Optical Engineering

OpticalEngineering.SPIEDigitalLibrary.org

Volumetric spatial feature representation for view-invariant human action recognition using a depth camera

Seong-Sik Cho
A-Reum Lee
Heung-Il Suk
Jeong-Seon Park
Seong-Whan Lee

SPIE.

Volumetric spatial feature representation for view-invariant human action recognition using a depth camera

Seong-Sik Cho,[a] A-Reum Lee,[a] Heung-Il Suk,[b] Jeong-Seon Park,[c] and Seong-Whan Lee[a,b,*]
[a]Korea University, Department of Computer Science and Engineering, Seongbuk-gu, Seoul 136-713, Republic of Korea
[b]Korea University, Department of Brain and Cognitive Engineering, Seongbuk-gu, Seoul 136-713, Republic of Korea
[c]Chonnam National University, Department of Multimedia, Yeosu, Jeollanam-do 550-749, Republic of Korea

Abstract. The problem of viewpoint variations is a challenging issue in vision-based human action recognition. With the richer information provided by three-dimensional (3-D) point clouds thanks to the advent of 3-D depth cameras, we can effectively analyze spatial variations in human actions. In this paper, we propose a volumetric spatial feature representation (VSFR) that measures the density of 3-D point clouds for view-invariant human action recognition from depth sequence images. Using VSFR, we construct a self-similarity matrix (SSM) that can graphically represent temporal variations in the depth sequence. To obtain an SSM, we compute the squared Euclidean distance of VSFRs between a pair of frames in a video sequence. In this manner, an SSM represents the dissimilarity between a pair of frames in terms of spatial information in a video sequence captured at an arbitrary viewpoint. Furthermore, due to the use of a bag-of-features method for feature representations, the proposed method efficiently handles the variations of action speed or length. Hence, our method is robust to both variations in viewpoints and lengths of action sequences. We evaluated the proposed method by comparing with state-of-the-art methods in the literature on three public datasets of $ACT4^2$, MSRAction3D, and MSRDailyActivity3D, validating the superiority of our method by achieving the highest accuracies. [DOI: 10.1117/1.OE.54.3.033102]

Keywords: view invariance; action recognition; depth camera; point clouds; volumetric spatial feature representation.

Paper 141699 received Nov. 3, 2014; accepted for publication Jan. 30, 2015; published online Mar. 3, 2015.

1 Introduction

Image-based human action recognition has been of great interest in computer vision for its potential applications in the real world; a few of them include intelligent video surveillance, human-computer interaction, and video games. In the past few decades, many research groups have proposed various methods for action recognition.[1,2] However, there still remain many challenges caused by variations in illumination, clothing, viewpoints, and self-occlusion that degrade recognition performance. In this paper, we focus on the problem of viewpoint changes which causes great intra-class variations.

Recently, there has been considerable research in feature recognition from images. A major limitation of the previous studies is that their methods or models were mostly trained using data obtained from a single, fixed viewpoint. Therefore, they were inevitably limited in their ability to handle image sequences recorded at different viewpoints. To resolve this problem, some research groups attempted to reconstruct three-dimensional (3-D) human body postures by estimating the body joints,[3] and then extracted features from the reconstructed 3-D human body postures.[4] Although the efficiency of these methods was demonstrated by experiments, 3-D human body posture reconstruction from two-dimensional (2-D) images is computationally intensive. Moreover, it is vulnerable to noise, thereby causing many errors in joint estimation. Consequently, the features extracted from reconstructed body postures are likely to be contaminated by errors, and thus result in performance degradation.

In this study, we propose a novel framework that directly uses the depth information of each pixel for view-invariant action recognition, rather than reconstructing 3-D body postures. Specifically, we propose a new method for spatio-temporal feature representations that initially involve building a point cloud density histogram, and subsequently constructing a spatio-temporal self-similarity matrix (SSM)[5] that represents the temporal variations of the spatial features. We then extract action descriptors from the SSM, represented by means of statistics via a bag-of-features method,[6] which is finally fed into a classifier. A framework based on our approach is illustrated in Fig. 1.

The main contributions of this work are threefold:

- We propose a volumetric spatial feature representation (VSFR) that is extracted from a depth sequence. The VSFR captures a density of 3-D point clouds in a 3-D grid, and expresses spatial variations in 3-D space.
- Our method recognizes human actions with no need of estimating and tracking 3-D body skeletons, or for computing spatio-temporal interest points (STIPs) in a frame. Skeleton tracking is a challenging issue by itself in the field of computer vision. Furthermore, STIPs are sensitive to viewpoint changes and are computationally very expensive, and prohibiting it from being applied to real-time applications.
- To measure temporal variations, we construct a spatio-temporal SSM that computes the dissimilarities between the VSFRs of two frames. From the SSM,

