OPEN ACCESS
Saudi Journal of Engineering and Technology
Abbreviated Key Title: Saudi J Eng Technol
ISSN 2415-6272 (Print) |ISSN 2415-6264 (Online)
Scholars Middle East Publishers, Dubai, United Arab Emirates
Journal homepage: http://scholarsmepub.com/sjet/

Original Research Article

Shape and Texture Features based Human Action Recognition Using Collaborative Representation Classification

Lasker Ershad Ali[1*], Md. Zahidul Islam[1], Biplab Madhu[1], Md. Farhad Bulbul[2] and Nazma Parveen[3]

[1]Mathematics Discipline, Khulna University, Khulna-9208, Bangladesh
[2]Department of Mathematics, Jessore University of Science and Technology Chaugachha Road, 7408, Bangladesh
[3]BISC, DOHS, Mohakhali, Dhaka Bangladesh

DOI:10.21276/sjeat.2019.4.7.2 | **Received:** 17.07.2019 | **Accepted:** 24.07.2019 | **Published:** 30.07.2019

***Corresponding author:** Lasker Ershad Ali

Abstract

This paper presents human action recognition by using shape and texture based DMM-Haar features where collaborative representation classifier is adopted for action classification. In this study, we have introduced effective feature extraction technique based on Depth Motion Maps (DMMs) and Haar wavelet transformation, where different actions can be represented with a range of features. Firstly, we have calculated three DMMs such as DMM front view, top view and side view from 3D action video sequences as the shape features. After that, we have utilized Haar wavelet on the DMMs generated images to extract texture information and concatenated all features as a feature matrix. We have utilized principal component analysis for reducing the feature dimensions of the feature matrix. Finally, l_2 normed based collaborative representative classification technique is adopted to classify different actions. For this research, we have analyzed the effects of the DMM-Haar features on experimental basis with DMM features based results. The performance study of the proposed method is comparable with the state-of-the-art methods to recognize human action on the publicly available Microsoft Research Action 3D dataset.

Keywords: Human action recognition, Depth motion map, Haar wavelet, Collaborative representation classification, MSR Action 3D.

INTRODUCTION

Human action recognition (HAR) is an interesting and effective research topic in the computer vision. There are many powerful applications in this research field, such as video analysis, surveillance systems, home care for elderly people, children and physically challenged people to prompt timely assistance, human computer interaction or robotics, computer gaming and so on [1]. Human action usually refers to a sequence of movements carried out by a person.

In the previous, research has generally based on learning and recognizing actions from image sequences taken by visible light cameras. For this reason, there are many latent limitations of this source of data type, it is impressible to color and revelation change of action, confinement, and also background clutters. But in the present time, as imaging technology improvements so that it has become possible to capture depth information. Compared with conventional images, depth maps are uncontrolled to changes in lighting conditions and can provide 3D information toward distinguishing that are difficult to characterize using conventional images. Figure 1 shows two examples consisting of nine depth maps of the action golf swing and action forward kick. Many research works have been carried out on human action recognition using depth images. Depth image means an image in which each pixel relates to a distance between image plane and the corresponding object in the RGB image. To get perspective computational efficiency, the problem of human action recognition from depth map sequences is to be examined. These images are captured by an RGBD camera. Particularly, the depth motion maps (DMMs) generated by accumulating motion energy of projected depth maps in three projections such as front view, side view, and top view and then used Haar wavelet which is used as feature descriptors. Compared with 3D depth maps, DMMs are 2D images that provide an encoding of motion characteristics of an action. Motivated by the success of sparse representation in face recognition [2-3] and image

classification [4], l_2 -regularized collaborative representation classifier is utilized which seeks a match of an unknown sample via a linear combination of training samples from all the classes.

In this paper, the main aim is to recognize human actions by using DMMs based Haar wavelet features and l_2-CRC classifier from the original depth images sequences so that we can achieve better recognition result as well as computational efficiency.

