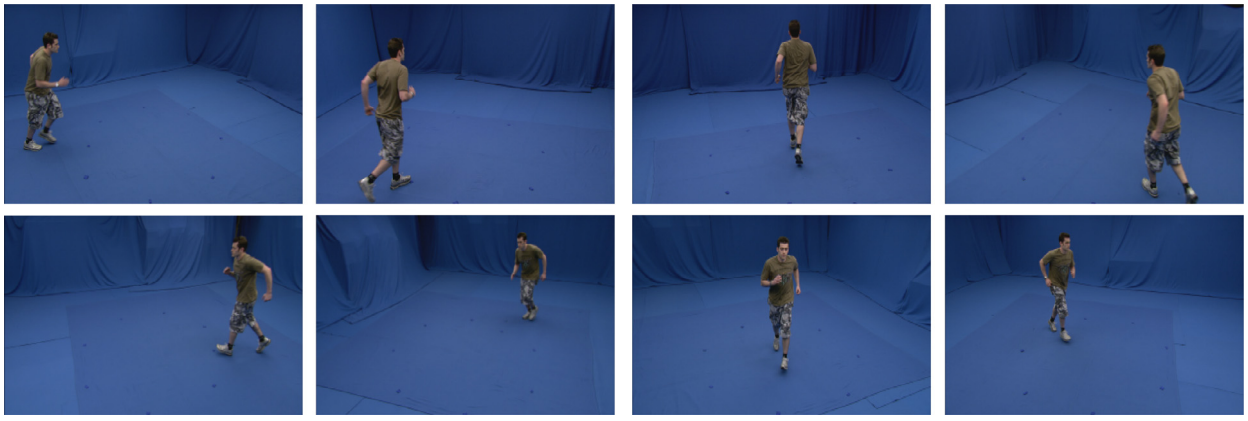

Fig. 15. Sample frames of id3Post data of 'Running' activity from eight different views.

angles hence total 64 videos for the single action. The sample frames of the dataset are as depicted in Fig. 15.

Inria Xmas Motion Acquisition Sequence (IXMAS) dataset: INRIA Xmas Motion Acquisition Sequences (IXMAS) database is introduced by Weinland et al. [46], which is one of widely used dataset for the multi-view/3D analysis of the human actions and recorded through 5 cameras at different positions. A total of 13 daily living activities are recorded and labelled as 'check watch', 'cross arms', 'scratch head', 'sit down', 'get up', 'turn and around', 'walk', 'wave', 'punch', 'kick', 'pointing', 'pickup' and 'throw'. These activities are performed 3 times by 12 actors surrounded with fixed 5 cameras, each capturing 23 fps with 390×291 spatial resolution. The actors performing the activities are different in body size, clothing, sex and execution rate. The sample frames of activities of the dataset are as shown in Fig. 16.

4.1. Performance evaluation on dataset

The performance of the proposed approaches is evaluated in terms of the average recognition accuracy (ARA) on five different benchmarked of the datasets. The recognition accuracy is computed through the proposed approach (SDEG+R) as well as the similar techniques like pyramid of histogram oriented gradients (PHOG) [5], Histogram of oriented gradients (HOG) [47], and $\Re$-Transform [39]. The ARA evaluated for each method is shown in Table 1.

Table 1
Shows the ARA (%) achieved with the similar methods on different benchmark of the datasets.

Method	Datasets				
	Weizmann	KTH	Ballet	i3dpost	IXMAS
PHOG	52.88	45.44	27.98	33.60	48.96
HOG + R	90.10	80.22	64.33	55.66	69.45
$\Re$-Transform	84.77	72.23	45.65	60.90	51.36
SDEG + $\Re$	***100***	***95.5***	***93.25***	***92.92***	***85.5***

As it is seen from Table 1, the ARA achieved through HOG + R is lower than the SDEG + R due to less structural information and more textural variations. Due to some similar representations for different actions HOG + R may produce false recognition for an action. The $\Re$-Transform provides good motion temporal information of the activity but when the activities are very similar

