A Robust Invariant Feature Descriptor for Human Action Recognition

Rafiat Oluwatosin Omisore, Olatunji Mumini Omisore*, and Abebe Yirga Alemu

***Abstract*—The dramatic progressions of studies in human action recognition are being attributed to inherent challenges of initial methods such as bag-of-words. As a result, researchers in the field of computer vision are still making efforts towards achieving structured interpretation of complex activities between multiple objects. This study proposes a new feature descriptor in Human Action Recognition (HAR). We connected images from Depth Motion Maps to Pyramid of Histogram Oriented Gradients (PHOG) for activity recognition. Different view orientations from depth maps were computed on three Cartesian planes, while PHOG is used to reflect their local shape and spatial layout. In our experiments, l_2-regularized CRC is adopted for activity classification. Performance evaluation shows that our novel Robust Invariant Feature Descriptor (RIFD) for activity recognition achieves better results.**

***Index Terms*— Human Action Recognition; Invariant Feature Descriptor; Depth Motion Maps; PHOG; MSR Daily Activities.**

I. INTRODUCTION

Human Action Recognition (HAR), simply described as labeling unknown actions, has gained much attention in the area of computer vision [1, 2]. The quest for HAR is spectacular in different research areas such as video analysis, human robotics training, and intelligent surveillance, to reflect more impacts for qualitative human lives [3]. Moreover, studies have focused more on recognizing actions in video streams [1]. In initial practice, video streams were captured from single RGB cameras that are only capable of representing agents as 2D images. Such image sequence possesses similar patterns which makes it difficult to classify actions in it. In addition, RGB camera produces intensity-based video images that are sensitive to lighting conditions and background clutters; these limit robustness of recognition results. These motivated design of 3D acquisition devices, such as Intel RealSense and Microsoft Kinect cameras [4, 5], lowered cost devices that considers 3D information from Depth Motion Maps (DMMs) for performance improvement. An example of depth map sequence is MSR Daily Activity 3D dataset shown in Fig. 1.

In the meantime, different methods were proposed for human action recognition. A carefully analysis shows that

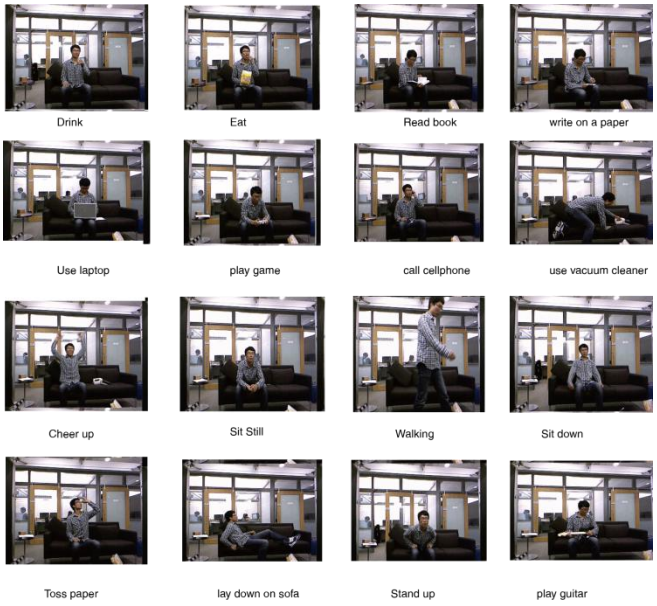

Fig. 1: MSR daily activity 3d dataset

shape-based features were commonly considered in activity recognition model, this could be due to low cost associated with extracting such features from video streams. However, since shape and geometry information of objects are truly reflected in depth maps, other features can be useful in discriminating and enhancing images for segmentation and object detection in video stream. Recently, Histogram of Oriented Gradients (HOG), and visual cues feature descriptors have been proposed to characterize human shapes in images, for instance, Dalal&Triggs [6] used HOG for object detection in image processing. Basically, HOG counts occurrences of gradient orientation in localized image, and it has been proved to outperform other descriptors in encoding human figures for action detection [1]. Also, Contourlet, Random, and Ridgelet transformation have been introduced to provide multi-scale and multi-directional image representations with edges and contours in images [7, 8]. In recent studies, these methods provide results on hybridization [9, 10].

