Human action representation and recognition: An approach to histogram od spatiotemporal templates

International Journal of Innovative
Computing, Information and Control
Volume **11**, Number **6**, December **2015**

 ISSN 1349-4198
pp. 1855–1867

HUMAN ACTION REPRESENTATION AND RECOGNITION: AN APPROACH TO A HISTOGRAM OF SPATIOTEMPORAL TEMPLATES

Sk. Md. Masudul Ahsan, Joo Kooi Tan, Hyoungseop Kim
and Seiji Ishikawa

Department of Control Engineering
Kyushu Institute of Technology
1-1 Sensui-cho, Tobata-ku, Kitakyushu-shi, Fukuoka 804-8550, Japan
{ ahsan; etheltan; ishikawa }@ss10.cntl.kyutech.ac.jp; kim@cntl.kyutech.ac.jp

Received April 2015; revised August 2015

Abstract. *The motion sequences of human actions have its own discriminating profile that can be represented as a spatiotemporal template like Motion History Image (MHI). A histogram is a popular statistic to present the underlying information in a template. In this paper a histogram oriented action recognition method is presented. In the proposed method, we use the Directional Motion History Images (DMHI), their corresponding Local Binary Pattern (LBP) images and the Motion Energy Image (MEI) as spatiotemporal template. The intensity histogram is then extracted from those images which are concatenated together to form the feature vector for action representation. A linear combination of the histograms taken from DMHIs and LBP images is used in the experiment. We evaluated the performance of the proposed method along with some variants of it using the renowned KTH action dataset and found higher accuracies. The obtained results justify the superiority of the proposed method compared to other approaches for action recognition found in literature.*
Keywords: Action recognition, MHI, DMHI, LBP, Histogram, Support vector machine

1. **Introduction.** Analyzing human motion and recognizing the performed action from a video sequence are very important and has been a well-researched topic in the field of computer vision. The reason behind such attention is its diverse applications in different domains like robotics, human computer interaction, video surveillance, controller-free gaming, video indexing, mixed or virtual reality, and intelligent environments. Such application domains have their individual demands, but in general, algorithms must be intelligent enough to detect and recognize various actions performed by different people with several possible body poses due to variation in viewing position. Moreover, for practical use, the designed methods must have the capability of real time recognition as well as the adaptability to various types of environments like illumination change or variation in clothing [1].

There are two main approaches of action representation: holistic and part-based representations. Holistic representation focuses on the whole human body and then tries to search distinctive characteristics such as contours or pose. Part-based representations typically search for Space-Time Interest Points (STIPs), and apply a robust description of the area around them and create a model [2]. Both approaches have their advantages and disadvantages. Holistic approaches use only pose or contour suffers in case of occlusion, but otherwise they provide better recognition accuracy, and they are generally simple and fast. Although part-based cases are better in case of partial occlusion, identifying

individual body parts may cause overhead in total recognition time. Therefore, choosing the method will depend on the problem at hand needed to be solved.

The proposed method generates spatiotemporal images of an action sequence and then goes with holistic approach for the feature vector generation. The details are presented in the rest part of the paper which is organized as follows. Section 2 describes some related works. Section 3 explains the proposed method of how an action can be represented. Experimental results are given in Section 4 followed by a conclusion in Section 5.

2. **Related Works.** There are several methods [3] introduced in the literature for learning and recognizing a broad class of motion or action patterns. Bobick et al. [4,5] used motion energy images (MEI) and motion history images (MHI) for the first time as temporal templates to represent human actions. Recognition was done by using seven Hu moments. They have developed a virtual aerobics trainer that can watch and respond to the user as she/he performs the workout. Weinland et al. [6] used multiple cameras to build motion history volumes and action classification was performed using Fourier analysis in the cylindrical coordinates. Related 3D approaches have been introduced by Blank et al. [7] and Yilmaz and Shah [8] who used time as the third dimension to form space-time volumes in the (x, y, t) space. Those volumes were matched using features from Poisson equations and geometric surface properties, respectively.

Ikizler and Duygulu [9] use a pose descriptor named as Histogram-of-Oriented- Rectangles (HOR) for representing and recognizing human actions. They represent each human pose in an action sequence by oriented rectangular patches extracted over the human silhouette, and form a spatial oriented histograms to represent the distribution of these rectangular patches. They used different classifiers like nearest neighbor, support vector machine, dynamic time warping for the matching purpose.

Kellokumpu et al. [1,10] extracted histogram of local binary pattern (LBP) from MHI and MEI as temporal templates to represent action. They also used another descriptor called LBP-TOP [11], which extracts LBP information from three orthogonal planes (xy, xt, and yt). They used HMMs to model the temporal behavior of action and hence to recognize them. Yau et al. [12] used MHI to represent the temporal information of the mouth, which is generated using accumulative frame differencing of the video. The MHIs are decomposed into wavelet sub-images using Discrete Stationary Wavelet Transform (SWT). Artificial neural network (ANN) with back propagation learning algorithm is used for classification. Kumar et al. [13] also used MHI and ANN for hand gesture classification.

In the basic MHI method, old motion information can be wiped out by new motion information (from the later motion sequence). This overwriting surely causes poor recognition rate for natural motions that have complex nature and overlapping motion (e.g., sitting down and then standing up). To counter the problem, Ahad et al. [14] employ a variant of MHI called Directional MHI (DMHI) to represent a human action. They also used Hu moments for the recognition purpose.

A histogram is a very popular statistic used in computer vision research. Many researchers [1,9,15] use histogram to represent their descriptors. Recently, Dalal and Triggs [16] used histograms of oriented gradients (HOGs) for human detection in images, which is found to be quite effective.

The methods described in [4,14] did not use any benchmark dataset to show the effectiveness of their descriptors. The HOR descriptor presented by Ikizler and Duygulu [9] is very simple, but it demands huge execution time. It takes approximately 1 second per frame only for the rectangle extraction phase, which is not suitable for real-time recognition. Kellokumpu et al. used a volume based descriptor [1,10] which is also inexpedient

for online recognition applications. In their descriptor, to extract the LBP of a particular frame they have to wait for the next few frames to arrive.

