SOCIETY FOR SCIENCE AND EDUCATION
UNITED KINGDOM

AIVP ADVANCES IN
IMAGE AND VIDEO PROCESSING

VOLUME 6, ISSUE 5
ISSN 2054-7412

Searching Human Action Recognition Accuracy from Depth Video Sequences Using HOG and PHOG Shape Features

Sadiya Tabussum, Javed Bin Al Amin, Mohammad Farhad Bulbul
Department of Mathematics,
Jessore University of Science and Technology, Jessore, Bangladesh
farhad@just.edu.bd

ABSTRACT

Research on human action recognition from depth video sequences are increasing day by day due to its vast application in automatic surveillance systems, entertainment environments, and healthcare systems etc. In our project, we improve human action recognition accuracy using shape features. We use Histogram of oriented gradients (HOG) and Pyramid Histogram of oriented gradients **(**PHOG) to extract shape features. The feature extraction algorithms are used to extract shape feature from dataset of different action videos. At first, depth motion maps (DMMs) are constructed from every action video. Then, the HOG and PHOG features are extracted from each DMMs. Using these features, actions are recognized by the l_2-regularized Collaborative Representation Classifier (l_2-CRC). In this paper, we evaluate our proposed method on MSR-Action 3D dataset. We divide this dataset into four action subsets such as AS1, AS2, AS3, and AS4 where each of them contains five actions. We compute the recognition accuracy of each action set using HOG and PHOG features respectively. Then, we take the comparison between the recognition accuracies of actions in every action set using HOG and PHOG features. Finally, we obtain the maximum recognition accuracy from most of the action subsets using PHOG feature. And the remaining subsets give poor results using HOG feature because of confusion between actions in those sets.

Keywords: Human action recognition, Depth motion maps, Histogram of oriented gradients, Pyramid Histogram of oriented gradients.

1 Introduction

Human action recognition is a significant area of computer vision research today. Computer vision tasks include methods for acquiring, processing, analyzing and understanding digital images, and extraction of high dimensional data from the real world in order to produce symbolic information. The goal of human action recognition is to automatically analyze ongoing actions from an unknown video. The human action recognition is the process of detecting & labeling of all occurring action from an input video [1]. There are different types of actions based on difficulties such as gestures, human actions, interactions, and group activities [2]. Gestures are basic movements of a person's body portion, and are the nuclear components describing the meaningful motion of a person. "Spreading an arm" and "moving a leg" are good examples of gestures. Human actions are activities by single-person that may be a collection of more than one gestures prepared temporally, such as "walking", "waving", and "punching". Interactions are human activities that include two persons and/or objects. For example, "two-person handshaking" is an interaction between two humans and "a person stealing a travel bag in an airport" is a human-object interaction. Finally, the activities which are done by

DOI: 10.14738/aivp.65.5340
Publication Date: 4th November, 2018
URL: http://dx.doi.org/10.14738/aivp.65.5340

groups made of numerous persons and/or objects “A group of a player playing a game”, “a group having a meeting,” and “two groups fighting” are typical examples. In this research, the main focus is given to improve the recognition accuracy of human actions from video sequences.

Nowadays, more and more people record their daily activities using digital cameras, and this brings the enrichment of video content on the internet, and also causes the problems of categorizing the subsisting video, and sorting new videos according to the action classes present. Categorizing these videos is a time-consuming task if it is done manually, and recognizing certain actions is impossible to accomplish through manual effort. For these causes, the area of human action recognition has interested considerable attention [1].

Previously, researches were dependent on recognizing human action from image sequences taken by RGB cameras and the typical RGB input devices are color TV, video cameras, image scanners, and digital cameras [3]. The image from RGB camera is called RGB image in which every color pixel is made of red, green & blue color. Various constraints relating to RGB cameras are responsible to background clutter, camera motion, occlusions and illumination variations. So, it has been a tough and difficult task to precisely recognize human actions [4, 5]. However, with the development of cost-effective RGB depth (RGB-D) camera sensors, the results from action recognition have improved, and they have become a point of consideration for many researchers. The Microsoft Kinect (see Figure: 1) is an example of Kinect sensor [4, 5]. It includes a RGB camera, 3D depth sensor cameras, a tilt motor, multi-array microphone and LED light [6]. Depth sensors help lessen and ease the complications found in RGB images, such as background subtraction and light variations. Also, depth camera can be beneficial for the entire range of day-to-day work, even at night. So, it has been a big challenge to utilize these data, together or independently, to present human behavior and to improve the accuracy of action recognition [5]. Depth sensor camera provides RGB color image, depth image/map, skeleton, cardiovascular, muscular, nervous information extraction, facial and voice recognition, virtual therapy, patient information, x-ray, MRI, CT scans etc [7]. Depth map is a 3D image formed of gray pixels defined by 0~255 values. Various research works on human action recognition have been carried out based on depth maps and we will discuss about these works in literature review. An example of a depth map sequence is shown in Figure 2.

Depth images also enable us to view and assess human skeleton joints in a 3D coordinate system. These 3D skeleton joints provide additional information to examine for recognition of action, which in turn increases the accuracy of the human–computer interface. However, though some of researches based on the skeleton information show high recognition performance, they are not suitable in the case where skeleton information is not available [4].