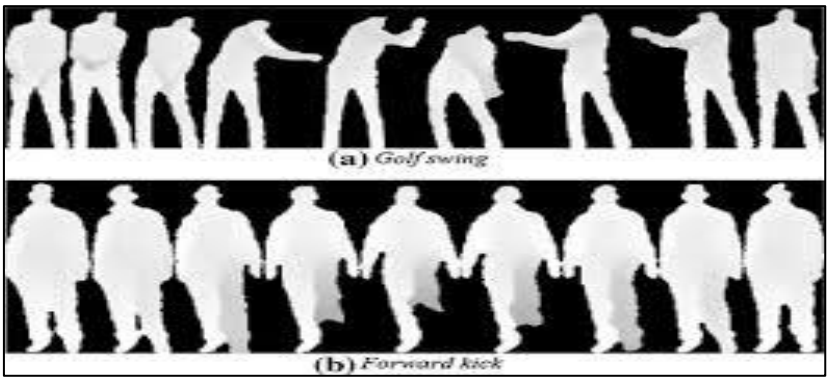

Fig-1: Depth map sequences for (a) Golf swing and (b) Forward kick action sequences

LITERATURE REVIEW

For the modern technology, a large number of techniques have been displayed to recognize human actions which are based on different methods such as position-time volumes, spatio-temporal features and trajectories. All these methods have widely distributed for human action recognition from video sequences captured by RGBD cameras. In addition, some recent works for human action recognition from depth video sequences have been discussed, including the wavelet features as well as features parameters. To obtain a good framework for human action recognition, spatio-temporal interest points coupled with an SVM classifier was proposed in [5]. Furthermore, Cubold feature descriptor was appointed for human action recognition [6]. In [7], SIFT-feature trajectories modeled into a hierarchy of three abstraction levels were adopted to recognize action from video sequences. See the work done by Chen, *et al*. for real-time human action recognition based on depth motion maps [1].

In the earlier, a large number of low quality features have been proposed for human action recognition. Li *et al*. proposed the most important framework for human action recognition from depth video sequences in 2010 [8]. They developed an action graph to build the dynamic model of actions and a collection of 3D points were used to characterize attitudes. In 2012, Viera *et al*. developed 3D action sequences based 4D shapes with random occupancy pattern (ROP) and sparse coding was fulfilled for further development of their proposed approach. They were used position-time control patterns as features which reserved local and temporary episodic information copying with intra-class variations and then adopted simple classifier based on the cosine distance for action recognition [9]. In 2013, Yang *et al*. calculated DMMs based HOG features and fed them into SVM classifier for recognizing 3D actions [10]. In the same year, Oreifej and Liu presented histogram of the oriented normal 4D surface (HON4D) as a new feature descriptor [11].

For the purpose of improving the accuracy of human action recognition, many researchers proposed different method which is based on feature extraction from the depth maps and RGB video sequences and classification techniques. Say for example, Luo *et al*. removed 3D joint features for each depth video and enhanced Central-Symmetric Motion Local Ternary Pattern (CS-Mltp) to extract both spatial and temporal features of the RGB sequences [12]. Various parameters have been used in that method. A large range of values, they proposed for each parameter instead of fixed value. So, many experiments should be conducted to get a relevant value for the corresponding parameter which is depends on each range that was an obstacle to the application of their proposed algorithm.

METHODOLOGY

In this research, we have proposed an effective feature extraction method by adopting DMM shape features and texture based Haar features for human action recognition. l_2-CRC is the basic classifier for this proposed framework. The details of the proposed framework are depicted in the given flowchart.