*Address all correspondence to: Seong-Whan Lee, E-mail: sw.lee@korea.ac.kr

0091-3286 /2015/$25.00

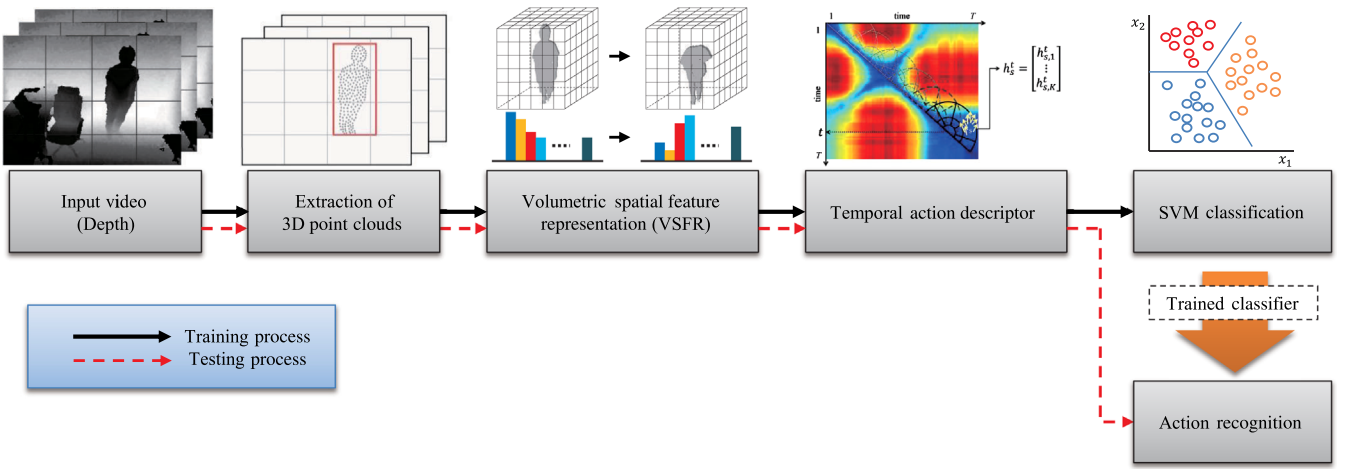

Fig. 1 The proposed framework for action recognition using spatial features and a self-similarity matrix.

we extract an action representation by constructing a code vector based on local descriptors that efficiently captures the spatial and temporal characteristics in an action. Therefore, for the same actions captured at different viewpoints, our action descriptors distribute close to each other in the feature space, while those of different actions are apart from each other. In this regard, our feature presentation method helps to improve the classification performance under viewpoint changes.

This paper is organized as follows. In Sec. 2, we briefly review the previous work in the literature. In Sec. 3, we describe the proposed volumetric spatial feature representation method. Section 4 explains the procedure for extracting local descriptors using the SSM of VSFRs. Section 5 outlines the building of an action descriptor and a classifier. In Sec. 6, we present and discuss our experimental results. Finally, our conclusions and future work are presented in Sec. 7.

2 Related Work

The variation in viewpoints is a challenging problem in action recognition because the same action can look different when the viewpoint of the action changes. In this section, we briefly describe the previous research studies that have looked into this issue.

The use of depth data or a 3-D joint model provides rich information for human actions. Huang et al. proposed a method for performance evaluation of SSMs of 3-D video sequences.[7] Their method handled viewpoint variations based on the human surface 3-D shape similarity. Weinland et al. proposed motion history volumes covering a variety of viewpoints that used a free-viewpoint representation of human actions based on a multicamera system.[3] Holte et al.'s method used a four-dimensional STIP detector in a multicamera system. However, this method required the reconstruction of 3-D human body poses.[4] Xia et al. presented a method that used 3-D joint locations as a representation of postures.[8] The 3-D skeletal joint locations were extracted from a depth sequence using Kinect. A major limitation of this method is that it requires accurate reconstruction of body joints. Roh et al. proposed a volume motion template and a projected motion template.[9] Their method extended the motion history image method to a 3-D space. The main drawback of their method is the requirement of prior knowledge for the start and end points of a sequence.

Recently, a comparative coding descriptor was proposed by Ref. 10. This descriptor was built using a 3-D cuboid and compared the depth value of a center point with the surrounding 26 points in a depth sequence. Although Cheng et al. demonstrated the efficiency and robustness of their method with respect to viewpoint changes in their experiments,[10] they did not consider the problem of varying action lengths. Furthermore, their method required a large amount of training data for training all views. Junejo et al. proposed a feature representation method from an SSM, which was constructed based on the distances between features for all frame pairs in a video sequence.[11,12] The main advantage of this method is that it does not require a large number of training samples for each view or a full body 3-D reconstruction. However, the joint information has to be manually marked in 2-D images. Hence, tracking techniques are required to track the joints in a human body.