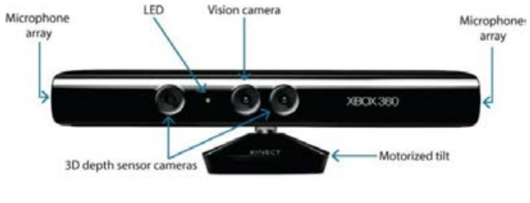
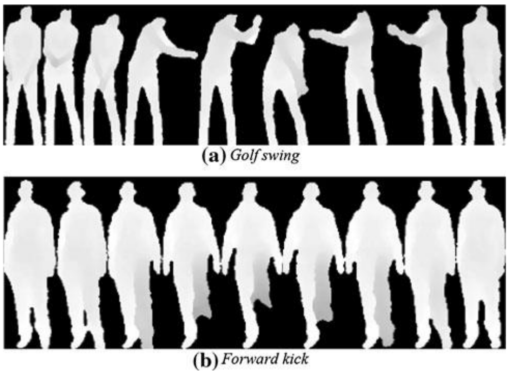


Figure 1: Kinect camera [8]



(a) Golf swing

(b) Forward kick

Figure 2: A depth map sequence for Golf swing & forward kick [9]

Figure 2: A depth map sequence for Golf swing & forward kick [9]

In this paper, Depth maps based action recognition system is proposed by representing an action with human shape in motion situation, representation and classification techniques. At first, all the video frames for each depth video are projected onto three orthogonal Cartesian planes so that the projected maps corresponding to three projection views (front, side, and top) to generated. For each projection sight, the addition of vital differences between consecutive projected maps forms Depth Motion Maps (DMMs) (i.e., DMM_f, DMM_s, and DMM_t) [4]. DMMs is the combination of depth maps. From this DMMs, we extract shape features through HOG (Histograms of Oriented Gradients) [10] and PHOG (Pyramid Histograms of Oriented Gradients) [11] descriptors. Then the dimension of this features is reduced by statistical procedure named Principal Component analysis (PCA) [12] and human action is recognized by using l_2-CRC (l_2-regularized collaborative representation classifier) [4] algorithm. The proposed approaches are primarily evaluated with MSR-action 3D dataset [13] which is specifically designed for human action recognition.

2 Related Work

In this section, we discuss about the recent related work for human action recognition from depth map sequences.

In 2010, Li et al. [14] presented a method to recognize human actions from sequences of depth maps. They engaged an action graph which model the temporal dynamics of actions, and used a combination of 3D points to characterize postures. This approach contained some limitations are the loss of spatial context information between interest points and computational inefficiency. To improve recognition accuracy, In 2012 Vieira et al. [15] presented Space-Time Occupancy Patterns (STOP), a new optical demonstration for 3D action recognition from depth motion maps. In the same year Wang et al. [16] represented 3-D action sequences as 4-D shapes and proposed Random Occupancy Pattern (ROP), and sparse coding was used to further improve the toughness of the proposed approach. To improve recognition rates, Yang et al. [17], used Histogram of Oriented Gradients (HOG) features from DMMs and SVM to classify action. In the same year, Oreifej et al. [18] presented a new descriptor called Histogram of Oriented 4D Normals (HON4D) for activity recognition from videos. Luo et al., in 2014 [19], extracted collection of 3D features such that both the spatial and temporary features of the RGB sequences for each depth video by using Centre-Symmetric Motion Local Ternary Pattern (CS-Mltp). Lu et al. [20] proposed binary range-sample feature descriptor in depth. In 2015 Chen et al. [21], used texture feature local binary patterns (LBPs) to recognize action. In the same year Farhad et al. [22], used Depth Motion Maps (DMMs), Con-tourlet Transform (CT) [23] and Histogram of Oriented Gradients (HOGs) in order to distinguishing actions. In the next year Chen et al. [24], presented an effective local spatio temporal descriptor, the local binary patterns (LBP) [25] descriptor and used kernel-based extreme learning machine classifier. Chen et al. [26], proposed action recognition method by using a distance-weighted Tikhonov matrix an $l2$-regularized collaborative representation classifier ($l2$-CRC). In 2012, Yang and Tian et al. [27], used Naive-Bayes-Nearest-Neighbor (NBNN) [28] classifier for multi-class action classification from human skeleton. To get more accuracy, in the same year, Xia et al. [29] used histograms of 3D joint locations (HOJ3D) as a solid representation of poses for recognizing human action. In 2013 Luo et al. [30], submitted Dictionary Learning (DL) method and used the Temporal Pyramid Matching (TPM) which keep the temporal information so that they can recognize human action. Wang et al [31], In 2011, proposed a method for action recognition using Pyramid Histogram of Orientation Gradient (PHOG) shape features and They adopted two state-space models, i.e., Hidden Markov Model (HMM) [32] and Conditional Random Field (CRF) [33] to model the dynamic human movement.

The rest of this paper is organized as follows. The proposed approach is presented in Section 3. The experimental results are demonstrated in Section 4. Finally, Section 5 contains a brief conclusion of this work.

3 Our approach

In this section, the approach for recognizing action is discussed. The whole recognition approach can be divided into two phases: the training phase and the testing phase shown in Figure 3.

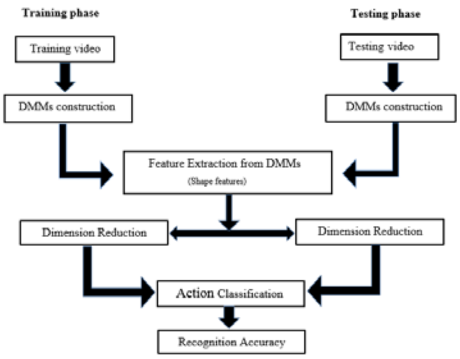


Figure 3: The proposed action recognition approach

During the training phase, DMMs (Depth motion maps) are constructed for each action, the shape feature is extracted from DMMs of each depth video obtained from the training video sequences. Histogram of oriented gradients (HOG) [10] and Pyramid Histogram of oriented gradients **(**PHOG) [11] descriptors are used to extract shape feature. Dimension of feature vector is reduced by Principal Component Analysis (PCA) [34].