Pyramid HOG (PHOG) is another descriptor employed to effectively encode local image shape of an object together with its spatial layout retrieval [11]. PHOG uses spatial pyramid kernel to characterize human figures for action recognition without requirement for extraction of silhouettes or contours thus, it reduces computation complexity requirement. The descriptor makes use of histogram of gradients on edge points at different spatial levels to generate a final descriptor.

***Rafiat Oluwatosin Omisore**,Department of Civil and Environmental Engineering, Federal University of Technology Akure,Nigeria (e-mail: rafiatagboola@aol.com;Phone:+2348038606655);*
***Olatunji Mumini Omisore**, University of Chinese Academy of Sciences, Beijing, China, (e-mail: tsorewilly@yahoo.com; Phone: +8613172482240);*
***Abebe Yirga Alemu**, University of Chinese Academy of Sciences, Beijing, China, (e-mail: yirga2007@gmail.com; Phone: +8613120030693);*

Previous works show that combining new features with shape information as feature descriptor could yield a better performance. Hence in this study, we propose a novel Robust Invariant Feature Descriptor (RIFD) for human action recognition using DMM and PHOG. The rest of this paper is structured that: Section 2 presents related works in HAR; Section 3 presents the novel RIFD for HAR, also with description of l2-regularized collaborative representation classifier; Experiments ran to evaluate the proposed method is presented in Section 4 with evaluation results; Finally, section 5presents some conclusive remarks and describes direction of future works.

II. RELATED WORKS

Studies in HAR began on video sequences captured with traditional RGB video cameras and local spatio-temporal features [8]. These methods were mainly based on low-level features extracted from depth images. Spatio-temporal methods interpret human actions by motions from key joints of human body. This is nontrivial to efficiently detect and track human actions in video streams. Conventional approach of action recognition is direct use of silhouettes and skeletons obtained from Motion Capture Systems to track body joints. For instance, Wang et al. [12] utilized skeleton tracking method by [13] to extract skeletal information of human. Advancements in imaging techniques motivated Microsoft Kinect which makes it feasible to capture color image sequences and depth maps in real time by RGBD sensors. Li et al. [14] presents a graphical model for learning and recognizing human actions encoded with weighted directed graph. Graph nodes modeled on Gaussian Mixture Models represent salient postures in human actions. The postures and action graph were automatically learned from training samples through unsupervised clustering. Consequently, dynamics of human actions were modeled with action graph and bag of 3D points in [15]. However, these adopt sparse pseudo-input Gaussian process: a low-ranked covariance approximation method usually based on unobserved pseudo-data points.

Kang &Szeliski [16] utilized DMM to preserve spatial and temporal contextual information between space-time cells rather than estimating from a single view of images. Hence, local points of all frames in a video sequence are molded as single image. Alternatively, Yang et al. [10] generates DMMs by projecting motion energy of depth maps onto three orthogonal Cartesian planes. Vieira et al. [17] represented depth sequence in a 4D space-time grid, and used a saturation scheme to enhance roles of sparse cells with moving parts of body since. This method is computationally efficient and outperforms conventional approaches nonetheless, DMM is inherent with complex computational scheme as demonstrated in [9]; it is inefficient if training data is huge. As a result, Chen et al [18] modified the computation procedure of DMM. Similarly, each depth frame in a depth video sequence is projected onto three orthogonal Cartesian planes. To eliminate this setback using HOG descriptors from DMMs, some transformation techniques have been proposed to characterize human actions by taking sum image pixels over a certain set of lines [8, 9, 19, 20].

Greater success in object recognition relies on HOG descriptors having histograms in non-Euclidean space [21]. Nowadays, dynamics of a scene cannot longer model temporal evolution with linear dynamical systems. Bosch et al. [11] earlier combined shape and motion information of human in video sequences to improve limitations of shape-based action recognition. A major limitation of motion-related approach is that it requires a large amount of training videos, as the activity to recognize gets more complex but efficient on short sub-sequence. As a result, we propose to represent each video frame DMM and PHOG descriptors for distinguishing human actions.

III. ROBUST INVARIANT FEATURE DESCRIPTOR

This section presents a stepwise procedure of how RIFD extracts precise features from video sequences, and utilizes l2-regularized Collaborative Representation Classifier (CRC) for human activities recognition. RIFD is a novel feature descriptor that combines depth motion images and spatial distribution of edges to represent image as vectors. The novelty is its ability to automatically identify and select features at region of interest in images. Our idea is to explore spatial distribution of shapes in images for action recognition, hence pyramid representations of the shape edges are considered for HOG of images.

