Hindawi
Mathematical Problems in Engineering
Volume 2018, Article ID 8201509, 7 pages
https://doi.org/10.1155/2018/8201509

Research Article

Group Sparse Regression-Based Learning Model for Real-Time Depth-Based Human Action Prediction

Meng Li, Liang Yan, and Qianying Wang

School of Mathematics and Statistics, Hebei University of Economics and Business, Shijiazhuang, Hebei 050061, China

Correspondence should be addressed to Meng Li; mli269-c@my.cityu.edu.hk

Received 24 July 2018; Revised 19 November 2018; Accepted 6 December 2018; Published 24 December 2018

Guest Editor: Ayed A. Salman

This paper addresses the problem of predicting human actions in depth videos. Due to the complex spatiotemporal structure of human actions, it is difficult to infer ongoing human actions before they are fully executed. To handle this challenging issue, we first propose two new depth-based features called pairwise relative joint orientations (PRJOs) and depth patch motion maps (DPMMs) to represent the relative movements between each pair of joints and human-object interactions, respectively. The two proposed depth-based features are suitable for recognizing and predicting human actions in real-time fashion. Then, we propose a regression-based learning approach with a group sparsity inducing regularizer to learn action predictor based on the combination of PRJOs and DPMMs for a sparse set of joints. Experimental results on benchmark datasets have demonstrated that our proposed approach significantly outperforms existing methods for real-time human action recognition and prediction from depth data.

1. Introduction

Predicting ongoing human actions based on incomplete observations plays an important role in many real-world applications such as surveillance, clinical monitoring, and human-robot interaction. Despite significant research efforts in the past decade, it is still a challenging task to represent human actions for action prediction due to the complex articulated essence of human movements performed under a variety of scenarios. In addition, some actions may include human-object interactions in the environment, which increases the difficulty of action representation.

Recently introduced cost-effective depth cameras largely ease the task of action representation due to the availability of 3D joint locations of human skeleton and depth map data describing actions. It has already run into a common view that knowing the 3D joint locations is helpful for describing the articulated nature of human actions. With the 3D locations of skeletal joints, skeletal action representation can be performed by characterizing their variations over time. The skeletal action representation has resulted in an interest in skeletal human action prediction.

Most of existing skeletal action prediction methods focus on predicting human actions using orientation of joint movements. However, these skeletal features model actions simply as the motion of individual joints, which is limited in capturing complex spatiotemporal relations among joints. Moreover, it is insufficient to use the 3D joint locations without local appearance to fully model a human action, especially when the action involves the interactions between human and external objects. Although many appearance features extracted from depth map data have been proposed in recent years, these features do not provide real-time processing times.

This paper presents a novel action prediction approach with a depth camera in real-time fashion. The flowchart of the proposed approach is illustrated in Figure 1 (left panel). We first propose two new depth-based features called pairwise relative joint orientations (PRJOs) and depth patch motion maps (DPMMs) extracted from skeletal and depth map data. The PRJOs and DPMMs are used to represent the relative movement between each pair of joints and local depth appearance of interactions between human and environmental objects over the duration of a human action. These two features complement each other as a bundle for each individual joints and are suitable for real-time prediction. Then, we associate these two features for each individual joint as a bundle and propose the sparse regression-based

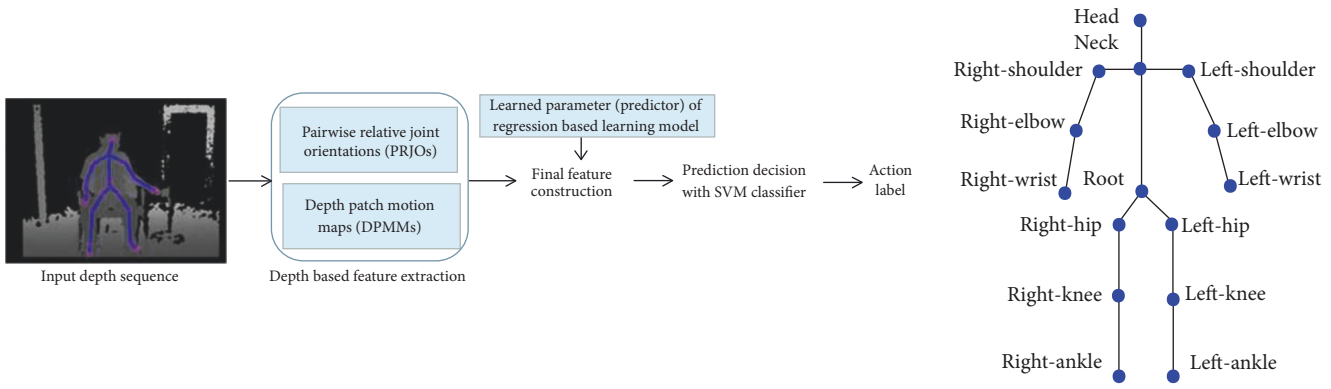

Figure 1: The flowchart of the proposed approach (left panel) and skeletal joints captured by the depth camera (right panel).

learning model which utilizes the group sparsity to select the associated features of active joints for each action class and utilize them to learn predictor for real-time action prediction.

