A Structured Multi-Feature Representation for Recognizing Human Action and Interaction

Bangli Liu, Zhaojie Ju, Honghai Liu*

Intelligent Systems and Biomedical Robotics Group, School of Computing, University of Portsmouth, UK

Abstract

Active research has been carried out for human action recognition using 3D human skeleton joints with the release of cost-efficient RGB-D sensors. However, extracting discriminative features from noisy skeleton sequences to effectively distinguish various human action or interaction categories still remains challenging. This paper proposes a structured multi-feature representation for human action and interaction recognition. Specifically, a novel kernel enhanced bag of semantic words (BSW) is designed to represent the dynamic property of skeleton trajectories. By aggregating BSW with the geometric feature, a GBSW representation is constructed for human action recognition. For human interaction recognition where the cooperation of each subject matters, a GBSWC representation is proposed via combining the GBSW feature with a correlation feature which addresses the intrinsic relationship between interactive persons. Experimental results on several human action and interaction datasets demonstrate the superior performances of the proposed features over the state-of-the-art methods.

*Corresponding author.
Email address: honghai.liu@port.ac.uk

September 10, 2018

Keywords:
Action recognition, Interaction recognition, RGB-D sensors, Skeleton joints, Multi-feature

1. Introduction

Human activity recognition has been an active research topic in computer vision and can be applied into many wide applications, such as video surveillance, virtual coaches, elderly care, and entertainment. Most early approaches concentrate on recognizing human activity in 2D data streams recorded by RGB cameras [1, 2]. Unfortunately, the high sensitivity to illumination conditions and texture variability of the RGB data makes it challenging to achieve accurate human activity recognition. The emergence of cost-efficient RGB-D sensors eases these difficulties and reveals a promising direction for human activity recognition by providing extra depth data. The depth information also enables 3D human skeleton joints to be easily estimated [3]. In this paper, human action denotes the activity performed by a single person, while human interaction means the interaction between two persons, and human activity includes both human action and human interaction.

A large number of research has been done for human action recognition using skeleton data. Various characteristics of skeleton joints, such as locations, angles, and geometric relationships, were utilized to model different human actions [4, 5, 6, 7, 8]. Many researchers showed that discovering features from a set of informative joints or focusing on discriminative features could improve the performance [9, 10, 11]. Recently, many convolutional

neural networks based methods have been proposed for action recognition from skeleton sequences [12, 13, 14]. Most of them focused on transforming the skeleton features into images and then adapted existing models for classification. However, it remains challenging to explore discirminative features from noisy skeleton sequences.

Although the development of RGB-D sensors has motivated considerable work conducted for human action recognition, research for human interaction is relatively unexplored. Unlike single person actions, human interaction is a behavior performed by more than one people, where the interaction relationship between people is of vital importance. Moreover, human interaction has large feature dimensions which consist of individual information as well as mutual relations. The mutual relations were typically represented by the distance between body parts in most existing methods [15, 9, 16]. The distance property could provide useful geometric information, however, it might be not effective enough to mine intrinsic characteristics embedded in diverse interaction classes. Thus, exploring high level or semantic information could help to enhance the performance of the traditional feature representation for human interaction recognition [17].

In a previous work [18], a histogram of 3D moving directions for each joint was constructed to represent the moving trend of skeleton joints using an effective histogram projection method. This feature could describe the specific tendency of skeleton joints in 3D space and was proven to be competitive in human action recognition. Based on this feature extracted from individuals, the moving similarity between body parts was proposed to describe the mutual relationship for human interaction recognition in [19]. However, sim-

ply projecting the skeleton trajectories to a histogram is not discriminative enough to represent the dynamic trajectory features. This paper proposes a novel kernel enhanced bag of semantic moving words (BSW) for both human action and interaction recognition. More specifically, a kernel function which consists of the discriminative directions weighting and the discriminative frames weighting is constructed to augment the discriminative ability of the original descriptor over non-linear skeleton representations by focusing more on salient features. The directions in BSW are grouped into semantic moving words, whose distribution over an activity sequence explicitly interprets the moving trend of skeleton joints. Based on this feature, a structured multi-feature representation is constructed for human action and interaction recognition. Experimental results on several human action and interaction datasets demonstrate the superior performance of the proposed features over the state-of-the-art methods.

The contributions of this paper are summarized as follows:

1) A novel kernel enhanced bag of semantic moving words (BSW) is designed to represent the dynamic property of skeleton trajectories. BSW augments the discriminative property of the feature via applying a dynamic weighting strategy to the extracted semantic moving words.

2) A structured multi-feature representation is proposed for human action and interaction recognition. In the proposed framework, a GBSW feature which aggregates BSW with the geometric feature of skeleton data is constructed for human action recognition and a GBSWC feature which further combines a correlation feature between body parts is built for human interaction recognition. Figure 1 shows the framework of the proposed structured

multi-feature representation.

3) Experimental results on three public datasets validate the outstanding performance of the proposed structured multi-feature representation for both human action and interaction recognition.

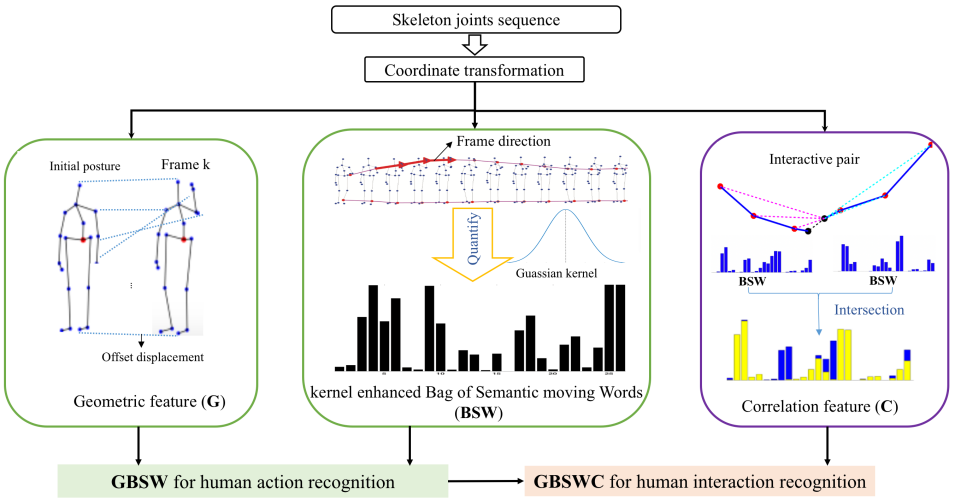


Figure 1: The framework of the proposed structured multi-feature representation. The Geometric feature (G) is built by using the offset displacement between the current frame and the initial frame. The Bag of Semantic moving Words (BSW) is calculated by accumulating the frame moving direction over the whole sequence. The Correlation feature (C) is described by computing the moving similarity between interactive pairs. GBSW which combines the G and BSW feature from individuals is developed for human action recognition, and GBSWC which further integrates the C feature between interaction pairs with GBSW is proposed for human interaction recognition.