A. DMM Computation

The concept of depth map was introduced to alienate limitations of 2D images. This shows ability to capture 3D view with shape information of image scenes. Yang et. al. [10] proposed DMM by simply projecting depth frames onto three orthogonal planes. The computation procedure got complex by applying threshold to inter-map variations for projected maps of different subject. Hence, a modified version presented in [18] is adopted due to computational simplicity.

Let the binary map of motion energy be deemed as regions where some actions take place in a given temporal interval and each video sequence have K frames, then a 3D frame in the sequence can be assumed as projection of three 2D maps on Cartesian planes with three possible projection views: top (*t*), side (*s*), and front (*f*). Assuming a view (*v*) from the projection map is represented as map_v^i , then DMM_v is obtained by stacking the motion energy (M_e) across an entire video sequence as in Eq. (1). Map_v^i is the projected map of i^{th} frame under projection view $v \in \{t, s, f\}$, and N is the number of frames in a video. Not all frames in a sequence are used to generate DMMs but only those in the bounded region are cropped out to form the final DMM. M_e is intra-class variation given as Eq. 2. For clarity, projection of DMM_vgenerated from an action video of Pickup and Throw is given in Fig. 2.

$$DMM_v = \sum_{i+1}^{N} M_e \qquad (1)$$

$$M_e = \left| map_v^i - map_v^{i-1} \right| \qquad (2)$$

This captures some occluded regions as a representation of shape and motion in Cartesian projection planes; however they are not always visible in reference image. Tricubic interpolation [22] is applied to resize the$DMM_v$to a fixed size in order to obtain values at arbitrary points in the 3D space.

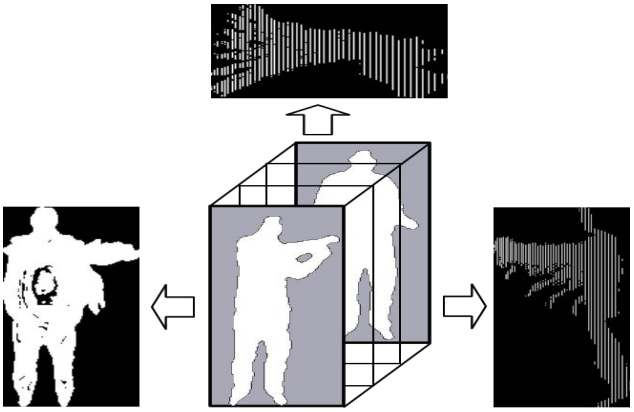

Fig. 2: $DMM_v$of depth action video sequences

B. DMM-PHOG Descriptor

PHOG is the representation of images by its local shape and spatial layout [23]. DMM of each view are used to obtain PHOG of images in video sequence, its local shape is captured by distribution over edge orientations in specific region, and the images are tiled to different regions at multiple resolutions for spatial layout. Distance between two PHOG image descriptors reflects the extent to which the images contain similar shapes and corresponding spatial layouts. Firstly for local shape representation, sub-regions of images are quantized into N bins with histogram of edge orientations. The bin stands for unique edge having orientations within certain angular range.

Spatial pyramid matching proposed in Ref. [23] is adopted for determining the spatial layout in frames. Assume a DMM_v has an input space X given as Eq. 3, where each point set is made up by k-dimensional feature vectors which are bounded by a sphere of diameter D.

$$X = \{x | x = \{[f_1^1, \ldots, f_k^1], \ldots, [f_1^{m_x}, f_k^{m_x}]\}\} \quad (3)$$

If the intra-vector distance is represented as $\frac{\sqrt{d}}{2}$, and m^xvaries across instances in X, then features from region of interest has an extraction function Ψ defined as Eq. 4. Where $L = [\mathrm{Log}_2 D]$, $x \in X$, and $H_i(x)$ is a histogram vector formed over data x using k-dimensional bins of side length 2^i, and $H_i(x)$ has a dimension of $r_i = \left(\frac{D}{2^i\sqrt{d}}\right)^d$. Ψ(x)isvector of concatenated histograms.The pyramid match kernel measures similarity between point sets based on implicit correspondences found within this multi-resolution histogram space. To compute the similarity between two input sets, we define pyramid match kernel K_Δ as in Eq. 5.