Our main contributions include three aspects: (1) We propose a group sparse regression-based learning model as a new way to learn action predictor using selected discriminative features for different action classes. (2) We propose a skeletal feature called pairwise relative joint orientations (PRJOs) to describe the relative movement between each pair of joints. Different from other existing skeletal features for real-time action prediction, the PRJTs can encode the complex spatiotemporal relations among joints. (3) We propose a depth appearance-based feature called depth patch motion maps (DPMMs) to characterize human-object interactions. The DPMMs are more efficiently computed than other common appearance features.

After a brief review of the related work in Section 2, the two depth-based features are described in Section 3. Section 4 presents the group sparse regression-based learning model and its learning method. Section 5 presents the experimental evaluations. The conclusions are provided in Section 6.

2. Related Work

We first review human action prediction methods based on RGB data. Then, we review existing feature representations extracted from depth videos.

2.1. Action Prediction in RGB Videos. Recent efforts on human action prediction are mainly focusing on predicting actions based on RGB videos. Hoai et al. [1] introduced an online Conditional Random Field method for human intent prediction. Ryoo et al. [2] presented a dynamic Bag-of-Words (BoW) method for action prediction. In this method, the entire BoW sequence is divided into subsegments to find the structural similarity between them. Based on a Nave–Bayes-Nearest-Neighbor classifier, Yang et al. [3] proposed an action classification approach which can achieve similar levels of accuracy after seeing only 15-20 frames of an action sequence as opposed to the full action observation. This method is in essence used to predict actions. Ryoo et al. [4] designed a method for early recognition of human actions from streaming videos. Wang et al. [5] developed a Markov-based method for early prediction of human actions, aimed at human-robot interaction. Cao et al. [6] predicted actions from unfinished videos based on a set of completely observed training video action samples. Kong et al. [7, 8] extended the SVM and built multiple temporal scale templates to predict actions. Xu et al. [9] intended to mine discriminative patches to autocomplete partial videos for action prediction. Kitani et al. [10] predicted destinations of pedestrian based on semantic scene understanding method. Li et al. [11] performed action prediction through capturing the causal relationships between constituent actions and the predictable characteristics of actions. Walker et al. [12] introduced an unsupervised approach to predict the possible change of scene with time. These methods predict human actions from RGB sequences. Although they have made significant advance in action prediction, they cannot capture rich spatiotemporal information of actions very well due to the limitation in capturing highly articulated motions from RGB data.

2.2. Action Analysis in Depth Videos. Recently, action analysis with depth cameras have attracted significant attention from many researchers. In the literature, how to mine a powerful depth-based feature representation for action analysis is one of the most fundamental research topics [13–16]. Depth-based features can be classified into two major classes. The first are skeletal features, which extract information from the provided 3D locations of joints on each frame of the depth sequence. The skeletal features make it easier to represent an articulated motion as a set of movements of body parts according to locations of joints. However, most existing skeletal features are extracted for action classification. Although Reily et al. [17] proposed skeletal features for action prediction, their skeletal features cannot model the complex structure among joints in motion. The readers are referred to [18] for a systematic review of action analysis methods based on skeletal representation, respectively. The other group consists of depth appearance features which are extracted proposed directly from depth map data. A lot of depth appearance features have been proposed in recent years [19–23]. These features mainly focus on off-line computation. Different from the previous depth-based features, in this paper, we propose the pairwise relative joint orientations (PRJOs) and depth patch motion maps (DPMMs) to characterize

the spatiotemporal relations among joints and the depth appearance of human-object interaction for real-time action prediction. Moreover, we associate the PRJOs and DPMMs into different feature groups according to different joints and learn the group weights based on group spare regulation, which was not considered in the previous work. The resulting group sparse weight matrices help to select the discriminative feature structures for real-time action prediction.

3. Depth-Based Feature Construction

In this section, a detailed description of two proposed depth-based features is given: the PRJTs and the DPMMs. These features can characterize the spatiotemporal relations among joints and the depth appearance of human-object interaction, respectively.