The key intent of the proposed study is to devise a simple and compact action representation method that can be applied to real-time recognition problems. To accomplish the objective, the proposed action descriptor is based on a basic idea that rather than detecting the exact body parts or analyzing each action sequence, human action can be represented by a collection of local texture patterns in spatiotemporal domain. Hence, we are only interested in the distribution of those spatiotemporal patterns. In this paper, we construct a concatenated block histogram for action representation which is extracted from a novel template, i.e., LBP images created from DMHI templates unlike [1,10]. We use a unique way of generating different LBP images which is described in Section 3.2. We have found by experiment that extracting LBP from DMHI gives a better recognition rate than using only DMHIs, and then blending them together provides even better results.

3. **The Proposed Method.**

3.1. **Spatiotemporal image.** The MHI $H_\tau(x, y, t)$ of an action sequence is computed from the following equations using an update function [4]:

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

$$\Psi(x,y,t) = \begin{cases} 1 & \text{if } D(x,y,t) \geq \text{threshold} \\ 0 & \text{otherwise} \end{cases} \tag{2}$$

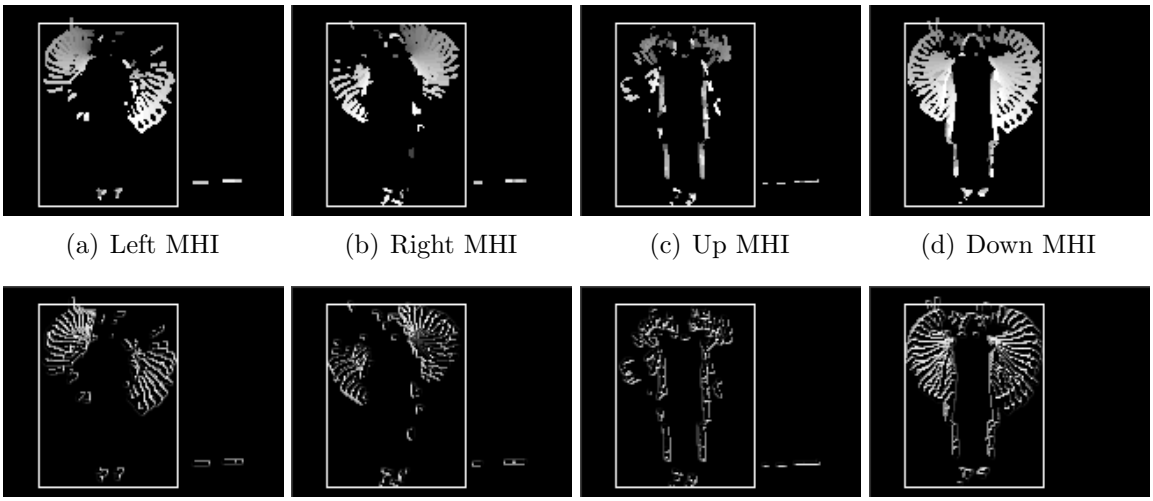
$$D(x,y,t) = |I(x,y,t) - I(x,y,t \pm \Delta)| \tag{3}$$

where x, y, and t show the position and time; the update function $\Psi(x, y, t)$ signals the presence of motion in the current video image $I(x, y)$; τ decides the temporal duration of MHI, e.g., in terms of the number of frames; δ is the decay parameter whose value was 1 in the original MHI [3], and $D(x, y, t)$ gives the absolute difference between pixels with step time difference Δ. This update function is called for every new video frame analyzed in the sequence. The result of this computation is a grayscale image where brighter pixels represent the more recent motion.

To generate spatiotemporal template we have used DMHI method [14,17] which extends the basic MHI method by considering motion directions. For DMHI, initially the moving region is tracked using a dense optical flow algorithm that generally produces a vector denoting the horizontal (x-direction) and vertical (y-direction) motion of an object. Each of these horizontal and vertical motions is further rectified to positive and negative directions, resulting in four update functions denoting directions right, left, up, and down. These update functions are used in Equation (1) to generate directional MHIs. Therefore, for DMHI, Equation (1) becomes the one as Equation (4), where the four different directions are denoted by $\ell \in \{\text{right}(+x), \text{left}(-x), \text{up}(+y), \text{down}(-y)\}$. First row of Figure 1 shows some examples of DMHIs of a hand waving action.

$$DMHI_\tau^\ell(x,y,t) = \begin{cases} \tau & \text{if } \Psi^\ell(x,y,t) = 1 \\ \max\left(0, DMHI_\tau^\ell(x,y,t-1) - \delta\right) & \text{otherwise} \end{cases} \tag{4}$$

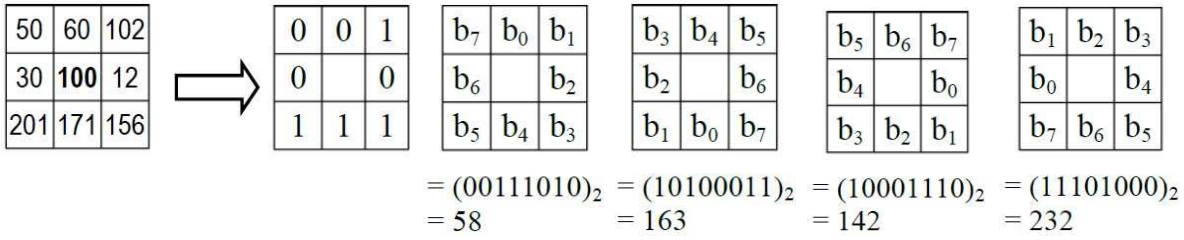
3.2. **Extraction of spatiotemporal texture.** The LBP operator becomes a popular approach in various applications [18,19] for its computational simplicity. LBP operator [20] describes the local texture pattern of an image with a binary code, which can be obtained by taking a threshold of neighboring pixels with the gray value of their center pixel. For each pixel, its intensity is compared to each of the neighbor pixel intensities. If the intensity of the neighbor pixel is greater than that of the center pixel with a certain



