J Multimodal User Interfaces (2016) 10:335–344
DOI 10.1007/s12193-016-0229-4

ORIGINAL PAPER

Action recognition based on binary patterns of action-history and histogram of oriented gradient

Md. Atiqur Rahman Ahad[1] · Md. Nazmul Islam[2] · Israt Jahan[2]

Received: 23 October 2015 / Accepted: 20 August 2016 / Published online: 7 September 2016

Abstract In this paper, we have focused on the view-based spatio-temporal template matching approach for human action detection and classification. We have proposed an approach for human activity modeling that describes human motions as a texture pattern. We have combined two relatively simple feature extractors for obtaining a system to get more accurate result. In this method, video sequences are converted into temporal templates called Motion History Image (MHI), which can preserve dominant motion information. The local features are described with Local Binary Pattern (LBP) and Histogram of Oriented Gradients (HOG) descriptors. LBP operator is texture operator that encodes the direction of motion from the non-monotonous areas of MHI images. HOG is used as feature descriptor and extracts the features from LBP. These descriptors are used to train with Support Vector Machine (SVM) classifier to recognize various action classes. This proposed method has been tested on the KTH Action Dataset (which is one of the most widely used benchmark datasets for human action classification), and on the Pedestrian Action Dataset. Our method has shown 86.67 % recognition rate in the 6-classes of KTH Action Dataset and 94.3 % accuracy in the 7-classes of Pedestrian Action Dataset. Based on the complexity of datasets, both the results are quite satisfactory.

Keywords Action recognition · MHI · HOG · LBP · SVM · Pedestrian recognition

✉ Md. Atiqur Rahman Ahad
atiqahad@yahoo.com

[1] University of Dhaka, Dhaka, Bangladesh

[2] Department of Electrical and Electronic Engineering, University of Dhaka, Dhaka 1000, Bangladesh

1 Introduction

The detection and recognition of human activities from video data has been widely researched in recent decades. The potential application of human activity analysis includes surveillance, robotics, rehabilitation, video indexing, biomedicine and biomechanics arena, sports analysis and so on [1–3]. A human motion detection algorithm works by converting a video with some sequences into a bunch of information. But developing algorithm for action recognition from a video should be not only accurate but also efficient in terms of computation. Recognizing human action is full of challenges because of the complexity of the scene where multiple interacting objects are moving, camera viewpoint, and different types of action schemes. While developing an algorithm for action recognition, the algorithm should be able to detect and recognize various activities. There are a number of benchmark datasets that are developed specifically to evaluate the systems in various experiments. The KTH (Kungliga Tekniska Högskolan) Human Action Dataset [4] is very widely compared dataset. Recently, a new dataset has been developed called Pedestrian Action Dataset [5], which we used along with KTH Action Dataset to evaluate our method.

This paper is organized as follows: Sect. 2 briefly describes the related works. In Sect. 3, we present our proposed method. Section 4 presents the experimental result and analysis and the data sets we have used in our experiment. Finally, we conclude the research with few future work guidelines in Sect. 5.

2 Related works

A great number of methods have been proposed for human action detection and recognition from videos. There are extensive works on human action detection and classification [1–13]. Different methodologies are used to detect human action such as background subtraction [14,15], frame subtraction [16,17] and optical flow [18,19], etc. Many algorithms have been proposed for feature extraction. For a practical pedestrian detection system, Gavrila [20] employ hierarchical template matching to find pedestrian candidates from incoming images. Mikolajczyk and Schmidt [21] show that the best matching results are obtained by the Scale Invariant Feature Transform (SIFT) descriptor. Dalal and Triggs [22] propose a human detection algorithm using histograms of oriented gradients (HOG), which are similar with the features used in the SIFT descriptor.

Another approach, called template matching becomes popular. For example, Bobick and Davis [23] developed Motion Energy Image (MEI) and Motion History Image (MHI) templates. These templates are matched for recognition by seven Hu moments. But there arose some problem with MEI and MHI such as occlusion and failure in detecting the subject in the video. A lot of works have been done to overcome the limitations of MHI and MEI. For example, Ahad et al. [1] propose Directional Motion History Image (DMHI) to solve overwriting problem of MHI. Menga et al. [24] propose Hierarchical Motion History Histogram (HMHH) that improves MHI to a great extend. Gait Energy Image (GEI) [25], Action Energy Image (AEI) [26], 3D MHI [42], and Multi-resolution Motion Energy Histogram (MRMEH) [27] are some modified version of MHI [28].

Recently, a simple feature descriptor Local Binary Pattern (LBP) has gained popularity due to its simplicity in calculation. The basic LBP is developed by Ojala et al. [29,30]. LBP has been used for various applications since then for texture analysis, motion detection, face and gender recognition, facial expression analysis and so on [29–36]. In mid of the 2000s, LBP has started using for motion analysis with a texture-based method [33,34] by removal of background object. With the development of few variants of LBP, LBP becomes more popular. It has been used for motion and activity analysis such as facial expression recognition using facial dynamics [36], face and gender recognition [35] and human action recognition [32]. In recent years, Histogram Oriented Gradient (HOG) has widely being used for human action detection and pedestrian detection [22,37–39]. For classification, we can implement Support Vector Machine (SVM), K-nearest neighbor classifier, etc. [1]. Support Vector Machine [4,38,40,41] has proved to be very efficient classifier and widely used in different methods.

