2014 22nd International Conference on Pattern Recognition

View-Invariant 3D Action Recognition using Spatiotemporal Self-Similarities from Depth Camera

A-Reum Lee, Heung-Il Suk, and Seong-Whan Lee
Department of Brain and Cognitive Engineering, Korea University,
145, Anam-Ro, Seongbuk-ku, Seoul 136-713, Korea
Email: {arlee, hisuk, swlee}@image.korea.ac.kr

***Abstract*—The problem of viewpoint changes is an important issue in the study of human action recognition. In this paper, we propose the use of spatial features in a spatiotemporal self-similarity matrix (SSM) based on action recognition that is robust in viewpoint changes from depth sequences. The spatial features represent a discriminative density of 3D point clouds in a 3D grid. We construct the spatiotemporal SSM for the spatial features that change along with frames. To obtain the spatiotemporal SSM, we compute the Euclidean distance of each spatial feature between two frames. The spatiotemporal SSM represents similarity of human action robust in viewpoint changes. Our proposed method is robust in viewpoint changes and various length of action sequence. This method is evaluated on $ACT4^2$ dataset containing the multi-view RGBD human action data, and MSRAction3D dataset. In the experimental validation, the spatiotemporal SSM is a good solution for the problem of viewpoint changes in a depth sequence.**

I. Introduction

Human action recognition is one of the best-known areas in computer vision research. In the last decade, many studies in this area have used 2D video sequence [1], [2]. Many algorithms in human-computer interaction (HCI) research should be able to analyze human action automatically. The action recognition task, however, can cause large changes in the action shape as we change the viewpoint. The problem of viewpoint change is a challenging issue in human action tracking and recognition tasks [3]–[6].

Recently, Microsoft launched a cheap RGBD camera, known as Kinect sensor. This camera provides depth information with high quality, high resolution, and high frame rate, and can track multiple human bodies. Many researchers expected that the Kinect sensor would lead to the growth of computer vision technologies. The tracking information obtained from the Kinect sensor framework has been used as joint information in many human action recognition studies. However, it has still limitations when the individual does not stand directly in front of the camera, or is blocked by other objects, including the individual him or herself. For this reason, Kinect's tracking algorithm may fail to track joints. So, we cannot trust the joint information in action recognition study using only the joint information. Therefore, the problem of viewpoint changes remains unsolved.

The previous approaches of view-invariant action recognition proposed novel features robust in viewpoint change. Weinland et al. proposed a motion history volumes (MHV) in a variety of viewpoints [7]. This method proposes a viewpoint-free representation for human action based on a multi-camera system. Holte et al. proposed novel action recognition based on the detection of 4D spatio-temporal interest points [8]. Xia et al. presented a method using 3D joint locations (HOJ3D) as a representation of postures [9]. The 3D skeletal joint locations were extracted from depth sequence. Roh et al. [10] proposed a volume motion template (VMT) and projected motion template (PMT). This method extended a motion history image (MHI) method to 3D space. Holte et al. [11] proposed 3D optical flow and a harmonic motion context. This method was represented by optical flow of 3D data in 3D space.

Recently, a comparative coding descriptor (CCD) which is a descriptor for action recognition robust in viewpoint changes, was proposed by Cheng et al. [12]. This descriptor compared the depth value of a center point with the surrounding 26 points in the depth sequence. This method is efficient and relatively robust in viewpoint changes. However, it does not solve problems such as recognition of various types of action duration. For training all views, it also requires a large amount of training data in each view. Junejo et al. [13] proposed a feature based on the self-similarity matrix (SSM). The SSM was calculated using distances between features for all frame pairs in a sequence [14], [15]. This method does not require preparing numerous training data for each view. Therefore, the algorithm is simple and fast. However, this method can be prepared from the joint information of a human body. To track joints in a human body, the system requires tracking techniques. This method needs a lot of execution time for processing sequence-level information. If we do not know the joint information of a human body, we cannot expect a good performance.

Therefore, we present a novel approach of spatial features that can be extracted without the use of joint information. The spatial features representing the density of 3D point clouds in each cell of a 3D grid. We measure the density of 3D point clouds using a histogram in each cell. Then, we construct a spatiotemporal SSM for the spatial features, and calculate the gradient orientations of neighborhood elements in the spatiotemporal SSM using the histogram of oriented gradient (HOG). Finally, we classify actions using a support vector machine (SVM). A framework of our approach is illustrated in Fig. 1.