(a) Left MHI (b) Right MHI (c) Up MHI (d) Down MHI

(e) Left LBP image (f) Right LBP image (g) Up LBP image (h) Down LBP image

FIGURE 1. Example of DMHIs and corresponding LBP images of a hand waving action



(a) Output of 3×3 LBP operator (b) Left (c) Right (d) Up (e) Down

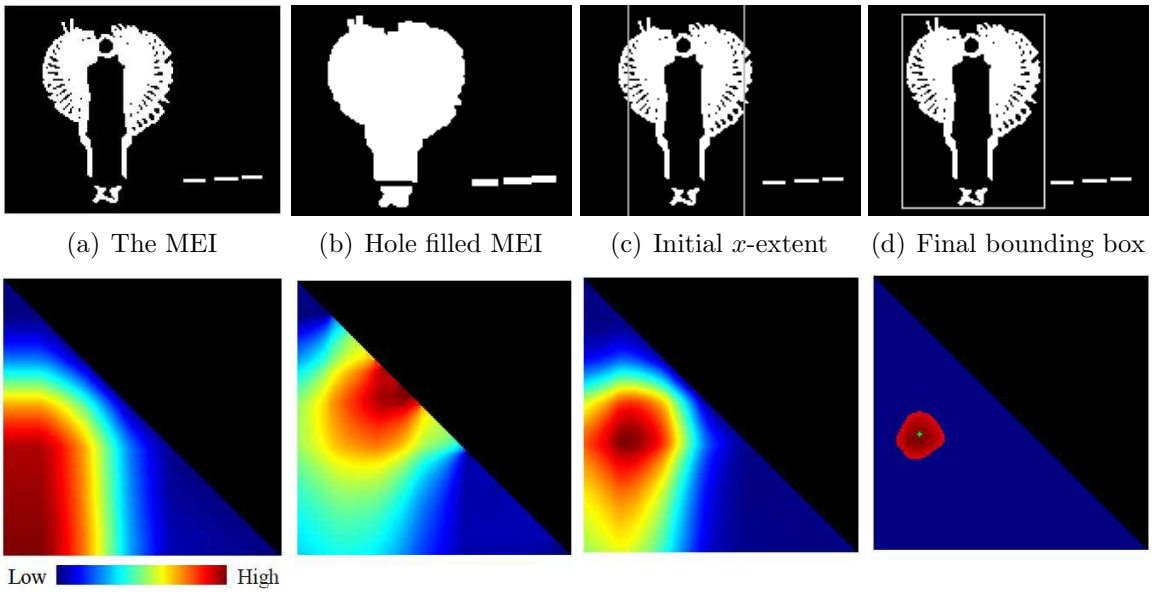
FIGURE 2. Illustration of different LBP bit arrangements

amount of threshold, write "1"; otherwise, write "0" or vice-versa. Mathematically, LBP operator can be written as Equation (5), and Equation (6), where g_c is the intensity of the center pixel (x, y) and g_i $(i = 0, 1, \ldots, p-1)$ are the intensities of the neighboring pixels.

$$LBP(g_c) = \sum_{i=0}^{p-1} S(g_i - g_c) \times 2^i \tag{5}$$

$$S(x) = \begin{cases} 1 & \text{if } x \geq \text{threshold} \\ 0 & \text{otherwise} \end{cases} \tag{6}$$

The neighborhoods can be of different sizes such as 3×3, 3×5, or 5×5 pixel. In this paper, we only use 3×3 neighborhood which has a different arrangement of LBP bit position for different DMHIs. This is to give more strength to the pattern of a particular direction. Figure 2(a) shows a basic LBP operator, whereas Figures 2(b)-2(e) show the LBP bit arrangements used for different DMHIs. Consider Figure 2(b): The arrangement of bit positions is chosen in such a way that it will give more emphasis on leftward motion. Other arrangements are chosen to have similar effects. From Figure 2 we can see that the binary output of the LBP operator (Figure 2(a)) can be assigned to a different decimal pattern number (Figures 2(b)-2(e)) by choosing a different starting and ending position for the least and most significant bit respectively. The LBP images corresponding to the DMHIs are shown in the second row of Figure 1.



(a) The MEI (b) Hole filled MEI (c) Initial x-extent (d) Final bounding box

(e) Information ratio distribution (f) Density distribution (g) Multiplied density-information distribution (h) Thresholded density-information highlighting the center

FIGURE 3. Images used to determine the action region

3.3. **Finding action region.** For better recognition results, the action region is determined, which also helps to alleviate the translational effect. To find a tight bounding area around the active region, we have used MEI. The MEI (Figure 3(a)) is deduced by accumulating all the DMHIs to a single image and then taking the image with a threshold equal to zero [4]. MEI sometimes contains holes within the action region, so holes are filled up. For hole filling, a simple algorithm is used. For a pixel in a hole, we search for four non-hole pixels in four different directions (straight-left, straight-right, straight-up, and straight-down). If they develop, there will be a change in the hole pixel to non-hole pixel, otherwise it is better to leave it as it was, and the details of the method can be found in [21]. Figure 3(b) shows the result after hole filling process. Then we try to find the minimum extent in the x-axis having the maximum ratio (which we call the information ratio) of pixels in that extent and the total number of pixels. For this purpose, we first determine the distribution of the information ratios for all possible extent combinations. Figure 3(e) shows the information ratio distribution of the MEI in the x-axis. The distribution of information ratio, Figure 3(e), is a square matrix where each row means start position and column stands for end position of the extent and their cross points represent the information ratio in that extent. As this matrix is a symmetric matrix, we only use the lower triangle of the matrix. The information ratio distribution is further divided by their corresponding extent length to get per unit information or information density of the extent. The density distribution is then normalized using min-max normalization (See Figure 3(f)) and multiplied to the information ratio distribution (Figure 3(g)). Then we take a threshold at the 90th percentile on the multiplied density-information distribution which results in an area with the maximum possible information with minimum extent. We find the center of that area (Figure 3(h)) and use one value as left (lp) and other as a right position (rp) of the extent. Figure 3(c) shows the initial extent. However, this extent sometimes loses some information. So we adjust the (lp, rp), which helps the utilized region to expand. However, this adjustment is within the 10 percent of the initially determined extent length on each side. We then determine the topmost and bottom most

pixels within that extent and use their ordinate (y) values as top (tp) and bottom position (bp) for the bounding rectangle. Figure 3(d) presents the rectangle.

