International Journal of Advanced Information Science and Technology (IJAIST) ISSN: 2319:2682
Vol.5, No.8, August 2016 DOI:10.15693/ijaist/2016.v5i8.30-35

An Approach to Human Action Recognition using the Histogram of Oriented Gradients

Shamama Anwar
Asst. Professor/Dept. of Computer Science & Engg.
Birla Institute of Technology, Mesra, Ranchi

G. Rajamohan
Asso.Professor / Dept. of Manufacturing Engg.
National Inst. of Foundry & Forge Tech., Ranchi

***Abstract* - A human action recognition method using the locally normalized histogram of oriented gradients with rectangular cells (R-HOG) is proposed. The HOG descriptors of images are stored in vector form. The HOG descriptor of input image containing the action to be classified is compared with that of images taken from different action classes of standard datasets and common HOG ratios are computed. The image belongs to that particular action class for which the highest common HOG ratio is obtained. The Weizmann and KTH datasets have been used to validate the proposed approach. Experimental results show that the accuracy of the proposed approach is very good. The main advantage of the proposed approach being that it classifies input actions based on common HOG ratios only and it does not need any classifier or training.**

Index terms -Action recognition, feature descriptors, histogram of oriented gradients.

I. INTRODUCTION

Enabling the computers to automatically recognize human actions or movements into certain predefined action classes is the primary objective of human action recognition systems. It is a challenging task to recognize the human actions from real time video clips due to variations in the way similar actions can be performed by different human appearances and variety of poses they can adopt. A good human action recognition system should be robust enough to deal with such variations. Human-computer interaction [1-3], human-robot interaction [4], surveillance [5-7], sports monitoring [8-9], military monitoring [10], etc. are some of the applications of human action recognition systems.

II. RELATED WORK

A variety of human action recognition methods, such as context based decision making methods, view based methods, shape based methods, flow based methods, etc. are found in the literature. Context-based methods need prior knowledge about the object being recognized and its background. Ayers and Shah used context based decisions method and applied low level computer vision techniques for tracking, skin and scene change detection for action recognition [11]. A similar approach by Intille, et al. starts with background subtraction, followed by the detection of blobs. A greedy approach is used finally to recognize the actions of children [12]. View-based recognition methods learn the appearance of the objects under different poses and condition. Davis and Bobick used a motion history images (MHI) along with motion energy images (MEI) for temporal template matching [13]. Template matching methods are computationally less intensive; however, they are more sensitive to variance in the duration of the movement [14]. Freeman and Roth developed a vision based approach for gesture recognition using the histograms of local orientation. They used the orientation histogram as a feature vector for gesture classification and interpolation and also explored the use of spatiotemporal histograms for the same. The method was tested with 10 different hand gestures [15]. Some of the researchers have used shape based methods for recognizing the actions. The human silhouette has been widely used as feature in these methods. Hsieh, et al. used it along with the polar coordinate system to recognize human actions [16]. A variation to the silhouette technique that used the triangular mesh technique along with depth-first-search (DFS) scheme has also been proposed for feature extraction [17-18]. Flow-based action recognition methods use the optical flow as motion descriptors [19-22]. The action recognition systems first extract the features from images/videos and then use a classifier based on extracted features to recognize or classify the actions. Some of the feature extraction methods include a component based approach using wavelets to extract features [23], rectangle filter method to detect motion filters [24], orientation histogram [15] and SIFT key points [25].

In the present work, a human action recognition method that uses the locally normalized histogram of oriented gradient with rectangular cells (R-HOG) [19] for the extraction of features has been proposed. A brief discussion on HOG is given in Section 2. The proposed method has been validated using the Weizmann and the KTH datasets. The Weizmann dataset consists of 10 action classes and 90 videos [26] while the KTH dataset consists of 6 action classes and 600 videos (192 for training, 192 for validating and 216 for testing) [27].

III. PROPOSED HUMAN ACTION RECOGNITION METHOD

The proposed human action recognition method, shown in Figure 1, is based on *histogram of oriented gradients* (*HOG*). The HOG is a feature descriptor, used in the areas of computer vision and image processing for the detection of objects from the images. The HOGdescriptor counts the occurrences of gradient orientation in localized portions of an image using the intensity gradient or edge direction to describe the shape of objects.A brief description of HOG and that of the proposed approach are outlined here. The steps shown within the box of

 DOI:10.15693/ijaist/2016.v5i8.30-35

dashed lines are the steps involved in computing the HOG descriptor for a given image, which may be the input image or the one from dataset images. These steps are shown in algorithmic form in Figure 3.