Our contributions are as follows: (1) We propose spatial features robust in viewpoint changes in a depth sequence. Changes in a human body represent specific spatial variation more than 2D sequence because the 3D point clouds include depth values. For this reason, this feature is robust in viewpoint changes. (2) Each viewpoint does not require training data. The

1051-4651/14 $31.00 © 2014 IEEE
DOI 10.1109/ICPR.2014.95

IEEE computer society

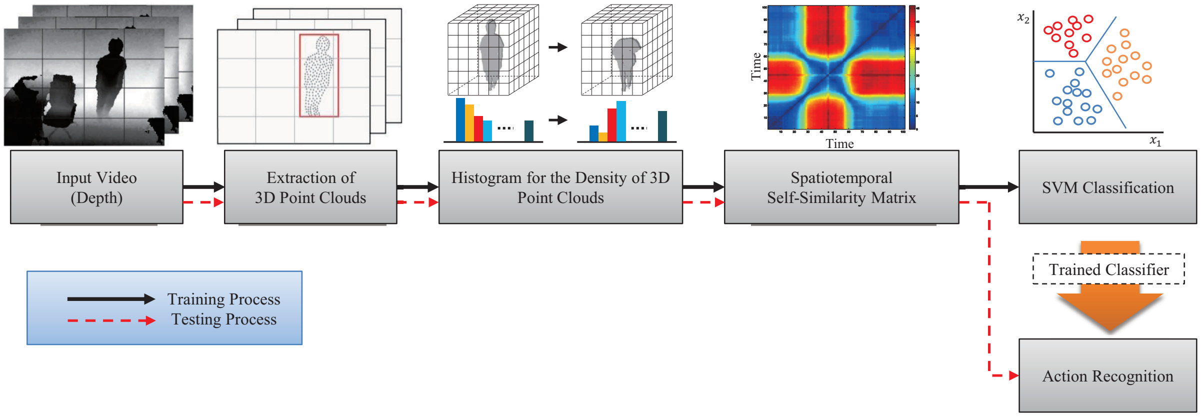


Fig. 1: The proposed framework for action recognition using spatial features and a spatiotemporal SSM.

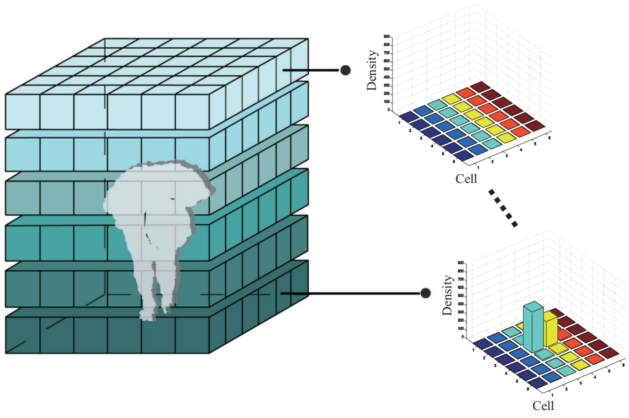


Fig. 2: Spatial features of each cell in the 3D grid.

spatiotemporal SSM represents similarity of the spatial and temporal variations. Consequently, the spatiotemporal SSM allows it to recognize actions across viewpoints.

II. Proposed Method

For view-invariant action recognition, we propose spatiotemporal SSM for spatial features. This method computes a similarity property of each spatial feature in depth sequence. The spatial features represent density of 3D point clouds in each cell of a 3D grid. To obtain the spatiotemporal SSM, we compute the mean Euclidean distance of each spatial feature between two frames. It represents similarity of human action robust in viewpoint changes. We finally evaluate the proposed method using a support vector machine (SVM).

A. Spatial Feature Extraction

Human body and 3D point clouds were extracted from the Microsoft Kinect sensor and the tracking system of OpenNI. The 3D point clouds represent the X, Y, and Z geometric coordinates of a human body.