3.4. **Feature vector generation.** An image texture can be described with its intensity histogram. We use the histogram of both the DMHI and LBP images as well as the MEI to represent an action.

For each of the DMHIs and LBP images, action region is partitioned into $p \times q$ disjoint blocks. For each block we compute weighted (see Equation (7)) intensity histogram dividing the entire intensity ranges (256 gray levels) into r equal sized bins. Rather than a global histogram, we calculate block histograms to have some spatial information lying on the template. These block histograms of each image are concatenated together in a raster scanning fashion to form an image histogram. All the image histograms are individually normalized using L2 norm, and then put together to form two final histograms for DMHIs and LBP images (abbreviated as *DMHI_H*, and *LBP_H*) which can be individually treated as a feature vector for recognition. However, we took the linear combination of those histograms using Equation (8) to generate the final feature vector for DMHI-LBP histogram representation (shortened as *DLBP_H*).

Moreover, we use the MEI to have some shape information of the performed action. In this case, we partition the action region of MEI into $2p \times 2q$ blocks, and find the non-zero pixel distribution of the blocks which yields MEI histogram (in short, *MEI_H*). The *MEI_H* is used to form a variant of *DLBP_H* named as *DLBP_MEI_H* by using Equation (9).

Figure 4 graphically illustrates the feature vector generation process explained above. Weight $w_i \geq 1$ for each block is calculated using Equation (7), where NP_{B_i}, NP_A are

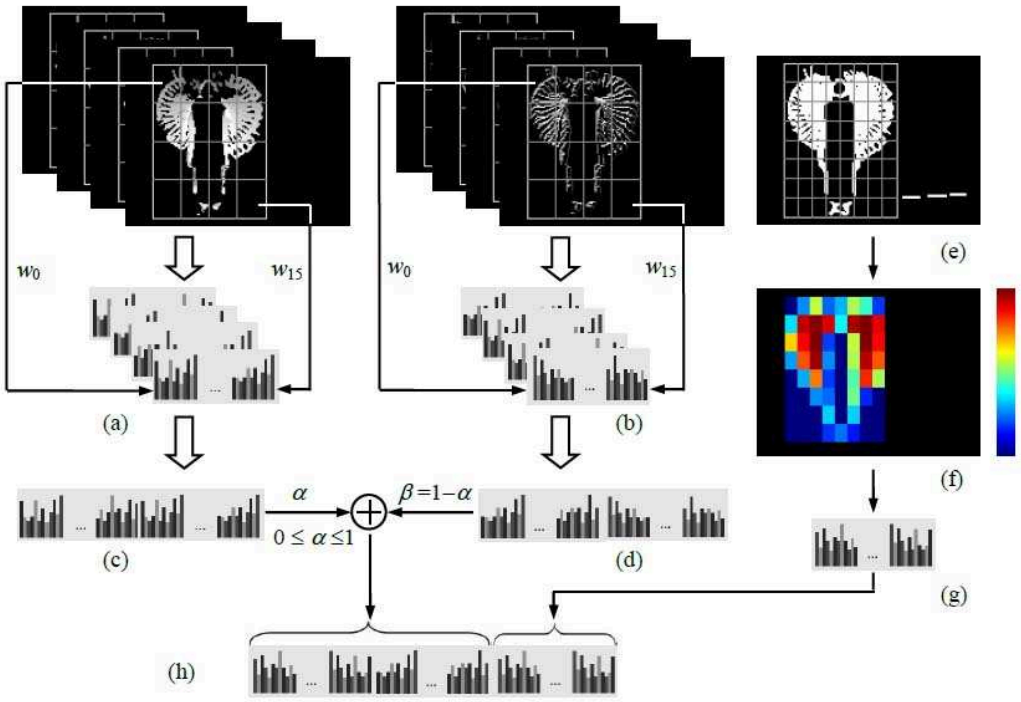


FIGURE 4. Generation of a feature vector: (a) and (b) block intensity histogram of DMHIs and LBP images, (c) DMHI histogram, (d) LBP histogram, (e) and (f) the MEI and the heat map of the blocks' non-zero pixel distribution, (g) linearized heat map – the MEI histogram, (h) the feature vector – linear combination of (c) and (d) concatenated with (g)

the number of non-zero pixels in block i, and in the action region, respectively, and ε is a constant (≈ 0) useful for no action scene. Equation (8) gives the feature vector of *DLBP_H*, where $0 \leq \alpha \leq 1$, and $\beta = 1 - \alpha$. Clearly, $\alpha = 0$ or 1 means *DLBP_H* becomes *LBP_H* or *DMHI_H*, respectively. Equation (9) gives *DLBP_MEI_H,* where $||$ is a concatenation operator, again for $\alpha = 0$ or 1, *DLBP_MEI_H* becomes *LBP_MEI_H* or *DMHI_MEI_H* respectively. Equations (10) and (11) give the number of dimensions in the feature vector.

$$w_i = \frac{NP_A + \varepsilon}{NP_A - NP_{B_i} + \varepsilon}, \quad i = 0, 1, \ldots, p \times q - 1 \tag{7}$$

$$DLBP\_H = \alpha \times DMHI\_H + \beta \times LBP\_H \tag{8}$$

$$DLBP\_MEI\_H = DLBP\_H||MEI\_H \tag{9}$$

$$Dim_{DLBP\_H} = 4 \times p \times q \times r \tag{10}$$

$$Dim_{DLBP\_MEI\_H} = 4 \times p \times q \times r + 2p \times 2q \tag{11}$$

4. **Experiments and Results.**

4.1. **Experimental setup.**

We measure the performance of the proposed representation method by experimenting with the popular benchmark dataset: the KTH action datasets [22-26], so that we can directly compare our results to other approaches reported in the literature. The KTH dataset consists of 600 sample videos of six different actions: Boxing, Hand waving, Hand clapping, Walking, Running, and Jogging. Actions are performed by 25 people in four different scenarios: outdoors (*scn0*), outdoors with zooming, i.e., scale variations (*scn1*), outdoors with variation in clothing (*scn2*), and indoors (*scn3*). For Walking, Running, and Jogging actions, we trim the video sequences so that the person always remains in view.