$$\Psi(x) = H_{-1}(x), H_0(x), \ldots, H_L(x) \quad (4)$$

$$K_\Delta = (\Psi(X_1), \Psi(X_2)) = \sum_{i=0}^{L} w_i * N_i \quad (5)$$

where$N_i$signifies number of newly matched pairs at level i. The robust invariance approach of PHOG is its ability to implicitlydetermine symmetries in point sets. For any two point sets, if they fall into the same histogram bin and share common feature at each resolution levels, they are likely categorized as one. To verify this, we will determine the overlap area between two histograms as Eq. (6), WhereXand Yare histograms with $k - bins$, and X(j)is the count of j^{th}bin of X. Thus, the generated PHOG feature descriptor represents shape and spatial features of depth video sequences in a dimension in Eq. (7) such that P_N is number of projection views, S_N is size of selected sub-bands while W_N is number of sliding windows in PHOG computation, and B_N is number of bins that was used.

$$\|\cap\|(X,Y) = \sum_{j=1}^{k} \min((X^j, Y^j)) \quad (6)$$

$$D = P_N * S_N * W_N * B_N \quad (7)$$

To compute gradient orientation histograms on a dense grid of uniformly spaced cells, we perform three local normalizations: L2-norm, L1-norm, L1-sqrt.We sampled20-by-10 non-overlapping cells and 8 gradient orientation bins hence each DMM_V generates a PHOG descriptor with dimension of 4800.

C. l_2-Regularized CRC

The final step in HAR with RIFD is to feed the PHOG descriptors into a recognition system. Bosch et al. [10] adopted Support Vector Machine for classifying images according to class on supervised learning, while the working mechanism of l_1-sparse representation based classification is not fully revealed in Zhang et al [24]. Performance of both l_1-sparse representation and SVM classifiers are not good as l_2 -regularized CRC, hence our motivation for adoption.Detailed explanation of l_2-regularized CRC can be found in [18]. For adoption, we give a brief description following its trend from sparse representation. Given a set of distinct classes C and n d-dimensional training sample in form of matrix$M = \{m\}_{i=1}^{n} \in \mathcal{R}^{d\,X\,n}$.A sparse representation y can be expressed with matrix Xgiven as Eq. (8), where$\propto$ is a coefficient vector of training samples in a sub-class C_i; $\forall \propto \in \mathcal{R}^{1\,X\,n}$. Sparsity representation cannot give a direct translation of $\propto$ in Eq. (8) because it is under-determined.

$$y = X \propto \quad (8)$$

Nevertheless, a solution can be obtained by solving an l_2-regularized minimization problem, for instance if λ is an independent regularization parameter for sparsity term inthe quadratic eq. (9). l_2 -regularized minimization is done via Tikhonov regularization earlier proposed in [25]. This has closed solution with $\widehat{\propto}$as in Eq. (10) such that $I \in \mathcal{R}^{n\,X\,n}$is an identity matrix. Therefore, for l_2-regularization, we translate Eq. (9) to Eq. (11) where $\Gamma$is a regularization matrix that allows imposition of prior knowledge on the solution.

$$\widehat{\propto} = \underset{\propto}{\mathrm{argmin}} ||y - X \propto||_2^2 + \lambda||\propto||_1 \quad (9)$$

$$\widehat{\propto} = (X^T X + \theta I)^{-1} X^T y \quad (10)$$

$$\widehat{\propto} = \underset{\propto}{\mathrm{argmin}} ||y - X \propto||_2^2 + \lambda||\Gamma \propto||_2^2 \quad (11)$$

Also, the diagonal matrix of $\Gamma$is usually considered as Eq. (12). Then, coefficient vector $\widehat{\propto}$ is calculated based on the approach in Golub et al. [26]. This is as given in Eq. (13); thus finally, $\widehat{\propto}$ can be partitioned such that class label of an unknown sample y has specific label following the

expression given in Eq. (14)

$$\not{\gamma} = \begin{pmatrix} ||y - x_1||_2 & \cdots & 0 \\ \vdots & \vdots & \vdots \\ 0 & \cdots & ||y - x_n||_2 \end{pmatrix} \quad (12)$$

$$\widehat{\propto} = \left(X^T X + \theta \not{\gamma}^T \not{\gamma}\right)^{-1} X^T y \quad (13)$$

$$class(y) = \underset{j \in \{1,\ldots,C\}}{argmin} ||y - X_j \widetilde{\propto}_j||_2 \quad (14)$$

IV. EXPERIMENTAL ANALYSIS

The proposed invariant model combines DMM and PHOG transforms as robust method of discriminating human actions. The procedure was observed in Matrix Laboratory (MATLAB) Version 2013b environment. A detailed description of MSR Action3D dataset used for testing the model, results gotten on applying to RIFD and analysis are documented. Evaluation was carried out on performance result of the proposed approach and this is as well discussed.