3D point clouds are generated by depth information of depth data. The 3D point clouds are noted by $D = \{d_n(x_n, y_n, z_n), n = 1, \cdots, N\}$. N is the number of the 3D point clouds. To construct a 3D grid of a human body, we calculate the body's center of gravity, which can be described as:

$$x_{cg} = \frac{\sum_{n=1}^{N} x_n m_n}{\sum_{n=1}^{N} m_n}$$
$$y_{cg} = \frac{\sum_{n=1}^{N} y_n m_n}{\sum_{n=1}^{N} m_n}$$
$$z_{cg} = \frac{\sum_{n=1}^{N} z_n m_n}{\sum_{n=1}^{N} m_n} \qquad (1)$$

where x_{cg}, y_{cg}, and z_{cg} are coordinates of the body's center of gravity. m_n is masses of particular 3D point clouds n-th, and x_n, y_n, and z_n are coordinates of particular 3D point clouds n-th.

We construct the 3D grid of the same size in order to measure the density of 3D point clouds in the human body. X and Z coordinates of a central point in the 3D grid sets up x_{cg} and z_{cg} coordinates of the body's center of gravity. It can be applied to a moving human body. The size of the 3D grid is greater than the width, height, and depth of the human body. Because, moving space of action is bigger than the human body. The size of the 3D grid is fixed by its initial size. Then the 3D grid is divided spatially at regular intervals surrounding the center of gravity in a human body. The cells are defined by the smallest unit of the 3D grid. At this time, the 3D grid includes an image into non-overlapping regions. Such a 3D grid is shown in Fig. 2. For the density of the 3D point clouds, all 3D point cloud data of a human body are located in each cell.

Then, we measure the density of the 3D point clouds using a histogram in each cell. Given cells $P = \{p_1, \cdots, p_c\}$. p_c are denoted c-th element of the cell array. If n-th 3D point cloud d_n belongs to c-th cell p_c in the i-th frame, the histogram h_i^c

bin is counted. h_i^c is computed as:

$$h_i^c = \begin{cases} h_i^c + 1, & \text{if } d_n \cap p_c \\ h_i^c, & \text{otherwise} \end{cases} \tag{2}$$

Then the histogram $H_i = \left[h_i^1, \cdots, h_i^c\right]$ is normalized as:

$$S_i = \frac{1}{N} H_i \tag{3}$$

Therefore, $S_i = \left[s_i^1, \cdots, s_i^c\right]$ is the spatial features of i-th frame. This method is illustrated in Fig. 2.

B. Spatiotemporal Self-Similarities Matrix

The spatiotemporal SSM is composed of matrices of the spatial features value difference between the current frame and the other frame [16]. The matrix is defined as:

$$[d_{ij}]_{i,j=1,\cdots,T} = \begin{bmatrix} 0 & d_{12} & \cdots & d_{1T} \\ d_{21} & 0 & \cdots & d_{2T} \\ \vdots & \vdots & \ddots & \vdots \\ d_{T1} & d_{T2} & \cdots & 0 \end{bmatrix} \tag{4}$$

The spatiotemporal SSM of depth sequence is a square symmetric matrix. T is a length of depth sequence. i and j are the numbers of frames. The matrix size T corresponds to the frame length of depth sequence $I = \{I_1, \cdots, I_T\}$. d_{ij} is the Euclidean distance between two frames, i and j. It is computed as:

$$d_{ij} = \frac{1}{C} \sum_{c=1}^{C} \left\| s_i^c - s_j^c \right\|_2 \tag{5}$$

The spatial features, s_i^c and s_j^c, indicate the histogram showing the density of 3D point clouds movement obtained in the c-th cell from the i frame and the j frame. That is, the histogram s_i^c and s_j^c are defined by the spatial features extracted from the i frame and the j frame. The spatial features indicate the histogram. This is shown in Fig. 3.

In the spatiotemporal SSM in Fig. 3, red color indicates high correlation and blue color indicates low correlation. The direction of movement of the 3D point clouds including spatial and temporal information correspond when the camera views of human action change. It is possible to recognize human robustness in viewpoint changes.

For analysis of the spatiotemporal SSM properties, we calculate the gradient orientations of neighborhood elements in the spatiotemporal SSM using the histogram of oriented gradient [17]. The HOG descriptors show association of overall elements of the spatiotemporal SSM.

C. Action Recognition

As previously described, our proposed method composes HOG descriptors of a spatiotemporal SSM using spatial features, which are the 3D point clouds movement obtained from depth sequence. We evaluate the HOG descriptors of the spatiotemporal SSM for changes in viewpoint. For action recognition, we classify actions using SVM [18].

We train the input data using non-linear SVM for multiclass classification in each different view. We use a linear kernel with multi-class SVM in OpenCV library.