During the testing phase, the same steps are followed to extract shape feature, build descriptors, reduce dimension as those done during the training phase. Then, $l2$-regularized Collaborative Representation Classifier ($l2$-CRC) [4] is adopted to train model and classify each testing sequence to the most probable action type.

3.1 DMMs construction

Depth Motion Maps (DMMs) was firstly proposed by Yang et al. [17]. In the feature extraction stage, by using generation techniques described in [26], DMMs are firstly constructed for each depth video sequence. Let F is the number of depth maps for each video sequence. The projection of each depth maps onto three orthogonal Cartesian planes provides three projected maps corresponding to the three projection views (front, side, and top). Let DMM_f ,DMM_s and DMM_t are the three projected maps. The equation of each DMM is given bellow:

$$DMM_x = \sum_{n=1}^{F-1} \left| map_x^{n+1} - map_x^{n} \right| \tag{1}$$

Where, n = the frame index.

$x \in \{f, s\ , t\}$ denotes the projection view and map_x represents the projection map [4]. Figure 4 represents an example of DMMs for a tennis serve action video sequence

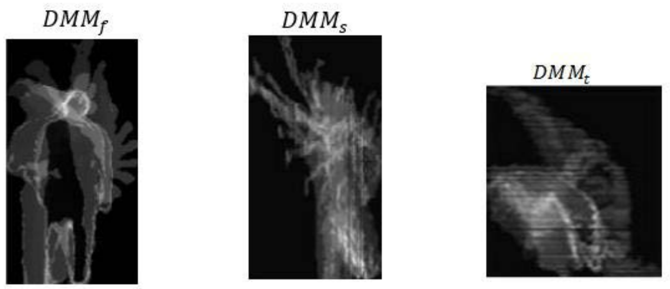


Figure 4: DMMs for a tennis serve action video sequence

3.2 Feature Extraction

We used Pyramid Histogram of oriented gradients (HOG) and Histogram of oriented gradients (PHOG) descriptor to extract shape features.

3.2.1 Histogram of oriented gradients (HOG)

The histogram of oriented gradients (HOG) is a hugely popular object descriptor. It has been shown to perform unpredictably well in human detection in still images as well as videos [35].The technique counts occurrences of gradient orientation in localized portion of an image-detection window [10].

The main goal of histogram of oriented gradients descriptor is to describe the local object shape feature within an image by the allocation of intensity gradients or edge directions. The gradient is a directional change in image intensity/color or measure of change in image function. The image is divided into small connected regions called cells, and for each cell a histogram of the centered horizontal and vertical gradient directions be computed for each pixel within the cell. Groups of neighboring cells are called blocks considered as spatial regions. Each cell is separated into angular bins according to the gradient orientation. Depending on the gradient magnitude is positive or negative, the histogram bins are equally distributed over 0 – 180 or 0 – 360. Finally, the edge orientations are quantized and the histograms of each block are normalized to compensate for brightness variation. Normalized group of histograms represents the block histogram. Then, the set of all normalized histograms obtained from all blocks represents the HOG descriptor [36]. The HOG on depth motion maps tennis serve is shown in figure 5.

Let, $f(X,Y)$ is the image function. Where, X represents the horizontal direction and Y represents the vertical direction of cells.

Centered derivative:

$$f' = \lim_{h \to 0} \frac{f(x+h)-f(x)}{h} \qquad (2)$$

$$f_x = \frac{\partial f}{\partial x} = f(x+1) - f(x) = \text{Change of intensity in X} \quad \text{direction.} \qquad (3)$$

$$f_y = \frac{\partial f}{\partial y} = f(y+1) - f(y) = \text{Change of intensity in Y direction.} \qquad (4)$$

Gradient Vector:

$$\nabla f = \left[\frac{\partial f}{\partial x}, \frac{\partial f}{\partial y}\right] = f_x\,\hat{\imath} + f_y\,\hat{\jmath} \qquad (5)$$

Gradient Magnitude:

$$\|\nabla f\| = \sqrt{f_x^2 + f_y^2} \qquad (6)$$

Gradient orientation:

$$\theta = tan^{-1}\left(\frac{f_y}{f_x}\right) \qquad (7)$$

Normalization:

$$\hat{u} = \frac{\nabla f}{|\nabla f|} \quad [37] \qquad (8)$$

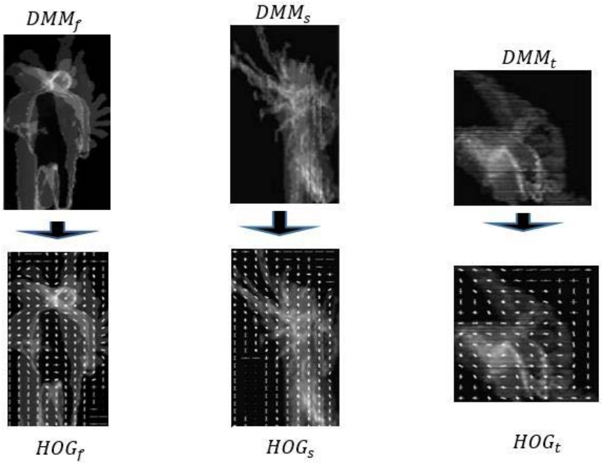


Figure 5: HOG on depth motion maps of tennis serve

3.2.2 Pyramid Histogram of oriented gradients (PHOG)

The pyramid of histogram of orientation gradients (PHOG) features are used to represent spatial shape descriptor. PHOG was proposed by Bosch et al. [38] and has been capably used in object classification. PHOG descriptor is used to represent an image by its local and global information of the shape. Pyramid histogram of gradients (PHOG) is an extension to HOG features. To find the PHOG feature image is partitioned into a sequence of small area of different resolutions by repeatedly doubling the division of area of interest at each level of the pyramid. Figure 6 shows the several pyramid levels. The small area of different levels called cells and a HOG feature vector are computed for every cell of different level. The combination of these HOG feature vectors represents the final PHOG descriptor. Hence, PHOG represents both edge direction and location. The histograms of oriented gradients (HOG) is mainly motivated the technique for extracting PHOG descriptor of Dalal and Triggs [39] to improve accuracy. Along both directions of 2D axis the image at level n is split into 2_n cells to build the pyramid. Therefore, level 0 is described by a