Please cite this article in press as: D.K. Vishwakarma, et al., A proposed unified framework for the recognition of human activity by exploiting the characteristics of action dynamics, Robotics and Autonomous Systems (2016), http://dx.doi.org/10.1016/j.robot.2015.11.013

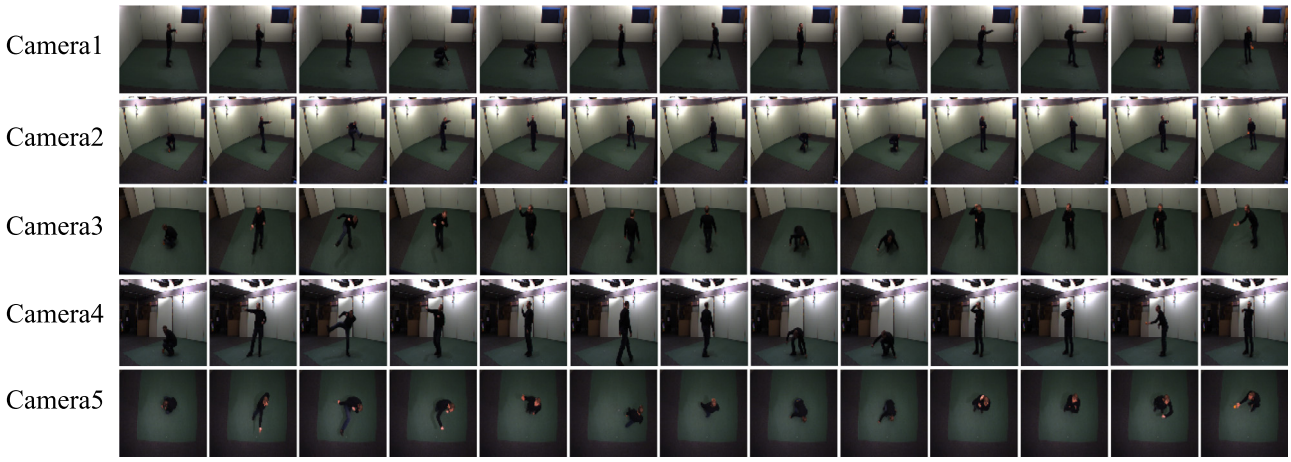

Fig. 16. Sample frames of IXMAS dataset of five different Camera.

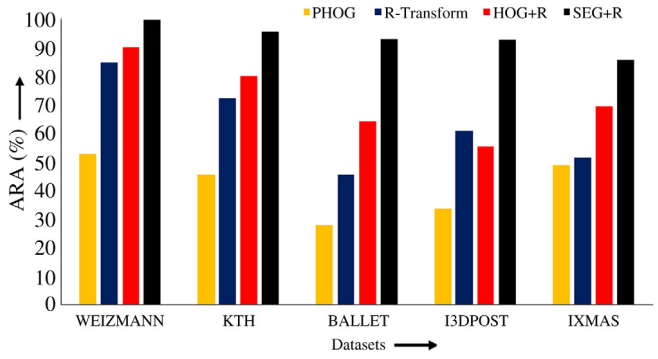

Fig. 17. Shows the comparison of ARA (%) with different approaches on different datasets.

in nature, it is unable to discern and thus, the recognition accuracy is inferior as compared to the SDEG + R. PHOG approach provides only the shape information of the object while for the activity recognition shape and temporal both are the essential information hence ARA is less as compared to the SDEG + R. Therefore, the high ARA shows that SDEG + R is an effective model for the recognition of human activity.

The ARA achieved through proposed approach is compared with the similar techniques on different datasets, which is shown in Fig. 17 and for each dataset, the ARA achieved through proposed method is highest in comparison with other approaches. Hence, the proposed approach is more effective in the recognition of human activities as verified under the variegated conditions presented in the datasets.

The ARA 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 silhouette is accomplished which favours in achieving 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 previous dataset. Hence, the less ARA (95.50%) is achieved as compared to Weizmann dataset.