3 Proposed method

The objective of this paper is to develop an action recognition strategy that will give good recognition rate. This strategy should have been worked in some complex actions. In this paper, we propose a template matching approach, exploiting Motion History Image (MHI), Local Binary Pattern (LBP) and Histogram of Oriented Gradient (HOG) approaches. Based on your knowledge, this combination has not been published. Our motive is to develop a method for automated detection of action performed by human subjects which can be used for intelligent surveillance. In order to detect the presence and direction of motion, we, in this paper, developed here a template matching procedure. The templates used for action representation are spatio-temporal templates called is Motion History Image (MHI). For accurately recognizing the slightly varying actions, Local Binary Pattern (LBP) and Histogram of Oriented Gradient (HOG) have been exploited for further feature extraction. Finally, Support Vector Machine (SVM) has been used for training and testing as classifier. The whole process is illustrated in the block diagram of Fig. 1.

3.1 Motion history image

As we have to analyze the motion occurring in a video within a given range of time, we need a method that allows capturing and representing motion information directly from video. Such static motion representation is Motion History Image (MHI) [23]. These methods work with the properties of observed motion at respective spatial frame location in the video sequences. MHI is actually a temporal template matching approach. Template matching approaches are the summation of immediate past successive images and the

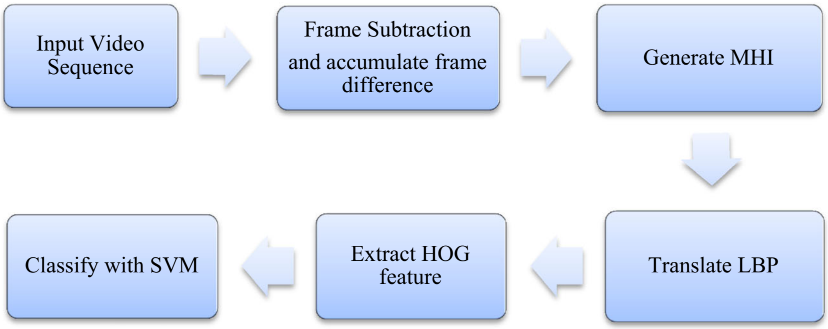

Fig. 1 Block diagram of action recognition process

weight intensity decays as time elapses. MHI is a cumulative gray scale images forms by spatial temporal motion information. MHI expresses the motion flow of a video sequence in a temporal manner. Details on MHI can be found in the book [2].

Bobick and Davis [23] proposed the MHI for recognition and representation. In MHI, the intensity of pixel of gray scale image is a function of motion density [1,2]. Usually, MHI is constructed from a binarized image, obtained from frame to frame subtraction, or background subtraction. From MHI image, we can see that older or past movement in a video sequence is relatively (step-wise) darker than any newer movement or moving information. Now, let's consider that two neighboring or consecutive images can be described by [2],

$$I\left(x,y,z\right)=b_t\left(x,y\right)+m_t\left(x,y\right)+n_t\left(x,y\right) \tag{3.1}$$

$$I\left(x,y,t+1\right)=b_{t+1}\left(x,y\right)+m_{t+1}\left(x,y\right)+n_{t+1}\left(x,y\right) \tag{3.2}$$

where, $b_{t+1}(x,y)$: static background t_{th} frame; $m_{t+1}(x,y)$: moving objects t_{th} frame; and $n_{t+1}(x,y)$; background noise t_{th} frame. Now, if we consider consecutive frame differencing approach for extracting moving objects, we can achieve difference as,

$$diff\left(x,y,t\right)=I\left(x,y,t+1\right)-I\left(x,y,t\right) \tag{3.3}$$

$$diff\left(x,y,t\right)=b\left(x,y\right)+md\left(x,y\right)+nd\left(x,y\right) \tag{3.4}$$

where, $b\left(x,y\right)$ is overlapped area in consecutive frames; $md\left(x,y\right)$ is the motion region and $nd\left(x,y\right)$ is the presence of noise. It is evident that the $diff\left(x,y,t\right)$ contains part of moving object, background aberration due to motion and noise that lead to incorrect results. Note that any motion at low speed may not be easily detected. To mitigate these constraints, we convert this gray-scale image $diff\left(x,y,t\right)$ into binary image, $diff\left(x,y,t\right)'$ [2]. Therefore, we define the $D\left(x,y,t\right)$ as,

$$D\left(x,y,t\right)=\left[diff\left(x,y,t\right)\times\tau\right]/255. \tag{3.5}$$

Then, by layering the successive $D\left(x,y,t\right)$, the Motion History Image can be produced. Therefore, an update function can be produced using difference of frames (DOF),

$$\Psi\left(x,y,t\right)\quad=\begin{cases}1 & if\,D\left(x,y,t\right)>\xi\\ 0 & otherwise\end{cases} \tag{3.6}$$

Here, $\Psi\left(x,y,t\right)$ is the binarization of the difference of frames by considering a threshold value ξ.The parameter ξ is the minimal intensity difference between two images for change detection [2]. The DOF of t_{th} frame with difference distance Δ is,

$$D\left(x,y,t\right)=\left|I\left(x,y,t\right)-I\left(x,y,t\pm\Delta\right)\right|. \tag{3.7}$$

So, an MHI template ($H\left(x,y,t\right)$) from the above update function $\Psi(x,y,t)$ can be computed in a recursive manner as follows,

$$H_\tau(x,y,t)=\begin{cases}\tau & if\,\Psi=1\\ \max\left(0,H_\tau\left(x,y,t-1\right)-\delta\right) & otherwise\end{cases} \tag{3.8}$$

where, (x,y) is pixel position; t is time that defines consecutive images in the video; τ decides the temporal duration of the MHI. Here, the value of the decay parameter (δ) can be 1 or more. In this manner, we can have a scalar-valued image called MHI image or MHI template, where more recently moving pixels are brighter and vice-versa.