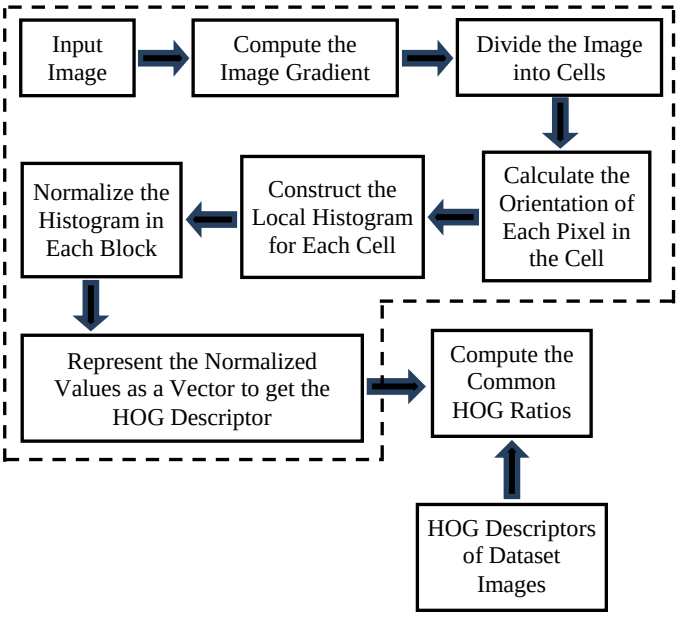


Figure 1. Proposed human action recognition method

The given image is first divided into blocks. The blocks are then sub-divided into a dense grid of rectangular (R-HOG) or radial (C-HOG)cells of 8×8 pixelseach. Each of these cells will contain a local histogram.For the pixels within each cell,a HOG is constructed and normalized by constructing larger overlapping blocks so as to achieve invariance to changes in illumination. A combination of these histograms forms the HOG descriptor that can be found by counting the number of occurrences of gradient orientations in the HOG cells. The gradient and magnitude of the given image are needed to find the gradient orientations. The gradient and magnitude of the given image *A* can be computed using (1).

$$|G| = \sqrt{A_x^2 + A_y^2}; \theta = tan^{-1}\left(\frac{A_y}{A_x}\right) \quad (1)$$

where, A_x and $A_y$respectively represent the *x* and *y* derivatives of the given image *A*, calculated using (2). The derivatives are obtained by sliding a convolution operator over the entire image. A sample image taken from the KTH dataset and its derivatives are shown in Figure 5.

$$A_x = A * D_x;\ A_y = A * D_y;$$
$$D_x = [-1\ 0\ 1];\ D_y = [1\ 0\ -1]^T \quad (2)$$

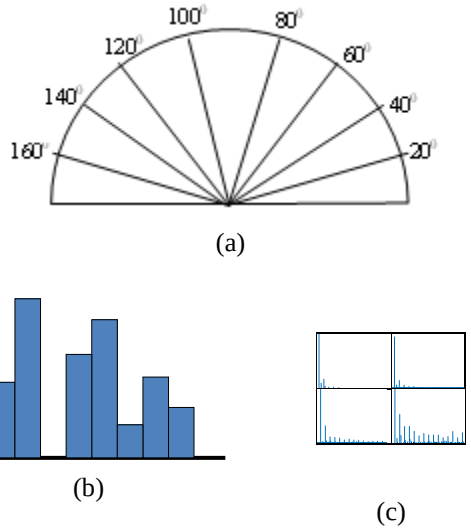
The histogram channels have been taken to be evenly distributed from 0° to 180° using 9 bins (Figure2(a)), as this configuration has been shown yield better results [19]. Each pixel contributes to the histogram according to its orientation by casting a *weighted vote*(V_w) into two neighboringbinsbased on its weight and center of the bin. The gradient magnitude, its square root or its squarecan be taken as the weightfor each pixel. In the present work, the gradient magnitude is taken as the weight. The center of the bin can be obtained as the mean of bin range. For example, if the orientation (θ) of a pixel is 75°, it would cast a weighted vote into the neighboring bins of 60°-80° and 80°-100° with bin centers at 70° and 90° respectively. The weighted vote can be calculated using (3).