The Ballet movement dataset is highly complex dataset in terms of intra-class dissimilarities like execution, speed, clothing, etc. The ARA achieved on this dataset is 93.25% which 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 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 $\mathfrak{R}$-transform but the multiple fusion of the orientation and spatial distribution gives significant improvement in the recognition accuracy. SDEG feature vector greatly improves the accuracy in the ballet dataset as it gives distinguishable features, which are more variant as compared to the $\mathfrak{R}$-Transform characteristics. For the computation of rotation feature, the silhouette is required and the extraction of silhouette is difficult in some cases where hands and body parts which are occluded.

The id3Post dataset is multi-view dataset in which the same action is recorded from 8 different angles. The ARA achieved on this dataset is 92.92%, verifying the robustness of proposed approach against the variation in viewing angle. The effect of $\mathfrak{R}$-transform coefficients is not much contrasting for the activities like Jump in Place, Sit–stand-up or Run and Run–Jump–Walk.

The recognition accuracy achieved on the IXMAS dataset for 13 activities from five Cameras (1, 2, 3, 4 and 5) is 83.23%, 81.49%, 78.01%, 85.80% and 74.84 respectively and the recognition accuracy achieved on the videos of all the cameras is significant, which shows the robustness of the proposed algorithms against the view variant. The higher recognition accuracy achieved by cameras 1 and 4 is due to the location of the camera and similarity of the actions but the variation of recognition accuracy is very minute.

4.2. Comparison of results

To know the effectiveness of the proposed approach in comparison to the techniques put forward by others, a comparative study is done and as presented in Tables 2–6 for different human action datasets. The classification strategies used in the different techniques are quite similar but named in the different form as leave-one-out cross validation (LOOCV), leave-one-sequence-out cross validation (LOSOCV), leave-one-video-out cross validation (LOVOCV) and leave-one-person-out cross validation (LOPOCV).

For the Weizmann dataset the experiment setting used in this experiment is similar to the Gorelick et al. [42] and other techniques in respect to the input conditions, classification model and classification strategies. Hence, the comparison given in Table 2 is reasonably fair.

The experimental setup used for the evaluation ARA on KTH dataset is similar to the original paper [43] on KTH dataset. The vital cause for achieving high ARA is the effective silhouette extraction, which is a tough task for this dataset because of the illumination change and zooming and zoom out of camera. The recognition accuracy achieved in [55] is very close to the proposed approach due to the effective textural features and the large amount of training.

Please cite this article in press as: D.K. Vishwakarma, et al., A proposed unified framework for the recognition of human activity by exploiting the characteristics of action dynamics, Robotics and Autonomous Systems (2016), http://dx.doi.org/10.1016/j.robot.2015.11.013

Table 2
Comparison of ARA with the techniques of others on Weizmann dataset.

Method	Input	Classifiers	Test scheme	ARA (%)
Gorelick et al. [42]	Silhouettes	KNN	LOOCV	97.5
Nibeles et al. [32]	Images	pLSA	LOOCV	90
Rahman et al. [48]	Silhouettes	NN	LOOCV	100
Chaaraoui et al. [49]	Silhouettes	KNN	LOSOCV	92.8
Wu and Shao [50]	Silhouettes	SVM	LOSOCV	97.78
Goudelis et al. [29]	Silhouettes	SVM	LOPOCV	95.42
Touati and Mignotte [51]	Silhouettes	KNN	LOOCV	92.3
Cai et al. [52]	Silhouettes	Max/sum pooling	LOOCV	97.85
Proposed method	*Silhouettes, image*	*SVM*	*LOOCV*	**100**

Table 3
Comparison of ARA with the techniques of others on KTH dataset.

Method	Input	Classifiers	Test scheme	ARA (%)
Sadek et al. [53]	Silhouettes	SVM	–	93.30
Saghafi and Rajan [54]	Silhouettes	KNN	LOOCV	92.6
Goudelis et al. [29]	Silhouettes	SVM	LOPOCV	93.14
Melfi et al. [55]	Silhouettes	SVM	LOOCV	95.25
Rahman et al. [56]	Silhouettes	KNN	LOOCV	94.49
Benmokhtar [57]	Images	SVM	LOOCV	92.5
Proposed method	*Silhouettes, images*	*SVM*	*LOOCV*	**95.50**

Table 4
Comparison of ARA with the techniques of others on Ballet dataset.