MHI templates preserve the dominant motion of information and the direction of the motion. To cover the motion information of the total action spanning the total range of time the variable values are selected as $\tau=255$, $\delta=2$ and the threshold value is set at 30. These selected values are chosen empirically. The main problem with MHI is occlusion. To solve the problem, two-layers of MHI template is created. The first layer retains the information of all the frames and the second layer covers information of partial frames. Figures 2 and 3 show some sample frames of Pedestrian Action Dataset [5] and KTH Action Dataset [4] and their corresponding MHI templates.

3.2 Local binary pattern (LBP)

Local binary pattern (LBP) texture operator converts the image into an array or an image of integer labels that describe small level changes in the image [29,30]. LBP is simple but effective as it can reconcile between traditional divergent statistical and structural model of texture analysis. LBP gained popularity due to its robustness to gray-scale changes caused by noises, such as illumination variance, and computational simplicity. The basic idea of LBP operator [29,30] was based on dividing 2D surface texture into two aspects: local spatial pattern and grayscale contrast.

Let, consider a gray scale image (x,y) and g_c denote the gray level of an arbitrary pixel (x,y), i.e., $g_c=I(x,y)$. Moreover, let g_c denote the gray value of a sampling point in an evenly spaced circular neighborhood of P sampling points and radius R around point (x,y). Consider that (z) is a thresholding step function,