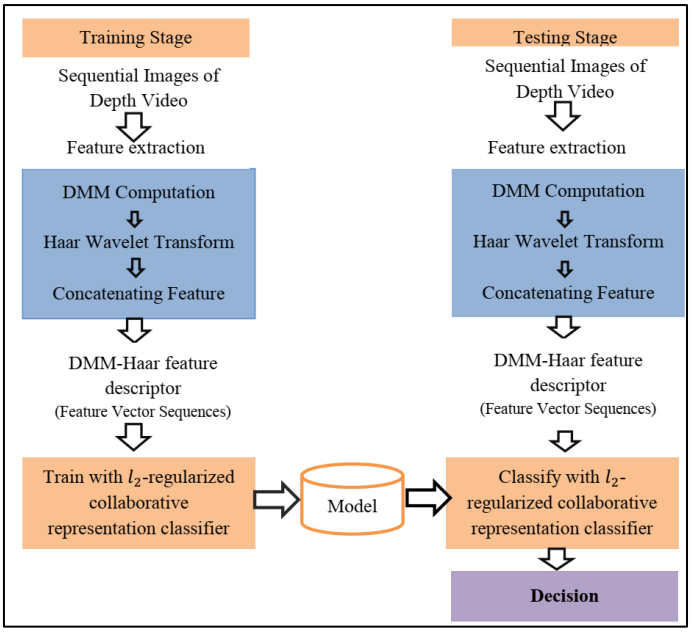

Fig-2: Proposed Framework

The whole framework can be separated by training stage and testing stage. For both training and test stages, we can compute the shape features from the sequential images of depth video as depth motion maps. After that, Haar wavelet information is captured from the depth images as the texture features. Finally, the concatenation of all shape feature based texture features are completed as the feature matrix for l_2 - collaborative representation classification. Shape feature based DMM computation; texture feature extraction and information from the depth video sequences. Yang et al. proposed depth frames onto three orthogonal Cartesian planes for the purpose of characterizing the motion of an action [10]. Due to the computational simplicity, the same approach is adopted in this work while the procedure to obtain DMMs is modified. More specifically, each 3D depth frame is used to produce three 2D projected maps corresponding to front (f), side (s), and top (t) views, denoted by map_v where, $v \in (f, s, t)$. For a point (x, y, z) in a depth frame with z denoting the depth value in the coordinate system, the pixel value in three projected maps is indicated by x and y respectively. For each projected map, the motion energy is calculated as the absolute difference between two consecutive maps without threshold. For a depth video sequence with N frames, DMM_v is obtained by stacking the motion energy across an entire depth video sequence as follows:

$$DMM_v = \sum_{i=a}^{b} \left| map_v^i - map_v^{i-1} \right| \quad (1)$$

where, i represents the frame index; map_v^i is the proposed map of the i^{th} frame under projection view $v; a \in \{2,3,...,N\}$ and $b \in \{2,3,...,N\}$ denote the starting frame and the end frame index. It should be noted that not all the frames in a depth video sequence are used to generate DMMs, a bounding box is then set to extract the non-zero region as the foreground in each DMM.

Texture Features Extraction

To generate DMM-Haar features, we have decomposed each generated 2D proposed maps corresponding to front, side, and top views by using Haar wavelet with first label decomposition. Actually, the Haar wavelet transform is

Shape Features based DMM Computation

Feature is an element or group of elements that establish a characteristic property or set of properties which is unique, assessable and differentiable. In this study, depth motion maps are used to capture the 3D structure and shape collaborative representation classification are briefly discussed in the following subsections.

one of most important part of the wavelet transforms which can capture the texture information from an image or a set of images [13]. Haar wavelets family for $t \in (0,1)$ is defined by

$$P(t) = \begin{cases} 1 & \text{if } 0 \le t < \frac{1}{2} \\ -1 & \text{if } \frac{1}{2} \le t < 1 \\ 0 & \text{otherwise} \end{cases} \quad (2)$$

and scaling function of $q(t)$ can be written as:

$$q(t) = \begin{cases} 1 & \text{if } 0 \le t < 1 \\ 0 & \text{otherwise} \end{cases} \quad (3)$$

After decomposition, we get 4 sub-band images for each 2D depth image and corresponding four features vectors. Then we have concatenated four feature vectors into a column vector for each DMM_v. DMM-Haar generated features from the tennis serves video sequences are shown in Fig.3. DMM-Haar from the three projection views effectively captures the characteristics of the motion in a distinguishable way. For this reason, we can use DMMs as feature descriptors for action recognition.