3.1. Pairwise Relative Joint Orientations. For a human action in depth video, suppose that S joint locations of a human body are detected by the skeleton detector provided by Shotton et al. [31]. Let $l_i(t) = (x_i(t), y_i(t), z_i(t))$ $(1 \le i \le S, 1 \le t \le T)$ be the 3D coordinates of i-th joint at frame t. The human body represented by 15 skeletal joints is shown in Figure 1 (right panel). The coordinates are normalized so that the motion is invariant to the absolute body position, the body size, and the initial body orientation. The trajectory of each joint in 3D space is spatially decomposed into three 2D joint trajectories, through projecting the original 3D joint trajectory onto orthogonal Cartesian planes. In our consideration, inspired by the observation of human skeletal actions, the relative movements between various skeletal joints provide a more meaningful description than their absolute movements (clapping is more intuitively described using the relative movements between the two hand joints). Hence, we describe the 2D trajectory of one joint relative to another instead of 2D trajectory of individual joint on each plane to capture the spatiotemporal variations between each joint pair. This relative joint trajectory is represented using a histogram of the oriented angles between temporally adjacent direction vectors. Let $d_i^j(t) = l_i(t) - l_j(t)$ be the direction vector of i-th joint l_i relative to j-th joint l_j at frame t in an orthogonal Cartesian plane; the oriented angles of temporally adjacent $d_i^j(t)$ is given by

$$\theta_i^j(t) = \arccos \frac{d_i^j(t) \cdot d_i^j(t+1)}{\left\| d_i^j(t) \right\| \left\| d_i^j(t+1) \right\|}, \quad t = 1, \ldots, T, \quad (1)$$

where $\theta_i^j(t) \in (-\pi, \pi]$ (Figure 2 (left panel)). Then, $\Theta_i(t) = \{\theta_i^j(t) \mid j \in \{1, \ldots, S\}\}$ is given as a histogram of the oriented angles calculated to represent spatiotemporal relations between joint l_i and the other joints. Moreover, in order to encode long-term temporal relationships, $\Theta_i(t)$ is processed using the Fourier temporal pyramid (FTP) proposed by Wang et al. [5]. As a result, we obtain the pairwise relative joint orientations (PRJOs) feature $\widehat{\Theta}_i$.

3.2. Depth Patch Motion Maps. While the PRJOs features can characterize the relative movement between joints, they cannot accurately record the interactions between human and object. As a result, another depth-based feature is designed to describe the depth appearance of human-object interaction. DMMs proposed by Yang et al. [10] can effectively encode the shape and motion cues of a depth sequence. In this paper, based on DMMs, we propose the depth patch motion maps (DPMMs) to describe the temporal dynamics of the depth appearances of human-object interaction according to 3D locations of joints.

First, in sake of computational simplicity, we project depth frames onto three orthogonal Cartesian planes as in Yang et al. [20]. More specifically, the three 2D projected maps correspond to front, side, and top views, denoted by map_v, where $v \in \{f, s, t\}$. Different from Yang et al. [20], each projected map is divided into different local patches according to the locations of joints on each frame (Figure 2 (right panel)), and the motion energy is computed without thresholding.

Then, for a depth sequence with T frames, the depth patch motion map (DPMM) of joint l_i under projection view v is given by stacking the motion energy across an entire depth sequence as follows:

$$\text{DMM}_v^i = \sum_{t=a}^{b} \left| \text{map}_v^i(t+1) - \text{map}_v^i(t) \right|, \quad (2)$$

where $a \in \{1, \ldots, T\}$ and $b \in \{1, \ldots, T\}$. DMM_v^i represents the temporal dynamics of the depth appearances around joint l_i. Since the HOG descriptors of the DPMMs are not calculated as done in [20] and image resizing process is applied to DPMMs but not to each projected map as done in [20], the computational complexity of the feature extraction is greatly reduced.

3.3. Combination of Depth-Based Features. For each joint l_i, we use the concatenation of the PRJOs feature and DPMMs feature $G_i = \{\widehat{\Theta}_i; \text{DMM}_v^i\}$ to denote the overall depth-based feature vector of its corresponding joint l_i, where $\widehat{\Theta}_i$ is the PRJOs feature and DMM_v^i is the DPMMs feature.

4. Group Sparse Regression-Based Learning Model

In this section, we propose a group sparse regression-based learning model for human action prediction with a depth camera.