$$s\left(z\right)=\begin{cases}0, z<0\\ 1, z\geq 0\end{cases} \tag{3.9}$$

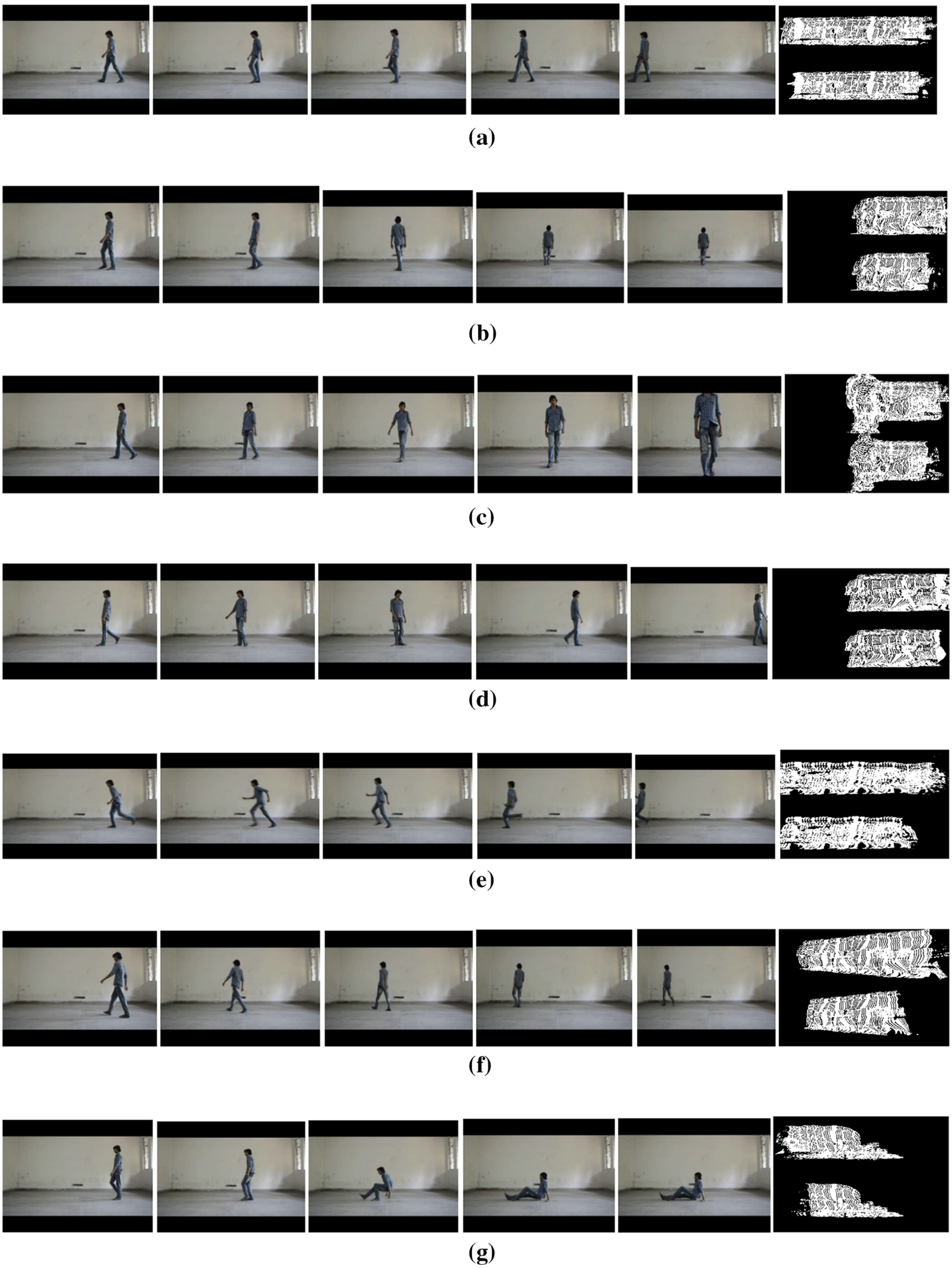

Fig. 2 Sample action sequence and corresponding MHI of the Pedestrian Action Dataset: **a** walk, **b** turn opposite to camera, **c** turn towards camera, **d** halfway return, **e** run, **f** cross-walk, and **g** fall down. Here two levels of MHI images are shown. The *upper* MHI covers the entire video sequence, whereas, the *lower part* of the MHI image ignores some initial frames and final frames

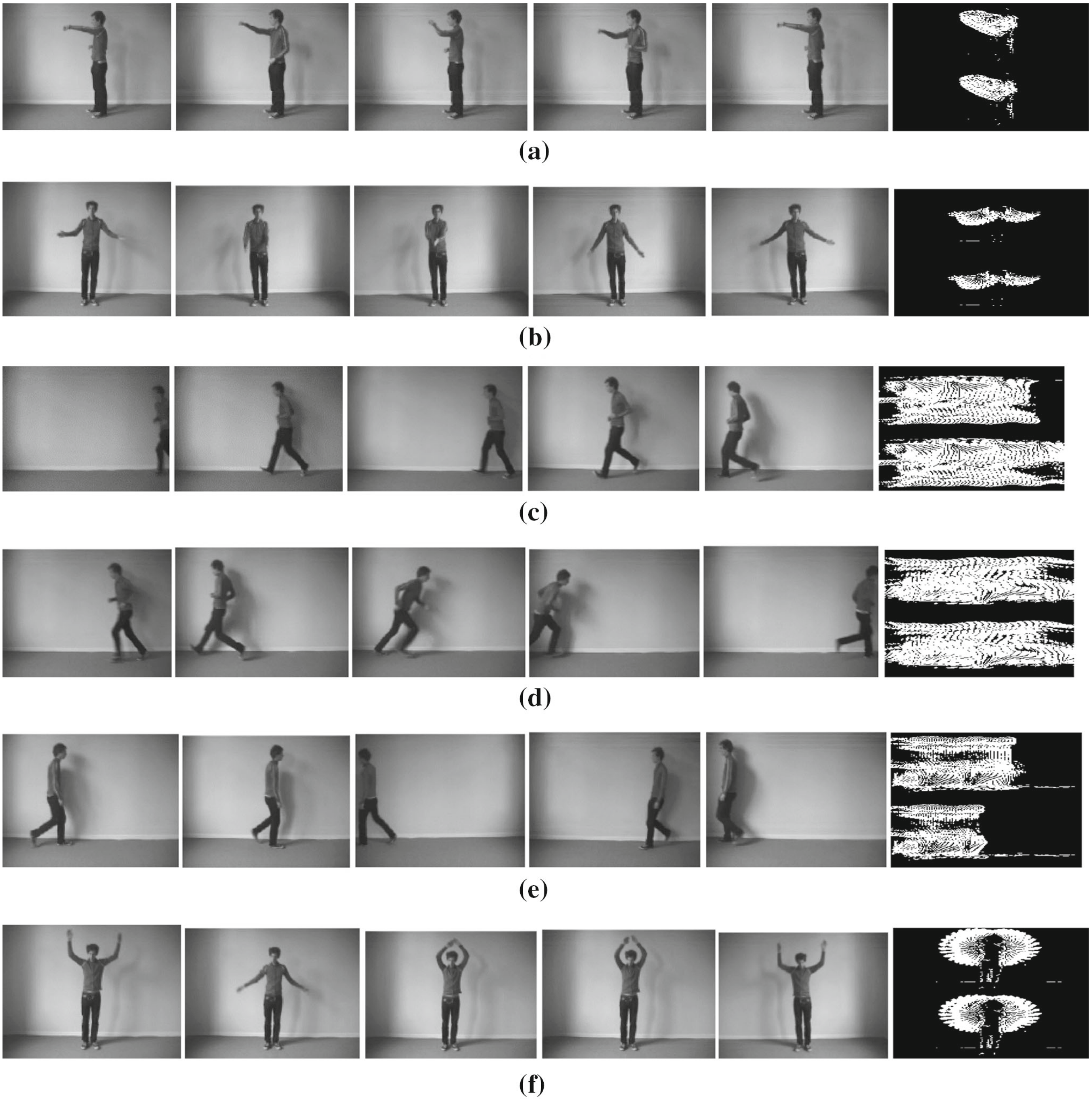

Fig. 3 Sample action sequence and corresponding MHI of the KTH Action Dataset: **a** boxing, **b** handclapping, **c** jogging, **d** running, **e** walking, and **f** hand-waving. Here two levels of MHI images are shown. The *upper* MHI covers the entire video sequence, whereas, the *lower part* of the MHI image ignores some initial frames and final frames