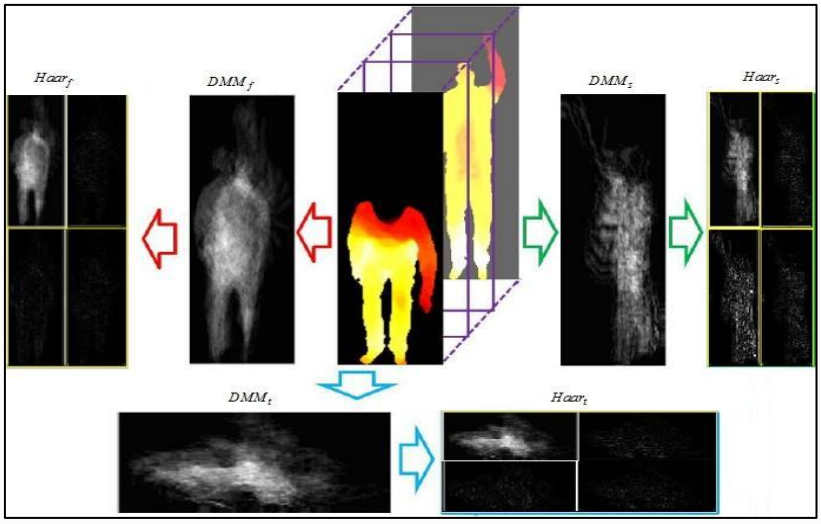

Fig-3: Proposed DMM-Haar feature extraction framework

Collaborative Representation Classification

Classification is a process to identify the class of a set of data that categories a new observation and also the process in which ideas and objects are predictable, distinguished and assumed. In this research, we have classified the data by using l_2-collaborative representation classifier (l_2-CRC). The classification process of the data using the following classifier is given below:

In machine learning, sparse representation (or sparse coding) has been an active research area due to its sample according to a small number of atoms sparsely picked out of an over-complete dictionary made by all the available training samples. Consider a dataset with M classes of training samples arranged column-wise $A = [A_1, A_2, ...A_M] \in R^{d \times n,}$, where, $A_j \ (j = 1, 2, ..., M)$ is the subset of the training samples associated with class *j*, *d* is the dimension of training samples and *n* is the total number of training samples from all the classes.

A test sample $g \in R^d$ can be represented as a sparse linear combination of the training samples, which can be formulated as

$$g = A\delta \quad (4)$$

Where $\delta = [\delta_1, \delta_2, ..., \delta_M]$ is a vector whose coefficients corresponding to all training samples and $\delta_j \left(j = 1, 2, ..., M \right)$ denotes the subset of the coefficients which is associated with the training samples from the j^{th} class. From a particular stand-point, one cannot directly solve for δ since (2) is usually under-determined [13]. The following norm minimization problem is used to reach a solution.

$$\hat{\delta} = \overset{\arg\min}{\delta} \left\{ \left\| g - A\delta \right\|_2^2 + \theta \left\| \delta \right\|_1 \right\} \tag{5}$$

where, θ is a scalar regularization parameter which branches the influence of the residual and sparsity term. The class label of *g* is then obtained via

$$class\left(g \right) = \overset{\arg\min}{j} e_j \tag{6}$$

where $e_j = \left\| g - A_j \hat{\delta}_j \right\|_2$. Each depth video sequence generates a feature vector $h \in R^{d \times n}$, therefore, the dictionary is $A = [h_1, h_2, ..., h_k]$ with K being the total number of available training samples from all the action classes.

Let $y_q \in R^{d \times n}$ denote the feature vector of an unknown action sample. Tikhonov regularization [8] is employed here to calculate the co-efficient vector according to

$$\hat{\delta} = \overset{\arg\min}{\delta} \left\{ \left\| y_q - A\delta \right\|_2^2 + \lambda \left\| L\delta \right\|_2^2 \right\} \tag{7}$$

where, *L* denotes Tikhonov regularization matrix and k is the regularization parameter. The term *L* allows the hassle of prior knowledge on the solution. Generally, *L* is considered as diagonal matrix *L* which is the following form:

$$\begin{bmatrix} \left\| y_q - h_1 \right\|_2 & & 0 \\ \cdots & \cdots & \cdots \\ 0 & & \left\| y_q - h_k \right\|_2 \end{bmatrix} \tag{8}$$

where, δ denote the coefficient vector which is calculated as follows [17]:

$$\hat{\delta} = \left(A^T A + \lambda L^T L \right)^{-1} A^T y_q \tag{9}$$

Finally, the class label for each unknown sample is then found from the equation (6).