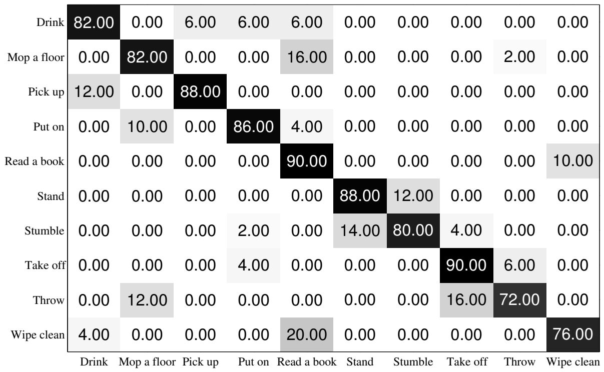
III. Experimental Results and Analysis

In this section, we demonstrated our approach on the $ACT4^2$ dataset [12] and MSRAction3D dataset [19]. We compare our approach with several recent algorithms. To extract the spatial features, we define that size of 3D grid sets 1.2 times more than the width, height, and depth of human body. Then, the 3D grid is divided by $6 \times 6 \times 6$ cells.

A. $ACT4^2$ Dataset

Our approach is evaluated by the performance on the $ACT4^2$ dataset [12]. It contains 14 types of human action classes. It also contains 24 subjects, each of which is performed twice per action in four viewpoints. We chose 10 actions (i.e., Drink, Mop floor, Pick up, Put on, Read a book, Stand, Stumble, Take off, Throw, and Wipe clean) in four views.

All viewpoints: This experiment shows the generality of our approach for all viewpoints. We used the depth sequence of 15 subjects as training data. The other sequences were used for testing. This experiment was also trained on all camera views using 10-fold cross-validation. The results are shown in Fig. 4. The total average accuracy for all four views is 83.4%.