The remainder of this paper is organized as follows: Section 2 reviews related work of human action and interaction recognition. Section 3 describes the structured multi-feature representation for human action and human interaction recognition. Section 4 presents experimental results as well as the

comparison with the state-of-the-art methods. Section 5 concludes the paper and discusses the future work.

2. Related Work

This section reviews work related to human action recognition and human interaction recognition using RGB-D data.

2.1. Human Action Recognition

A large number of approaches have been proposed for human action recognition in recent years. Existing approaches can be grouped into depth-based methods and skeleton-based methods.

The depth images can reduce the difficulty in subtracting 3D human motion information from the cluttered background by providing the distance information. Some scholars projected the 3D information onto three 2D orthogonal planes corresponding to the front, side, and top view, and then calculated features from these planes to describe actions [20, 21, 22, 23, 24, 25, 26]. For example, 3D discriminative points were selected from the body silhouette on each plane to depict body postures in [20]. Similarly, the spatial relationship among selected joints in depth sequences with discriminative shape and movement was used to build the depth context descriptor for final action recognition in [25]. The limitation of this approach is that detecting interest regions through the whole depth sequence requires extra computational cost and the spatial information may be lost. Bulbul *et al.* [22] modified DMMs [21] by providing a multi-scale and multi-directional analysis via contourlet transform. They improved the recognition accuracy by strengthening the shape characteristic of DMMs. The shape of a human body was described

delicately from various viewpoints in these DMMs, however, they did not consider the relation of 3D point neighborhood which contains useful spatial information. Liang *et al.* [26] divided each depth sequence into a set of subsequences whose shape information and motion features were extracted by the DMMs-based gradient local auto-correlations. Instead of selecting features from 2D planes, some methods captured features in a 4D space by adding the depth and time domain. With the 4D space, body shapes and movements were jointly explored from the surface normal in [27, 28, 29]. For example, Oreifej *et al.* [27] built the histogram of oriented 4D (HON4D) with the surface normal from spatiotemporal cells to capture the change of body shape and motion. Similarly, Yang *et al.*[28] proposed to group local hypersurface normals to create the super normal vector (SNV), which further preserved the correlation among local normals in the polynormal and achieved a better recognition rate compared to HON4D.

Feature representations in skeleton-based category mainly utilized different joints' information, such as joint trajectories or postures to represent action sequences. The effectiveness of sequence relationship is beneficial in identifying dynamic scences[30]. Characteristics of the spatiotemporal trajectory of skeleton joints were explored to identify actions in [4, 5, 31]. For instance, Qiao *et al.* [5] acquired trajectories in a short temporal range and then proposed a trajectorylet based on local feature representation to express both static and dynamic features. Ofli *et al.* [31] firstly captured few informative joints of each action within an instant time according to the mean or variance of joint angles. The sequences of these selected joints were then used to model human actions. Body postures are also advantageous repre-

sentations of human actions as supported by [32, 6, 7, 33]. Xia *et al.*[32] used histograms of 3D joint locations (HOJ3D) to represent key postures. Furthermore, they learned a set of visual words based on the key postures and obtained the temporal evolution of these visual words via a discrete hidden markov model. Pazhoumand *et al.* [6] depicted body poses using the angles between joints and simultaneously used the relative movement to describe their relationships in the time domain. Similarly, the 3D geometric information of a human skeleton was modeled using rotations and translations among body parts, with which actions were translated as curves in the Lie group in [33]. Instead of using the movement from all skeleton joints, only relevant joints were encoded into postures to describe their recurrent pattern, and then human activity was represented by a sequence of postures in [7]. Motivated by the great achievement of convolutional neural networks in image classification tasks [34, 35, 36], some researchers [14, 37] proposed to code the spatial and temporal information of skeleton sequences into image which were then fed into pre-trained convolutional neural network models for classification. For example, Ke *et al.* [14] utilized the spatial structural feature between the skeleton joints and four reference joints to build gray images which were then fed into a pre-trained VGGNet [38] for classification.

2.2. Human Interaction Recognition

Compared to human action recognition, human interaction has a larger feature space due to the involvement of mutual relationship between people. Some researchers decomposed human interaction into individual actions for recognition. [39, 40, 41]. In [39], each player's action was addressed separately and the final classification was achieved by applying a decision level

fusion strategy in a computer gaming scenario. Wang *et al.* [41] proposed a two-stream recurrent neural network architecture to jointly model the spatial and temporal dynamic of skeleton joints on each person, then averaged the scores as the final recognition result. These methods rely on the successful separation of subjects which is challenging due to occlusions between subjects, and ignore the relationship of human interaction by conducting action recognition on each person. Alternatively, some approaches utilized features extracted from two subjects to exhibit the spatial and motion relation over the time [15, 42, 16, 9, 43]. Yun *et al.* [15] used joint features, such as the joint movement and velocity, to represent human interaction. Ji *et al.* [9] learned both intra-frame and inter-frame features from active body part pairs studied by the contrast mining. With these features, a contrastive feature distribution model was then built to improve the recognition performance. Compared to features extracted from all joints proposed in [15], these representations are more discriminative and not computationally expensive. The interdependent relation between interactive persons is mainly represented by low-level features (e.g. the distance between body parts) in most existing methods, which might not be effective enough to reflect the intrinsic interaction pattern.

3. Structured Multi-Feature Representation

This section firstly introduces the data pre-processing which makes the extracted features invariant to different locations and orientations to the sensor. Then the structured GBSW feature is introduced for human action recognition. Finally, the correlation feature that characterizes the relation-

ship between interactive persons is further combined with GBSW to build the GBSWC feature for human interaction recognition.

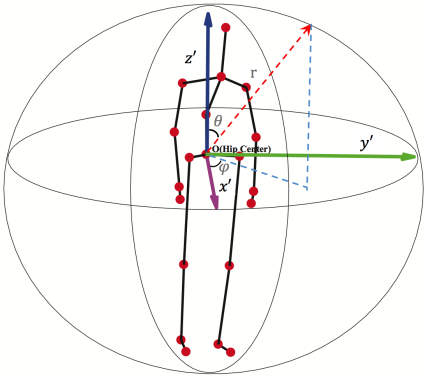


Figure 2: Person-centric coordinate system.

3.1. Data Pre-processing

Figure 2 shows the proposed person-centric coordinate system. It is defined as follows: the z' axis can be calculated using the vector from the hip center to the spine joint, and its unit vector is denoted as (a_7, a_8, a_9); the x' axis is the normal vector of a plane constructed by the spine point, left hip point and right hip point, and its unit normal vector is represented by (a_1, a_2, a_3); finally, the y' axis can be determined by the dot product of above two unit vectors, and the value of its unit y' is (a_3, a_4, a_5). Consequently, the transformation of coordinates is calculated using the following equation:

$$P = R * P' + T \tag{1}$$

where P and P' denote the original coordinate and the transformed coordinate, respectively, and T is the coordinate of the hip center $[x_h, y_h, z_h]^{-1}$. R

is the rotation matrix:

$$R = \begin{bmatrix} a_1 & a_2 & a_3 \\ a_4 & a_5 & a_6 \\ a_7 & a_8 & a_9 \end{bmatrix}^{-1}$$

This coordinate transformation makes features invariant to various locations and orientations by extracting them in the relative position rather than the absolute position.