Method	Input	Classifiers	Test scheme	ARA (%)
Fathi and Mori [44]	Silhouettes	SVM	–	51
Wang and Mori [58]	Silhouettes	KNN	LOO	91.3
Guha and Ward [59]	Silhouettes	SVM	LOO	91.1
Ming et al. [60]	Silhouettes	SVM	LOO	90.8
Iosifidis et al. [61]	Silhouettes	KNN	LOVOCV	91.1
Vishwakarma and Kapoor [62]	Silhouettes	SVM	LOOCV	92.75
Proposed method	*Silhouettes, images*	*SVM*	*LOOCV*	**93.25**

Table 5
Comparison of ARA with the techniques of others on Multi-view i3dPost dataset.

Method	Actions	ARA (%)
Gkalesis et al. [45]	5 single actions (walk, run, jump in place, jump forward and bend)	90.00
Holte et al. [63]	5 single actions (excluding run)	97.00
	10 actions (single + combined)	80.00 (3D MC)
Iosifidis et al. [64]	5 single actions (excluding run)	97.80
	8 actions(6 single + 2-interactions)	96.34
Proposed method	*5 (single actions) excluding run*	**97.50**
	6 (single actions) including run	**95.56**
	10 actions (single + combined)	**92.92**

Table 6
Comparison of ARA with the techniques of others on IXMAS multi-view dataset.

Method	Camera	Number of actions	Classifier/test scheme	ARA (%)
Chaaraoui et al. [49]	5	12	KNN/LOSOCV	*85.86*
Wu et al. [65]	4	12	SVM/LOSOCV	*89.4*
Weinland et al. [46]	5	11	PCA + M/LOOCV	*93.33*
Yang et al. [66]	5	13	SVM/10-fold	84.55
Zhang et al. [67]	5	11	SVM/LOACOCV	83.5
Proposed method	**5**	**13**	*SVM/LOOCV*	**85.80**

The Ballet dataset is considered to be toughest dataset set in respect of the complexity of human activity performed, but the silhouette extraction for the dataset is reasonably easy due to stable environmental conditions. The experimental setting used in this experiment is as used by creator [44] of the dataset. The ARA achieved in [62] is close to the proposed method due to the similar approach used for the silhouette extraction, though in the proposed approach ARA is slightly higher than the Vishwakarma and Kapoor [62] due to the additional information of rotation of human silhouette.

Table 5, shows the comparison of recognition accuracy achieved on i3dPost dataset with the techniques of others. The comparison is quite limited due to less number of the work reported on this dataset. The recognition accuracy computed on this dataset is in three different categories and these are created on the basis of similarities between the different classes. The recognition accuracy reported in the three categories is varying due to the similarity between the classes and when highly similar classes are clubbed together then the recognition accuracy achieved is the highest. The action classes which have high similarity are 'Run', 'Walk',

Please cite this article in press as: D.K. Vishwakarma, et al., A proposed unified framework for the recognition of human activity by exploiting the characteristics of action dynamics, Robotics and Autonomous Systems (2016), http://dx.doi.org/10.1016/j.robot.2015.11.013

'Jump', 'Run–Jump–Walk', and 'Run–fall' and for these classes the performance of SDEG feature vector is less effective as compared to the $\Re$-transform. The overall ARA achieved for all ten classes is less as compared to the grouped similar classes but in comparison with the techniques of others, the achieved ARA is still encouraging.

Table 6, presents the comparison of the ARA achieved on IXMAS dataset by the proposed approach with the quite similar approaches. The comparison is marginally different in respect of the number action classes, number camera, classifier and test strategy used. The per class recognition accuracy achieved by proposed approach is reasonably high as compared to other techniques. Recognizing 13 actions is difficult as compared to recognizing 11 actions due to more confusion, while in this work the 13 actions are classified.