The LBP P, R operator is defined as:

$$\mathrm{LBP}_{\mathrm{P,R}}\left(x_c - y_c\right) = \sum\nolimits_{p=0}^{p-1} s\left(g_p - g_c\right) 2^p \tag{3.10}$$

Hence, the signs of the differences in a neighborhood are interpreted as a P-bit binary number, resulting in 2^p distinct values for the LBP code. Based on this concept, lets apply local binary pattern in every 3 × 3 pixel blocks of an image. Each bock is thresholded by its central pixel value. An LBP code is produced bymultiplying the thresholded values by the power of two and summed up to get the result. As the neighborhood consists of 8 pixels, 28 = 256 different labels can be used to describe the texture. This method give great performance in case of unsupervised texture segmentation. Though a number of variants of LBP are developed, the core

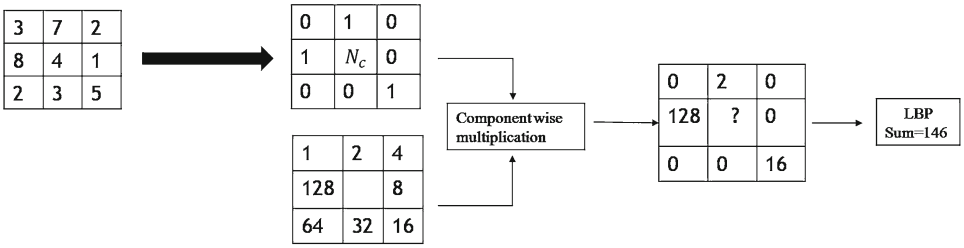

Fig. 4 Example of an LBP computation

idea remains the same, i.e., neighborhood pixels are binarized by thresholding. Figure 4 demonstrates an example of LBP computation for a 3 × 3 cell. Hence, the central value(4) will be replaced by the computed value (146), and this process will continue for an entire image, for every 3 × 3 cell, as raster scan manner.

LBP operator was initially designed for grayscale still images. But with the development of area of texture analysis, LBP is now used widely in the modern field of computer vision for detection of face and facial expression, object recognition, background subtraction, visual speech recognition and human action detection.

At present the use of spatiotemporal space has become popular for motion analysis. As mentioned above,

Motion History Image (MHI) is used for spatiotemporal analysis [13]. However, MHI overlaps the actions occupy in the same space and the new observation removes the previous one (which can be called as overlapping or motion overwriting). To solve these, LBP operators are used to characterize MHI which gave a new texture based descriptor for human movement. We only consider MHI template, as MEI image is just a binary image to demonstrate the presence of motion in the scene. LBP codes encode the direction of motion from MHI. LBP is restricted only in the non-monotonous area of MHI [27]. Nonmoving pixels are omitted in calculation of LBP code. LBP histogram only gives information of spatial structure, not the overall structure of motion. To preserve the structure of motion MHI is divided into sub regions. The choice of the division is not restricted. By increasing the number of divisions, the resolution of descriptor can be increased and more activities can be specified.

3.3 Histogram of oriented gradients

In this paper, Histogram of Oriented Gradient (HOG) is used as feature descriptor based on the LBP codes. The technique first localizes some portions of an image and then counts the occurrences of gradient orientation. This descriptor is computed on the uniformly local histogram of image gradient orientation cells in a dense grid and for high accuracy overlapping local contrast normalization is used [22]. The fundamental idea is that local object appearance and with the distribution of local intensity gradients or edge detections the shape can often be distinguished rather well [39]. HOG is implemented by dividing the image window into small spatial regions (cells), for each cell accumulating a local 1-D histogram of gradient directions or edge orientations over the pixels of the cell. The combined histogram entries form the representation [39].

In image pre-processing, the first step of calculation is evaluating several input pixel representations including grayscale, RGB and LAB color spaces with power law (gamma) equalization. These normalizations have very small effect on performance [39]. The next step is to compute the gradient values. However, the detector performance is sensitive to the way in which gradients are computed.

Gradients are computed by using Gaussian smoothing followed by one of several 1-D points discrete derivative masks (e.g., un-centered [−1; 1], centered [−1; 0; 1], cubi-corrected, Sobel masks of 3 × 3, diagonal masks). Note that simple 1-D [−1; 0; 1] masks work the best.

The next step is the fundamental nonlinearity of the descriptor. Each pixel calculates a weighted vote for an edge orientation histogram channel based on the orientation of the gradient element centered on it. The votes are then accumulated into spatial or orientation bins over local spatial regions called cells. A cell can be either rectangular or radial (log-polar sector). The orientation bins are evenly spaced over 0° to 180° (‘unsigned’ gradient) [22]. A vote is a function of the gradient magnitude at the pixel. It can be either the magnitude itself, or its square, or its square root, or a clipped form of the magnitude representing soft presence or absence of an edge at the pixel. Practically, using the magnitude itself gives the best result [22].

4 Results and analysis

For classification, we exploited the Support Vector Machine (SVM) classifier as a baseline classifier throughout the study, which is a state-of-the-art large margin classifier [4,42]. SVM is a set of supervised learning method for classification and outlier detection. The concept of Support Vector Machine is a decision plan that is created by decision boundaries. The classification is done by constructing hyperplanes in a multi-dimensional space to separate cases of different class labels.