EXPERIMENTAL RESULTS AND DISCUSSIONS

In this paper, we collected data from the publicly available MSR-Action3D dataset [8] which consists of 20 different action categories performed by 10 subjects. These subject are high wave, horizontal wave, hammer, hand catch, forward punch, high throw, draw x, draw circle, hand clap, two hand wave, side boxing, bend, forward kick, side kick, jogging, tennis swing, tennis serve, golf swing, and pick up throw. For the experimental evaluation and comparison with the other state-of-the-art research, we have divided the action set into three action subsets which are listed in the following table:

Table-1: Three action subsets which are used in our experiments

Action Set 1 (AS1)	Action Set 2 (AS2)	Action Set 3 (AS3)
Horizontal wave	High wave	High throw
Hammer	Hand catch	Forward kick
Forward Punch	Draw x	Side kick
High throw	Draw tick	Jogging
Hand clap	Draw circle	Tennis swing
Bend	Two hand wave	Tennis serve
Tennis serve	Forward kick	Golf swing
Pickup throw	Side boxing	Pickup throw

The size of each depth maps is 320×240. There are three different tests for each subset, i.e, test one, test two, and cross-subject test. In test one, 1/3 of the subset is used as the training and rest of the subjects for testing, in test two, 2/3 of the subset is used as training and the rest as testing, for cross subject test, half subjects are used for training and the rest ones used for testing. The principal component analysis (PCA) is applied to reduce the dimensionality for the training feature set and the test feature set. We calculate the PCA transform matrix by using the training feature set and then applied to the test feature set. We have used this dimensionality reduction step to get computational efficiency for the classification. In, l_2 -regularized collaborative representation classifier, a key parameter is λ which controls the relative effect of the Tikhonov regularization term in the optimization stated in equation (6). To find an optimal value of λ, we can examine a set of values for 5-fold cross validation which give us more accurate results in test data set.

We have compared the proposed method with some other existing methods on the Microsoft Action3D dataset. The comparison results are shown in Table 2. In this table, the bold faces indicate the best recognition results. The average recognition rates of our method outperform all the other methods except cross subject test. In cross subject test, the performance of Vieira *et al*. is relatively better than all other methods. We have noticed that Vieira *et al*. used several parameters for the proposed dictionary learning, and the values of all parameter were not fixed. In fact, our proposed method has improved the recognition accuracy compared with some previous scholarly articles.

Table-2: Recognition rates (%) comparison of fixed tests for MSR-action3D dataset

	Li, *et al*. [8]	**Vieira, *et al*. [9]**	**Yang, *et al*. [10]**	**Our method**
Test one				
AS1	89.5	**98.2**	94.7	96.0
AS2	89.0	94.8	95.4	**95.5**
AS3	96.9	97.4	97.3	**98.6**
Average	91.6	**96.8**	95.8	**96.8**
Test two				
AS1	93.4	**99.1**	97.3	98.6
AS2	92.9	97.0	**98.7**	98.6
AS3	96.3	98.7	97.3	**99.9**
Average	94.2	98.3	97.8	**99.3**
Cross subject test				
AS1	72.9	**84.7**	74.5	84.0
AS2	71.9	**81.3**	76.1	70.7
AS3	79.2	88.4	**96.4**	80.3
Average	74.7	**84.8**	82.3	78.3

All other methods, they mentioned a range of values for each parameter instead of a fixed value. So, it's very clear that we should conduct the experiment on each range for many times to find a suitable value for the corresponding parameter, which is an obstacle to the application of their algorithm. Moreover, if we don't use fixed values for all parameters (the DMM size, the number of decomposition levels and the selection of the five sub-bands), our method is also capable of showing expected higher performance, like the method reported in [9]. The average results for three tests of three methods as well as our proposed method are ploted in following Fig. 3.