The experimental result and comparison with the similar state-of-the-art methods show few interesting observations, which are as follows:

- As the number of levels for the computation of SDEG features increases, the dimension of SDEG feature vector increases but the recognition accuracy does not increase significantly.
- As the number of key poses increases for the computation of temporal information, the recognition accuracy increases slightly, but it results in high feature vector dimension. This may lead to system complexity and increases in cost.
- It is also observed that $\Re$-transform is most effective for those activities which have high orientation like bending, crouching, etc.

5. Conclusion

In this paper, human activity recognition based upon the fusion of SDEG of human poses and orientation of key poses of human silhouettes is presented, which is executed separately but sequentially. SDEG is computed using a single frame representing the 2D posture of the activity while $\Re$-transform is used for the computation of orientation features providing the temporal information of human silhouettes. A single frame of the activity is extracted from a video sequence using histogram distances between the key frames. The SDEG is computed at different levels and orientations of bins. As the number of levels increases, a better quality of spatial distribution is achieved, but the complexity of the system increases because of the increase in the number of vectors. It is also inferred that with an increase in the levels, the magnitude of the spatial gradient vectors across the degrees decreases, which further diminishes the significance of discriminating characteristics of features vectors due to lower values. $\Re$-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. It is also inferred that increasing the number of frames not only increases the robustness, but also increases the computation time significantly. The advantage of this fusion approach is to offer a numerous distinctive feature vector, which leads us to robust and noise free action modelling.

In future, the number of key poses used for orientation estimation, the number of levels and orientations bins used for description of still poses may be optimized to get better results in terms of the recognition rate, and faster computation with less complexity. Another possible direction of future work is to incorporate a dynamic model having multiple objects perform simultaneously and in front of a dynamic crowded background since it is an action classification method.

For the application point of view using this approach an intelligent system can be developed for the purpose of surveillance, prohibited area notification, gait recognition, face animation, abnormal activity monitoring of elderly people, and to provide assistance for physical exercise, etc.