3 Proposed Volumetric Spatial Feature Representation

Let D^t denote the t'th input depth image, $D^t(x,y)$ denotes a depth value in the pixel (x,y), and W and H denote, respectively, the width and height of an image. After detecting a human subject by applying background subtraction,[13] we compute the center of gravity of the detected subject as follows:

$$\hat{x}_c^t = \frac{\sum_{y=1}^{H}\sum_{x=1}^{W} x \cdot I^t(x,y)}{\sum_{y=1}^{H}\sum_{x=1}^{W} I^t(x,y)}$$

$$\hat{y}_c^t = \frac{\sum_{y=1}^{H}\sum_{x=1}^{W} y \cdot I^t(x,y)}{\sum_{y=1}^{H}\sum_{x=1}^{W} I^t(x,y)}$$

$$\hat{z}_c^t = \frac{\sum_{y=1}^{H}\sum_{x=1}^{W} D^t(x,y) \cdot I^t(x,y)}{\sum_{y=1}^{H}\sum_{x=1}^{W} I^t(x,y)}, \quad (1)$$

where I^t is an indicator matrix whose element is set to one if the corresponding depth value in D^t is positive; otherwise, it is set to zero.

Based on the center of gravity, we then define a volumetric bounding box $\mathbf{B}^t$ of the t'th frame as follows:

$$\mathbf{B}^t \equiv \begin{bmatrix} B_x^t \in [\hat{x}_c^t - \Delta x, \hat{x}_c^t + \Delta x] \\ B_y^t \in [Y - \Delta y, Y + \Delta y] \\ B_z^t \in [\hat{z}_c^t - \Delta z, \hat{z}_c^t + \Delta z] \end{bmatrix}, \tag{2}$$

where $[a, b]$ denotes a range between the values of a and b, and Δx, Δy, and Δz denote half the size of the bounding box in each dimension. Y is fixed as the height of a subject detected in the first frame, $\hat{x}_c^t$ and $\hat{z}_c^t$ are obtained for each separate frame. In the ensuing process, we consider only the voxels in the bounding box $\mathbf{B}^t$, in which a human subject is in motion. In this manner, we can greatly reduce the computational burden by focusing only on the specified area. Here, we should note that the location of the bounding box in a frame changes over time in accordance with the motion of the actor.

Unlike the previous methods that reconstruct a 3-D skeleton of a body by inferring joint information; in this work, we propose a method that directly uses the observed depth values in D^t and builds a 3-D volumetric binary image $P^t(x, y, z)$, called a "3D point cloud," as follows:

$$P^t(x, y, z) = \begin{cases} 1 & \text{if } z - 1 < D^t(x, y) < z + 1, \\ 0 & \text{otherwise} \end{cases}, \tag{3}$$

where $(x, y, z) \in \mathbf{B}^t$.

Using the 3-D volumetric image, we propose a feature representation mechanism that can efficiently describe spatial motion characteristics in an action. We first divide a 3-D volumetric bounding box into equally spaced nonoverlapping blocks, and then compute a voxel density for each block. Hereafter, we refer to this block as a "cell." Without loss of generality, we can label the cells with a sequential index $\{1,2, \cdots, N\}$, where N denotes the total number of cells in the bounding box. We define a voxel density in a cell as follows:

$$C_k^t = \sum_{(x,y,z) \in \mathbf{b}_k^t} P^t(x, y, z), \tag{4}$$

where $\mathbf{b}_k^t$ denotes the k'th cell in $\mathbf{B}^t$ and (x, y, z) denotes a locational index in $\mathbf{B}^t$. From the voxel densities of cells in the bounding box, we construct a histogram $\mathscr{H}^t$ that collectively represents the spatial characteristics for an action occurring in the t'th frame as follows:

$$\mathscr{H}^t = [C_1^t, C_2^t, \cdots, C_N^t]. \tag{5}$$

Figure 2 illustrates the examples of cell densities and the corresponding histograms of the frames. By taking into account the varying number of voxels in 3-D volumetric images, we normalize the histogram with the total number of nonzero point clouds in $\mathbf{B}^t$ as follows:

$$\hat{\mathscr{H}}^t = \frac{1}{\sum_{(x,y,z) \in \mathbf{B}^t} P^t(x, y, z)} [C_1^t, C_2^t, \cdots, C_N^t]. \tag{6}$$

We call this normalized histogram a "volumetric spatial feature representation."

Note that although the proposed VSFR cannot capture the information of hidden or occluded body parts, the main advantage of our method is that it can efficiently represent the 3-D spatial information for action recognition in spite of this limitation. Furthermore, from a computational point of view, our method requires simple logic and sum operations. This means that when compared to the previous 3-D skeleton-based methods that require huge computational resources and time, our approach is more suitable for real-time applications.

4 Temporal Action Descriptor

For detecting temporal motion characteristics, we use an SSM that can graphically represent temporal features in a sequence of images and extract a local descriptor for each frame. Specifically, given a sequence of T depth images $\mathbf{D} = \{D^1, D^2, \cdots, D^T\}$, we define the SSM as follows:

$$\mathbf{M} = [M_{ij}]_{i,j=1,\cdots,T}, \tag{7}$$

where $M_{ij} = (1/N)\|\mathscr{H}^i - \mathscr{H}^j\|_2$, $\|\ \|_2$ denotes an ℓ_2-norm of a vector, and N is the number of cells in a bounding box. The SSM graphically depicts the dissimilarities between a pair of frames in a video sequence, thus setting the diagonal elements to zero, which corresponds to self-dissimilarity. To measure the similarity between frames, we use the squared Euclidean distance between the proposed VSFRs of the respective frames.