3.2. GBSW for Action Recognition

Given a human action skeleton sequence, the geometric feature (G) and the kernel enhanced bag of semantic moving words (BSW) are extracted to represent the spatial and temporal information. Then the GBSW is constructed by aggregating the G and BSW feature for human action recognition.

3.2.1. Geometric Feature

Aiming to describe the temporal evolution of each joint, the geometric feature is represented by the offset displacement between the current frame and the initial frame. The coordinates of each joint are first transformed to the person-centric coordinate system by using Eq. 2:

$$(p_t^i)' = R^{-1} * (p_t^i - T) \tag{2}$$

where p_t^i and $(p_t^i)'$ represent the original position and the relative position of the $i-th$ joint at time t, respectively. We define the offset displacement of

$(p_t^i)'$ as $\triangle d_t^i : (\triangle x_t^i, \triangle y_t^i, \triangle z_t^i)$ which can be calculated as follows:

$$\begin{cases} \triangle x_t^i = (x_t^i)' - (x_1^i)', \\ \triangle y_t^i = (y_t^i)' - (y_1^i)', \\ \triangle z_t^i = (z_t^i)' - (z_1^i)', \end{cases} \tag{3}$$

where $((x_1^i)', (y_1^i)', (z_1^i)')$ and $((x_t^i)', (y_t^i)', (z_t^i)')$ are three transformed coordinates with respect to the *first* and $t-th$ person-centric coordinate system, respectively. The frame geometric feature is defined as follows:

$$g(t) = \{\triangle d_t^1, ..., \triangle d_t^n\} \tag{4}$$

Consequently, the sequence $G(k) = \{g(1), ..., g(f)\}$ denotes the geometric property of an action k, which can represent the temporal evolution of each joint. The cubic spline interpolation [33] is utilized to rescale the feature to cope with the different duration problem. Furthermore, the extracted geometric feature is normalized to make it scale-invariant by:

$$G_n(k) = G(k) / \parallel G(k) \parallel \tag{5}$$

3.2.2. Kernel Enhanced Bag of Semantic Moving Words

To augment the discriminative information of skeleton joints, the kernel enhanced bag of semantic moving words (BSW) is proposed, where features

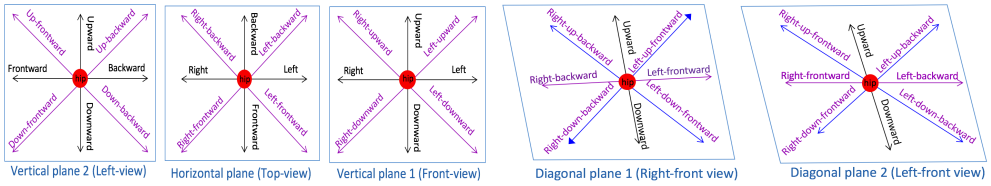
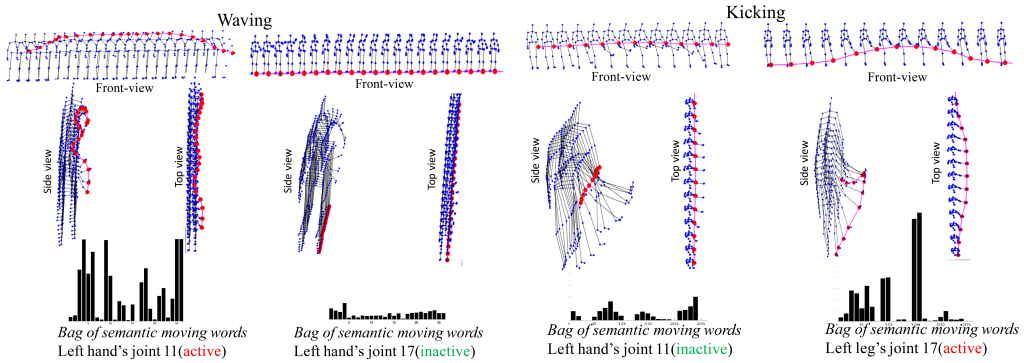


Figure 3: Samples of semantic moving words (taking $m = 26$ for example). The inclination angle θ and azimuth angle φ in this modified coordinate system are used to represent directions in the space.

in both spatial and temporal domain are jointly weighted based on their contribution to an activity. The directions in 3D space are divided into several semantic moving words $\mathbf{V_w}$ (as shown in Figure 3), and a distribution of movements over these semantic moving words is captured to interpret the moving trend of an activity sequence. As shown in Figure 4, the moving trend of the active joint is more apparent than that of the inactive joint, while the moving trend of the same joint in different classes is also diverse. Thus, it is reasonable to use the moving trend of skeleton joints to discriminate different action categories.



(a) (b)

Figure 4: The comparison of moving trends of skeleton joints. The moving trend of joints is captured from the front, side and top view. The moving characteristic for the joints in the left hand (11) and leg (17) captured from the action *waving* and *kicking* are shown in (a) and (b) respectively. Joint 11 with apparent moving property is regarded as the active one in the action *waving*, where joint 17 with few move is inactive. While joint 11 is inactive in the action *kicking* where joint 17 is opposite. To conclude, the same joint could have a different moving trend in different actions and different joints could have various moving trends in the same action.

Given a joint sequence:

$$P = \{p_1, ..., p_t, ..., p_f\} \tag{6}$$

where f is the frame number of the sequence, and p_t denotes three transformed coordinates x', y', z' using Eq. 2. The 3D direction $\mathbf{v_t}$ is captured from p_t and p_{t-1}:

$$\mathbf{v_t} = \{x'_{p_t} - x'_{p_{t-1}}, y'_{p_t} - y'_{p_{t-1}}, z'_{p_t} - z'_{p_{t-1}}\} \tag{7}$$

So the frame moving direction sequence could be defined as:

$$\mathbf{V} = \{\mathbf{v_1}, \mathbf{v_2}, ..., \mathbf{v_f}\} \tag{8}$$

To quantitatively describe the moving degree of joints in each direction, the *cosine* similarity which is an effective technique to measure the similarity between two vectors is utilized to encode the frame moving direction to predefined directions. Different from this work, this paper improves the discriminative ability of the extracted features by building a Gaussian kernel function to dynamically enhance the informative moving directions. On top of this, the BSW feature is constructed by encoding skeleton movements $\mathbf{V}$ to semantic words $\mathbf{V_w}$. The $cos\langle \mathbf{v_t}, \mathbf{v_w^i} \rangle$ is computed by:

$$cos\langle \mathbf{v_t}, \mathbf{v_w^i} \rangle = \frac{\mathbf{v_t} \cdot \mathbf{v_w^i}}{\|\mathbf{v_t}\| \|\mathbf{v_w^i}\|}, i \in [1, m] \tag{9}$$

where $\mathbf{v_t}$ is the frame direction and $\mathbf{v_w^i}$ is the $i-th$ semantic moving word. m is the number of semantic moving words.