K-vector analogues to the K bins of the histogram, level 1 by a $4K$ -vector and so on. Hence, the PHOG descriptor of the entire image is a vector of size $K * \sum_{n \in N} 4^n$.

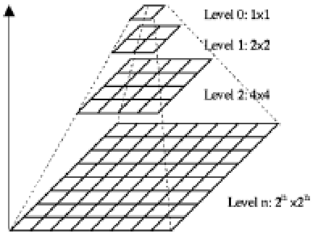


Figure 6: The several pyramid levels [47]

For example, for levels up to N = 3 and K = 9 bins the PHOG descriptor dimension will be $(9 * \Sigma 4_n)$ 3 $_{n=0} = 765$. For any application, it is necessary to limit the number of levels of the pyramid to N = 3 to prevent over fitting [40].

3.3 Action classification with $\boldsymbol{l}_2$-regularized Collaborative Representation Classifier ($\boldsymbol{l}_2$- CRC)

$\boldsymbol{l}_2$-CRC is widely used classifier to recognize human action [9, 22]. To explain $\boldsymbol{l}_2$-CRC in details, with C classes let us consider a data set. An over-complete dictionary $S = [S_1, S_2, \ldots\ldots\ldots, S_C] = [s_1, s_2, \ldots\ldots\ldots, s_n] \in \mathbb{R}_{d\times n}$ can be obtained by arranging the training samples column wise.

Where,

$\boldsymbol{S}_j \in \mathbb{R}_{d\times m_j}, (j = 1,2, \ldots\ldots\ldots, C)$ = Subset of the training samples related to class

$\boldsymbol{s}_i \in \mathbb{R}_d, (i = 1,2, \ldots\ldots\ldots, n)$ = Single training sample

d = dimension of training samples

n = total number of training samples from all classes

Using the matrix $\boldsymbol{S}$, let $\boldsymbol{X} \in \mathbb{R}_d$ can be expressed as any unknown sample. Which can be formulated as,

$$x = \boldsymbol{S}\beta, \tag{9}$$

Where, $\beta = [\beta_1, \beta_2, \ldots\ldots\ldots, \beta_j]$ is $n \times 1$ vector of coefficients analogous to all the training samples and β_j $(j = 1,2, \ldots\ldots\ldots, C)$ is the subset of the coefficients related with the training samples from the class j.

Effectively, one can not directly solve Equation (9) since it is naturally under-determined. Then the solution is obtained by solving the following norm minimization problem.

$$\hat{\beta} = \arg\min_{\beta}\{\|x - S\beta\|_2^2 + \alpha\|M\beta\|_2^2\}, \tag{10}$$

Where, α= the regularization parameter

$\boldsymbol{M}$= The Tikhonov regularization matrix [41]

By applying the approach described in [42-44], the term related with $\boldsymbol{M}$ approve the assessment of preceding knowledge of the solution. Where, the weight of the training samples which are extremely different from a test sample are less than the training samples that are extremely similar. Particularly, the form of the matrix $\boldsymbol{M}\in\mathbb{R}^{d\times n}$ is represented as follows:

$$M = \begin{pmatrix} \|x - s_1\|_2 & 0 & \cdots & 0 \\ 0 & \|x - s_2\|_2 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \cdots & \|x - s_n\|_2 \end{pmatrix} \tag{11}$$

According to [45] the coefficient vector β is calculated as follows,

$$\hat{\beta} = (\boldsymbol{S}^T\boldsymbol{S} + \alpha\boldsymbol{M}^T\boldsymbol{M})^{-1}\,\boldsymbol{S}^T\boldsymbol{V} \tag{12}$$

^

$\hat{\beta}$ can be partitioned into C subsets $\hat{\beta} = [\hat{\beta}_1; \hat{\beta}_2; \dots \dots; \hat{\beta}_C]$ with $\hat{\beta}_j(j = 1,2, \dots \dots C)$

By using the class labels of all the training samples. After portioning β the class label of the unknown sample $\boldsymbol{x}$ is then calculated as follows:

$$class(\mathbf{x}) = \arg\min_{j\in\{1,2,\dots,C\}}\{e_j\} \tag{13}$$

Where, $e_j = \|x - S_j\ \hat{\beta}\|_2$ [4].

4 Experimental Results and Discussion

In this section, we first evaluate our proposed method on publicly available MSR- Action 3D dataset then compare between the results obtained by using HOG and PHOG descriptor

4.1 MSR-Action 3D Dataset & Setup

MSR-Action 3D dataset is an action dataset of depth map sequences recorded by a depth camera. This dataset contains 20 action types performed by 10 different subjects and every subject performs each action 2 or 3 times. There are 557 depth map sequences in total. The resolution of each map is 320x240. The data was a depth sensor similar to the Kinect device. The 20 actions of this datasets are: “High arm wave”, “Horizontal arm wave”, “Hammer”, “Hand catch”, “Forward punch”, “High throw”, “Draw x”, “Draw tick”, “Draw circle”, “Hand clap”, “Two hand wave”, “Side-boxing”, “Bend”, “Forward kick”, “Side kick”, “Jogging”, “Tennis swing”, “Tennis serve”, “Golf swing”, “Pickup & throw” [46].