A. MSR Action3D Dataset

MSR action3D dataset is an action dataset of depth sequences captured by a depth camera. The dataset includes 20 different actions taken from 10 subjects. Each subject demonstrates an action with 3 repetitions making a total 567 sequences. Intra-class variations of the actions performed by singular subjects are also kept.For benchmark analysis, the actions were partitioned into three subsets as shown in Table I. Three different test cases were performed on each action subset in a way that in Test 1, 33.3% of samples from each subset are utilized for training while the rest are taken for test. Subsequently in Test 2, 66.6% of samples from each subset are used for training and also, the remaining samples for test. Lastly or cross subject test was performed in Test 3, that is training and testing are done on 50% of the subjects.

Table 1: Actions in MSR Action3d Dataset

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

B. Result and Analysis

To have a fair comparison with existing methods, a similar experimental setting with [9, 18] was observed. Accordingly, we resized the depth images to different sizes of 32×32, 64×64, 128×128, and used the most representative key frame of each video shot. Example of frames gotten from video sequence is given in Fig. 3. The top view, front view, and side view, views observed on projecting a frame of Tennis Serve action in AS1 onto the three orthogonal planes, in Fig 4. Then, edge detector of each view is obtained by applying Canny edge detector which transforms the views to their edge representations shown in Fig. 5.

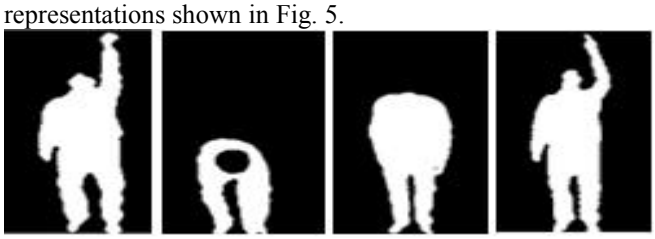

Fig. 3: Frames from video sequence for DMM analysis

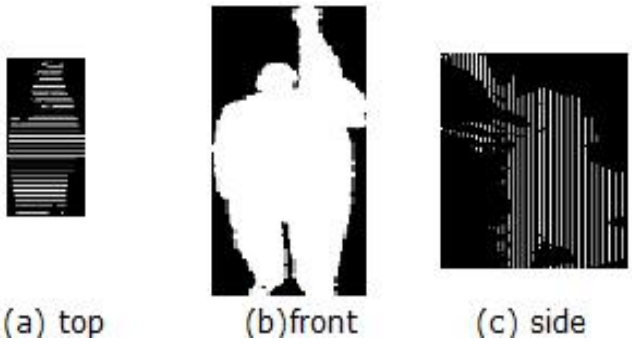

Fig. 4: View of DMM representation for tennis throw

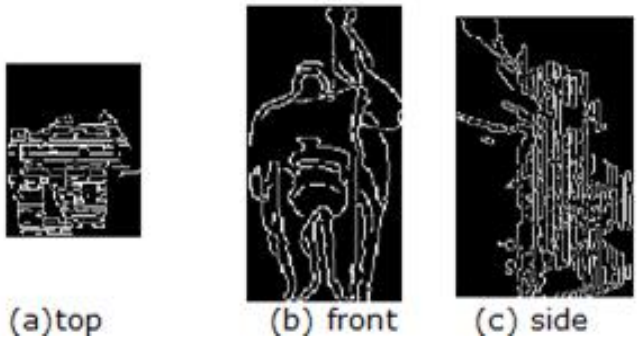

Fig. 5: Edge representation of tennis throw

Finally, PHOG was applied onthe shapes (image edges) and spatial motionin DMM of each view to obtain pyramid representation of region ofinterest.PHOG representations of view are shown as in respective views of Fig. 6. A Combination of these is used by l_2 -regularized CRC for discriminating activities.