The Gaussian kernel function using the *cosine* similarity as variable is defined as follows:

$$K(cos(\mathbf{v}_t, \mathbf{v}_w^i)) = \frac{1}{\sqrt{2\pi}\delta} exp(-\frac{(cos(\mathbf{v}_t, \mathbf{v}_w^i)) - \mu)^2}{2\delta^2}) \tag{10}$$

where μ and δ are the mean and standard deviation of *cosine* values, respectively. Here, the mean of the Gaussian function is 1 due to the trait of *cosine* values.

A soft-assignment strategy where a frame direction is distributed to multiple most relevant word candidates is achieved using a $1 \times f$ vector S. The soft voting degree is controlled by using a parameter k to determine the elements in S. For example, if $k = 3$, the frame direction is encoded to the 3 words with the top 3 similarities. Thus, the elements of S satisfy:

$$S_t = \begin{cases} 1 & \textit{if the value } cos(\mathbf{v}_t, \mathbf{v}_w^i) \textit{ belongs to} \\ & \quad k \textit{ biggest similarities} \\ 0 & \textit{otherwise} \end{cases} \tag{11}$$

To weight the frames which make bigger contributions to the whole sequence, the frame displacement $Dis(t) = \|\mathbf{v_t}\|$ is added during a quantization process. Therefore, the frame weight function can be achieved as follows:

$$w(t) = Dis(t) * K(cos(\mathbf{v}_t, \mathbf{v}_w^i)) \tag{12}$$

The final representation of each word is built by accumulating the movement through the action sequence:

$$BSW(\mathbf{v}_w^i) = \sum_{t=1}^{f} S_t * w(t) \tag{13}$$

Compared to our previous work [18], the discriminative directions weighting and discriminative frames weighting are proposed in the semantic moving words encoding. These two improvements can make the proposed method focus more on salient features of different action classes, thus improve the recognition performance.

3.3. GBSWC for Human Interaction Recognition

Apart from features from individuals, the mutual relationship between people needs to be considered for human interaction recognition. Inspired by the histogram intersection proposed in [44], this paper calculates the correlation feature between body parts from the same subject and from two subjects (referred to as intra-similarity and inter-similarity, respectively) based on the BSW.

Assuming each body part includes n joints, its moving trend feature could be summarized by traversing n joints:

$$\{BSW_1^p, BSW_2^p, ..., BSW_n^p\}, \tag{14}$$

where p denotes the body part.

The correlation feature between the corresponding word w from BSW_i (joint i) and from BSW_j (joint j) is denoted as follows:

$$\begin{aligned} &SoW\left(BSW(\mathbf{v}_w^i), BSW(\mathbf{v}_w^j)\right) \\ &\qquad = min\left(|BSW(\mathbf{v}_w^i)|, |BSW(\mathbf{v}_w^j)|\right) \end{aligned} \tag{15}$$

The histogram of semantic moving words is interpolated into the same number of frames (N). By doing this, each bin in BSW having the same dimension. Thus, the revised BSW with an $N \times n$-dimensional vector could be defined as follows:

$$\widehat{BSW} = (\overbrace{1, ..., 1}^{BSW(\mathbf{v}_1)}, \underbrace{0, ..., 0}_{N-BSW(\mathbf{v}_1)}, ..., \overbrace{1, ..., 1}^{BSW(\mathbf{v}_w)}, \underbrace{0, ..., 0}_{N-BSW(\mathbf{v}_w)}) \tag{16}$$

With Eq. 15 and Eq. 16, the intersection between two $BSMs$ is equal to the inner product between their corresponding $\widehat{BSW}$:

$$SoJ(BSW_i, BSW_j) = \widehat{BSW}_i \cdot \widehat{BSW}_j \tag{17}$$

Following Eq.17, the similarity between body parts is denoted as follows:

$$\begin{aligned} SoP_{type} &= \sum_{p=1}^{8}\sum_{q=1}^{8}\sum_{i=1}^{n}\sum_{j=1}^{n} SoJ\left(BSW_i^p, BSW_j^q\right) \\ &= \sum_{p=1}^{8}\sum_{q=1}^{8} \widehat{BSW}^p \cdot \widehat{BSW}^q \end{aligned} \tag{18}$$

where p and q are body parts, i and j are joints, and SoP_{type} could be SoP_{intra} or SoP_{inter}, which means intra-similarity or inter-similarity, respectively. $\widehat{BSW}^p$ and $\widehat{BSW}^q$ are the histogram concatenation of joints from the body part p and q, respectively. The final correlation feature (C) of body parts for each sequence is the concatenation of all body part pairs:

$$C = \{SoP_{intra1}, SoP_{inter1}, ...\} \tag{19}$$

Finally, a GBSWC representation is constructed by combining the individual GBSW feature with this correlation feature between interactive subjects for human interaction recognition.

4. Experiment Results

4.1. Datasets and Settings

This subsection introduces three adopted RGB-D datasets which are commonly used to compare the performance of human action or interaction recognition algorithms and their relative evaluation criteria. A linear SVM [45] algorithm with default parameters is applied to achieve all the recognition results of the proposed features.

4.1.1. MSR-Action3D Dataset[20]

This dataset has 20 action categories and each action is performed by 10 subjects for 2 or 3 times. The actions are grouped into three sets *AS1*, *AS2* and *AS3*, as shown in Table 1. Actions in *AS1* and *AS2* are similar, while actions in *AS3* are complex. Each action set has three tests: *Test One* (1/3 of the samples for training), *Test Two* (2/3 of the samples for training) and *Cross Subject Test* (samples from half of the subjects for training).To carry out a fair comparison, we follow two different protocols from [20] and [11] in *Cross Subject Test*. Compared to the *Cross Subject Test*, the *Test One* and the *Test Two* are less challenging since the training set contains all variations of individuals' performing styles. Many state-of-the-art methods only adopted the challenging *Cross Subject Test*. To systematically evaluate the proposed method, this paper conducts all the three evaluations.

Table 1: Three action sets of *MSR-Action3D* dataset.

AS1	*AS2*	*AS3*
Horizontal Wave	High Wave	High Throw
Hammer	Hand Catch	Forward Kick
Forward Punch	Draw X	Side Kick
High Throw	Draw Tick	Jogging
Hand Clap	Draw Circle	Tennis Swing
Bend	Hands Wave	Tennis Serve
Tennis Serve	Forward Kick	Golf Swing
Pickup & Throw	Side Boxing	Pickup & Throw

4.1.2. Florence3D-Action Dataset[46]

This dataset has 9 actions: *wave, drink from a bottle, answer phone, clap, tight lace, sit down, stand up, read watch and bow,* performed by 10 subjects for 2 or 3 times. Most of the actions, such as *answer phone* and

drink a bottle, have a great similarity. For this dataset, we follow the test setting from [33] for *Cross Subject Test*, where actions performed by half of the subjects (1,3,5,7,9) are used for training and actions performed by the remaining subjects (2,4,6,8,10) are used for testing.