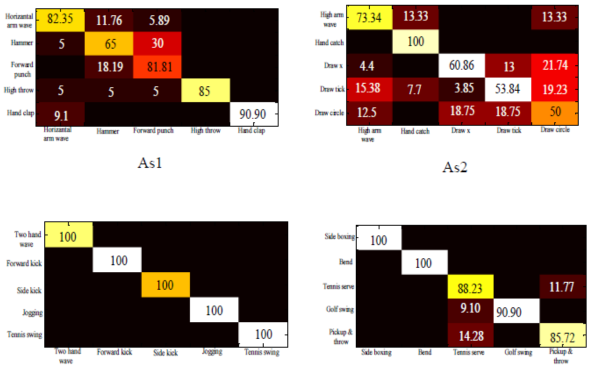
Table 1: Four action subsets of MSR-Action 3D dataset

Action Label	Action set 1 (AS1)	Action Label	Action set 2 (AS2)	Action Label	Action set 3 (AS3)	Action Label	Action set 4 (AS4)
2	Horizontal arm wave	1	High arm wave	11	Two hand wave	12	Side boxing
3	Hammer	4	Hand catch	14	Forward kick	13	Bend
5	Forward punch	7	Draw x	15	Side kick	18	Tennis serve
6	High throw	8	Draw tick	16	Jogging	19	Golf swing
10	Hand clap	9	Draw circle	17	Tennis swing	20	Pickup & throw

This dataset is very challenging dataset. To find out recognition accuracy using these 20 actions together, we have to use a computer with high configuration. But we don't have such computer. So, we divide these 20 actions into four action subsets, those are AS1, AS2, AS3 and AS4 showed in Table 1. Each of them contains 5 actions. Two different test cases were performed on each action subset. For test one, 1/3 samples of each subset are used for training and the remaining samples for test; For test two, 2/3 samples of each subset are used for training and the remaining samples for test.

4.2 Action Classification result

To compute accuracy we use DMMs of size 320x240 to extract HOG & PHOG feature vectors. Dimension of these HOG & PHOG feature vectors are 216 & 2104 respectively. These dimensions are reduced by PCA. So, after dimensionality reduction, the new dimension of HOG & PHOG is 15 & 11. Then final feature vectors are fed into $\boldsymbol{l}_2$-CRC to recognize human action and the key parameter α is a set as α=0.0001 in $\boldsymbol{l}_2$-CRC. We find out our action classification accuracy of sets AS1, AS2, AS3 and AS4 of MSR-Action 3D dataset by using HOG and PHOG shape features and then classification accuracies of the each action set is compared obtained by using HOG and PHOG. We obtained our result using confusion matrix on the test data. Figure (7,8) represents the confusion matrix of action sets for test one and test two using HOG feature descriptor and Figure (9,10) represents the confusion matrix of action sets for test one and test two using PHOG feature descriptor.



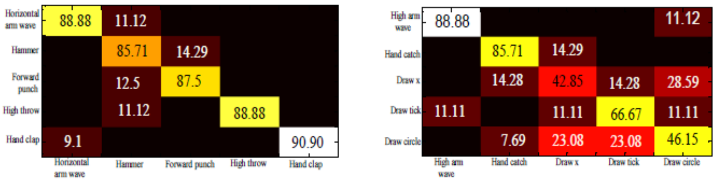
As1

As2

As3

As4

Figure 7: The confusion matrix of action sets for test one using HOG feature descriptor



As1

As2

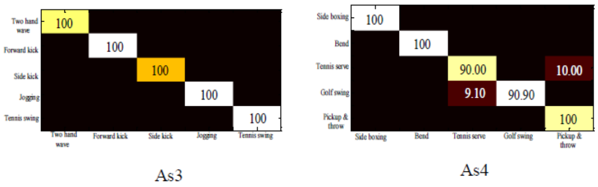


Figure 8: The confusion matrix of action sets for test two using HOG feature descriptor

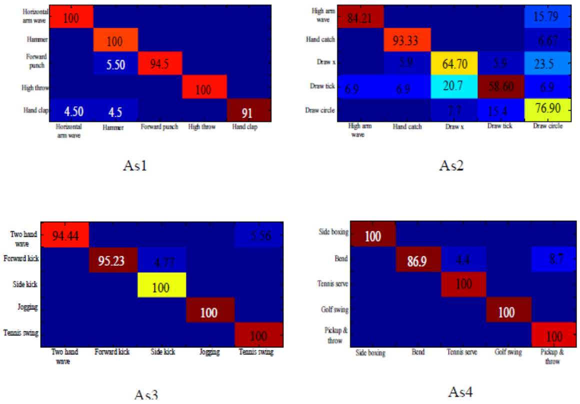


Figure 9: The confusion matrix of action sets for test one using PHOG feature descriptor

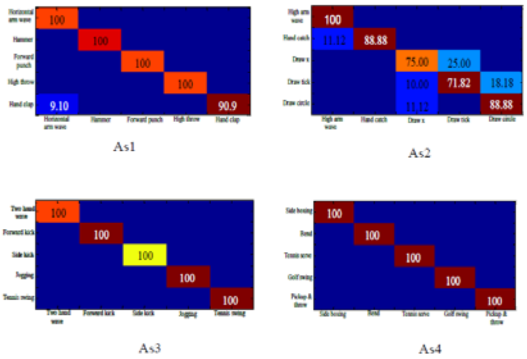


Figure 10: The confusion matrix of action sets for test two using PHOG feature descriptor

We compare between the recognition accuracies obtained by using HOG & PHOG descriptor to find out which one is the best. Figure 11-18 represent the comparison between the accuracies obtained by HOG and PHOG shape feature descriptor.