References

[1] P. Turaga, R. Chellappa, V.S. Subrahmanian, O. Udrea, Machine recognition of human activities: A survey, IEEE Trans. Circuits Syst. Video Technol. 18 (11) (2008) 1473–1488.
[2] D.K. Vishwakarma, R. Kapoor, An efficient interpretation of hand gestures to control smart interactive television, Int. J. Comput. Vis. Robot. (2015) 1–18. http://www.inderscience.com/info/ingeneral/forthcoming.php?jcode=ijcvr.
[3] M.B. Holte, C. Tran, M.M. Trivedi, Human pose Estimation and activity recognition from multi-view videos: Comparative explorations of recent developments, IEEE J. Sel. Top. Sign. Proces. 6 (5) (2012) 538–552.
[4] J. Agrawal, M. Rayoo, Human activity analysis: A review, ACM Comput. Surv. (2011) 16–43.
[5] A. Bosch, A. Zisserman, X. Munoz, Representing shape with a spatial pyramid kernel, in: ACM International Conference on Image and Video Retrieval, Amsterdam, 2012.
[6] Y. Wang, H. Jiang, M. Drew, Z. Li, G. Mori, Unsupervised discovery of action classes, in: IEEE Conference on Computer Vision and Pattern Recognition, 2006.
[7] P. Li, J. Ma, What happening in a still picture? in: IEEE Asian conference on pattern Recognition, 2011.
[8] J.L. Li, L. Fei-Fei, What, where and who? Classifying events by scene and object recognition, in: IEEE Conference on Computer Vision, Rio de Janeiro, Brazil, 2007.
[9] A. Lopes, E. Santos, E. Valle, J. Almeida, A. Araujo, Transfer learning for human action recognition, in: International Conference on Graphics Patterns and Images, 2011.
[10] Y. Zheng, Y. Zhang, X. Li, B. Liu, Action Recognition in still images using a combination of human pose and context information, in: 19th International Conference on Image Processing, ICIP, 2012.
[11] G. Guo, A. Lai, A survey on still image based human action recogntion, Pattern Recognit. 47 (2014) 3343–3361.
[12] H. Zhang, Z. Liu, H. Zhao, Recognizing human activities by key frame in video sequence, J. Softw. 5 (8) (2010) 818–825.
[13] Z. Khan, W. Sohn, Abnormal human activity recogntion system based on R-Transform and Kernel Discriminant Technique for Eldely Home Care, IEEE Trans. Consum. Electron. 57 (4) (2011) 1843–1850.
[14] A. Jalal, M. Uddin, T. Kim, Depth video based human activity recognition system using translation and scaling invariant features for life logging at smart home, IEEE Trans. Consum. Electron. 58 (3) (2012) 863–871.
[15] M. Singh, M. Mandal, A. Basu, Pose recognition using the radon transform, in: 48th Midwest Symposium on Circuits and Systems, 2005.
[16] D. Zhao, L. Shao, X. Zhen, Y. Liu, Combining appearence and structural features for human action recogntion, Neurocomputing 113 (3) (2013) 88–96.
[17] M. Bregonzio, T. Xiang, S. Gong, Fusing appearance and distribution information of interst points for action recognition, Pattern Recognit. 45 (3) (2012) 1220–1234.
[18] J. Dou, J. Li, Robust human action recognition based on spatio-temporal descriptors and motion temopral templates, Optik 125 (7) (2014) 1891–1896.
[19] D.K. Vishwakarma, P. Rawat, R. Kapoor, Human activity recognition using gabor wavelet transform and ridgelet transform, Procedia Comput. Sci. 57 (2015) 630–636.
[20] S. Althloothi, M.H. Mahoor, X. Zhang, R.M. Voyles, Human activity recognition using multi-features and multiple kernel learning, Pattern Recognit. 47 (5) (2014) 1800–1812.
[21] D.K. Vishwakarma, R. Kapoor, Integrated approach for human action recognition using edge spatial distribution, direction pixel, and R-transform, Adv. Robot. (2015) http://dx.doi.org/10.1080/01691864.2015.1061701.
[22] B. Yao, A. Khosla, L. Fei-Fei, Combining randomization and discrimination for fined-grained image categorization, 2011.
[23] C. Thurau, V. Hlavac, Pose primitive based human action recognition in videos or still images, in: IEEE Conference on Computer Vision and Pattern Recognition, 2008.
[24] J. Hu, W. Zheng, J. Li, S. Gong, T. Xiang, Recognizing human–object interaction via exemplar based modelling, in: IEEE International Conference on Computer Vision, 2013.
[25] A.F. Bobick, J.W. Davis, The recognition of human movement using temporal templates, IEEE Trans. Pattern Anal. Mach. Intell. 23 (3) (2001) 257–267.
[26] A.A. Efros, A.C. Berg, G. Mori, J. Malik, Recognizing action at a distance, in: IEEE International Conference on Computer Vision, 2003.
[27] L. Shao, L. Ji, Y. Liu, J. Zhang, Human action segmentation and recognition via motion and shape analysis, Pattern Recognit. Lett. 33 (2012) 438–445.
[28] L. Shao, X. Zhen, D. Tao, X. Li, Spatio-temporal Laplacian pyramid coding for action recognition, IEEE Trans. Cybern. 44 (6) (2014) 817–827.
[29] G. Goudelis, K. Karpouzis, S. Kollias, Exploring trace transform for robust human action recognition, Pattern Recognit. 46 (12) (2013) 3238–3248.
[30] I. Laptev, On space–time interest points, Int. J. Comput. Vis. 64 (2/3) (2005) 107–123.
[31] I. Laptev, M. Marszałek, C. Schmid, B. Rozenfeld, Learning realistic human actions from movies, in: IEEE Conference on Computer Vision & Pattern Recognition, 2008, pp. 1–8.
[32] J. Nibeles, H. Wang, L. Fei-Fei, Unsupervised learning of human action categories using spatial–temporal words, Int. J. Comput. Vis. 79 (3) (2008) 299–318.