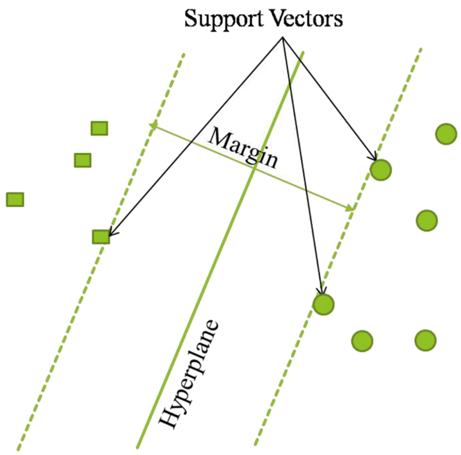

Fig. 5 SVM with two classes. Samples on the margin are called the support vectors

Assume that SVM is data of one class to the other. The best hyperplane is the one with the largest 'margin' between two used for two classes of data. SVM classifier classifies the data by best hyperplane to separate classes, whereas margin means the slab parallel to the hyperplane that has no interior data points. The 'support vectors' are the data points on the boundary of slab that are closest to the separating hyperplane. Figure 5 demonstrates margin, hyperplane, and support vectors to understand the SVM. It has 2 classes: square denotes one class and circle stands for another class.

This method becomes very popular among visual pattern recognition [43,44] due to its effectiveness in high dimension spaces even when the dimension is higher than the number of samples. As they use subset of training points in the decision function, it is memory effective. For this work, a two-layer MHI template for each action sequence has been extracted. The first layer contains the information of all the frames in the video and the second layer contains information from partial frames. After extracting Motion History Image (MHI) feature templates, these have been converted into Local Binary Pattern (LBP) codes. Before feeding the codes to the classifier Histogram of Oriented Gradient (HOG) features have been extracted. Leave-one-out cross-validation (LOOCV) scheme is considered to split the dataset for analysis. If the algorithm fails to classify the action, or if there is any input which it fails to recognize properly, the recognizer algorithm leaves it as unrecognized in some cases (instead of showing an false answer).

In our experiment, we have used two datasets. The first one is the Pedestrian Action Dataset [5], which is based on the range of possible pedestrian actions that might take place while crossing roads. It is an indoor dataset. In this dataset, seven actions [namely, walk, turn opposite to camera (TrnOp), turn towards camera (TrnTo), half-way return (HafRt), Run, crosswalk (CrsWk) and fall down (FallDw)] are performed by 20 different people in indoor environment with a static background. Subjects are taken from both sexes and different heights. At first each video sequence is converted into a double layer MHI template. Then the templates are converted into LBP codes. Before feeding to the SVM classifier HOG features are extracted from LBP codes. For classification the features are classified into seven different actions. 19 features are used to train each class of action. LOOCV scheme is used for classification. If the input doesn't match with any of the training images, it is considered as unrecognized. A confusion matrix is shown in Table 1 for the proposed method.

The recognition rate that has been obtained from our MHI-LBP-HOG-SVM method is 94.3 %. In 5.72 % cases, this approach has failed to identify any class, while false recognition rate has been obtained from the experiment is 1.42 %. Due to the close similarity between the MHI templates of actions 'Walk' and 'Run', the action 'Walk' achieved lowest recognition rate due to more confusion between these

Table 2 Recognition results with the pedestrian action Dataset

Method	Recognition rate (%)
MHI-HOG-SVM	91.42
MHI-HU-SVM	80
MHI-HOG-KNN	78.66
Proposed method	94.28

Table 1 Confusion matrix of the pedestrian action dataset

	Walk (%)	TrnOp (%)	TrnTo (%)	HafRt (%)	Run (%)	CrsWk (%)	FallDw (%)	Unrecognized (%)
Walk	80							20
TrnOp		80						20
TrnTo			100					
HafRt				100				
Run					100			
CrsWk						100		
FallDw							100	

two. We have also some other approaches such as MHI-Hu-SVM (where feature extractions are done by using seven Hu moment, similar to [23]) and MHI-HOG-KNN (where K-nearest neighbor classifier is used to classify actions from HOG features). But in both approaches, the recognition rates are poor (Table 2). Hence, we did not apply these approaches in our next experiment with KTH dataset. In the Pedestrian Action Dataset, our method found the available best recognition results.

The second dataset we have used is the most widely used benchmark dataset called KTH Action Dataset [35]. In this dataset, six natural actions such as walking, running, jogging, boxing, hand-clapping (HndClp) and hand-waving (HndWav) are performed by 25 different persons both in indoor and outdoor. The videos are in gray-scale, taken with a static camera with 25 fps, having resolution 160 × 120 pixels. The main problem arises while working KTH Action Dataset is the repetition of one action in the video which causes self-occlusion or motion overwriting in the MHI templates [2]. To solve this problem, we have implemented two-layer MHI templates for partial frames of the video. Once we have got MHI templates, we have encoded the LBP codes from the templates. HOG features have been extracted from the LBP codes. Then the classification has been done by using SVM. All the features are classified into six different actions. Leave-one-out-cross-validation scheme is used. 24 features are used to train each class of action. If the test feature does not match with any training image, it is considered as unrecognized. The confusion matrix of KTH Action Dataset using our proposed method has given in Table 3. Table 4 shows comparative results. Though our result is not optimum, but it is comparable to most of the other methods.