In the literature, KTH dataset has been treated either as one large set with strong intra-subject variations, or as four independent scenarios, which are trained and tested separately (i.e., four visually dissimilar databases, which share the same classes). We use both alternatives in our experiment.

4.2. **Method.**

We use support vector machine (SVM) [27,28] for the classification purpose. SVMs are state-of-the-art large margin classifiers which have recently become very popular for visual pattern recognition [29] and many other applications. As a classifier for k action classes, we train a bank of k linear one-vs-rest and ${}^kC_2$ linear one-vs-one binary SVMs, each with identical weights. Since the dimension of the feature vector is sufficiently large, rather than non-linear, we use a linear SVM classifier. We apply a 10 fold cross validation method to train and test the classifier. For the cross validation purpose, we stratify each partition of data, i.e., each partition resembles the global distribution of the dataset. Finally, an action is labeled with the class having maximum votes assigned by those classifiers.

Each action video sequence is preprocessed to extract the foreground. In this phase, we extract the morphological gradient of the current frame which serves as a foreground mask. This mask generally contains lots of noise which is minimized by using the information of the previous frame. The method used here is a simplified version of the method presented in [30] which provides enough information for reasonable recognition. The extracted foreground is further tracked by dense Gunnar Farneback [31] algorithm to measure the optical flow that serves as an update function to create DMHIs. We use $\tau = 0.9$ in Equation (4), i.e., only 0.9 second frames are used to create DMHI templates. Each DMHI is used to calculate the LBP image, where 3×3 neighborhood and *threshold* =

1 are used for Equations (5) and (6). We used $p = q = 2, 4, 6, 8$ and $r = 8, 16, 32$ for a feature vector generation.

4.3. **Recognition results.** Figure 5 shows the correct classification rate (averaged on 6 actions) of the proposed method (*DLBP_MEI_H*) for different representation with various numbers of blocks and bins where one large set is considered as the data set. The results are averaged on 3 runs of experiment, where, one run means performing the experiment (executing the program for recognition) once. Abscissa of the graphs denotes different action representation methods with discrete values of α with a step difference 0.1 of Equation (8), i.e., the leftmost results are for *LBP_MEI_H*, the rightmosts are for *DMHI_MEI_H*, and the in-between results are for *DLBP_MEI_H* representation. In Figure 5, for all cases *LBP_MEI_H* produces better accuracy than its corresponding (same p, q, r values) *DMHI_MEI_H* representation. However, for *DLBP_MEI_H* representation, the curves show that its accuracy increases up to a certain value of α (0.3-0.5) and then goes down again. LBP operator highlights the patterns or texture lying in DMHI but loses the recency of motion. Hence, mixing a certain amount of *DMHI_H* information to the *LBP_H* gives better accuracy than *LBP_H* alone.

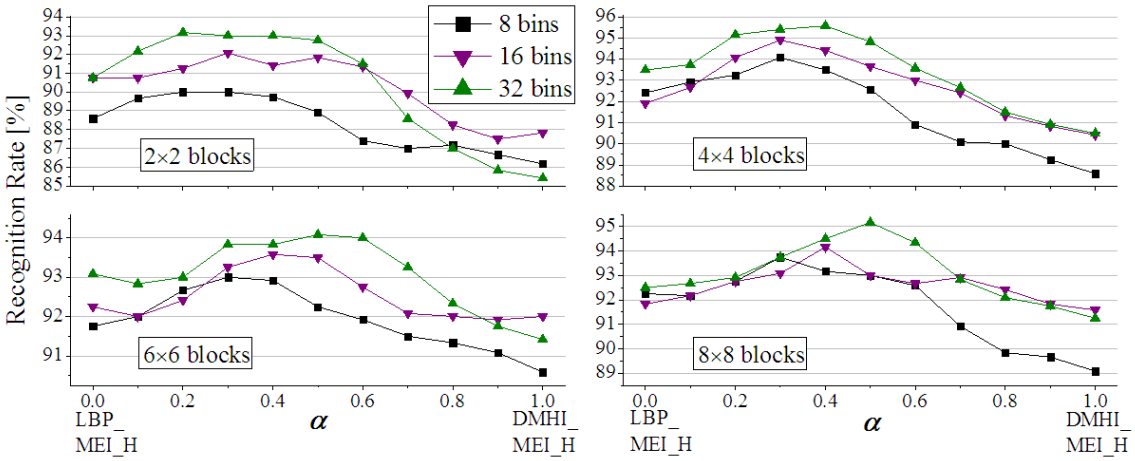


FIGURE 5. Correct recognition rate for different representation with various numbers of blocks and bins on KTH dataset