$$V_w = w \times \left(\frac{B_{center} - \theta}{h}\right) \quad (3)$$

where, w is the weight, θ is the pixel's orientation,B_{center} and $h$are the centerand width of the bin respectively. The bin width h has been taken to be 20. The votes are accumulated over the pixels of each cell and a histogram is constructed for each cell (Figure 2(b)). The cells are now combined to form blocks of 2×2 size in such a way that 50% of the block's regions overlap. Figure 2(c) represents the local histogram constructed from a 2×2 size, random block.The histograms ofthe cells within the blocks are now normalized using L_2 norm as:

$$f = \frac{v}{\sqrt{||v||_2^2 + e^2}} \quad (4)$$

where,v represents the vector of n elements containing all the histograms in a given block, e is a small constant. The v value can be computed using (5).



(a)

(b) (c)

Figure 2. (a) The 9 bin representation of histogram channels for 0° to 180°; (b) Local histogram for a cell; (c) Local histogram of block consisting of 2×2 cells

$$||v||_2 = \sqrt{\sum_{i=1}^{n} |v_i|^2} \quad (5)$$

The obtained *f*values are the feature vectors representing the HOG values. The feature vector of the input image is compared with those of dataset images in order to find the common feature descriptors or common HOG ratios, which are used for recognizing the action contained in the given

 DOI:10.15693/ijaist/2016.v5i8.30-35

input image. The common HOG ratio will be *high for similar action classes* and *low for dissimilar action classes*.

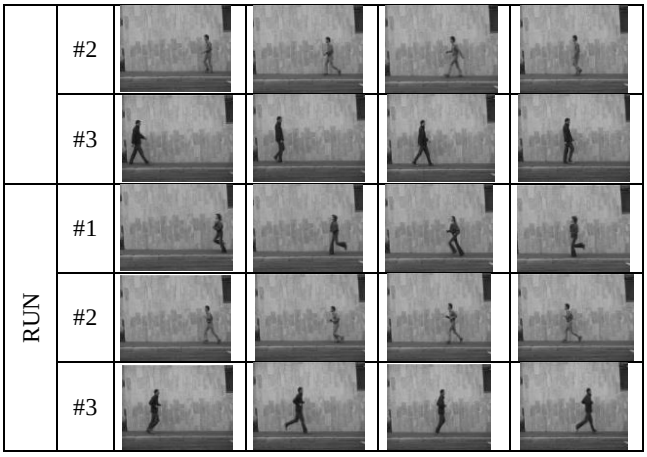
Input Given*ImageA*
Output *HOG descriptor of Given Image*
Steps

- Obtain *x* and *y* derivatives of the image and calculate the magnitude of the gradient using (1)
- Divide the image into cells containing 8×8 pixels each
- For *each cell*
  For *each pixel* in the *cell*
  Find the *magnitude and orientation of gradient* using (1)
  Calculate the *pixel's vote* for the *histogram bins* using (3)
  End
  Construct the *histogram*
  End
- Form the*blocks* by grouping 2×2 *cells* such that they overlap
- For *each block*
  Normalize the histogram in the *block* using (4)
  End
- Represent the values as *Feature Vector* representing *HOG Descriptor*

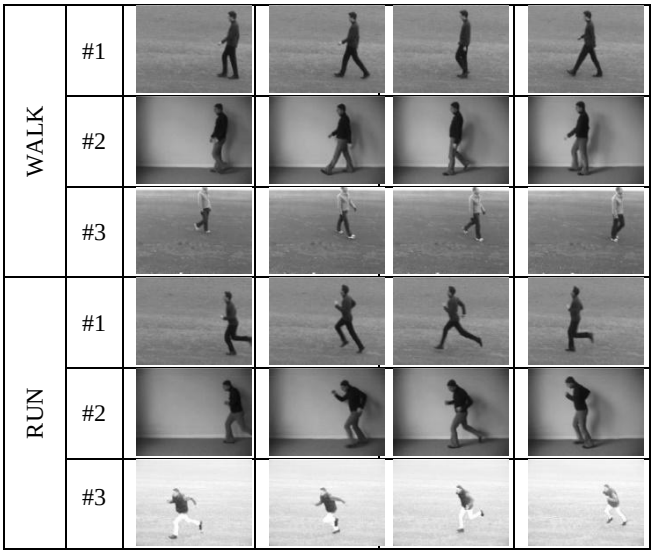
Figure 3. Algorithm for calculating the HOG Descriptor for a given image