4.1. Model Formulation. To train a human action predictor, we divide each completed training sequence into M segments as in [7, 8]. Let $\text{X}_m = [\text{x}_{m,1}, \ldots, \text{x}_{m,N}]$ be the feature matrix based on the concatenation of PRJOs and DPMMs for N training samples of the m-th segment, in which $\text{x}_{m,n} \in \mathfrak{R}^d$ and operator ϕ partitions each $\text{x}_{m,n}$ into S parts according to the number of joints in each frame. Labeling a subsequence as the label of its full sequence could make confusing. To solve this problem, we learn a label for each subsequence and define the label of the m-th subsequence of full sequence. The corresponding labels of $\text{x}_{m,n}$ for C action classes are

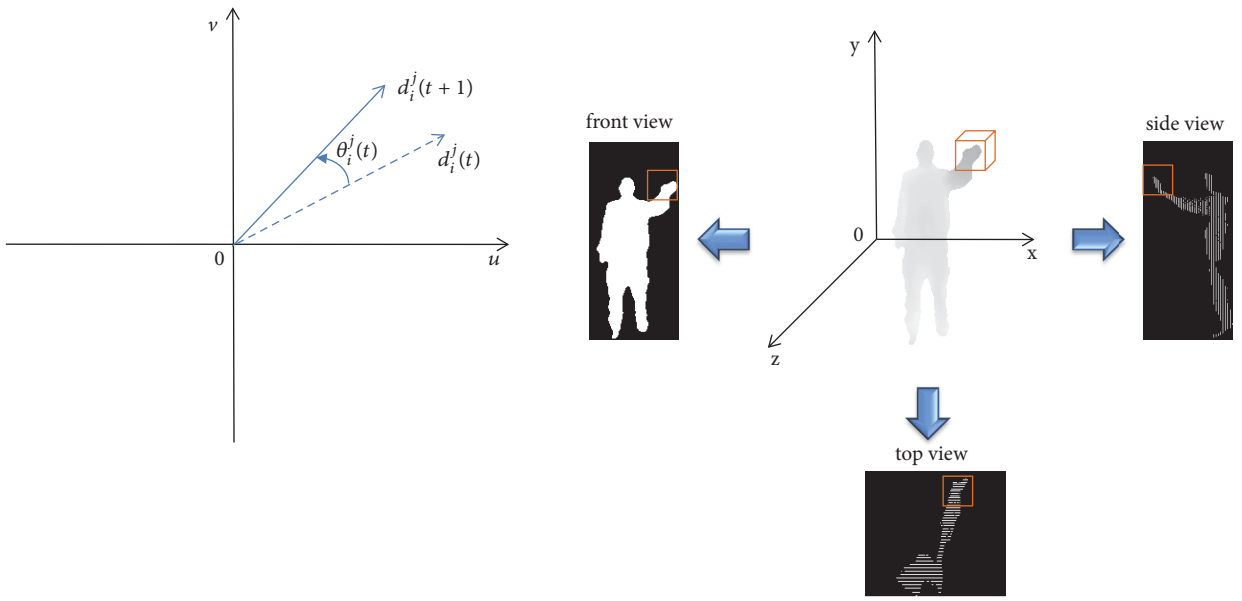

Figure 2: The oriented angle $\theta_i^j(t)$ in the $u-v$ orthogonal Cartesian plane ($\{u, v\} \subset \{x, y, z\}$) (left panel) and the local patches according to the locations of hand joint in the three 2D projected depth maps of the 3D depth map (right panel).

Input: $\lambda, \mathrm{X}_m \in \mathfrak{R}^{d\times n}, \mathrm{Y}_m \in \mathfrak{R}^{C\times n}$.
Output: W_m
1: Let t=1. Initialize $\mathrm{W}_m(t) \in \mathfrak{R}^{d\times C}$.
2: **while** not converge **do**
3: Calculate the block diagonal matrix $\mathrm{D}_m(t)$, where the ϕ_s-th diagonal block of $\mathrm{D}_m(t)$ is $(1/2\|\mathrm{w}_m^{\phi_s}(t)\|_2)\mathrm{I}_{\phi_s}$.
4: Update W_m with $\mathrm{W}_m(t+1) = (\mathrm{X}_m\mathrm{X}_m^T + \lambda\mathrm{D}_m(t))^{-1}\mathrm{X}_m\mathrm{Y}^T$.
5: $t = t + 1$
6: **end while**

Algorithm 1: Our approach for optimizing the objective function in (3).

$\mathrm{Y}_m = [\mathrm{y}_{m,1}, \ldots, \mathrm{y}_{m,C}]$ with $\mathrm{y}_{m,c} \in \{0, 1\}^N$ and $\forall n : \sum_{c=1}^{C} y_{m,c}^n = 1$. To obtain the weight matrix set $\{\mathrm{W}_1, \ldots, \mathrm{W}_M\}$ as action predictor based on mining the discriminative features of each type of input samples with M segments, we propose a group sparse regression-based learning model as follows:

$$\min_{\{\mathrm{W}_m\}} \sum_{m=1}^{M}\sum_{c=1}^{C}\sum_{n=1}^{N} \left\|\mathrm{w}_{m,c}^T x_{m,n} - y_{m,c}^n\right\| + \lambda \sum_{m=1}^{M}\sum_{c=1}^{C} \left\|\mathrm{w}_{m,c}\right\|_{2,1|\phi}. \quad (3)$$