Please cite this article in press as: D.K. Vishwakarma, et al., A proposed unified framework for the recognition of human activity by exploiting the characteristics of action dynamics, Robotics and Autonomous Systems (2016), http://dx.doi.org/10.1016/j.robot.2015.11.013

[33] G. Somasundaram, A. Cherian, V. Morellas, N. Papanikolopoulos, Action recognition using global spatio-temporal features derived from sparse representations, Comput. Vis. Image Underst. 123 (2014) 1–13.
[34] K. Tran, I. Kakadiaris, S. Shah, Part-based motion descriptor image for human action recognition, Pattern Recognit. 45 (7) (2012) 2562–2572.
[35] L. Shao, R. Gao, Y. Liu, H. Zhang, Transform based spatio-temporal descriptors for human action recognition, Neurocomputing 74 (6) (2011) 962–973.
[36] D.K. Vishwakarma, A. Dhiman, M. Rockey, R. Kapoor, Human motion analysis by fusion of silhouette orientation and shape features, Procedia Comput. Sci. 57 (2015) 438–447.
[37] P. Zeng, Z. Chen, Perceptual quality measure using JND model of the human visual system, in: EEE International Conference on Electric Information and Control Engineering, 2011.
[38] R. Haralick, K. Shanmugam, I. Dinstein, Textural features for image classification, IEEE Trans. Syst. Man Cybern. 6 (1973) 610–621.
[39] S. Tabbone, L. Wendling, J.P. Salmon, A new shape descriptors defined on the Randon Transform, Comput. Vis. Image Underst. 102 (1) (2006) 42–51.
[40] J. Chen, Y. Liu, Local linear embedding: a survey, Artif. Intell. Rev. 6 (1) (2011) 29–48.
[41] I.T. Jolliffe, Principal Component Analysis, Springer-Verlag, New York, 2002.
[42] L. Gorelick, M. Blank, E. Shechtman, M. Irani, R. Basri, Actions as space–time shapes, IEEE Trans. Pattern Anal. Mach. Intell. 29 (12) (2007) 2247–2253.
[43] C. Schuldt, I. Laptev, B. Caputo, Recognizing human actions: a local SVM approach, in: Proc. of the International conference on Pattern Recognition, 2004.
[44] A. Fathi, G. Mori, Action recognition by learning mid-level motion features, in: Proc. of IEEE conference on Computer Vision and Pattern Recognition, 2008.
[45] H. Gkalesis, H. Kim, A. Hilton, N. Nikolaidis, I. Pitas, The i3D Post multi-view and 3D human action/interaction, in: Proc. CVMP, 2009.
[46] D. Weinland, R. Ronfard, E. Boyer, Free viewpoint action recognition using motion history volumes, Comput. Vis. Image Underst. 104 (2–3) (2006) 249–257.
[47] N. Dalal, B. Triggs, Histograms of oriented gradients for human detection, in: IEEE Conference onComputer Vision and Pattern Recognition, CPVR, San Diego, CA, USA, 2005.
[48] S.A. Rahman, S.-Y. Cho, M.K.H. Leung, Recognising human actions by anlysing negative spaces, IET Comput. Vis. 6 (3) (2012) 197–213.
[49] A. Chaaraoui, P.C. Perez, F. Revuelta, Sihouette-based human action recognition using sequences of key poses, Pattern Recognit. Lett. 34 (2013) 1799–1807.
[50] D. Wu, L. Shao, Silhouette analysis-based action recognition via exploiting human poses, IEEE Trans. Circuits Syst. Video Technol. 23 (2) (2013) 236–243.
[51] R. Touati, M. Mignotte, MDS-based multi-axial dimensionality reduction model for human action recognition, in: Proc. of IEEE canadian Conference on Compter and Robot Vision, 2014.
[52] J.X. Cai, X. Tang, G.C. Feng, Learning pose dictionary for human action Recognition, in: International Conference on Pattern Recognition, 2014.
[53] S. Sadek, A.A. Hamadi, M. Elmezain, B. Michaelis, U. Sayed, Human action recognition via affine moment invariants, in: 21st International conference on Pattern Recognition, 2012.
[54] B. Saghafi, D. Rajan, Human action recognition using Pose-based disriminant embedding, Signal Process., Image Commun. 27 (2012) 96–111.
[55] R. Melfi, S. Kondra, A. Petrosino, Human activity modeling by spatio temporal textural appearance, Pattern Recognit. Lett. 34 (2013) 1990–1994.
[56] S. Rahman, I. Song, M.K.H. Leung, I. Lee, Fast action recognition using negative space features, Expert Syst. Appl. 41 (2014) 574–587.
[57] R. Benmokhtar, Robust human action recognition scheme based on high-level feature fusion, Multimedia Tools Appl. 69 (2) (2014) 253–275.
[58] Y. Wang, G. Mori, Human action recognition by semi-latent topic models, IEEE Trans. Pattern Anal. Mach. Intell. 31 (10) (2009) 1762–1764.
[59] T. Guha, R.K. Ward, Learning sparse representations for human action recognition, IEEE Trans. Pattern Anal. Mach. Intell. 34 (8) (2012) 1576–1588.
[60] X.L. Ming, H.J. Xia, T.L. Zheng, Human action recognition based on chaotic invariants, J. South Cent. Univ. 20 (2014) 3171–3179.
[61] A. Iosi?dis, A. Tefas, I. Pitas, Discriminant bag of words based representation for human action recognition, Pattern Recognit. Lett. 49 (2014) 185–192.
[62] D.K. Vishwakarma, R. Kapoor, Hybrid classifier based human activity recognition using the silhouette and cells, Expert Syst. Appl. 42 (20) (2015) 6957–6965.
[63] M. Holte, T. Moeslund, N. Nikolaidis, I. Pitas, 3D human action recognition for multi-view cameras systems, in: 3DIMPVT, 2011.
[64] A. Iosifidis, A. Tefas, N. Nikolaidis, I. Pitas, Multi-view movement recognition based on fuzzy distances and linear disriminant analysis, Comput. Vis. Image Underst. 116 (3) (2012) 347–360.
[65] X. Wu, D. Xu, L. Duan, J. Luo, Action recognition using context and appearance distribution features, in: IEEE Conference on Computer Vision and Pattern Recognition, CVPR, Providence, RI, 2011.
[66] W. Yang, Y. Gao, L. Cao, M. Yang, Y. Shi, mPadal: a joint local-and-global multi-view feature selection method for activity recognition, Appl. Intell. 41 (3) (2014) 776–790.
[67] Z. Zhang, C. Wang, B. Xiao, W. Zhou, S. Liu, Cross-view action recognition using contextual maximum margin clustering, IEEE Trans. Circuits Syst. Video Technol. 24 (10) (2014) 1663–1668.