In Fig. 2, we present examples of SSMs for an action observed from different viewpoints. In this figure, red and blue colors denote high and low dissimilarities, respectively. Note that despite the different viewpoints, the resulting SSMs look very similar to each other. This means that the proposed VSFR effectively represents 3-D motion information even under varying viewpoints. Hence, the SSM computed with the VSFR can be a useful tool for view-invariant action recognition.

To extract an action descriptor from each frame, we use Junejo et al.'s local descriptor.[14] Specifically, we first extract a block descriptor at each frame by using a histogram of oriented gradients[14] with a log-polar block structure, as illustrated in Fig. 3. Note that due to the symmetric nature of an SSM, it is sufficient to consider only one side of the off-diagonal in SSM. To represent a block descriptor at time t, we define the diameter of our block structure with consecutive $m + 1$ frames (in our experiments, we set m to 28 by following Junejo et al.'s work,[14] although they further considered a larger number of frames to capture varying temporal information) centered at the t'th frame. For a block s in our log-polar structure, we build a normalized histogram of oriented gradients with K bins by evenly spacing orientations over 0 deg to 360 deg. For each pixel in a block s, a vote is made to the bin, to which the orientation of a pixel corresponds. Let $\mathbf{h}_s^t = [h_{s,1}^t, h_{s,2}^t, \cdots, h_{s,K}^t]$ denotes a normalized K-bin histogram for a block s, where $s \in \{1,2, \cdots, S\}$, S denotes the number of blocks in a log-polar structure (S was set to 11 as presented in Fig. 3, based on Junejo

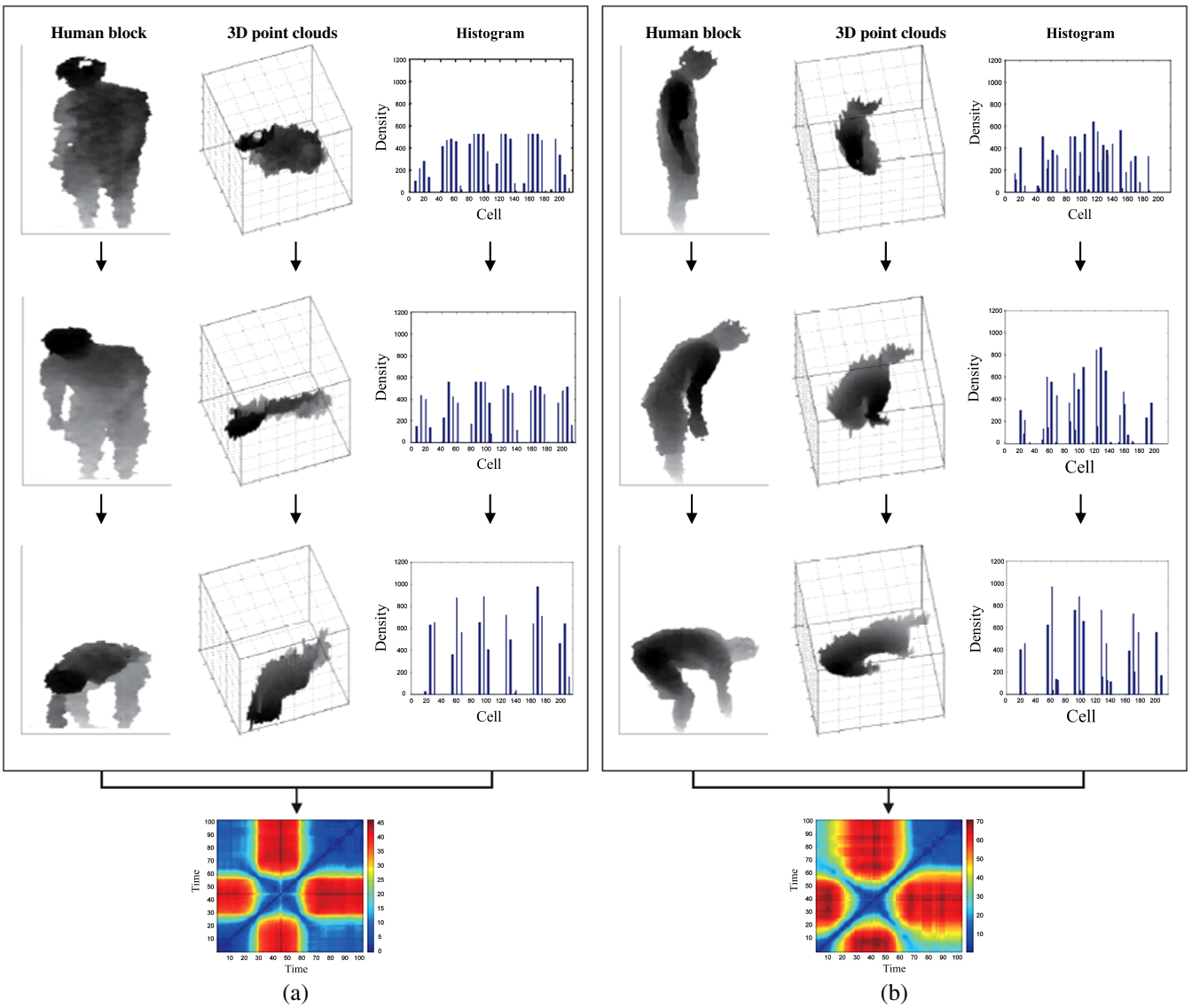

Fig. 2 Examples of self-similarity matrices (SSMs) for different views on the *ACT4*[2] dataset: (a) front view (b) side view.