4.2. Model Optimization. To optimize predictor W_m, we can obtain

$$\mathrm{W}_m = \left(\mathrm{X}_m\mathrm{X}_m^T + \lambda\mathrm{D}_m\right)^{-1}\mathrm{X}_m\mathrm{Y}^T \quad (4)$$

in which D_m is a block diagonal matrix with the ϕ_s-th diagonal block as $(1/2\|\mathrm{w}_m^{\phi_s}\|_2)\mathrm{I}_{\phi_s}$, I_{ϕ_s} is an identity matrix, and $\mathrm{w}_m^{\phi_s}$ is the s-th part of W_m obtained by operator ϕ in which $s \in \{1, \ldots, S\}$. Since $\mathrm{w}_m^{\phi_s}$ is dependent on W_m, we give an iterative algorithm described in Algorithm 1.

4.3. Activity Prediction. Given an ongoing action sequence, we first extracted the depth-based feature x based on the PRJOs and DPMMs. Then, based on $W^T x$ as final feature representation, a linear SVM classifier is employed to make the final prediction decision.

5. Experiments

In this section, we evaluated our approaches on three public benchmarks. Throughout our experiments, we apply the LIBSVM software provided by Chang et al. [32] with our final feature representation to train our linear SVM classifier.

5.1. Experimental Setting. According to its intraclass variations and choices of action classes, MSR-Daily Activity dataset [5] is one of the most challenging benchmarks for human action recognition. This dataset contains 16 types of actions: drink, eat, read book, call cell phone, write on a paper, use laptop, use vacuum cleaner, cheer up, sit still, toss paper, play electronic game, lie down on sofa, walk, play a guitar, stand up, and sit down. A skeleton has 20 joint positions. The total number of the action samples is 320. Most of the actions involve human-objective interactions. UTKinect-Action dataset [27] consists of depth sequences

Table 1: Comparisons with other real-time existing methods.

MSR-Daily Activity dataset	
Dynamic temporal warping [5]	54.0
Actionlet ensemble [5] (skeletal feature only)	68.0
Fourier temporal pyramid [5]	78.0
Distinctive canonical poses [24]	65.7
Relative position of joints [25]	70.0
Moving pose [26]	73.8
BIPOD representation [17]	79.7
Our approach (skeletal feature only)	85.6
Our approach	88.2
UTKinect-Action dataset	
HOJ3D [27]	90.9
Histogram of Direction vectors [28]	92.0
BIPOD representation [17]	92.8
Our approach (skeletal feature only)	94.3
Our approach	95.1
SYSU 3D HOI dataset	
ST-LSTM(Tree)+Trust Gate [29]	76.5
Part-aware LSTM [30]	76.9
BIPOD representation [17]	77.3
Our approach (skeletal feature only)	79.1
Our approach	80.7

captured using a single stationary Kinect. The 3D locations of 20 joints are provided with the dataset. This dataset contains action types: walk, sit down, stand up, pick up, carry, throw, push, pull, wave hands, and clap hands. There are 10 subjects; each subject performs each action twice. The SYSU 3D HOI dataset [33] is a new challenging action recognition dataset. This dataset contains 12 action types: drinking, pouring, calling phone, playing phone, wearing backpacks, packing backpacks, sitting chair, moving chair, taking out wallet, taking from wallet, mopping, and sweeping. The 3D locations of 20 joints are associated with each frame of the human action sequence.

For the MSR-Daily Activity dataset, UTKinect-Action dataset, and SYSU 3D HOI dataset, we first investigate the performance of our approach for recognizing human actions in real-time fashion using complete observations. We follow the same experiment setting as other related works. For the three datasets, we use half of the subjects for training and the other half for testing. Then, we perform evaluation of our approach for real-time action prediction on the three datasets.

5.2. Experimental Results. As shown in Table 1, for the MSR-Daily Activity dataset, UTKinect-Action dataset, and SYSU 3D HOI dataset, the proposed approach achieves high accuracies, which are much better than the reported results of other real-time state-of-the-art methods. Besides, it is clear that, only using skeletal feature, our approach can also perform better than other methods, since our PRJOs feature can capture the spatiotemporal relations among joints and our group sparse learning model can mine discriminative features according to the sparse joint set. Although deep learning models have achieved great progress in action recognition, they cannot model the spatial complex structure among skeletal joints very well. Moreover, the experimental results also show the benefit of the combination of our skeletal feature and depth appearance feature.