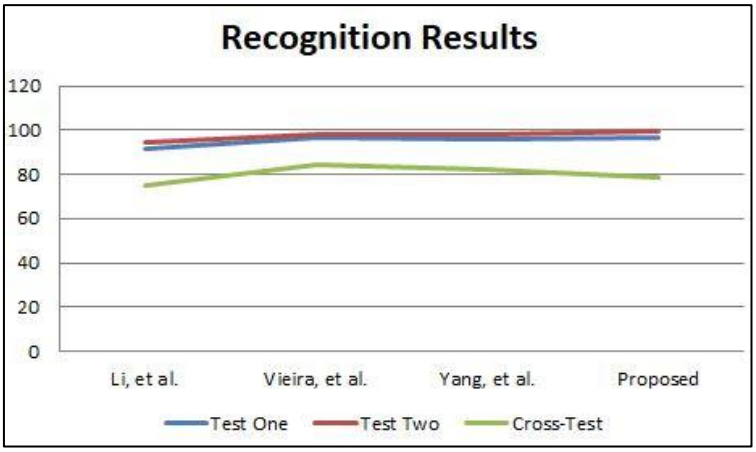

Fig-3: Average Recognition Results

From the above Fig. 3, we can observe that, for test one our recognition result is better than Li, *et al*. and Yang, *et al*. but same as the Veira, *et al*. because they used space-time occupancy patterns for recognition. In test two, our recognition result is better than all other existing method.

CONCLUSIONS

We have examined the human action recognition problem in our paper and established a modified feature descriptor called DMM-Haar feature extraction. The performance study of the proposed method is relatively better than the state-of-the-art methods to recognize human action on the publicly available Microsoft Research Action 3D dataset for test one and test two only. In future, we will try to improve the recognition accuracies for cross subject test by using different feature extraction or classification techniques.

REFERENCES

1. Chen, C., Liu, K. & Kehtarnavaz, N. (2013). Real-Time Action Recognition Based on Depth Motion Maps. Journal of Real-Time Image processing, 12(1), 155-163.
2. Wright, J., Yang, A., Ganesh, A. Sastry, S. & Ma, (2009). Robust face recognition via sparse representation. IEEE Transaction on Pattern Analysis and Machine Intelligence, 31(2), 210-227.
3. GAO, S., Tsang, I. W-H. & Chia, L. (2010). Kernel sparse representation for image classification and face recognition. In Proceedings of IEEE European Conference on Computer Vision, 1-14.
4. Ni, B., Wang, G. & Moulin, P. (2013). Rgbd-Hudaact: A Color-Depth Video Database for Human Daily Activity Recognition. In Proceedings of the Consumer Depth Cameras for Computer Vision, 193-208.
5. Schuldt, C., Laptev, I. & Caputo, B. (2004). Recognition human actions: a local SVM approach. In Proceedings of IEEE International Conference on Pattern Recognition, 3, 32-36.
6. Dollar, P., Rabaud, V., Cottrell, G. & Belongie, S. (2005). Behaviour recognition via sparse spatio-temporal features. IEEE International Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance, 65-72.
7. Sun, J., Wu, X., Yan, S., Cheong, L. F., Chua, T. & Li, J. (2009). Hierarchical spatio-temporal context modeling for action recognition. In Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition, 2004-2011.
8. Li, W., Zhang, Z. & Liu, Z. (2010). Action Recognition Based on a Bag of 3D Points. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, 9-14.
9. Viera, A. W., Nascimento, E. R., Oliveira, G. L., Liu, Z. & Campos, M. M. (2012). STOP: Space-Time Occupancy Patterns for 3D Action Recognition from Depth Map Sequences, In Proceedings of the progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications, Springer, 7441, 252-259.
10. Yang, J., Yu, K., Gong, Y. & Huang, T. (2009). Linear spatial pyramid matching using sparse coding for image classification. In Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 1794-1801.
11. Oreifej, O. & Liu, Z. (2013). Hon4d: Histogram of oriented 4d normal for activity recognition from depth sequences. In Proceedings of the 2013 IEEE Conference on Computer Vision and Pattern Recognition, 716-723.
12. Luo, J., Wang, W. & Qi, H. (2014). Spatio-Temporal Feature Extraction and Representation for RGB-D Human Action Recognition. Pattern Recognition Letters, 139-148.
13. Stuzik, Z. R. & Seibes, A. (1999). The Haar Wavelet Transform in the Time Series Similarity Paradigm. In Proceeding of European Conference on Principles of Data Mining and Knowledge Discovery, 12-22.