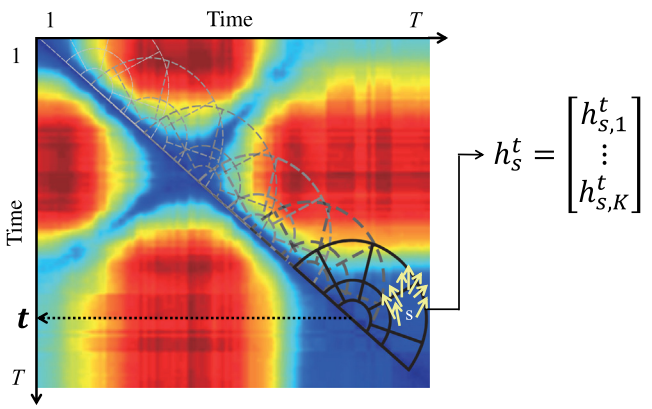

Fig. 3 An illustration of extracting local descriptors from an SSM.

et al.'s work[14]), $t \in \{1,2, \cdots, T\}$, and T is the total number of frames in a video sequence. Here, we should note that the boundaries that fall outside the SSM are simply represented by zeros as in Junejo et al.'s work.[14] Finally, a local descriptor $\mathbf{F}^t$ at time t is represented by concatenating the histograms of S blocks as follows:

$$\mathbf{F}^t = [\mathbf{h}_1^t, \mathbf{h}_2^t, \cdots, \mathbf{h}_S^t]' \in \mathbb{R}^{KS}. \quad (8)$$

5 Action Representation and Classifier Learning

Motivated by the successful application of a bag-of-features method to various pattern classification problems,[15,16] we represent the local descriptors of a video sequence with a set of quantized visual features that are learned by means of a clustering method. In the clustering method, we first learn L clusters using a k-means[6] algorithm offline and subsequently use each of the learned clusters as a visual vocabulary.[6] Given local descriptors $[\mathbf{F}^t]_{t=1,\cdots,T}$ of a video sequence, we perform a vector quantization by finding the most similar visual

vocabulary for each of the descriptors based on the Euclidean distance, and then describe a T-length video sequence in terms of a histogram by regarding the learned visual vocabularies as bins. Note that the resulting histogram summarizes the spatio-temporal information in a statistical manner. Concretely, the local descriptors $[\mathbf{F}^t]_{t=1,\cdots,T}$ of a video sequence are now converted to an L-bin histogram, for which each of the trained visual vocabularies is a histogram bin. To handle the problem of varying numbers of frame sequences, we further normalize the histogram with the total number T of frames in a video sequence. This normalized histogram is finally fed into a classifier, for which we use a linear support vector machine (SVM) that has been successfully used in various fields.[17,18] Regarding the multi-class classification, since an SVM is essentially a binary classifier, we apply a "one-against-all" strategy due to its small computational burden compared to a "one-against-one" strategy.[19]

6 Experimental Results and Analysis

In this section, we demonstrate the effectiveness of the proposed method by conducting experiments on three public datasets, namely, $ACT4^2$, MSRAction3D, and MSRDailyActivity3D.

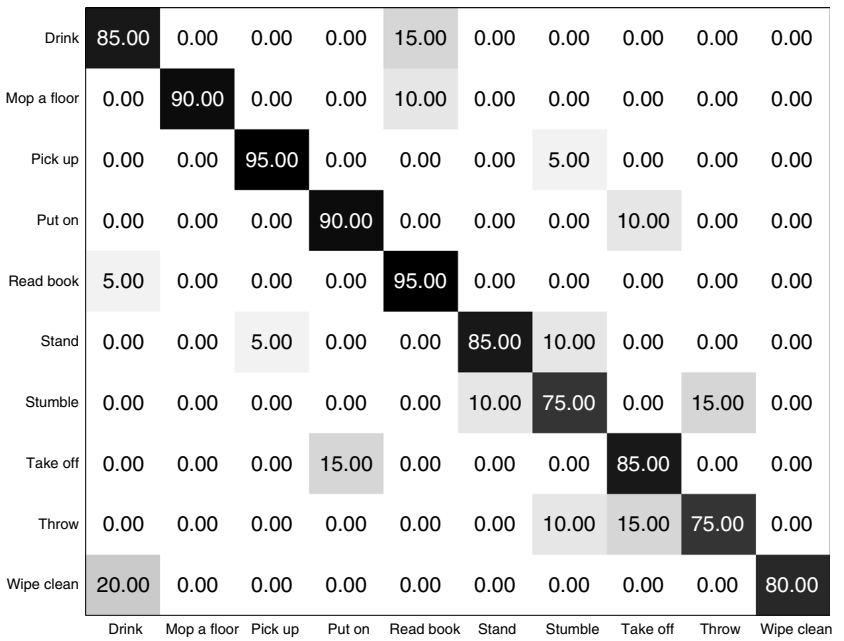

Fig. 4 Confusion matrix of action recognition under different experimental scenarios. (a) All-view and (b) cross-view.