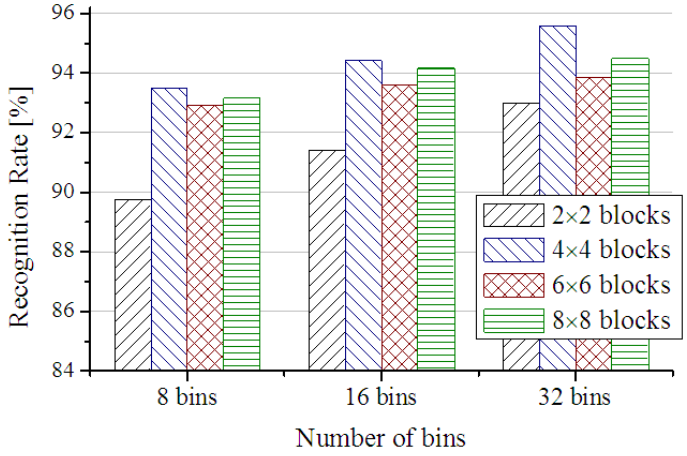


FIGURE 6. Relation between the number of blocks and histogram bins for *DLBP_MEI_H* with $\alpha = 0.4$

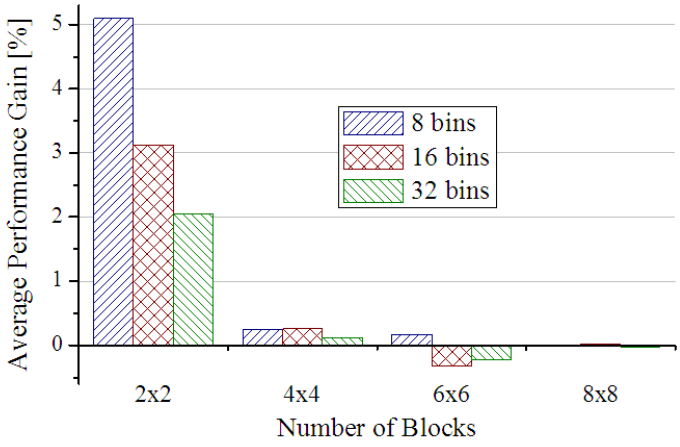


FIGURE 7. Average performance gain of *DLBP_MEI_H* over *DLBP_H*

Also in Figure 5, almost all cases, for a constant number of blocks, the higher number of bins produces better results. However, the reverse is not true. It can be better understood from Figure 6 which shows the relation between the number of blocks and histogram bins for *DLBP_MEI_H* with only for a specific $\alpha = 0.4$.

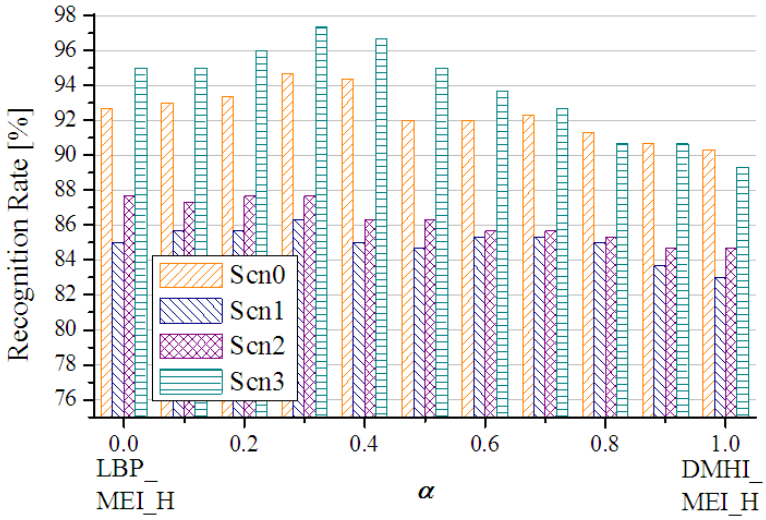
Figure 7 presents the average performance increase or gain in classification accuracy of *DLBP_MEI_H* over *DLBP_H* representation. Keeping each p, q, r value constant, the performance gain is measured using the following equations (Equations (12) and (13)). We can see from Figure 7 that, using the *MEI_H* along with *DLBP_H* improves the performance significantly only in a lower number of blocks. However, if we increase the number of blocks, there is no substantial gain or sometimes it is even negative.

$$Gain_{DLBP\_MEI\_H} = \frac{1}{11} \sum_{\alpha=0,0.1,\ldots,1} \left(Accuracy_{DLBP\_MEI\_H_\alpha} - Accuracy_{DLBP\_H_\alpha} \right) \tag{12}$$

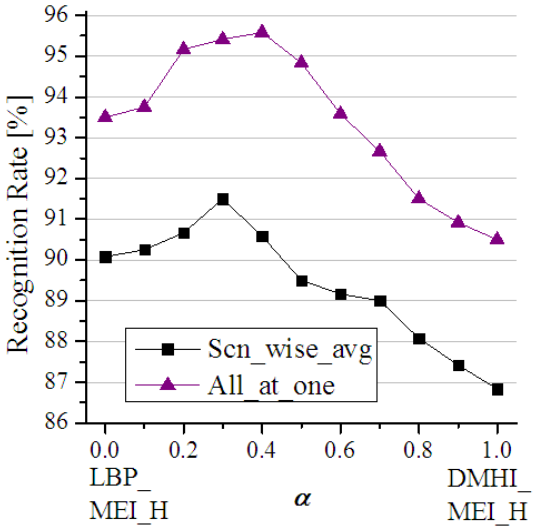
$$Accuracy = \frac{No.\ of\ correct\ classification}{No.\ of\ total\ test\ data} \tag{13}$$

Figure 8(a) presents the recognition results for different representation methods when the classifier is trained and tested on action performed at different scenarios with p, $q = 4$, $r = 32$. We find that the indoor (*scn3*) and the outdoor (*scn0*) scenarios provide better results than the outdoor with zooming (*scn1*) and clothing (*scn2*) variations. It is obvious that using different clothing may hide some part of the body and consequently the action pose becomes different (e.g., long overcoat may hide some leg portion). Similarly, zooming the camera while performing an action, sometimes overwrites some motion information. However, the results are still above 85% for *scn1* and *scn2* in the best case ($\alpha = 0.3$). Figure 8(b) shows the comparison of the average scenario-wise results with the result found from the experiment taking all scenario-wise datasets as a large single dataset for the same p, q, r values. Clearly, the classifier provides better performance for the large single dataset.