Dinesh Kumar Vishwakarma received the Bachelor of Technology (B.Tech.) from Dr RML Avadh University, Faizabad, Uttar Pradesh, India, in 2002 and the Master of Technology (M.Tech.) 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 activity and hand gesture recognition. He is also reviewer of Elsevier, Springer, and IEEE/IET journals.

Rajiv Kapoor received his B.E., M.E. degree from Delhi University, India and Ph.D. from Punjab University, India in the field of Electronics & Communication. He is currently Professor in the Department of Electronics & Communication, Delhi Technological University, Delhi, India. He is Editor of ST Micro Electronics International Journal in the field of Electronics Design. He has authored more than 70 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.

Ashish Dhiman received the B.E. degree in Electronics and Communication Engineering from Panjab University, Chandigarh and completed Master's in Signal Processing & Digital Design from Delhi Technological University (Formerly Delhi College of Engineering) in 2011 and 2014, respectively. He is currently working as Junior Research Fellow in IIT Gandhi Nagar at Department of Electrical Engineering. His research interest includes human action/activity analysis from images and video, signal processing, computer vision, image processing, pattern recognition.

Please cite this article in press as: D.K. Vishwakarma, et al., A proposed unified framework for the recognition of human activity by exploiting the characteristics of action dynamics, Robotics and Autonomous Systems (2016), http://dx.doi.org/10.1016/j.robot.2015.11.013