IV. RESULTS AND DISCUSSION

The proposed to human action recognition method has been implemented in MATLAB language. The *classification accuracy* has been taken as the measure to evaluate the effectiveness of the proposed approach. To achieve this purpose, Weizmann and KTH datasets have been used. The Weizmann dataset consists of 10 action classes, viz. walk, run, bend, side gallop, jump, wave by one hand, wave by two hands, jump in place, jumping jack and skip. These actions have been performed by different agents. The agents are essentially different people performing the same action but in their own style. A total of 126 images have been taken (first 7 *action classes* performed by 3 *agents* and 6 *images per agent*) from this dataset. The KTH dataset consists of 6 action classes performed by different agents, viz. walk, run, jog, hand waving, hand clapping and boxing. A total of 108 images (3 *agents* and 6 *images per agentper action class*) have been taken from this dataset. Figures 4(a)and 4(b) show the partial list of images from Weizmann and KTH datasets respectively. The numbers in Column 2 indicate the Agent. Different agents performing same action adds some variability in the manner in which the action is performed and it may increase the efficiency of action recognition.



(a) Images extracted from Weizmann dataset videos



(b) Images extracted from KTH dataset videos

Figure 4. Partial list of images extracted from Standard Datasets

Since the dataset consists of videos, images have been extracted for computing the HOG descriptors. The algorithm in Fig. 3 has been used to compute the HOG descriptors for all images taken from the standard datasets corresponding to the action classes considered. For representing the entire class of actions using one feature representation, irrespective of the agent, the HOG descriptors are averaged. For the Weizmann dataset, HOG descriptors are averaged for all the 6 images for each agent thereby representing one feature value for that agent. Hence, each action class has 3 HOG descriptors, one for each agent. These values are again averaged to represent the HOG descriptor for the entire action class. The averaging of HOG descriptors makes the variability of the action being performed insignificant for the classification purpose. A similar approach is also used for the KTH dataset. The computed HOG descriptors of images from the dataset are stored in vector format.

 DOI:10.15693/ijaist/2016.v5i8.30-35

Some example images containing the actions to be classified, their *x* and *y* derivativesand magnitude of gradient are shown in Figure 5. The HOG descriptors for these images have also been calculated using the algorithm in Figure 3. As given in the algorithm, the derivatives are needed to compute the magnitude of the gradient.As per the algorithm, imagesare then divided into a dense grid of cells each containing 8×8 pixels, followed by the construction of the histograms for each cell. The cells are now combined to form blocks of 2×2 cells, followed by making half of theregions of two consecutive blocks overlap. These intermediate processing steps in the computation of HOG descriptor for the example image of RUN action in Figure 5 are illustrated in Figure 6. The overlapping blocks are normalized using (4). The output of normalization is a vector representing the HOG descriptors.A pictorial representation of the HOG descriptorsfor the RUN image is shown in Figure 7.A sample HOG descriptor for this image in Figure 5 is as given below:

[0.34480.21140.3448 0.3448 0.0 0.0 ……. 0.3218 0.3448]