4.1.3. SBU Interaction Dataset [15]

This dataset contains examples of eight different interaction classes: *approaching, departing, kicking, punching, pushing, hugging, shaking hands, exchanging something*. All the videos were collected in the same laboratory environment from a third-person perspective. The majority of the interactions involve acting-reacting relation. 21 sets consist of 7 subjects perform each category 1 or 2 times. The evaluation on this dataset contains *Test One* (1/3 of the samples for training), *Test Two* (2/3 of the samples for training) and *Cross Subject Test* (samples from half of the subjects for training).

4.2. Evaluation of the Structured Multi-feature Representation

To show the superior performance of the structured multi-feature representation, the comparison between its recognition result and that of the monotonous features on three datasets is listed in Table 2. It can be seen

Table 2: Recognition accuracy (%) of different features on *MSR-Action3D* dataset, *Florence3D-Action*, and *SBU Interaction dataset*

Feature type	*MSR-Action3D*			*Florence3D-Action*	*SBU Interaction*		
	AS1	*AS2*	*AS3*	*Cross subject*	*Test one*	*Test two*	*Cross subject*
G	50	79.5	92.4	85.9	87.68	83.33	87.67
BSW	92.4	85.7	93.3	88.0	52.17	61.11	56.16
Structured feature	**93.4**	**94.9**	**98.4**	**93.6**	**92.75**	**91.67**	**93.84**

that the structured feature could improve the performance in each experimental setting, which proves that the specific information from different types of features can complement each other. For example, the geometric feature seems to be complementary in term of spatial information to the motion feature in BSW, which enables the hybrid representation to be more discriminative among different activity categories.

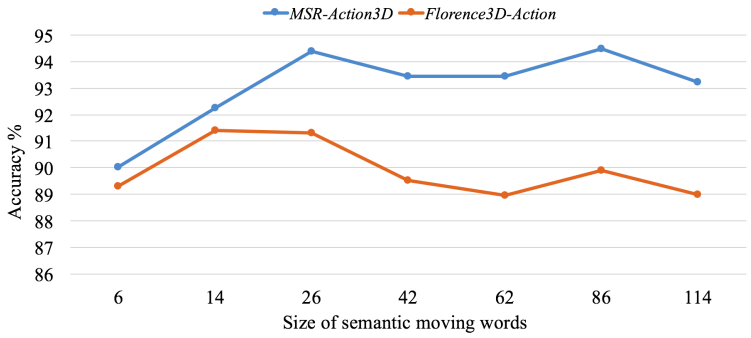


Figure 5: The average recognition accuracy of the proposed feature representation versus the size of semantic moving words on *MSR-Action3D* and *Florence3D-Action* dataset.

In addition, to evaluate the effect of the number of semantic moving words (n_s), we test the recognition performance of the proposed method with $n_s = 6, 14, 26, 42, 62, 86, 114$. The selection criterion is based on whether the constructed moving words can equally divide the 3D space and they are representative for the moving direction. Figure 5 shows the recognition accuracy when using different sizes of moving words. The accuracy increases till $n_s = 26$, while it decreases when n_s is over 26. This is because the rising number of semantic moving words augments the ambiguous moving trend between actions, which influences the discriminating capacity of the feature. Based on this finding, this paper utilizes $n_s = 26$ to get the following

performance.

Table 3: Recognition Accuracy (%) of *Test One* and *Test Two* on *MSR-Action3D*. Notation: S-skeleton; D-depth.

		Test One				*Test Two*			
Method	Feature	*AS1*	*AS2*	*AS3*	**Average**	*AS1*	*AS2*	*AS3*	**Average**
Bag of 3D Points [20]	D	89.5	89.0	96.3	91.6	93.4	92.9	96.3	94.2
DMM-HOG[47]	D	97.3	92.2	98.0	95.8	98.7	94.7	98.7	97.4
STOP[48]	S+D	98.2	94.8	97.4	96.8	**99.1**	97.0	98.7	98.3
Jalal *et al.* [49]	S+D	96.9	**98.3**	**98.7**	97.9	97.1	98.6	98.9	98.2
HOJ3D[32]	S	**98.5**	96.7	93.5	96.2	98.6	97.9	94.9	97.2
EigenJoints[50]	S	94.7	95.4	97.3	95.8	97.3	**98.7**	97.3	97.8
3GMTG[18]	S	94.7	95.0	96.8	95.5	98.5	97.8	**99.1**	98.5
GBSW	S	97.9	98.2	98.5	**98.2**	98.2	**98.7**	**99.1**	**98.7**

4.3. Comparison with State-of-the-art Methods

Evaluations have been conducted for both human action and interaction recognition. The following subsections present a comparison of the proposed method with the state-of-the-art methods in terms of recognition accuracy.

4.3.1. Action Recognition Results

The experiment on *MSR-Action3D* dataset and *Florence3D-Action* dataset aims to demonstrate the comparative results of the proposed GBSW on human action recognition.

Table 3 reports the results of *Test One* and *Test two* on the *MSR-Action3D* dataset. It can be seen that our structured representation obtains the highest average recognition rates in both cases and could achieve over 98% accuracies in most individual sets. The proposed method outperforms 3DMTG owing to the use of the kernel-based dynamic weighting algorithm. Although the best performance of the *AS2* and *AS3* in *Test One* is achieved

Table 4: Average accuracy (%) of *Cross Subject Test* on *MSR-Action3D* dataset.

Protocol from [20] (1,3,5,7,9 subjects as training)						
Method	Feature Type	Classifier	*AS1*	*AS2*	*AS3*	**Average(%)**
Bag of 3D Points (2010)[20]	Depth	-	72.9	71.9	79.2	74.7
DMM-HOG (2012)[47]	Depth	structural SVM	96.2	84.1	94.6	91.6
SNV (2014)[28]	Depth	SVM	-	-	-	93.1
STOP (2014)[48]	Depth	SVM	91.7	72.2	98.6	87.5
ROP (2012)[51]	Depth	SVM	-	-	-	86.5
DSTIP (2013)[52]	Depth	SVM	-	-	-	89.3
Liu *et al.* (2016)[25]	Depth	SVM	-	-	-	94.28
LASC (2017)[26]	Depth	CRC	-	-	-	94.6
Jalal *et al.* (2017)[49]	Skeleton+Depth	HMM	90.8	93.4	95.7	93.3
HOJ3D (2012)[32]	Skeleton	HMM	72.9	85.5	63.5	79.0
EigenJoints (2012)[50]	Skeleton	Bayes	74.5	76.1	96.4	82.3
Actionlets Ensemble (2012)[53]	Skeleton	SVM	-	-	-	88.2
Vemulapalli *et al.* (2014)[33]	Skeleton	SVM	95.3	83.8	98.2	92.5
Devanne *et al.* (2015)[54]	Skeleton	kNN	-	-	-	92.1
LM^{3}TL (2017)[55]	Skeleton	MTL	-	-	-	90.53
MIMTL (2017)[10]	Skeleton	MTL	-	-	-	93.6
3DMTG (2016)[18]	Skeleton	SVM	92.4	93.8	97.1	94.4
Lillo *et al.* (2017)[8]	Skeleton	LSSVM	94.3	92.9	99.1	95.4
DSRF (2018)[56]	Skeleton	SVM	-	-	-	95.24
GBSW	Skeleton	SVM	93.4	**94.9**	98.4	**95.6**
Procotol from [11] (1,2,3,4,5 subjects as training)						
Pose set (2013)[57]	Skeleton	SVM	-	-	-	90.2
Moving Pose (2013)[11]	Skeleton	kNN	-	-	-	91.3
3DMTG (2016)[18]	Skeleton	SVM	87.50	95.8	94.7	92.7
Lillo *et al.* (2017)[8]	Skeleton	LSSVM	-	-	-	93.0
GBSW	Skeleton	SVM	**88.9**	**96.2**	**95.5**	**93.5**

by Jalal *et al.* [49] (98.3% and 98.7% respectively), the difference of the performance is tiny since our method also achieves an accuracy of 98.2% and 98.5% respectively. The highest accuracy of *AS1* in *Test Two* achieved by STOP [48], which jointly utilizes the skeleton and depth information, indicates that the recognition performance of similar actions might be improved