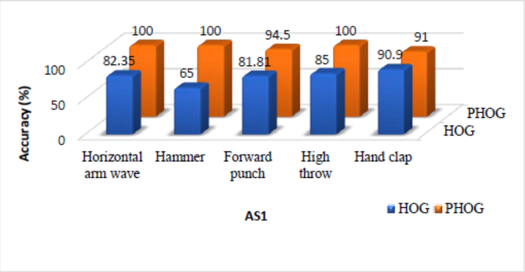


Figure 11: Comparison graph of actions in AS1 for test one using HOG and PHOG

In this case, PHOG is the best feature descriptor for AS1 test one comparing with HOG. Three actions in AS1 gives maximum accuracy for test one using PHOG feature. But the remaining two actions give poor result, because the action Forward punch is confused with Hammer and Hand clap is confused with Horizontal arm wave and Hammer.

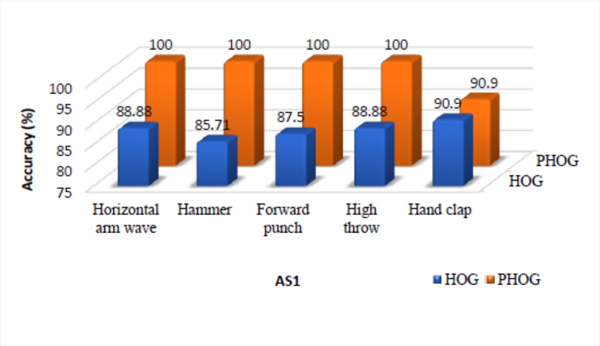


Figure 12: Comparison graph of actions in AS1 for test two using HOG and PHOG feature

In this case, PHOG is the best feature descriptor for AS1 test two comparing with HOG. Four actions in AS1 gives maximum accuracy for test two using PHOG feature. But, the remaining action Hand clap gives poor result, because this action is confused with Horizontal arm wave

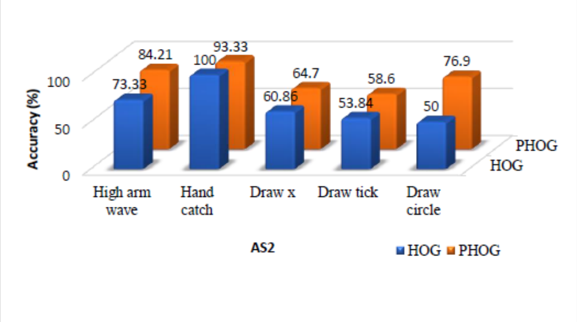


Figure 13: Comparison graph of actions in AS2 for test one using HOG and PHOG

In this case, Although HOG is the best feature descriptor for AS2 test one comparing with PHOG, only one action Hand catch in AS2 gives maximum accuracy for test one. And the remaining four actions gives poor result, because High arm wave is confused with Draw circle, Hand catch and Draw x are confused with each other, Draw x and Draw tick are confused with each other, Draw x is confused with Draw circle, and Draw tick is confused with High arm wave

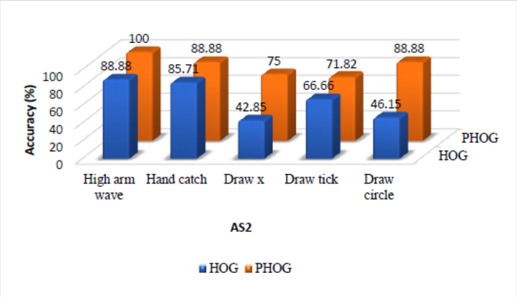


Figure 14: Comparison graph of actions in AS2 for test two using HOG and PHOG feature

In this case, Although PHOG is the best feature descriptor for AS2 test two comparing with HOG, only one action High arm wave in AS2 gives maximum accuracy for test two. And the remaining four actions gives poor result, the action Hand catch is confused with High arm wave, Draw x and Draw tick are confused with each other, Draw tick is confused with Draw circle, and Draw circle is confused with Draw x.

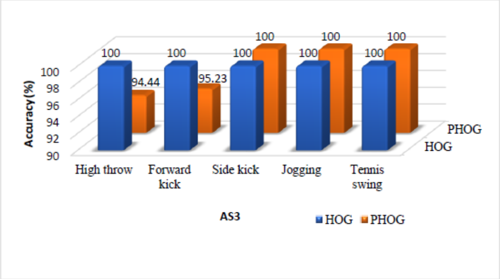


Figure 15: Comparison graph of actions in AS3 for test one using HOG and PHOG feature

In this case, HOG is the best feature descriptor for AS3 test one comparing with PHOG, because all actions in AS3 gives maximum accuracy. But, only three actions give maximum accuracy using PHOG feature

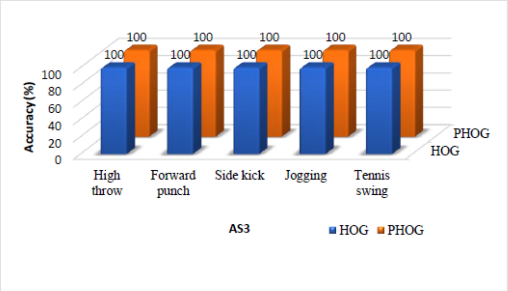


Figure 16: Comparison graph of actions in AS3 for test two using HOG and PHOG feature

In this case, all actions in AS3 gives maximum accuracy for test two using HOG and PHOG feature. So, HOG & PHOG both are the best feature descriptors for AS3 test two comparing each other.