6.1 *Image Processing and Experimental Settings*

To detect foreground silhouettes of an actor in a sequence, we sequentially applied background subtraction and morphological operations with erosion and dilation.[13] We then defined a bounding box, based on which the following feature extraction was performed by setting the size to be 1.2 times larger than the width, height, and depth of a minimal box that completely surrounds the detected actor. Subsequently, the 3-D bounding box was evenly divided into $6 \times 6 \times 6$ cells, the size of which was empirically determined. The diameter of a log-polar block structure for local descriptors was set to 28 based on Junejo et al.'s work.[14] The local descriptors were then used to define visual vocabularies by means of k-means clustering. In all our experiments, k was set to 1000. In our implementation, we used an OpenCV library[20] for SVM classifier learning and testing.

6.2 *ACT4*^{2} Dataset

This dataset contains 14 kinds of human action classes, namely, "Collapse, Drink, Make phone call, Mop floor, Pick up, Put on, Read book, Sit down, Sit up, Stumble, Take off, Throw away, Twist open, and Wipe clean," performed by 24 subjects and recorded at four viewpoints. Of the 14 actions, we chose 10 actions, i.e., "Drink, Mop floor, Pick up, Put on, Read book, Stand, Stumble, Take off, Throw away, Wipe clean," for performance comparison with the competing methods in the literature.

We considered the following two experimental scenarios:

- (All-view) We combined all video sequences captured at different viewpoints into one big dataset, i.e., video samples of all viewpoints were included in a training set. In this scenario, we measured the classification performance by means of a "leave-one-subject-out" manner. That is, we used video sequences of 23 subjects for training and those of the remaining subject for testing. This procedure was repeated 24 times. In this scenario, we aim to show the generality of our approach to all viewpoints.
- (Cross-view) We considered video sequences of all subjects captured at one specific viewpoint for training and used the remaining video sequences of all subjects captured at the other viewpoints. The experiments were repeated as many times as the total number of viewpoints, i.e., 4. With this scenario, we aim to show the validity of our method for view invariance in action recognition.

The classification performance is summarized with a confusion matrix in Fig. 4. On average, we achieved accuracies of 85.50 ± 7.24 and 82.00 ± 8.88 in the scenarios of all-view and cross-view, respectively. Overall, the averaged classification accuracy in the cross-view scenario was lower than that in the all-view scenario, as much as 3.50%, although the accuracies of "Mop a floor, Put on, Take off" actions were higher in the cross-view scenario.

We also compared our method with the competing methods in terms of modality (color or depth),[10] feature type (histogram of oriented gradients and histogram of optical flow: HOGHOF,[10] comparative coding descriptor: CCD[21]), in the literature. Figure 5 illustrates the performance of four different methods on the *ACT4*^{2} dataset. It is remarkable that the proposed method clearly outperforms the three competing methods.

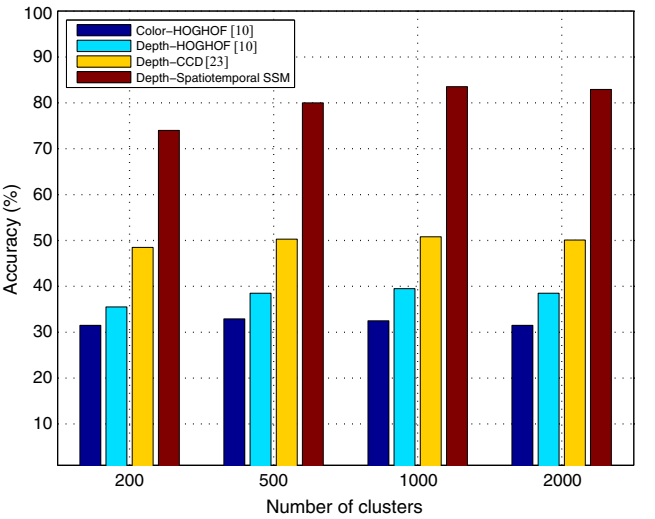

Fig. 5 Performance comparison with the previous methods for cross view.

In Table 1, we also compared the classification performance between two different modalities of color and depth, but using the same descriptors, i.e., VSFR, proposed in this paper. The accuracies of color-VSFR in all-view and cross-view were 63.6% and 51.6%, respectively, while the accuracies of depth-VSFR in all-view and cross-view were 85.5% and 82%, respectively. These results show that the proposed method achieved 21.9% and 30.4% higher accuracies than the color-VSFR method in the all-view and cross-view scenarios, respectively.