Figure 3 shows the confusion matrices for the MSR-Daily Activity dataset (left panel), the UTKinect-Action dataset (middle panel), and the SYSU 3D HOI dataset (right panel). We can see that our approach works very well. The confusions occur when the two actions are highly similar to each other like "drinking" and "calling phone" in the case of SYSU 3D HOI dataset (right panel), or the similar actions with slight movements such as "sit still" and "play electronic game" in the case of MSR-Daily Activity dataset (left panel).

Figure 4 shows the accuracy rates for early prediction of human actions for our proposed method and BIPOD representation method based on the MSR-Daily Activity dataset (left panel), the UTKinect-Action dataset (middle panel), and the SYSU 3D HOI dataset (right panel). From Figure 4, it is clear that our proposed method has pretty good performance in early action prediction. This is because our regression model makes use of the segments that contains partial action executions for obtaining a reliable predictor.

6. Conclusions

This paper presents a novel sparse regression learning approach for real-time depth-based action prediction. We

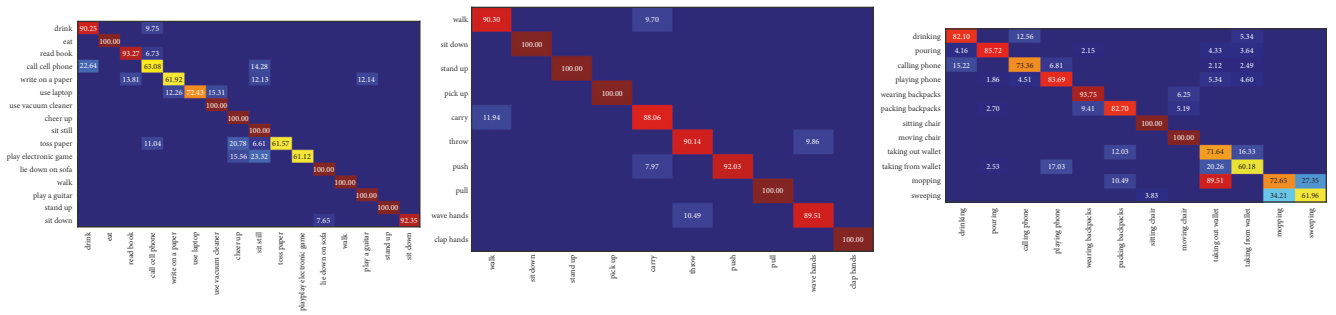

FIGURE 3: The confusion matrices for the MSR-Daily Activity dataset (left panel), the UTKinect-Action dataset (middle panel), and the SYSU 3D HOI dataset (right panel).

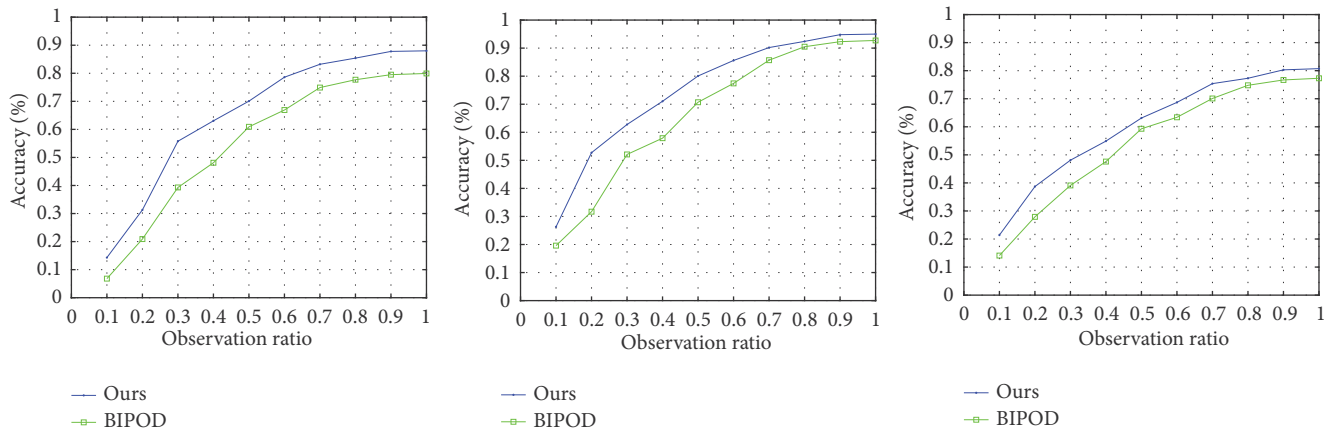

FIGURE 4: The accuracy rates showed for early prediction of human actions for our proposed method and BIPOD representation on the MSR-Daily Activity dataset (left panel), the UTKinect-Action dataset (middle panel), and the SYSU 3D HOI dataset (right panel).

first introduce the pairwise relative joint orientations (PRJOs) and depth patch motion maps (DPMMs) to construct the associated depth-based feature for describing each individual joints. Then, a group sparse regression-based learning model is proposed to learn action predictor by mining a sparse combination of the associated depth-based features for discriminatively representing all the available human action classes. Finally, an SVM classifier is trained for action prediction decision based on the learned feature representation. State-of-the-art results are achieved in different experiments, which shows the effectiveness of our proposed approach.