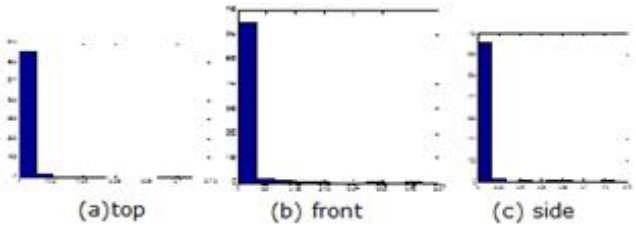

Fig. 6: Pyramid of histogram edge orientationfor tennis throw

C. 4.3 Model Evaluation

Evaluation of the model is specifically comparison of our method against studies [9, 10, 14, 27] since their scopes are similar and MSR Action 3D dataset was used for experimenting all methods. The results given in Table 2 shows that the performance values of our methods are in good range with more recent works in state-of-the-art and across test sets. Best results from the different test sets are emboldened for easy spotting. It can be clearly seen that

RIFD slightly outperforms states-of-the-art under Test One and Cross Subject cases and also, it is relatively within range of performance grades attained by other methods, in fact RIFD have almost the same accuracy level with Ref. [10] after beating down the others.

For reduced dimension of feature vector, we chose just appropriate values for normalization, non-overlapping cells, and gradient orientation bins resulting in a total of 4800 features. This is lower compared to methods reported in Refs. [10][18], but more than the method in Ref. [9]. The better performance of our proposal would not only be associated with reduced dimensional feature vector used, however adoption of pyramid based descriptors. The optimization together with spatial pyramid based descriptors played important roles in action discrimination. RIFD improves the average recognition rate of human actions in previous studies by2.3%, 1.15%, and 3.05% in the first test, second, and cross subject test respectively, however has close recognition rate with Ref. [10]. In the first and cross subject test, RIFD performs better than the later but otherwise in the second test. This can be attributed to more reduced feature set achieved over transformation.

Table 2: Evaluation of recognition accuracies based on MSR data

Methods	Test one	Test two	Cross Subject
Proposed Method	98.6	99.3	92.6
Farhad et al.[9]	98.5	99.6	92.3
Yang et al.[10]	95.8	97.4	91.6
Li et al.[14]	91.6	94.2	74.7
Luo et al.[27]	96.8	98.9	87.5

V. CONCLUSION

The recent attention gained by Human Action Recognition has been attributed to design of lowered cost 3D cameras. These devices capture depth information in addition with the 2D-dimension. Several approaches were proposed from the bag-of-words and the later however, major limitations include lost in context of spatial information between points of interest, and computational complexity together with sub-optimal accuracies. In this study, we proposed a RIFD to extract precise features from video sequences and utilize a regularized collaborative representation classifier for human activity recognition. The novel feature descriptor combines depth images and spatial distribution of edges to represent image vectors. Finally, good recognition rate of RIFD shows that shape and spatial pyramid based descriptors harmonize well. We hereby recommend that feature works considers this yet with some transformation techniques to find summations of image pixels over certain set of lines, this could yield better results.

REFERENCES

[1] C. Chen, R. Jafari, N. Kehtarnavaz, “Improving Human Action Recognition Using Fusion of Depth Camera and Inertia Sensors”, IEEE Transactions on Human-Machine Systems, 2015.

[2] M.O. Omisore, Y. A. Abebe, F Isinkaye, B. O. Ojokoh, N. A. Azeez, and L. Wang, “A Salient Invariant Feature Descriptor for Human Action Recognition”, 27th National Conference, Digital Inclusion Opportunities, Challenges and Strategies, Ibadan, Nigeria, 2018.

[3] J. Padilla-López, A. Chaaraoui, and F. Flórez-Revuelta, “A discussion on the validation tests employed to compare human action recognition methods using the MSR Action3D dataset”, Computer Vision and Pattern Recognition, 1-16, 2015.

[4] A. Baldominos, Y. Saez, and C. García del Pozo, “An Approach to Physical Rehabilitation using State-of-the-Art Virtual Reality and Motion Tracking Technologies”, Conference on Health and Social Care Information Systems and Technologies, Procedia Computer Science, 64: 10-16 ,2015

[5] B. Galnaa, G. Barrya, D. Jackson, D. Mhiripiria, P. Olivierb, and L. Rochestera, “Accuracy of the Microsoft Kinect Sensor for Measuring Movement in People with Parkinson's Disease”, Gait and Posture, 39(4): 1062–68, 2014

[6] N. Dalal, and B. Triggs, “Histograms of Oriented Gradients for Human Detection” IEEE Conference on Computer Vision & Pattern Recognition, 1:886-893, 2005

[7] S. Chen, T. Chen, W. Chen, and Y. Lee, “Human Action Recognition using Star Skeleton”, ACM International Workshop on Video Surveillance and Sensor Networks, Santa Barbara, USA, 2006.