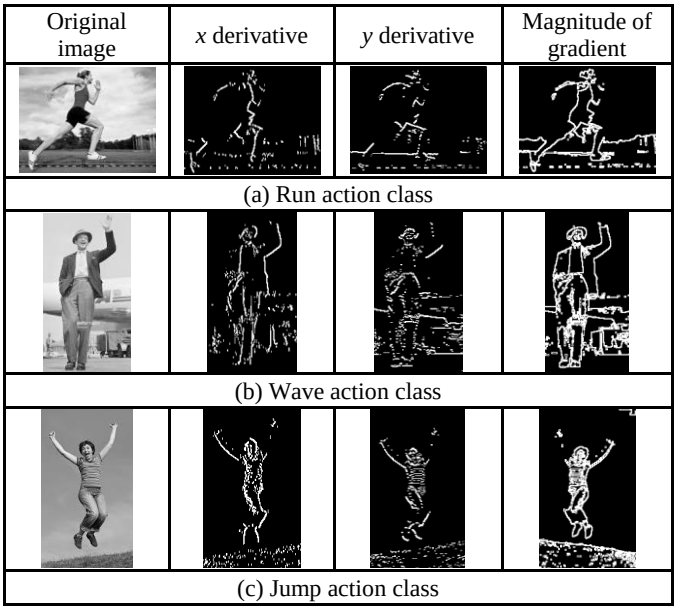
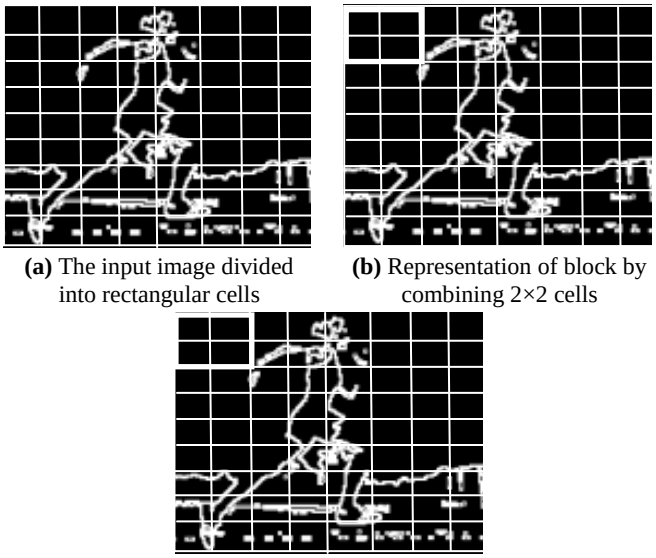


Figure 5.Some example images used for validation

The next step is to calculate the *Common HOG Ratio* for all dataset imagesby pairing each image with the input image. TheHOG descriptor of input image is then compared with that of the imagefrom dataset and numbers of similar HOG descriptors are counted. The common HOG ratio is computed as the ratio of common HOG descriptors and the length of HOG descriptor. The images are similar with larger common HOG values and vice versa.



(a) The input image divided into rectangular cells **(b)** Representation of block by combining 2×2 cells

(c) Representation of the next overlapping block

Figure 6. Processing steps for first example image

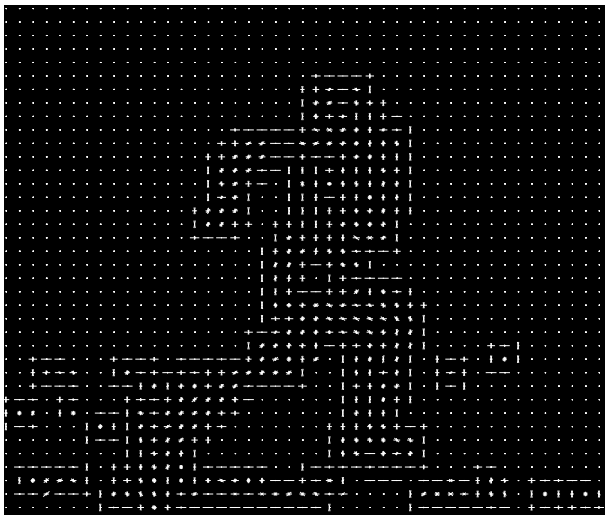


Figure 7. Pictorial representation of HOG descriptor for the Run imagein Figure 5

The proposed approach has been tested on 40 different input images takenfrom both Weizmann and KTH datasets. The common HOG ratios for an input image of *run action class* taken from the Weizmann dataset has been shown in Table 1. It may be observed from this table that the common HOG ratio is the highest for run action class. Similar results shown in Table 2 for the KTH dataset also has the highest common HOG ratio for run action class. Therefore, it may be concluded that the action contained in the input image has been correctly classified in both the datasets. The classification accuracy is represented using the confusion matrix (Table 3 for Weizmann dataset and Table 4 for KTH dataset), created by testing of 40 different input images per action class considered. The entries in the confusion matrix are

basically the ratio of number of correct classifications to the total number classifications attempted.