Table 5: Average accuracy(%) of *Cross Subject Test* on *Florence3D-Action*. dataset

Methods	Feature type	Classifier	Accuracy
Multi-Part Bag-of-Poses (2013) [46]	Skeleton	Nearest-neighbor	82.0
Full skeleton (2015)[54]	Skeleton	kNN	85.9
Body part (2015)[54]	Skeleton	kNN	87.0
Vemulapalli *et al.* (2014)[33]	Skeleton	SVM	90.9
3DMTG (2016)[18]	Skeleton	SVM	91.3
GBSW	Skeleton	SVM	**93.6**

by an effective fusion of the depth information.

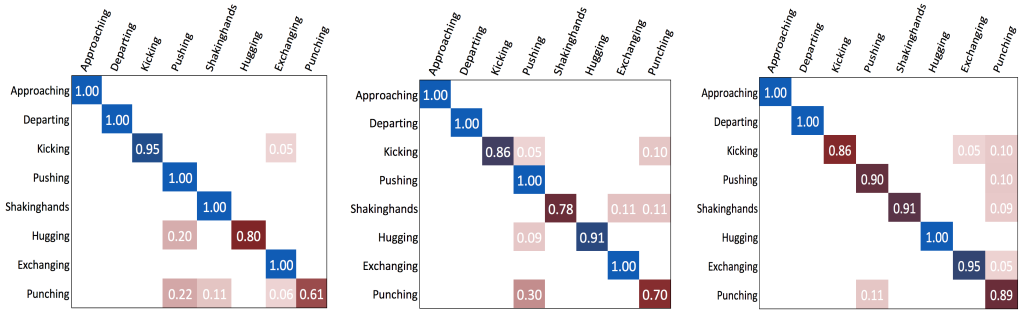
In Table 4, we list the performance of leading methods on the *MSR-Action3D* dataset in terms of *Cross Subject Test*. To conduct a fair comparison, the considered methods are grouped according to the protocol from [20] and [11]. The table shows that the highest average recognition accuracies (95.6% and 93.5%) using both procotols are achieved by the proposed GBSW method. Specifically, GBSW obtains recognition rates over 90% on *AS1*, *AS2* and *AS3* with the procotol from [20] and the rates are over 95% on *AS2* and *AS3* with the procotol from [11].

In addition, Table 5 records the *Cross Subject Test* performance of different methods on *Florence3D-Action* dataset. Some actions in this dataset are quite confused with each other, for example, the body movement in *answer phone* and *drink a bottle* is similar. The table shows that our feature descriptor performed 93.6% recognition accuracy, which improved the performance of [46] and [33] by 11.6% and 2.7%, respectively.

By combining data from Table 3 to Table 5, it can be seen that the proposed method achieves better recognition performance than our previous work (3DMTG) on both *MSR-Action3D* dataset and *Florence3D-Action* dataset, which indicates that the discriminative information weighted using

the kernel function can help the proposed method improve the ability to distinguish different activity classes.

4.3.2. Interaction Recognition Results



(a) *Test One* (b) *Test Two* (c) *Cross Subjects Test*

Figure 6: Confusion Matrices on *SBU Interaction dataset*

To test the recognition ability of the proposed GBSWC on human interaction recognition, the confusion matrices that indicate the confusion among activity categories and the comparison to the state of the art are conducted on the *SBU Interaction dataset*. In Figure 6, it can be seen that our method is able to successfully classify *approaching* and *departing* under various settings. The most common confusion is between *pushing* and *punching* in all tests due to their similar poses.

Table 6 compares the recognition accuracy of the proposed GBSWC with state-of-the-art methods. There are three types of methods involved in this table, namely, skeleton based, skeleton+depth based, and skeleton+RGB based. Although higher accuracies (93.08% and 94.1%) are achieved in [63] and [62], the requirement of combining the skeleton information with the depth / RGB information introduces extra computational complexity. It can

Table 6: Recognition Accuracy (%) on *SBU Interaction dataset*.

	Methods	Feature type	Classifier	Accuracy(%)
State-of-the-art	Velocity features (2012)[15]	Skeleton	SVM	48.4
	Plane features (2012)[15]	Skeleton	SVM	73.8
	Joint features (2012)[15]	Skeleton	SVM	80.3
	Ji *et al.* (2014)[16]	Skeleton	SVM	86.9
	CFDM (2015)[9]	Skeleton	SVM	89.4
	CHARM (2015)[58]	Skeleton	MCP	83.9
	HBRNN (2015)[59]	Skeleton	RNN	80.35
	Co-occurrence LSTM (2016)[60]	Skeleton	LSTM	90.4
	STA-LSTM (2017)[61]	Skeleton	LSTM	91.51
	Baradel *et al.* (2017)[62]	Skeleton	LSTM	90.5
	RHI (2015)[63]	Skeleton+Depth	SVM	93.08
	Baradel *et al.* (2017)[62]	Skeleton+RGB	LSTM	94.1
GBSWC	*Test One*	Skeleton	SVM	**92.75**
	Test Two			**91.67**
	Cross Subjects Test			**93.84**
	Average			**92.75**

be seen that the GBSWC method achieves an average rate of 92.75%, which outperforms the best performance (91.51%) [61] of the listed skeleton-based methods. This indicates that the correlation feature explored in our method could extract high-level information from the movement of skeleton joints, thus helps to reinforce the performance of discriminating complex human interactions.

5. Conclusion

In this paper, a structured multi-feature representation for human action and interaction recognition was proposed. The proposed BSW, which highlighted the discriminative moving trend of each activity category via a kernel-based dynamic encoding algorithm, was aggregated with the ge-

ometric feature in GBSW for human action recognition. The correlation feature between body parts which can represent the intrinsic relationship between interactive subjects was further combined in GBSWC for human interaction recognition. Experimental results on three public datasets have provided compelling recognition results of our approach (e.g., 95.6% on MSR-Action3D, 93.6% on Florence3D-Action, and 92.75% on SBU Interaction). This outperforming performance is owed to the semantic representation and the complementary effect of the aggregation of different types of features.