The confusion matrix of *DLBP_MEI_H* representation for the best result (for $p = q = 4$, $r = 32$, $\alpha = 0.4$) is presented in Table 1. The recognition rate is above 90% in all actions except jogging. Jogging action has very subtle difference between walking and running, and it varies with the person performing the action. Therefore, the classifier puts some jogging actions as walking or running action. However, accuracy of jogging is among the top compared to some other methods [22].



(a)



(b)

FIGURE 8. Scenario-wise recognition for p, $q = 4$, $r = 32$: (a) recognition rate for different scenarios, (b) comparison of the average scenario-wise results with that of all data taken as one

TABLE 1. Confusion matrix of *DLBP_MEI_H* representation

	Box	HW	Clap	Walk	Run	Jog
Box	99	0	0	1	0	0
HandW	0	100	0	0	0	0
Clap	0	0	100	0	0	0
Walk	1	0	0	97.5	0	1.5
Run	0	0	0	0	90.5	9.5
Jog	0	0	0	1	12.5	86.5

TABLE 2. Comparison of the proposed method to other methods reported on KTH dataset

	Reference	**Result [%]**
Proposed	*DMHI_MEI_H*	92.0
	LBP_MEI_H	93.5
	DLBP_MEI_H	95.6
Archived in literature	Schuldt et al. [22]	71.7
	Masumitsu and Fuchida [32]	79.9
	Ke et al. [23]	80.9
	Dollar et al. [33]	81.2
	Niebles et al. [25]	81.5
	Ikizler and Duygulu [9]	89.4
	Schindler and Gool [34]	90.1
	Wong et al. [26]	91.6
	Ahsan et al. [35]	93.3
	Maninis et al. [36]	93.5
	Kellokumpu et al. [1]	93.8
	Kim et al. [24]	95.3

TABLE 3. Evaluation of the per frame execution time (in milliseconds) of the proposed method over KTH dataset

Template creation time			Feature vector generation time	Testing time	Total time
DMHI	LBP image	MEI & action region			
19.49	6.67	3.83	10.0	5.95	45.94

Table 2 summarizes the best experimental results of different action representation along with the best result found by other methods on KTH dataset. Here, *DMHI_MEI_H/ LBP_MEI_H* results are for $p = q = 6/4$, $r = 16/32$. It is worthy to mention that, these results are not directly comparable, since different authors used different classifiers even different testing methods such as leave-one-out or different dataset splitting technique. Overall, the result found by the proposed method is among the top-listed results that have been reported in the literature regarding KTH dataset.

4.4. **Computational time.** The run time of the method can be parted in two phases, with the first being the time to create the spatiotemporal templates, and the second being the feature vector generation and testing time, which depend on the values of p, q, r, and α. Table 3 shows the average of the per frame computational times (in milliseconds) of the different phases for KTH dataset. The feature vector generation and testing times in Table 3 are for best recognition rate parameters $p = q = 4$, $r = 32$, $\alpha = 0.4$. It should be noted that the experiment is done on a machine with a processor Intel® Core™ i7-3770, speed 3.40 GHz, and memory 8GB. We implement the program in Microsoft Visual Studio 2010, and OpenCV version 2.4.8 without applying any code optimization method.

5. **Conclusion.** In this paper, we have targeted the problem of human action recognition, and proposed a novel descriptor that uses a fusion of the block intensity histogram of DMHI and LBP images along with MEI histogram to represent human action. We experienced that correctly determining the action region in the scene greatly affects the overall performance. We did not incorporate any direct mechanism for scale and rotation invariance or view point changes. However, our method successfully recognizes the actions

with scale and view point changes which are present in KTH dataset. We always partition the action region in a constant number of blocks rather than fixed size blocks just to have some benefit for scale variation. We observed that most of the misperception, especially in the KTH dataset, occurs due to the imperfect extraction of foregrounds. However, even with the noisy foreground information, our method reaches higher recognition rates, which mean that the method is robust to noise. Correct classification rate found for the proposed action representation method justifies that the method is good enough for practical use.

The significance of the proposed study is that without constructing any complex model, the proposed method represents an action in a simple and compact distribution of texture patterns that provide robust recognition rate. The work is done with a dataset having generic action classification problem. However, the method can easily be incorporated into applications like gaming or human computer interaction without using any controller such as mouse, trackball, joystick. Moreover, the potential of the proposed representation method can be applied to other relevant fields like a patient monitoring system or automatic labeling of video sequences in a video dataset.

Acknowledgment. This study was supported by JSPS KAKENHI Grant Number 2535 0477, which is greatly acknowledged. We wish to express our sincere appreciation to Prof. R. Long for his help in preparation of the paper.

REFERENCES

[1] V. Kellokumpu, G. Zhao and M. Pietikäinen, Recognition of human actions using texture descriptors, *Machine Vision and Applications*, vol.22, no.5, pp.767-780, 2009.

[2] R. Mattivi and L. Shao, Human action recognition using LBP-TOP as sparse spatio-temporal feature descriptor, *Proc. of the 13th International Conf. on Computer Analysis of Images and Patterns*, Münster, Germany, pp.740-747, 2009.

[3] R. Poppe, A survey on vision-based human action recognition, *Image and Vision Computing*, vol.28, no.6, pp.976-990, 2010.

[4] A. F. Bobick and J. W. Davis, The recognition of human movement using temporal templates, *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol.23, no.3, pp.257-267, 2001.

[5] J. Davis and G. Bradski, Real-time motion template gradients using Intel CVLib, *Proc. of the IEEE ICCV Workshop on Framerate Vision*, pp.1-20, 1999.

[6] D. Weinland, R. Ronfard and E. Boyer, Free viewpoint action recognition using motion history volumes, *Computer Vision and Image Understanding*, vol.104, nos.2-3, pp.249-257, 2006.

[7] M. Blank, L. Gorelick, E. Shechtman, M. Irani and R. Basri, Actions as space-time shapes, *Proc. of the 10th IEEE International Conf. on Computer Vision*, Beijing, China, pp.1395-1402, 2005.