6.3 *MSRAction3D and MSRDailyActivity3D Datasets*

The MSRAction3D dataset contains 20 action types performed by 10 subjects and recorded by using a depth camera at a frontal view. We compared our method with the previous methods of Vieira et al.,[22] Wang et al.,[23] Lu and Aggarwal,[24] and Oreifej and Zicheng[25] in Table 2. From the table, we can see that the proposed method achieved the best performance among the competing methods with an accuracy of 91.3%.

The MSRDailyActivity3D dataset, which collected daily activities in a more realistic setting having background objects and persons appearing at different distances from the camera, contains 16 types of human actions captured in a frontal view. We compared our method with the previous methods of Wang et al.,[26] Dollar et al.,[27] Laptev et al.,[28] and Lu and Aggarwal[24] by considering only the front view

Table 1 Performance comparison with different modalities on *ACT4*^{2} dataset.

	Accuracy (%)	
Method	All-view	Cross-view
Color-VSFR [22]	63.6	51.6
Depth-VSFR (proposed)	**85.5**	**82.0**

Note: The boldface denotes the best accuracy for each case.

Table 2 Comparison of the classification accuracy on MSRAction3D dataset.

Method	Accuracy (%)
STOP[22]	84.8
ROP[23]	86.5
DCSF[24]	89.3
HON4D[25]	88.9
Ours	**91.3**

Note: The boldface denotes the best accuracy.

Table 3 Comparison of the classification accuracy on MSRDaily Activity3D dataset.

Method	Accuracy (%)
Wang et al.[26]	68.0
Dollar et al.[27]	73.6
Laptev et al.[28]	79.1
Lu and Aggarwal[24]	83.6
Ours	**89.7**

Note: The boldface denotes the best accuracy.

from the depth sequence. As presented in Table 3 (the performance of all the competing methods was obtained from Lu and Aggarwal's work[24]), the proposed method significantly outperformed the competing methods.

7 Conclusion

In this paper, we proposed a VSFR method for view-invariant action recognition. Specifically, we constructed a histogram that presented the density of 3-D point clouds in a 3-D grid. Our method has the capability to recognize human actions robust to viewpoint changes because movement of 3-D point clouds includes spatial and temporal information. Moreover, it is faster and produces more accurate measurements than other methods that extract the skeletal joint information. We then calculated the similarity of the VSFR using SSM. The robustness of the SSM for viewpoint changes indicates that our method does not require additional training data for each viewpoint. Finally, we extracted the local descriptors of the SSM and used a linear SVM for classification with a one-versus-all strategy.

Acknowledgments

This work was partly supported by the ICT R&D program of MSIP/IITP [14-824-09-005, Development of Predictive Visual Intelligence Technology] and also supported by the Implementation of Technologies for Identification, Behavior, and Location of Human based on Sensor Network Fusion Program through the Ministry of Trade, Industry and Energy (Grant No. 10041629).