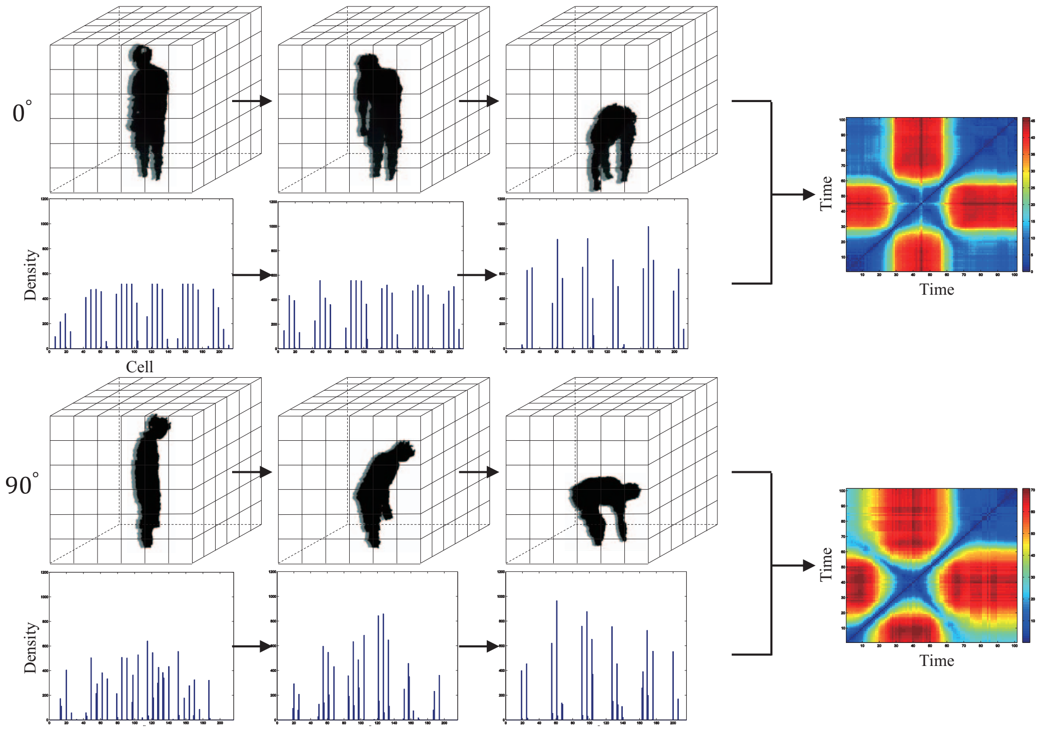


	Drink	Mop a floor	Pick up	Put on	Read a book	Stand	Stumble	Take off	Throw	Wipe clean
Drink	82.00	0.00	6.00	6.00	6.00	0.00	0.00	0.00	0.00	0.00
Mop a floor	0.00	82.00	0.00	0.00	16.00	0.00	0.00	0.00	2.00	0.00
Pick up	12.00	0.00	88.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Put on	0.00	10.00	0.00	86.00	4.00	0.00	0.00	0.00	0.00	0.00
Read a book	0.00	0.00	0.00	0.00	90.00	0.00	0.00	0.00	0.00	10.00
Stand	0.00	0.00	0.00	0.00	0.00	88.00	12.00	0.00	0.00	0.00
Stumble	0.00	0.00	0.00	2.00	0.00	14.00	80.00	4.00	0.00	0.00
Take off	0.00	0.00	0.00	4.00	0.00	0.00	0.00	90.00	6.00	0.00
Throw	0.00	12.00	0.00	0.00	0.00	0.00	0.00	16.00	72.00	0.00
Wipe clean	4.00	0.00	0.00	0.00	20.00	0.00	0.00	0.00	0.00	76.00

Fig. 4: Confusion matrix of recognition results for all viewpoints.

Cross viewpoints: This experiment shows the robustness of our approach for cross viewpoints. We experimented with action recognition between different camera views. We chose a target view for measuring the performance of recognition, and used other views for training. We used the sequences of 15 subjects for training, and used the others for testing with 10-fold cross-validation. The results are shown in Fig. 5. The total average accuracy for cross viewpoints is 81.2%.

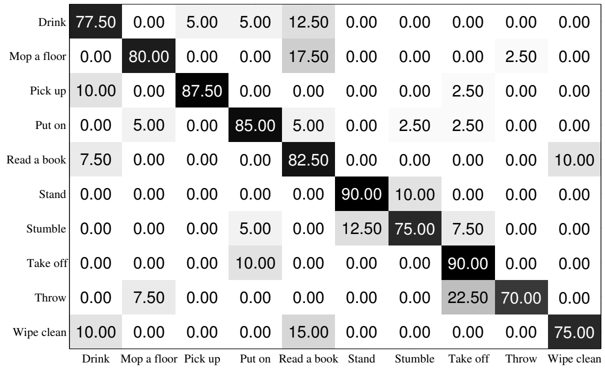
The accuracy of the cross viewpoints recognition is 2.2% less than in the first experiment. When the viewpoints of the actions change, these actions tend to be irregular on the $ACT4^2$ dataset. In the case of the fourth viewpoint, a shape of a human body shows the back. For this reason, the accuracy of the fourth viewpoint is low. If we do not include the fourth viewpoint, the accuracy of the cross viewpoints will increase in all viewpoints.



(a) Samples of the spatial features

(b) Samples of the spatiotemporal SSM

Fig. 3: Examples of spatiotemporal SSM for different views of a $ACT4^2$ dataset. (a) Samples of changing of spatial features. Columns 1, 2, and 3 represent the change of actions the change in the density of 3D point clouds, whereas rows 1 and 3 represent different views. Then, rows 2 and 4 represent the spatial features. (b) Samples of the spatiotemporal SSM in different views. Column 4 represents the spatiotemporal SSM obtained from the spatial features of depth sequence.



	Drink	Mop a floor	Pick up	Put on	Read a book	Stand	Stumble	Take off	Throw	Wipe clean
Drink	77.50	0.00	5.00	5.00	12.50	0.00	0.00	0.00	0.00	0.00
Mop a floor	0.00	80.00	0.00	0.00	17.50	0.00	0.00	0.00	2.50	0.00
Pick up	10.00	0.00	87.50	0.00	0.00	0.00	0.00	2.50	0.00	0.00
Put on	0.00	5.00	0.00	85.00	5.00	0.00	2.50	2.50	0.00	0.00
Read a book	7.50	0.00	0.00	0.00	82.50	0.00	0.00	0.00	0.00	10.00
Stand	0.00	0.00	0.00	0.00	0.00	90.00	10.00	0.00	0.00	0.00
Stumble	0.00	0.00	0.00	5.00	0.00	12.50	75.00	7.50	0.00	0.00
Take off	0.00	0.00	0.00	10.00	0.00	0.00	0.00	90.00	0.00	0.00
Throw	0.00	7.50	0.00	0.00	0.00	0.00	0.00	22.50	70.00	0.00
Wipe clean	10.00	0.00	0.00	0.00	15.00	0.00	0.00	0.00	0.00	75.00

Fig. 5: Confusion matrix of recognition results in cross viewpoints.