The recognition rate that has been obtained from KTH indoor dataset is 86.7 %. In 5.1 % cases, the method failed to recognize the action, whereas rest of the cases false recognition has been obtained. Due to the close similarities between 'Jogging' and 'Running', there has been repeated confusion between them. But taking into account the challenges the result is satisfactory. KTH Action Dataset being one of the most popular benchmark datasets, many established methods used them for testing. Some of them are given in Table 4.

Table 4 Comparison of recognition rate of KTH action dataset

Method	Recognition rate (%)
Local Jets-SVM [4]	71.72
Histogram of 3D Oriented Gradient [45]	91.80
Hierarchical Mined [46]	94.50
PLSA model [47]	81.50
Spat-Temp [48]	81.20
NNMF Detector [49]	80.99
Relative Motion Descriptor [50]	84
MHI-HOG-SVM [5]	78
Proposed method	86.67

Though some methods show better result than our proposed method, still it has shown better result than most of the methods. Considering the complexity in KTH Action Dataset, the recognition rate of our method is quite satisfactory.

5 Conclusion

In this paper, we proposed method for human action classification. We have used global descriptor for representation and to capture information. This is a view-based spatio-temporal template matching approach. Motion History Image (MHI) has been used to convert video sequences into temporal templates for preprocessing stages, while dominant motion information is preserved. We have combined two simple methods for feature extraction. First, we have used Local Binary Pattern (LBP) to convert the MHI template into binary code. Then local features have been extracted from LBP code using Histogram of Oriented Gradient descriptor. Finally, the extracted features have been used to train Support Vector Machine classifier to recognize the action classes. We have applied our method in Pedestrian Action Dataset that contains 140 video sequences of 7 classes, by 20 subjects. Our method has shown an excellent result of 94.3 % in this dataset. We have also used a benchmark dataset KTH Action Dataset.

KTH indoor dataset contains 150 videos of 6 different classes, by 25 subjects. We have gained 86.7 % accuracy on

Table 3 Confusion matrix of KTH action dataset

	Boxing (%)	HndClp (%)	Jogging (%)	Running (%)	Walking (%)	HndWav (%)	Unrecognized (%)
Boxing	92	4					4
HndClp	4	88					8
Jogging			84	12	4		
Running			16	80			4
Walking			4		80		16
HndWav						96	4

that dataset. The KTH Action Dataset has the challenges of close similarities between actions; and repetitiveness of similar patterns and self-occlusion, which are evident from MHI templates. Yet, the result we have achieved is satisfactory. It takes almost one and half minute to process and classify an input video sequence including all processing steps. However, our experiments are limited to only one person in the view, like most of the researches on action recognition. Hence, in future, we intend to concentrate on two person interactions and segmenting individual person in the video for separate recognition.

Acknowledgements Authors are thankful to the Center for Natural Science and Engineering Research (CNSER) and UGC, Bangladesh.