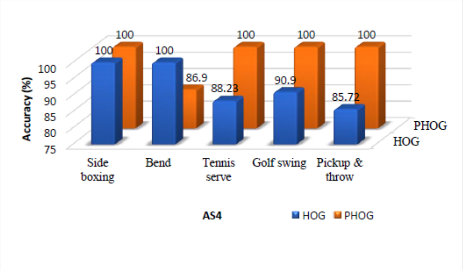


Figure 17: Comparison graph of actions in AS4 for test one using HOG and PHOG feature

In this case, PHOG is the best feature descriptor for AS4 test one comparing with HOG. Four actions in AS4 gives maximum accuracy for test one using PHOG feature. But the remaining action gives poor result, because this action Bend is confused with the action Tennis serve.

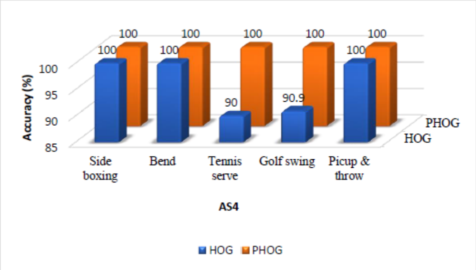


Figure 18: Comparison graph of actions in AS4 for test two using HOG and PHOG feature

In this case, PHOG is the best feature descriptor for AS4 test two comparing with HOG, because all actions in AS4 gives maximum accuracy for test two using PHOG.

5 Conclusion

In this paper, depth map based human action recognition method using Histogram of Oriented Gradients (HOG) and Pyramid Histogram of Oriented Gradients (PHOG) as feature descriptors is introduced. We use HOG because, it gives local feature but we also used PHOG so that we can get both the global and local information. Our experiment has been carried out on MSR-action 3D dataset that both the HOG and PHOG features are extracted from DMMs of every action of four action set AS1, AS2, AS3, and AS4 of MSR-action 3D dataset. The implementation has been done exclusively using MATLAB version R2014a on Lenovo with 8 GB RAM, CORE i5 processor, and win10 system. Then the experimental results obtained by using HOG feature and those of using PHOG feature for AS1, AS2, AS3, and AS4 are compared. Using HOG feature the classification accuracy of AS2 for test one is 100% for one action, AS3 for test one and test two is 100% for all actions. And using PHOG feature the classification accuracy of AS1 for test one is 100% for three actions and for test two is 100% for four actions, AS2 for test two is 100% for one action, AS3 for test two is 100% for all actions and AS4 for test one is 100% for four actions and test two is100% for all actions. But using PHOG, maximum accuracy of AS2 for test one is 93.33% for one action where HOG gives 100% accuracy for one action and AS3 for test one is 100% for three actions where HOG gives 100% accuracy for all actions. So, PHOG with HOG feature, AS2 for test one and AS3 for test one gives the worst result, because in AS2 the action High arm wave is confused with Draw circle, Hand catch is confused with Draw circle, Draw x is confused with Hand catch, Draw tick and Draw circle, Draw tick is confused with High wave, Hand catch and Draw x; and in AS3 the action Two hand wave is confused with Tennis swing, and forward kick is confused with side kick. This confusion occurs because, the local features of subjects can’t be detected properly. Overall, PHOG gives the best result to recognize human action.

REFERENCES

[1] Umakanthan, Sabanadesan. *Human action recognition from video sequences*. Diss. Queensland University of Technology, 2016.

[2] Aggarwal, Jake K., and Michael S. Ryoo. "Human activity analysis: A review." *ACM Computing Surveys (CSUR)* 43.3 (2011): 16.

[3] https://en.wikipedia.org/wiki/RGB_color_model

[4] Bulbul, Mohammad Farhad, Yunsheng Jiang, and Jinwen Ma. "Real-Time Human Action Recognition Using DMMs-Based LBP and EOH Features." *International Conference on Intelligent Computing*. Springer, Cham, 2015.

[5] Farooq, Adnan, and Chee Sun Won. "A survey of human action recognition approaches that use an RGB-D sensor." *IEIE Transactions on Smart Processing & Computing* 4.4 (2015): 281-290.

[6] Chen, Chen, Roozbeh Jafari, and Nasser Kehtarnavaz. "A real-time human action recognition system using depth and inertial sensor fusion." *IEEE Sensors Journal* 16.3 (2016): 773-781.

[7] https://www.hongkiat.com/blog/innovative-uses-kinect/

[8] https://www.researchgate.net/figure/Architecture-of-Microsoft-Kinect-sensor_fig10_282477283

[9] Chen, Chen, Kui Liu, and Nasser Kehtarnavaz. "Real-time human action recognition based on depth motion maps." *Journal of real-time image processing* 12.1 (2016): 155-163.

[10] https://en.wikipedia.org/wiki/Histogram_of_oriented_gradients

[11] http://www.robots.ox.ac.uk/~vgg/research/caltech/phog.htm

[12] ftp://statgen.ncsu.edu/pub/thorne/molevoclass/AtchleyOct19.pdf

[13] http://users.eecs.northwestern.edu/~jwa368/my_data.html

[14] Li, Wanqing, Zhengyou Zhang, and Zicheng Liu. "Action recognition based on a bag of 3d points." *Computer Vision and Pattern Recognition Workshops (CVPRW), 2010 IEEE Computer Society Conference on*. IEEE, 2010.

[15] Vieira, Antonio W., *et al*. "Stop: Space-time occupancy patterns for 3d action recognition from depth map sequences." *Iberoamerican congress on pattern recognition*. Springer, Berlin, Heidelberg, 2012.