Table I compares the performance of the proposed method with that of the competing methods on the $ACT4^2$ dataset. The accuracy for the cross viewpoints recognition is 31.20% superior to CCD accuracy, which was reported as 50.8%. The performance of our proposed method is better than those of Cheng et al. [12] and Laptev et al. [20]. These results show that action recognition using the spatiotemporal SSM is robust in viewpoint changes.

TABLE I: Comparison of results on $ACT4^2$ dataset.

Method	Accuracy (%)	
	All-view	Cross-view
Depth-HOGHOF [20]	74.5	39.8
Depth-CCD [12]	76.2	50.8
Ours	**83.4**	**81.2**

B. MSRAction3D Dataset

MSRAction3D dataset [19] is an action dataset using a depth sequence. It contains 20 types of human action classes. It has only a front view. Therefore, we compare our approach with previous methods [19]–[23] for the front view from depth sequence.

Table II shows the comparison of our proposed method with the previous methods on the MSRAction3D dataset. The performance of our proposed method is not superior to the previous method. However, the accuracy of our proposed

TABLE II: Comparison of results on MSRAction3D dataset.

Method	Accuracy (%)
Li et al. [19]	74.7
STOP [20]	84.8
ROP [21]	86.5
DCSF [22]	89.3
HON4D [23]	88.9
Ours	**86.1**

method shows good performance. Consequently, the proposed method is a good solution robust in viewpoint changes in a depth sequence.

IV. Conclusion and Future Work

In this paper, we proposed a spatial feature extraction method robust in viewpoint changes. For extracting the spatial features, we suggested the histogram for computing the density of 3D point clouds of each cell in the 3D grid. Therefore, it has a capability to recognize human actions robust in viewpoint changes because movement of 3D point clouds includes spatial and temporal information. It is also faster and produces more accurate measurements than obtaining joint information. We calculated similarity of the spatial features using spatiotemporal SSM. The robustness of the proposed method for viewpoint changes means that our method does not require additional training data for each viewpoint. Then, we extracted the HOG descriptors of the spatiotemporal SSM and used a multi-class SVM as a classifier. We demonstrated that the spatiotemporal SSM is a good solution for viewpoint changes in depth sequence.

In this work, we only considered a single-person activity recognition. For real applications, it will be our forthcoming research issue to develop a method for multiple-human behavior analysis under a multi-RGBD camera environment.

Acknowledgment

This work was supported by the 2014 R&D Program for S/W Computing Industrial Core Technology through the MSIP (Ministry of Science, ICT and Future Planning)/KEIT (Korea Evaluation Institute of Industrial Technology). (Project No. 2014-044-023-001)

References

[1] H.-I. Suk, B.-K. Sin, and S.-W. Lee, “Hand gesture recognition based on dynamic Bayesian network framework,” *Pattern Recognition*, Vol. 43, No. 9, September 2010, pp. 3059-3072.

[2] H.-I. Suk, A. K. Jain, and S.-W. Lee, “ A Network of Dynamic Probabilistic Models for Human Interaction Analysis,” *IEEE Transactions on Circuits and Systems for Video Technology*, Vol. 21, No. 7, July 2010, pp. 932-945.

[3] D. Weinland, M. Ozuysal, and P. Fua, “Making Action Recognition Robust to Occlusions and Viewpoint Changes,” *Proc. 11th European Conference on Computer Vision*, Heraklion, Greece, September 5-11, 2010, pp. 635-648.

[4] I. N. Junejo, E. Dexter, I. Laptev, and P. Perez, “Cross-View Action Recognition from Temporal Self-Similarities,” *Proc. 10th European Conference on Computer Vision*, Marseille, France, October 12-18, 2008, pp. 293-306.

[5] M. Ahmad and S.-W. Lee, “Human Action Recognition using Shape and CLG-Motion Flow from Multi-View Image Sequences,” *Pattern Recognition*, Vol. 41, No. 7, July 2008, pp. 2237-2252.