References

1. R. Poppe, "A survey on vision-based human action recognition," *Image Vis. Comput.* **28**(6), 976–990 (2010).
2. D. Weinland, R. Ronfard, and E. Boyer, "A survey of vision-based methods for action representation, segmentation and recognition," *Comput. Vis. Image Und.* **115**(2), 224–241 (2011).
3. J. Shotton et al., "Real-time human pose recognition in parts from a single depth image," in *Proc. 2012 IEEE Computer Society Conf. on Computer Vision and Pattern Recognition Workshops*, IEEE (2011).
4. M. Holte et al., "A local 3-D motion descriptor for multi-view human action recognition from 4-D spatio-temporal interest points," *IEEE J. Sel. Topics Signal Process.* **6**(5), 553–565 (2012).
5. I. Junejo et al., "View-independent action recognition from temporal self-similarities," *IEEE Trans. Pattern Anal. Mach. Intell.* **33**(1), 172–185 (2011).
6. T. Kanungo et al., "An efficient k-means clustering algorithm: analysis and implementation," *IEEE Trans. Pattern Anal. Mach. Intell.* **24**(7), 881–892 (2002).
7. P. Huang, A. Hilton, and J. Starck, "Shape similarity for 3D video sequences of people," *Int. J. Comput. Vis.* **89**(2–3), 362–381 (2010).
8. L. Xia, C.-C. Chen, and J. Aggarwal, "View invariant human action recognition using histograms of 3D joints," in *Proc. 2012 IEEE Computer Society Conf. on Computer Vision and Pattern Recognition Workshops*, pp. 20–27, IEEE, Providence (2012).
9. M.-C. Roh, H.-K. Shin, and S.-W. Lee, "View-independent human action recognition with volume motion template on single stereo camera," *Pattern Recognit. Lett.* **31**(7), 639–647 (2010).
10. Z. Cheng et al., "Human daily action analysis with multi-view and color-depth data," *Lec. Notes Comput. Sci.* **7584**, 52–61 (2012).
11. C. Benabdelkader, R. Cutler, and L. Davis, "Gait recognition using image self-similarity," *EURASIP J. Adv. Signal Process.* **2004**(4), 572–585 (2004).
12. E. Shechtman and M. Irani, "Matching local self-similarities across images and videos," in *Proc. 2007 IEEE Computer Society Conf. on Computer Vision and Pattern Recognition*, pp. 1–8, IEEE, Minneapolis (2007).
13. R. C. Gonzalez and R. E. Woods, *Digital Image Processing*, 3rd ed., Prentice-Hall, Inc., Upper Saddle River, New Jersey (2006).
14. N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proc. 2005 IEEE Computer Society Conf. on Computer Vision and Pattern Recognition*, Vol. 1, pp. 886–893, IEEE, San Diego (2005).
15. J. Niebles, H. Wang, and L. Fei-Fei, "Unsupervised learning of human action categories using spatial-temporal words," *Int. J. Comput. Vis.* **79**(3), 299–318 (2008).
16. I. Laptev, B. Caputo, and T. Lindeberg, "Local velocity-adapted motion events for spatio-temporal recognition," *Comput. Vis. Image Und.* **108**(3), 207–229 (2007).
17. H.-I. Suk and S.-W. Lee, "A novel Bayesian framework for discriminative feature extraction in brain-computer interfaces," *IEEE Trans. on Pattern Anal. Mach. Intell.* **35**(2), 286–299 (2013).
18. C.-C. Chang and C.-J. Lin, "LIBSVM: a library for support vector machines," *ACM Trans. on Intell. Syst. Technol.* **2**(3), 1–27 (2011).
19. J. Milgram, M. Cheriet, and R. Sabourin, "'One Against One' or 'One Against All': Which one is better for handwriting recognition with SVMs?," in *Proc. Tenth Int. Workshop on Frontiers in Handwriting Recognition*, SuviSoft, La Baule, France (2006).
20. G. Bradski, "OpenCV," *Dr. Dobb's J. Software Tools* (2000).
21. H. Wang et al., "Evaluation of local spatio-temporal features for action recognition," in *Proc. British Machine Vision Conf.*, pp. 124.1–124.11, London, UK (2009). .
22. A. Vieira et al., "STOP: Space-time occupancy patterns for 3D action recognition from depth map sequences," *Lec. Notes Comput. Sci.* **7441**, 252–259 (2012).
23. J. Wang et al., "Robust 3D action recognition with random occupancy patterns," *Lec. Notes Comput. Sci.* 872–885, (2012).
24. X. Lu and J. Aggarwal, "Spatio-temporal depth cuboid similarity feature for activity recognition using depth camera," in *Proc. 2013 IEEE Computer Society Conf. on Computer Vision and Pattern Recognition*, pp. 2834–2841, IEEE, Portland (2013).
25. O. Oreifej and Z. Liu, "HON4D: Histogram of oriented 4D normals for activity recognition from depth sequences," in *Proc. 2013 IEEE Computer Society Conf. on Computer Vision and Pattern Recognition*, pp. 716–723, Portland (2013).
26. J. Wang et al., "Mining actionlet ensemble for action recognition with depth cameras," in *Proc. 2012 IEEE Computer Society Conf. on Computer Vision and Pattern Recognition*, pp. 1290–1297, IEEE, Providence (2012).
27. P. Dollar et al., "Behavior recognition via sparse spatio-temporal features," in *Proc. 2nd Joint IEEE Int. Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance, 2005*, pp. 65–72, IEEE, Washington (2005).
28. I. Laptev et al., "Learning realistic human actions from movies," in *Proc. 2008 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 1–8, IEEE, Anchorage (2008).

Seong-Sik Cho received the BS degree in information and communication engineering from Pai Chai University, Daejeon, Republic of Korea, in 2006 and the MS degree in computer science and engineering from Korea University, Seoul, Republic of Korea, in 2008. Currently, he is a PhD candidate at Korea University, Republic of Korea. His research interests include human behavior analysis, sign language recognition, and gesture recognition.

A-Reum Lee received the MS degree in computer science and engineering, Korea University, Republic of Korea, in 2014. Her research interests include computer vision and pattern recognition.

Heung-Il Suk received the PhD degree in computer science and engineering, Korea University, Republic of Korea, in 2012. From 2012 to 2014, he was a postdoctoral research associate in the University of North Carolina at Chapel Hill, USA. Since March 2015, he has been an assistant professor in the Department of Brain and Cognitive Engineering at Korea University. His research interests include machine learning, neuroimaging analysis, and brain-computer interfaces.

Jeong-Seon Park received the PhD degree in the Department of Computer Science from Korean University, Seoul, Republic of Korea, in 2005. Since 2005, she has been a professor in the Department of Multimedia, Chonnam National University, Chonnam, Republic of Korea. Her research interests include image processing, pattern recognition, and IT convergence application.

Seong-Whan Lee received the PhD degree in computer science from Korea Advanced Institute of Science and Technology, Seoul, in 1989. From February 1989 to February 1995, he was an assistant professor at Chungbuk National University, Cheongju, Republic of Korea. In March 1995, he joined the faculty of the Department of Computer Science and Engineering at Korea University, Republic of Korea, and currently, he is the Hyundai-Kia Motor chair professor and the head of the Department of Brain and Cognitive Engineering at Korea University. He is a Fellow of IEEE.