[16] Wang, Jiang, *et al*. "Robust 3d action recognition with random occupancy patterns." *Computer vision–ECCV 2012*. Springer, Berlin, Heidelberg, 2012. 872-885.

[17] Yang, Xiaodong, Chenyang Zhang, and YingLi Tian. "Recognizing actions using depth motion maps-based histograms of oriented gradients." *Proceedings of the 20th ACM international conference on Multimedia*. ACM, 2012.

[18] Oreifej, Omar, and Zicheng Liu. "Hon4d: Histogram of oriented 4d normals for activity recognition from depth sequences." *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2013.

[19] Luo, Jiajia, Wei Wang, and Hairong Qi. "Spatio-temporal feature extraction and representation for RGB-D human action recognition." *Pattern Recognition Letters* 50 (2014): 139-148.

[20] Lu, Cewu, Jiaya Jia, and Chi-Keung Tang. "Range-sample depth feature for action recognition." *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2014.

[21] Chen, Chen, Roozbeh Jafari, and Nasser Kehtarnavaz. "Action recognition from depth sequences using depth motion maps-based local binary patterns." *Applications of Computer Vision (WACV), 2015 IEEE Winter Conference on*. IEEE, 2015.

[22] Bulbul, Mohammad Farhad, Yunsheng Jiang, and Jinwen Ma. "Human action recognition based on DMMs, HOGs and Contourlet transform." *Multimedia Big Data (BigMM), 2015 IEEE International Conference on*. IEEE, 2015.

[23] https://en.wikipedia.org/wiki/Contourlet

[24] Chen, C., Liu, M., Zhang, B., Han, J., Jiang, J., & Liu, H. (2016, July). 3D Action Recognition Using Multi-Temporal Depth Motion Maps and Fisher Vector. In *IJCAI* (pp. 3331-3337).

[25] https://en.wikipedia.org/wiki/Local_binary_patterns

[26] Chen, Chen, Kui Liu, and Nasser Kehtarnavaz. "Real-time human action recognition based on depth motion maps." *Journal of real-time image processing* 12.1 (2016): 155-163.

[27] Yang, Xiaodong, and Ying Li Tian. "Eigenjoints-based action recognition using naive-bayes-nearest-neighbor." *Computer vision and pattern recognition workshops (CVPRW), 2012 IEEE computer society conference on*. IEEE, 2012.

[28] Timofte, Radu, Tinne Tuytelaars, and Luc Van Gool. "Naive bayes image classification: beyond nearest neighbors." *Asian Conference on Computer Vision*. Springer, Berlin, Heidelberg, 2012.

[29] Xia, Lu, Chia-Chih Chen, and Jake K. Aggarwal. "View invariant human action recognition using histograms of 3d joints." *Computer vision and pattern recognition workshops (CVPRW), 2012 IEEE computer society conference on*. IEEE, 2012.

[30] Luo, Jiajia, Wei Wang, and Hairong Qi. "Group sparsity and geometry constrained dictionary learning for action recognition from depth maps." *Proceedings of the IEEE international conference on computer vision*. 2013.

[31] Wang, Jin, *et al.* "Human action recognition based on pyramid histogram of oriented gradients." *Systems, Man, and Cybernetics (SMC), 2011 IEEE International Conference on*. IEEE, 2011.

[32] https://en.wikipedia.org/wiki/Hidden_Markov_model

[33] https://en.wikipedia.org/wiki/Conditional_random_field

[34] https://en.wikipedia.org/wiki/Principal_component_analysis

[35] https://www.quora.com/What-is-a-histogram-of-gradient-directions-in-computer-vision

[36] https://scc.ustc.edu.cn/zlsc/tc4600/intel/2017.0.098/ipp/common/ipp_manual/GUID-83B0EE35-E5EF-4C23-96A4-F0918DFA826B.htm

[37] https://www.youtube.com/user/Udacity

[38] Bosch, Anna, Andrew Zisserman, and Xavier Munoz. "Representing shape with a spatial pyramid kernel." *Proceedings of the 6th ACM international conference on Image and video retrieval*. ACM, 2007.

[39] Dalal, Navneet, and Bill Triggs. "Histograms of oriented gradients for human detection." *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*. Vol. 1. IEEE, 2005.

[40] http://www.robots.ox.ac.uk/~vgg/research/caltech/phog.htm

[41] Tikhonov, A., Arsenin, V.: Solutions of ill-posed problems. Math. Comput. 32(144), 1320–1322 (1978)

[42] Chen, C., Tramel, E.W., Fowler, J.E.: Compressed-sensing recovery of images and video using multi-hypothesis predictions. In: Proceedings of the 45th Asilomar Conference on signals, Systems, and Computers, pp. 1193–1198 (2011)

[43] Chen, C., Li, W., Tramel, E.W., Fowler, J.E.: Reconstruction of hyperspectral imagery from random projections using multi-hypothesis prediction. IEEE Trans. Geosci. Remote Sens. 52(1), 365–374 (2014)

[44] Chen, C., Fowler, J.E.: Single-image Super-resolution Using Multi-hypothesis Prediction. In: Proceedings of the 46th Asilomar Conference on Signals, Systems, and Computers, pp. 608–612 (2012)

[45] Golub, G., Hansen, P.C., O'Leary, D.: Tikhonov-regularization and total least squares SIAM J. Matrix Anal. Appl. 21(1), 185–194 (1999)

[46] http://users.eecs.northwestern.edu/~jwa368/my_data.html

[47] http://fourier.eng.hmc.edu/e161/lectures/canny/node2.html