Data Availability

Previously reported MSR-Daily Activity dataset and UTKinect-Action dataset were used to support this study and are available at [DOI: 10.1109/TPAMI.2013.198 and DOI: 10.1109/CVPRW.2012.6239233]. These prior studies and datasets are cited at relevant places within the text as references [5, 27].

Conflicts of Interest

There are no conflicts of interest.

Acknowledgments

This work is supported by the Foundation of Heibei Department of Human Resources and Social Security (no. C201810 "河北省引进留学人员资助项目(课题)" (in Chinese)), the National Science Foundation of China (no. 61602148), and the Foundation of Heibei Educational Department (no. QN2018018).

References

[1] M. Hoai and F. De la Torre, "Max-Margin early event detectors," in *Proceedings of IEEE Internationall Conference Computer Vision and Pattern Recognition*, pp. 2863–2870, 2012.

[2] M. S. Ryoo, "Human activity prediction: Early recognition of ongoing activities from streaming videos," in *Proceedings of the 2011 IEEE International Conference on Computer Vision, ICCV 2011*, pp. 1036–1043, Spain, November 2011.

[3] L. Bi, X. Yang, and C. Wang, "Inferring driver intentions using a driver model based on queuing network," in *Proceedings of the 2013 IEEE Intelligent Vehicles Symposium, IEEE IV 2013*, pp. 1387–1391, Australia, June 2013.

[4] M. S. Ryoo, T. J. Fuchs, L. Xia et al., "Robot-centric activity prediction from first-person videos: What will they do to me?" in *Proceedings of the tenth annual ACM/IEEE international conference on human-robot interaction*, pp. 295–302, 2015.

[5] J. Wang, Z. Liu, and Y. Wu, "Learning Actionlet Ensemble for 3D Human Action Recognition," in *Human Action Recognition with Depth Cameras*, SpringerBriefs in Computer Science, pp. 11–40, Springer International Publishing, Cham, 2014.

[6] Y. Cao, D. Barrett, A. Barbu et al., "Recognize human activities from partially observed videos," in *Proceedings of the 26th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2013*, pp. 2658–2665, USA, June 2013.

[7] Y. Kong, D. Kit, and Y. Fu, "A Discriminative Model with Multiple Temporal Scales for Action Prediction," in *Computer Vision – ECCV 2014*, vol. 8693 of *Lecture Notes in Computer Science*, pp. 596–611, Springer International Publishing, Cham, 2014.

[8] Y. Kong and Y. Fu, "Max-margin action prediction machine," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 38, no. 9, 2015.

[9] Z. Xu, L. Qing, and J. Miao, "Activity auto-completion: Predicting human activities from partial videos," in *Proceedings of the 15th IEEE International Conference on Computer Vision, ICCV 2015*, pp. 3191–3199, Chile, December 2015.

[10] K. Kitani, B. Ziebart, J. Bagnell, and M. Hebert, "Activity Forecasting," in *Proceedings of the European Conference Computer Vision*, pp. 201–214, 2012.

[11] K. Li and Y. Fu, "Prediction of human activity by discovering temporal sequence patterns," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 36, no. 8, pp. 1644–1657, 2014.

[12] J. Walker, A. Gupta, and M. Hebert, "Patch to the future: Unsupervised visual prediction," in *Proceedings of the 27th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2014*, pp. 3302–3309, USA, June 2014.

[13] A. Collet, M. Martinez, and S. S. Srinivasa, "The MOPED framework: object recognition and pose estimation for manipulation," *International Journal of Robotics Research*, vol. 30, no. 10, pp. 1284–1306, 2011.

[14] D. Gehrig, P. Krauthausen, L. Rybok et al., "Combined intention, activity, and motion recognition for a humanoid household robot," in *Proceedings of the 2011 IEEE/RSJ International Conference on Intelligent Robots and Systems: Celebrating 50 Years of Robotics, IROS'11*, pp. 4819–4825, USA, September 2011.

[15] A. Pieropan, C. H. Ek, and H. Kjellstrom, "Functional object descriptors for human activity modeling," in *Proceedings of the 2013 IEEE International Conference on Robotics and Automation, ICRA 2013*, pp. 1282–1289, Germany, May 2013.