Table 1. The common HOG ratios for an input image of *run class* and the Weizmann dataset

Action Class	Common HOG ratio
Walk	0.765
Run	0.802
Bend	0.526
Jump	0.475
Side Gallop	0.558
Wave 1	0.693
Wave 2	0.709

Table 2.The common HOG ratios for an input image of *run class* and the KTH dataset

Action Class	Common HOG ratio
Walk	0.795
Run	0.817
Jog	0.812
Hand Waving	0.637
Hand Clapping	0.591
Boxing	0.724

Table 3. Confusion matrix for Weizmann dataset

	Walk	Run	Bend	Jump	Side gallop	Wave1	Wave2
Walk	32	8	0	0	0	0	0
Run	10	30	0	0	0	0	0
Bend	0	0	38	2	0	0	0
Jump	0	0	0	40	0	0	0
Side gallop	0	0	0	0	40	0	0
Wave1	0	0	0	0	0	36	4
Wave2	0	0	0	0	0	3	37

Table 4. Confusion matrix for KTH dataset

	Walk	Run	Jog	Hand waving	Hand clapping	Boxing
Walk	32	3	5	0	0	0
Run	2	26	12	0	0	0
Jog	3	10	27	0	0	0
Hand Waving	0	0	0	38	2	0
Hand Clapping	0	0	0	0	40	0
Boxing	0	0	0	0	0	40

It may be seen from these tables that the classification accuracy is about 90% for the Weizmann dataset and about 84.5% for the KTH dataset, which are quite good given the fact that no classifiers are used.

V. CONCLUSION

In the present work, histogram of oriented gradients (HOG) has been used as a feature descriptor for human action recognition purposes. The existing action recognition methodsuse a classifier like the SVM or Neural Network for training, but such training is not required for the proposed approach as only a common HOG ratio measure has been used for the classification. The performance of the proposed measure has been verified using images taken from different action classes belonging to Weizmann and KTH datasets. The accuracy obtained using the Weizmann dataset is about 90%, while that for KTH is about 84.5%. As the proposed method does not require training,it will be computationally faster than the existing methods.

REFERENCES

[1]. J. Alon, V. Athitsos, Q. Yuan and S. Sclaroff, "A unified framework for gesture recognition and spatiotemporal gesture segmentation", IEEE Trans. on Pattern Analysis &Machine Intelligence, Vol. 31, No. 9, pp. 1685-1699, 2009.

[2]. D. Kim, J. Song and D. Kim, "Simultaneous gesture segmentation and recognition based on forward spotting accumulative HMMs", Pattern Recognition, Vol. 40, No. 11, pp. 3012-3026, 2007.

[3]. T. Kirishima, K. Sato and K. Chihara, "Real - time gesture recognition by learning and selective control of visual interest points", IEEE Trans. on Pattern Analysis and Machine Intelligence, Vol. 27, No. 3, pp. 351-364, 2005.

[4]. C. D. Kidd and C. Breazeal, "Robots at home: Understanding long-term human-robot interaction," Proc. of IEEE/RSJ Int. Conference on Intelligent Robots and Systems, 2008.

[5]. L. Havasi, Z. Szlavik and T. Sziranyi, "Detection of gait characteristics for scene registration in video surveillance system", IEEE Trans. on Image Processing, Vol. 16, No. 2, pp. 503-510, 2007.

[6]. K. Huang, L. Wang, T. Tan and S. Maybank, "A real-time object detecting and tracking system for outdoor night surveillance", Pattern Recognition, Vol. 41, No. 1, pp. 432-444, 2008.

[7]. M. T. Lopez, A. F. Caballero, M. A. Fernandez, J. Mira and A. E. Delgado, "Visual surveillance by dynamic visual attention method", Pattern Recognition, Vol. 39, No. 11, pp. 2194-2211, 2006.

[8]. M. H. Hung and C. H. Hsieh, "Event detection of broadcast baseball videos", IEEE Trans. on Circuits and Systems for Video Technology, Vol. 18, No. 12, pp. 1713-1726, 2008.

[9]. E. Ramasso, M. Rombaut and D. Pellerin, "State filtering and change detection using TBM conflict application to human action recognition in athletics videos", IEEE Trans. on Circuits and Systems for Video Technology, Vol. 17, No. 7, pp. 944-949, 2007.

[10]. T. McKenna, "Video surveillance and human activity recognition for anti-terrorism and force protection", Proc. of IEEE Conference on Advanced Video and Signal based Surveillance, 2003.