Although the structured feature descriptor was only evaluated in human action and interaction recognition, it can be easily extended to group activity where multiple persons are involved. The limitation of the proposed method is that its performance will decrease if the skeleton joints are not accurate or missing. Future work will therefore focus on fusing the information from depth or RGB modality for better recognition performance.

Acknowledgments

This work was supported in part by the EU Seventh Framework Programme (No. 611391, Development of Robot-Enhanced therapy for children with AutisM spectrum disorders (DREAM)) and China Scholarship Council.

References

[1] J. C. Niebles, L. Fei-Fei, A hierarchical model of shape and appearance for human action classification, in: Proc. IEEE Conf. Comput. Vis. Pattern Recog., 2007, pp. 1–8.

[2] J. C. Niebles, H. Wang, L. Fei-Fei, Unsupervised learning of human action categories using spatial-temporal words, Int. J. Comput. Vis. 79 (3) (2008) 299–318.

[3] J. Shotton, T. Sharp, A. Kipman, A. Fitzgibbon, M. Finocchio, A. Blake, M. Cook, R. Moore, Real-time human pose recognition in parts from single depth images, Commun. ACM 56 (1) (2013) 116–124.

[4] M. A. Gowayyed, M. Torki, M. E. Hussein, M. El-Saban, Histogram of oriented displacements (hod): describing trajectories of human joints for action recognition, in: Proc. Int. joint Conf. Artificial Intell., 2013, pp. 1351–1357.

[5] R. Qiao, L. Liu, C. Shen, A. van den Hengel, Learning discriminative trajectorylet detector sets for accurate skeleton-based action recognition, Pattern Recog. 66 (2017) 202–212.

[6] H. Pazhoumand-Dar, C.-P. Lam, M. Masek, Joint movement similarities for robust 3d action recognition using skeletal data, J. Vis. Commun. Image Represent. 30 (2015) 10–21.

[7] S. Gaglio, G. L. Re, M. Morana, Human activity recognition process using 3-d posture data, IEEE Trans. Human-Mach. Syst. 45 (5) (2015) 586–597.

[8] I. Lillo, J. C. Niebles, A. Soto, Sparse composition of body poses and atomic actions for human activity recognition in rgb-d videos, Image Vis. Comput. 59 (2017) 63–75.

[9] Y. Ji, H. Cheng, Y. Zheng, H. Li, Learning contrastive feature distribution model for interaction recognition, J. Vis. Commun. Image Represent. 33 (2015) 340–349.

[10] Y. Yang, C. Deng, S. Gao, W. Liu, D. Tao, X. Gao, Discriminative multi-instance multitask learning for 3d action recognition, IEEE Trans. Multimedia 19 (3) (2017) 519–529.

[11] M. Zanfir, M. Leordeanu, C. Sminchisescu, The moving pose: An efficient 3d kinematics descriptor for low-latency action recognition and detection, in: Proc. IEEE Int. Conf. Comput. Vis., 2013, pp. 2752–2759.

[12] M. Liu, H. Liu, C. Chen, Enhanced skeleton visualization for view invariant human action recognition, Pattern Recog. 68 (2017) 346–362.

[13] Y. Hou, Z. Li, P. Wang, W. Li, Skeleton optical spectra based action recognition using convolutional neural networks, IEEE Trans. Circuits Syst. Video Technol. 28 (3) (2018) 807–811.

[14] Q. Ke, M. Bennamoun, S. An, F. Sohel, F. Boussaid, A new representation of skeleton sequences for 3d action recognition, in: Proc. IEEE Conf. Comput. Vis. Pattern Recog., 2017, pp. 4570–4579.

[15] K. Yun, J. Honorio, D. Chattopadhyay, T. L. Berg, D. Samaras, Two-person interaction detection using body-pose features and multiple instance learning, in: Proc. IEEE Conf. Comput. Vis. Pattern Recog. Workshops, 2012, pp. 28–35.

[16] Y. Ji, G. Ye, H. Cheng, Interactive body part contrast mining for human interaction recognition, in: IEEE Int. Conf. Multimedia and Expo Workshops, 2014, pp. 1–6.

[17] B. Ni, Y. Pei, P. Moulin, S. Yan, Multilevel depth and image fusion for human activity detection, IEEE Trans. Cybern. 43 (5) (2013) 1383–1394.

[18] B. Liu, H. Yu, X. Zhou, D. Tang, H. Liu, Combining 3d joints moving trend and geometry property for human action recognition, in: IEEE Int. Conf. Syst. Man, Cyber., 2016, pp. 000332–000337.

[19] B. Liu, H. Cai, X. Ji, H. Liu, Human-human interaction recognition based on spatial and motion trend feature, in: IEEE Int. Conf. Image Processing, IEEE, 2017, pp. 4547–4551.

[20] W. Li, Z. Zhang, Z. Liu, Action recognition based on a bag of 3d points, in: Proc. IEEE Conf. Comput. Vis. Pattern Recog. Workshops, 2010, pp. 9–14.

[21] X. Yang, C. Zhang, Y. Tian, Recognizing actions using depth motion maps-based histograms of oriented gradients, in: Proc. 20th ACM Int. Conf. Multimedia, 2012, pp. 1057–1060.

[22] M. F. Bulbul, Y. Jiang, J. Ma, Human action recognition based on dmms, hogs and contourlet transform, in: Proc. IEEE Int. Conf. Multimedia Big Data, 2015, pp. 389–394.

[23] P. Wang, W. Li, Z. Gao, J. Zhang, C. Tang, P. O. Ogunbona, Action recognition from depth maps using deep convolutional neural networks, IEEE Trans. Human-Mach. Syst. 46 (4) (2016) 498–509.

[24] C. Chen, R. Jafari, N. Kehtarnavaz, Improving human action recognition using fusion of depth camera and inertial sensors, IEEE Trans. Human-Mach. Syst. 45 (1) (2015) 51–61.

[25] M. Liu, H. Liu, Depth context: A new descriptor for human activity recognition by using sole depth sequences, Neurocomputing 175 (2016) 747–758.

[26] C. Liang, L. Qi, Y. He, L. Guan, 3d human action recognition using a single depth feature and locality-constrained affine subspace coding, IEEE Trans. Circuits Syst. Video Technol.

[27] O. Oreifej, Z. Liu, Hon4d: Histogram of oriented 4d normals for activity recognition from depth sequences, in: Proc. IEEE Conf. Comput. Vis. Pattern Recog., 2013, pp. 716–723.

[28] X. Yang, Y. Tian, Super normal vector for activity recognition using depth sequences, in: Proc. IEEE Conf. Comput. Vis. Pattern Recog., 2014, pp. 804–811.

[29] R. Slama, H. Wannous, M. Daoudi, Grassmannian representation of motion depth for 3d human gesture and action recognition, in: Proc. Int. Conf. Pattern Recog., 2014, pp. 3499–3504.

[30] Y. Huang, X. Cao, Q. Wang, B. Zhang, X. Zhen, X. Li, Long-short term features for dynamic scene classification, IEEE Trans. Circuits Syst. Video Technol., (2018),doi:10.1109/TCSVT.2018.2823360.