[8] M. Do and M. Vettterli, “The Contourlet Transform: an Efficient Directional Multi-Resolution Image Representation,” IEEE Transaction on Image Processing, 14(12): 2091–2106, 2005.

[9] M. Farhad, Y. Jiang, J. Ma, “Human Action Recognition based on DMMs, HOGs and Contourlet Transform”, IEEE International Conference on Multimedia Big Data, pp. 389–394, 2015.

[10] X. Yang, C. Zhang, and Y. Tian, “Recognizing actions using depth motion maps-based histograms of oriented gradients”, 20th International conference on Multimedia, Nara, Japan, 1057-1060, 2012.

[11] A. Bosch, A. Zisserman, and X. Munoz, “Representing Shape with a Spatial Pyramid Kernel”, ACM International Conference on Image and Video Retrieval, Amsterdam, Netherlands, July 9 – 11, 2007.

[12] J. Wang, Z. Liu, J. Chorowski, Z. Chen, and Y. Wu, “Robust 3D Action Recognition with Random Occupancy Patterns”, 12th European Conference on Computer Vision, Firenze, Italy, 2012.

[13] J. Shotton, A. Fitzgibbon, M. Cook, T. Sharp, M. Finocchio, R. Moore, A. Kipman, and A. Blake, “Real-time Human Pose Recognition in Parts from Single Depth Images”, 24th IEEE Conference on Computer Vision and Pattern Recognition, Colorado Springs, USA, 2011.

[14] W. Li, Z. Zhang, and Z. Liu, “Action Recognition based on a Bag of 3D Points”, IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, San Francisco, 2010.

[15] W. Li, Z. Zhang, and Z. Liu, “Expandable Data-Driven Graphical Modeling of Human Actions Based on Salient Postures”, Circuits and Systems for Video Technology, 18(11):1499-1510, 2008.

[16] B. Kang, and R. Szeliski “Extracting View – Dependent Depth Maps from a Collection of Images”, International Journal of Computer Vision, 58(2): 139-163, 2004.

[17] A. Vieira, E. Nascimento, G. Oliveira, Z. Liu, and M. Campos, “Stop: Space-Time Occupancy Patterns for 3D Action Recognition from Depth Map Sequences”, Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications, Buenos Aires, Argentina, 2012.

[18] C. Chen, K. Liu, and N. Kehtarnavaz, “Real-Time Human Action Recognition based on Depth Motion Maps”, Journal of Real-Time Image Processing, 2013.

[19] Y. Wang, K. Huang, and T. Tan, “Human Activity Recognition based on R Transform”, IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Minnesota, USA, 2007.

[20] A. Ouanane, and A. Serir, “Fingerprint Compression by Ridgelet Transform”, IEEE International Symposium on Signal Processing and Information Technology, Sarajevo, Bosnia, 2008.

[21] J. Aggarwal, and Q. Cai “Human Motion Analysis: A Review”, Computer Vision and Image Understanding, 73:90-102, 1999.

[22] F. Lekien, and J. Marsden, "Tricubic interpolation in three dimensions", International Journal for Numerical Methods in Engineering, 63:455–471, 2005.

[23] K. Grauman, and T. Darrell, “The Pyramid Match Kernel-Discriminative Classification with Sets of Image Features”, Proceedings of IEEE International Conference on Computer Vision, China, 2005.

[24] L. Zhang, M. Yang, and X. Feng, “Sparse Representation or Collaborative Representation: Which Helps Face Recognition?” 13th International Conference on Computer Vision, Spain, 2011.

[25] R. Chaudhry, A. Ravichandran, G. Hager, and R. Vidal, “Histograms of Oriented Optical Flow and Binet-Cauchy Kernels on Nonlinear Dynamical Systems for the Recognition of Human Actions”, IEEE Conference on Computer Vision and Pattern Recognition, Miami, USA, 2009.

[26] Golub, G., Hansen, P.C., O’Leary, D. “Tikhonov-Regularization and Total Least Squares”, Journal on Matrix Analysis and Applications, 21(1), 185–194, 1999.

[27] J. Luo, W. Wang, and H. Qi, "Spatio-Temporal Feature Extraction and Representation for RGB-d Human Action Recognition," Pattern Recognition Letters, 2014.