[16] C. L. Teo, Y. Yang, H. Daumé III, C. Fermuller, and Y. Aloimonos, "Towards a Watson that sees: Language-guided action recognition for robots," in *Proceedings of the Robotics and Automation (ICRA), 2012 IEEE International Conference*, pp. 374–381, 2012.

[17] B. Reily, F. Han, L. E. Parker, and H. Zhang, "Skeleton-based bio-inspired human activity prediction for real-time human–robot interaction," *Autonomous Robots*, pp. 1–18, 2017.

[18] F. Han, B. Reily, W. Hoff, and H. Zhang, "Space-time representation of people based on 3D skeletal data: A review," *Computer Vision and Image Understanding*, vol. 158, pp. 85–105, 2017.

[19] M. Ye, Q. Zhang, L. Wang, J. Zhu, R. Yang, and J. Gall, "A Survey on Human Motion Analysis from Depth Data," in *Time-of-Flight and Depth Imaging. Sensors, Algorithms, and Applications*, vol. 8200 of *Lecture Notes in Computer Science*, pp. 149–187, Springer Berlin Heidelberg, Berlin, Heidelberg, 2013.

[20] X. Yang, C. Zhang, and Y. Tian, "Recognizing actions using depth motion maps-based histograms of oriented gradients," in *Proceedings of the 20th ACM International Conference on Multimedia, MM 2012*, pp. 1057–1060, Japan, November 2012.

[21] O. Oreifej and Z. Liu, "HON4D: histogram of oriented 4D normals for activity recognition from depth sequences," in *Proceedings of the 26th IEEE Conference on Computer Vision and Pattern Recognition (CVPR '13)*, pp. 716–723, IEEE, June 2013.

[22] H. Rahmani, A. Mahmood, D. Q. Huynh, and A. Mian, "Hopc: Histogram of oriented principal components of 3d pointclouds for action recognition," in *Proceedings of the European Conference on Computer Vision*, pp. 742–757, 2014.

[23] J. Zhang, W. Li, P. O. Ogunbona, P. Wang, and C. Tang, "RGB-D-based action recognition datasets: A survey," *Pattern Recognition*, vol. 60, pp. 86–105, 2016.

[24] C. Ellis, S. Z. Masood, M. F. Tappen, J. J. Laviola Jr., and R. Sukthankar, "Exploring the trade-off between accuracy and observational latency in action recognition," *International Journal of Computer Vision*, vol. 101, no. 3, pp. 420–436, 2013.

[25] L. Seidenari, V. Varano, S. Berretti, A. Del Bimbo, and P. Pala, "Recognizing actions from depth cameras as weakly aligned multi-part Bag-of-Poses," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW '13)*, pp. 479–485, Portland, Ore, USA, June 2013.

[26] M. Zanfir, M. Leordeanu, and C. Sminchisescu, "The moving pose: an efficient 3D kinematics descriptor for low-latency action recognition and detection," in *Proceedings of the 14th IEEE International Conference on Computer Vision (ICCV '13)*, pp. 2752–2759, Sydney, Australia, December 2013.

[27] L. Xia, C.-C. Chen, and J. K. Aggarwal, "View invariant human action recognition using histograms of 3D joints," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPRW '12)*, pp. 20–27, Providence, RI, USA, June 2012.

[28] A. Chrungoo, S. S. Manimaran, and B. Ravindran, "Activity Recognition for Natural Human Robot Interaction," in *Social Robotics*, vol. 8755 of *Lecture Notes in Computer Science*, pp. 84–94, Springer International Publishing, Cham, 2014.

[29] J. Liu, A. Shahroudy, D. Xu, A. C. Kot, and G. Wang, "Skeleton-Based Action Recognition Using Spatio-Temporal LSTM Network with Trust Gates," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, no. 12, pp. 3007–3021, 2018.

[30] Y. Tang, Y. Tian, J. Lu, P. Li, and J. Zhou, "Deep Progressive Reinforcement Learning for Skeleton-Based Action Recognition," in *Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5323–5332, Salt Lake City, UT, USA, June 2018.

[31] J. Shotton, A. Fitzgibbon, M. Cook et al., "Real-time human pose recognition in parts from single depth images," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR '11)*, pp. 1297–1304, June 2011.

[32] S. Zhang, X. Liu, and J. Xiao, "On geometric features for skeleton-based action recognition using multilayer LSTM networks," in *Proceedings of the 17th IEEE Winter Conference on Applications of Computer Vision, WACV 2017*, pp. 148–157, USA, March 2017.

[33] J.-F. Hu, W.-S. Zheng, J. Lai, and J. Zhang, "Jointly learning heterogeneous features for RGB-D activity recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015*, pp. 5344–5352, USA, June 2015.