[8] A. Yilmaz and M. Shah, Action sketch: A novel action representation, *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*, pp.984-989, 2005.

[9] N. Ikizler and P. Duygulu, Histogram of oriented rectangles: A new pose descriptor for human action recognition, *Image and Vision Computing*, vol.27, no.10, pp.1515-1526, 2009.

[10] V. Kellokumpu, G. Zhao and M. Pietikäinen, Human activity recognition using a dynamic texture based method, *Proc. of the British Machine Vision Conference*, United Kingdom, pp.885-894, 2008.

[11] G. Zhao and M. Pietikäinen, Dynamic texture recognition using local binary patterns with an application to facial expressions, *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol.29, no.6, pp.915-928, 2007.

[12] W. Yau, D. Kumar, S. Arjunan and S. Kumar, Visual speech recognition using image moments and multiresolution wavelet images, *Proc. of the International Conf. on Computer Graphics, Imaging and Visualisation*, Sydney, Australia, pp.194-199, 2006.

[13] S. Kumar, D. Kumar, A. Sharma and N. McLachlan, Classification of hand movements using motion templates and geometrical based moments, *Proc. of the International Conf. on Intelligent Sensing and Information Processing*, pp.299-304, 2004.

[14] M. A. R. Ahad, T. Ogata, J. K. Tan, H. S. Kim and S. Ishikawa, View-based human motion recognition in the presence of outliers, *International Journal of Biomedical Soft Computing and Human Sciences*, vol.13, no.1, pp.71-78, 2008.
[15] W. T. Freeman and M. Roth, Orientation histograms for hand gesture recognition, *Proc. of the International Workshop on Automatic Face and Gesture Recognition*, Switzerland, pp.296-301, 1995.
[16] N. Dalal and B. Triggs, Histograms of oriented gradients for human detection, *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*, USA, pp.886-893, 2005.
[17] M. A. R. Ahad, J. K. Tan, H. S. Kim and S. Ishikawa, Motion history image: Its variants and applications, *Machine Vision and Applications*, vol.23, no.2, pp.255-281, 2012.
[18] M. Pietikäinen, A. Hadid, G. Zhao and T. Ahonen, *Computer Vision Using Local Binary Patterns*, Springer, 2011.
[19] D. Huang, C. Shan, M. Ardabilian, W. Yunhong and C. Liming, Local binary patterns and its application to facial image analysis: A survey, *IEEE Trans. Systems, Man, and Cybernetics, Part C: Applications and Reviews*, vol.41, no.6, pp.765-781, 2011.
[20] T. Ojala, M. Pietikainen and T. Maenpaa, Multiresolution gray-scale and rotation invariant texture classification with local binary patterns, *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol.24, no.7, pp.971-987, 2002.
[21] M. Baranwal, M. T. Khan and C. W. D. Silva, Abnormal motion detection in real time using video surveillance and body sensors, *International Journal of Information Acquisition*, vol.8, no.2, pp.103-116, 2011.
[22] C. Schuldt, I. Laptev and B. Caputo, Recognizing human actions: A local SVM approach, *Proc. of the 17th International Conf. on Pattern Recognition*, pp.32-36, 2004.
[23] Y. Ke, R. Sukthankar and M. Hebert, Spatio-temporal shape and flow correlation for action recognition, *IEEE Conf. on Computer Vision and Pattern Recognition*, Minneapolis, MN, USA, pp.1-8, 2007.
[24] T. K. Kim, K. Y. K. Wong and R. Cipolla, Tensor canonical correlation analysis for action classification, *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*, Minneapolis, MN, USA, pp.1-8, 2007.
[25] J. C. Niebles, H. Wang and L. Fei-Fei, Unsupervised learning of human action categories using spatial-temporal words, *International Journal of Computer Vision*, vol.79, no.3, pp.299-318, 2008.
[26] S. F. Wong, T. K. Kim and R. Cipolla, Learning motion categories using both semantic and structural information, *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*, Minneapolis, MN, USA, pp.1-6, 2007.
[27] N. Cristianini and J. S. Taylor, *An Introduction to Support Vector Machines and Other Kernel-based Learning Methods*, Cambridge University Press, New York, 2000.
[28] V. Vapnik, *Statistical Learning Theory*, Wiley, New York, 1998.
[29] C. Wallraven, B. Caputo and A. Graf, Recognition with local features: The kernel recipe, *Proc. of the 9th IEEE International Conf. on Computer Vision*, Nice, France, pp.257-264, 2003.
[30] W.-C. Hu, Real-time on-line video object segmentation based on motion detection without background construction, *International Journal of Innovative Computing, Information and Control*, vol.7, no.4, pp.1845-1860, 2001.
[31] G. Farnebäck, Two-frame motion estimation based on polynomial expansion, *Proc. of the 13th Scandinavian Conf., LNCS*, Halmstad, Sweden, vol.2749, pp.363-370, 2003.
[32] K. Masumitsu and T. Fuchida, A proposition of human action recognition method considering co-occurrence of corner trajectories, *Proc. of the 19th International Symposium on Artificial Life and Robotics*, Beppu, Japan, 2014.
[33] P. Dollar, V. Rabaud, G. Cottrell and S. Belongie, Behavior recognition via sparse spatio-temporal features, *Proc. of the IEEE International Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance*, pp.65-72, 2005.
[34] K. Schindler and L. van Gool, Action snippets: How many frames does human action recognition require? *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*, Anchorage, AK, USA, pp.1-8, 2008.
[35] S. M. M. Ahsan, J. K. Tan, H. Kim and S. Ishikawa, Histogram of DMHI and LBP images to represent human actions, *Proc. of the IEEE International Conf. on Image Processing*, Paris, France, pp.1440-1444, 2014.
[36] K. Maninis, P. Koutras and P. Maragos, Advances on action recognition in videos using an interest point detector based on multiband spatio-temporal energies, *Proc. of the IEEE International Conf. on Image Processing*, Paris, France, pp.1490-1494, 2014.