[31] F. Ofli, R. Chaudhry, G. Kurillo, R. Vidal, R. Bajcsy, Sequence of the most informative joints (smij): A new representation for human skeletal

action recognition, J. Vis. Commun. Image Represent. 25 (1) (2014) 24–38.

[32] L. Xia, C.-C. Chen, J. Aggarwal, View invariant human action recognition using histograms of 3d joints, in: Proc. IEEE Conf. Comput. Vis. Pattern Recog. Workshops, 2012, pp. 20–27.

[33] R. Vemulapalli, F. Arrate, R. Chellappa, Human action recognition by representing 3d skeletons as points in a lie group, in: Proc. IEEE Conf. Comput. Vis. Pattern Recog., 2014, pp. 588–595.

[34] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proc. IEEE Conf. Comput. Vis. Pattern Recog., 2016, pp. 770–778.

[35] Q. Wang, J. Gao, Y. Yuan, Embedding structured contour and location prior in siamesed fully convolutional networks for road detection, IEEE Trans. Intell. Transp. Syst. 19 (1) (2018) 230–241.

[36] Q. Wang, J. Gao, Y. Yuan, A joint convolutional neural networks and context transfer for street scenes labeling, IEEE Trans. Intell. Transp. Syst. 19 (5) (2017) 1457–1470.

[37] Y. Du, Y. Fu, L. Wang, Skeleton based action recognition with convolutional neural network, in: IAPR Asian Conf. Pattern Recog., 2015, pp. 579–583.

[38] K. Simonyan, A. Zisserman, Very Deep Convolutional Networks for Large-Scale Image Recognition, in: Int. Conf. Learning Representations, 2015. arXiv:1409.1556, doi:10.1016/j.infsof.2008.09.005.

[39] V. Bloom, V. Argyriou, D. Makris, Hierarchical transfer learning for online recognition of compound actions, Comput. Vis. Image Understanding 144 (2016) 62–72.

[40] T. Hu, X. Zhu, W. Guo, K. Su, Efficient interaction recognition through positive action representation, Math. Problems in Eng. 2013.

[41] H. Wang, L. Wang, Modeling Temporal Dynamics and Spatial Configurations of Actions Using Two-Stream Recurrent Neural Networks, in: Proc. Conf. Comput. Vis. Pattern Recog., 2017, pp. 499–508.

[42] T. Huynh-The, O. Banos, B.-V. Le, D.-M. Bui, S. Lee, Y. Yoon, T. Le-Tien, Pam-based flexible generative topic model for 3d interactive activity recognition, in: IEEE Int. Conf. Advanced Technol. Commun., 2015, pp. 117–122.

[43] S. Saha, A. Konar, R. Janarthanan, Two person interaction detection using kinect sensor, Facets of Uncertainties and Applicat. (2015) 167–176.

[44] M. J. Swain, D. H. Ballard, Color indexing, Int. J. Comput. Vis. 7 (1) (1991) 11–32.

[45] C.-C. Chang, C.-J. Lin, Libsvm: A library for support vector machines, 2001. Software available at http://www.csie.ntu.edu.tw/ cjlin/libsvm.

[46] L. Seidenari, V. Varano, S. Berretti, A. Del Bimbo, P. Pala, Recognizing actions from depth cameras as weakly aligned multi-part bag-of-poses, in: Proc. IEEE Conf. Comput. Vis. Pattern Recog. Workshops, 2013, pp. 479–485.

[47] Y. Yang, S. Baker, A. Kannan, D. Ramanan, Recognizing proxemics in personal photos, in: Proc. IEEE Conf. Comput. Vis. Pattern Recog., 2012, pp. 3522–3529.

[48] A. W. Vieira, E. R. Nascimento, G. L. Oliveira, Z. Liu, M. F. Campos, On the improvement of human action recognition from depth map sequences using space–time occupancy patterns, Pattern Recog. Lett. 36 (2014) 221–227.

[49] A. Jalal, Y.-H. Kim, Y.-J. Kim, S. Kamal, D. Kim, Robust human activity recognition from depth video using spatiotemporal multi-fused features, Pattern Recog. 61 (2017) 295–308.

[50] X. Yang, Y. Tian, Eigenjoints-based action recognition using naive-bayes-nearest-neighbor, in: Proc. IEEE Conf. Comput. Vis. Pattern Recog. Workshops, 2012, pp. 14–19.

[51] J. Wang, Z. Liu, J. Chorowski, Z. Chen, Y. Wu, Robust 3d action recognition with random occupancy patterns, in: Proc. Eur. Conf. Comput. Vis., 2012, pp. 872–885.

[52] L. Xia, J. Aggarwal, Spatio-temporal depth cuboid similarity feature for activity recognition using depth camera, in: Proc. IEEE Conf. Comput. Vis. Pattern Recog., 2013, pp. 2834–2841. doi:10.1109/cvpr.2013.365.

[53] J. Wang, Z. Liu, Y. Wu, J. Yuan, Mining actionlet ensemble for action recognition with depth cameras, in: Proc. IEEE Conf. Comput. Vis. Pattern Recog., 2012, pp. 1290–1297.

[54] M. Devanne, H. Wannous, S. Berretti, P. Pala, M. Daoudi, A. Del Bimbo, 3-d human action recognition by shape analysis of motion trajectories on riemannian manifold, IEEE Trans. Cybern. 45 (7) (2015) 1340–1352.

[55] Y. Yang, C. Deng, D. Tao, S. Zhang, W. Liu, X. Gao, Latent max-margin multitask learning with skelets for 3-d action recognition, IEEE Trans. Cybern. 47 (2) (2017) 439–448.

[56] Y. Guo, Y. Li, Z. Shao, Dsrf: A flexible trajectory descriptor for articulated human action recognition, Pattern Recog. 76 (2018) 137–148.

[57] C. Wang, Y. Wang, A. L. Yuille, An approach to pose-based action recognition, Proc. IEEE Conf. Comput. Vis. Pattern Recog. (2013) 915–922.

[58] W. Li, L. Wen, M. Choo Chuah, S. Lyu, Category-blind human action recognition: a practical recognition system, in: Proc. IEEE Int. Conf. Comput. Vision, 2015, pp. 4444–4452.

[59] Y. Du, W. Wang, L. Wang, Hierarchical recurrent neural network for skeleton based action recognition, in: Proc. IEEE Conf. Comput. Vis. Pattern Recog., 2015, pp. 1110–1118.

[60] W. Zhu, C. Lan, J. Xing, W. Zeng, Y. Li, L. Shen, X. Xie, et al., Co-occurrence feature learning for skeleton based action recognition using regularized deep lstm networks., in: AAAI, Vol. 2, 2016, p. 8.

[61] S. Song, C. Lan, J. Xing, W. Zeng, J. Liu, An end-to-end spatio-temporal attention model for human action recognition from skeleton data., in: AAAI, 2017, pp. 4263–4270.

[62] F. Baradel, C. Wolf, J. Mille, Pose-conditioned spatio-temporal attention for human action recognition, arXiv preprint arXiv:1703.10106, 2017.

[63] I. Gori, J. Aggarwal, M. Ryoo, Building unified human descriptors for multi-type activity recognition, CoRR, abs/1507.02558.