[11]. D. Ayers and M. Shah, "Recognizing human actions in a static room", Proc. of 4thIEEE Workshop on Applications of Computer Vision, 1998.

[12]. S. S. Intille, J. W. Davis and A. F. Bobick, "Real-time

closed-world tracking", Proc. of IEEE Conference on Computer Vision and Pattern Recognition, 1996.
[13]. A. F. Bobick and J. W. Davis, "The recognition of human movement using temporal templates", IEEE Trans. on Pattern Analysis and Machine Intelligence, Vol. 23, No. 3, pp. 257-267, 2001.
[14]. D. Das, "Activity recognition using histogram of oriented gradient pattern history" , Int. Journal of Computer Science, Engineering and Information Technology, Vol. 4, No. 4, pp. 23-31, 2014.
[15]. W. T. Freeman and M. Roth, "Orientation histograms for hand gesture recognition", Proc. of Int. Conference on Automatic Face and Gesture Recognition, 1995.
[16]. C. H. Hsieh, P. S. Huang and M. D. Tang, "Human action recognition using silhouette histogram", Proc. of 34thAustralasian Computer Science Conference, 2011.
[17]. C. H. Chuang, J. W. Hsieh, L. W. Tsai and K. C. Fan, "Human action recognition using star templates and Delaunay triangulation", Proc. of Int. Conference on Intelligent Information Hiding and Multimedia Signal Processing, 2008.
[18]. J. W. Hseih, Y. T. Hsu, H. Y. M. Liao and C. C. Chen, "Video-based human movement analysis and its application to surveillance", IEEE Trans. on Multimedia, Vol. 10, No. 3, pp. 372-384, 2008.
[19]. N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection", Proc. of Int. Conference on Computer Vision and Pattern Recognition, 2005.
[20]. A. A. Efros, A. C. Berg, G. Mori and J. Malik, "Recognizing action at a distance", Proc. of 9thInt. Conference on Computer Vision, 2003.
[21]. C. Liu, Y. Yang and Y. Chen, "Human action recognition using sparse representation", Proc. of IEEE Int. Conference on Intelligent Computing and Intelligent Systems, 2009.
[22]. J. Liu, S. Ali and M. Shah, "Recognizing human actions using multiple features", Proc.of IEEE Conference on Computer Vision and Pattern Recognition, 2008.
[23]. A. Mohan, C. Papageorgiou and T. Poggio, "Example-based object detection in images by components", IEEE Trans. on Pattern Analysis and Machine Intelligence, Vol. 23, No. 4, pp. 349-361, 2001.
[24]. P. Viola, M. J. Jones and D. Snow, "Detecting pedestrians using patterns of motion and appearance", Proc. of 9thInt. Conference on Computer Vision, 2003.
[25]. D. G. Lowe, "Distinctive image features from scale-invariant key points", Int. Journal of Computer Vision, Vol. 60, No. 2, pp. 91-110, 2004.
[26]. L. Gorelick, M. Blank, E. Shechtam, M. Irani and R. Basri, "Actions as space-time shapes", Proc. of 10thIEEE Int. Conference on Computer Vision (ICCV), 2005.
[27]. C. Schuldt, I. Laptev and B. Caputo, "Recognizing human actions: a local SVM approach", Proc.of 17thInt. Conference on Pattern Recognition (ICPR), 2004.

Authors Profile

Shamama Anwar received her **M. Sc.** (Information Technology) and **M. Tech.** (Computer Science) degrees from Birla Institute of Technology, Mesra, Ranchi.At present, sheis working as an Assistant Professor inthe Department of Computer Science & Engineering at Birla Institute of Technology, Mesra, Ranchi. Her research interests include image processing, artificial intelligence and soft computing applications.

G. Rajamohanreceived his **B. E.** (Mechanical Engineering) degree from the College of Engineering, Chennai, **M. E.** (Manufacturing Technology) degree from Regional Engineering College, Trichy and **Ph. D.** (Mechanical Engineering) degree from the Indian Institute of Technology Madras. He is now working as Associate Professor in the Dept. of Manufacturing Engineering at the National Institute of Foundry and Forge Technology, Ranchi. His research interests include metrology and computer aided inspection, machining, image processing and applications of computers in design and manufacturing.