[6] D.-C Hur, H.-I. Suk, C. Wallraven, and S.-W. Lee, “Biased Manifold Learning for View Invariant Body Pose Estimation,” *International Journal of Wavelets, Multiresolution and Information Processing*, Vol. 10, No. 6, November 2012, pp. 1250058.1-1250058.24.

[7] D. Weinland, R. Ronfard, and E. Boyer, “Free Viewpoint Action Recognition using Motion History Volumes,” *Computer Vision and Image Understanding*, Vol. 104, No. 2-3, November 2006, pp. 249-257.

[8] M. B. Holte, B. Chakraborty, J. Gonzalez, and T. B. Moeslund, “A Local 3-D Motion Descriptor for Multi-View Human Action Recognition from 4-D Spatio-Temporal Interest Points,” *IEEE Journal on Selected Topics in Signal Processing*, Vol. 6, No. 5, 2012, pp. 553-565.

[9] L. Xia, C. C. Chen, and J. K. Aggarwal, “View Invariant Human Action Recognition using Histograms of 3D Joints,” *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, Providence, USA, June 16-21, 2012, pp. 20-27.

[10] M.-C. Roh, H.-K. Shin, and S.-W. Lee, “View-Independent Human Action Recognition with Volume Motion Template on Single Stereo Camera,” *Pattern Recognition Letters*, Vol. 31, No. 7, May 2010, pp. 639-647.

[11] M. B. Holte, T. B. Moeslund, and P. Fihl, “View-Invariant Gesture Recognition using 3D Optical Flow and Harmonic Motion Context,” *Computer Vision and Image Understanding*, Vol. 114, No. 12, December 2010, pp. 1353-1361.

[12] Z. Cheng, L. Qin, Y. Ye, Q. Huang, and Q. Tian, “Human Daily Action Analysis with Multi-view and Color-Depth Data,” *Proc. 12th European Conference on Computer Vision*, Florence, Italy, October 7-13, 2012, pp. 52-61.

[13] I. N. Junejo, E. Dexter, I. Laptev, and P. Perez, “View-Independent Action Recognition from Temporal Self-Similarities,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 33, No. 1, 2011, pp. 172-185.

[14] C. Benabdelkader, R. Cutler, and L. Davis, “Gait Recognition using Image Self-Similarity,” *EURASIP Journal on Applied Signal Processing*, Vol. 2004, No. 1, April 2004, pp. 572-585.

[15] E. Shechtman and M. Irani, “Matching Local Self-Similarities across Images and Videos,” *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Minneapolis, USA, June 17-22, 2007, pp. 1-8.

[16] S. Lele, “Euclidean Distance Matrix Analysis (EDMA): Estimation of Mean Form and Mean Form Difference,” *Mathematical Geology*, Vol. 25, No. 5, July 1993, pp. 573-602.

[17] N. Dalal and B. Triggs, “Histograms of Oriented Gradients for Human Detection,” *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, San Diego, USA, June 25-25, 2005, pp. 886-893.

[18] H.-I. Suk and S.-W. Lee, “A Novel Bayesian Framework for Discriminative Feature Extraction in Brain-Computer Interfaces,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 35, No. 2, February 2013, pp. 286-299.

[19] W. Li, Z. Zhang, and Z. Liu, “Action Recognition Based on a Bag of 3D Points,” *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, San Francisco, USA, June 13-18, 2010, pp. 9-14.

[20] I. Laptev, “On Space-Time Interest Points,” *International Journal of Computer Vision*, Vol. 64, No. 2-3, September 2005, pp. 107-123.

[21] J. Wang, Z. Liu, J. Chorowski, Z. Chen, and Y. Wu, “Robust 3d action recognition with random occupancy patterns,” *Proc. 12th European Conference on Computer Vision*, Florence, Italy, October 7-13, 2012, pp. 872-885.

[22] L. Xia and J. K. Aggarwal, “Spatio-Temporal Depth Cuboid Similarity Feature for Activity Recognition Using Depth Camera,” *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, Portland, USA, June 23-28, 2013, pp. 2834-2841.

[23] O. Oreifej and Z. Liu, “HON4D: Histogram of Oriented 4D Normals for Activity Recognition from Depth Sequences,” *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, Portland, USA, June 23-28, 2013, pp. 716-723.