References

1. Ahad MAR (2011) "Computer vision and action recognition: a guide for image processing and computer vision community for action understanding", 1st edn. Atlantis Ambient and Pervasive Intelligence and Springer, Berlin
2. Ahad MAR (2013) "Motion history images for action recognition and understanding", 1st edn. Springer, Berlin
3. Ahad MAR, Tan JK, Kim H, Ishikawa S (2008) "Human activity recognition: various paradigms". International Conference in Control, Automation and Systems, pp 1896–1901
4. Schuldt C, Laptev I, Caputo B (2004) "Recognizing human actions: a local SVM approach". International Conference on Pattern Recognition, pp 32–36
5. Mueid RM, Ahmed C, Ahad MAR (2015) "Pedestrian activity classification using patterns of motion and histogram of oriented gradient". J Multimodal User Interfaces 1–7. doi:10.1007/s12193-015-0178-3
6. Turaga P, Chellappa R, Subrahmanian VS, Udrea O (2008) Machine recognition of human activities: a survey. IEEE Trans Circuits Syst Video Technol 18(11):1473–1488
7. Moeslund TB, Granum E (2001) A survey of computer vision-based human motion capture. Comput Vis Image Understand 106:90–126
8. Wang L, Hu W, Tan T (2003) Recent developments in human motion analysis. Pattern Recognit 36:585–601
9. Poppe R (2007) Vision-based human motion analysis: an overview. Comput Vis Image Understand 108:4–18
10. Poppe R (2010) A survey on vision-based human action recognition. Image Vis Comput 28:976–990
11. Moeslund TB, Hilton A, Kruger V (2006) A survey of advances in vision-based human motion capture and analysis. Comput Vis Image Understand 104(2–3):90–126
12. Gavrila DM (1999) "The visual analysis of human movement: a survey". Comput Vis Image Understand 73:82–98
13. Aggarwal JK, Cai Q (1999) Human motion analysis: a review. Comput Vis Image Understand 73(3):428–440
14. Bobick A, Davis J (1996) "An Appearance-based Representation of Action". International Conference on Pattern Recognition, pp 307–312
15. Davis JW (1996) "Appearance-based motion recognition of human actions", M.I.T. Media Lab Perceptual Computing Group Tech. Report No. 387, p 51
16. Anderson C, Bert P, Wal GV (1985) Change detection and tracking using pyramids transformation techniques. SPIE-Intell Robot Comput Vis 579:72–78
17. Lipton AJ, Fujiyoshi H, Patil RS (1998) "Moving Target classification and tracking from real-time video". IEEE Workshop on Applications of Computer Vision, pp 8–14
18. Ahad MAR, Tan JK, Kim H, Ishikawa S (2009) Temporal motion recognition and segmentation approach. Int J Imaging Syst Technol 19:91–99
19. Ahad MAR, Ogata T, Tan JK, Kim H, Ishikawa S (2008) A complex motion recognition technique employing directional motion templates. Int J Innov Comput, Inf Control 4(8):1943–1954
20. Gavrila DM (2000) Pedestrian detection form a moving vehicle, vol 1843. Springer, Berlin. pp 37–49
21. Mikolajczyk K, Schmid C (2005) A performance evaluation of local descriptors. IEEE Trans on PAMI 27(10):1615–1630
22. Dalal N, Triggs B (2005) "Histograms of oriented gradients for human detection". International Conference on Computer Vision and Pattern Recognition, pp 886–893
23. Bobick A, Davis J (2001) The recognition of human movement using temporal templates. IEEE Trans Pattern Anal Mach Intell 23(3):257–267
24. Meng H, Pears N, Bailey C (2007) "A human action recognition system for embedded computer vision application". In: Workshop on Embedded Computer Vision (with CVPR), pp 1–6
25. Han J, Bhanu B (2006) Individual recognition using gait energy image. IEEE Trans PAMI 28(2):316–322
26. Chandrashekhar V, Venkatesh KS (2006) "Action energy images for reliable human action recognition". In: Asian Symposium on Information Display (ASID), pp 484–487
27. Liu J, Zhang N (2007) "Gait history image: a novel temporal template for gait recognition". IEEE International Conference on Multimedia and Expo, pp 663–666
28. Ahad MAR, Tan J, Kim H, Ishikawa S (2012) Motion history image: its variants and applications. Mach Vis Appl 23(2):255–281
29. Ojala T, Pietikainen M, Harwood D (1994) Performance evaluation of texture measures with classification based on kullback discrimination of distributions. Int Conf Pattern Recognit 1:582–585
30. Ojala T, Pietikainen M, Harwood D (1996) A comparative study of texture measures with classification based on feature distributions. Pattern Recognit 29(1):51–59
31. Pietikainen M, Zhao G, Hadid A, Ahonen T (2011) "Computer vision using local binary pattern", 1st edn. Springer, Berlin
32. Kellokumpu V, Zhao G, Pietikainen M (2008) Texture based description of movements for activity analysis. Int Conf Comput Vis Theory Appl 1:206–213
33. Heikkila J, Pietikainen M (2006) A texture-based method for modeling the background and detecting moving objects. IEEE Trans PAMI 28(4):657–662
34. Heikkila J, Pietikainen M (2004) "A texture-based method for detecting moving objects". British Machine Vision Conference, pp 187–196
35. Hadid A, Pietikainen M (2009) Combining appearance and motion for face and gender recognition from videos. Pattern Recognit 42(11):2818–2827
36. Zhao G, Pietikainen M (2007) Dynamic texture recognition using local binary patterns with an application to facial expressions. IEEE Trans PAMI 29(6):915–928
37. Huang C, Hsieh C, Lai K, Huang WY (2011) "Human action recognition using histogram of oriented gradient of motion history image". In: International conference on instrumentation, measurement, computer, communication and control, pp 353–356
38. Bertozzi M, Broggi A, Del Rose M, Felisa M, Rakotomamonjy A, Suard F (2007) "A pedestrian detector using histograms of oriented gradients and a support vector machine classifier". In: IEEE intelligent transportation systems conference, pp 143–144

39. Dalal N, Triggs B, Schmid C (2006) “Human detection using oriented histograms of flow and appearance”. In: European Conference on Computer Vision, pp 428–441
40. Osuna E, Freund R, Girosi F (1997) “Training support vector machines: an application to face detection”. CVPR, pp 130–136
41. Vapnik VN (1999) An overview of statistical learning theory. IEEE Trans neural Netw 10(5):988–999
42. Weinland D, Ronfard R, Boyer E (2006) Free viewpoint action recognition using motion history volumes. Comput Vis Image Understand 104:249–257
43. Wolf L, Shashua A (2003) “Kernel principal angles for classification machines with applications to image sequence interpretation”. CVPR, pp 635–640
44. Wallraven C, Caputo B, Graf A (2003) “Recognition with local features: the kernel recipe”. ICCV, pp 257–264
45. Klaser A, Marszalek M, Schmid C (2008) “A spatio-temporal descriptor based on 3D gradients”. British machine vision conference
46. Gilbert A, Illingworth J, Bowden R (2011) Action recognition using mined hierarchical compound features. IEEE Trans Pattern Anal Mach Intell 33(5): 883–897. doi:10.1109/TPAMI.2010.144
47. Niebles J, Wang H, Fei-Fei L (2006) “Unsupervised learning of human action categories using spatial-temporal words”. In: British machine vision conference
48. Dollar P, Rabaud V, Cottrell G, Belongie S (2005) “Behavior recognition via sparse spatiotemporal features”. In: International workshop on visual surveillance and performance evaluation of tracking and surveillance, pp 65–72
49. Wong S, Cipolla R, “Extracting spatio-temporal interest points using global information”. In: International conference on computer vision
50. Bregonzio M, Li J, Gong S, Xiang T (2010) “Discriminative topics modeling for action feature selection and recognition”. In: British machine vision conference