Mining Mid-level Features for Action Recognition Based on Effective Skeleton Representation

Pichao Wang, Wanqing Li, Philip Ogunbona and Zhimin Gao
University of Wollongong, Wollongong, NSW, Australia, 2522
pw212@uowmail.edu.au, {wanqing, philipo}@uow.edu.au, {zg126}@uowmail.edu.au
Hanling Zhang
Hunan University, P. R. China, jt_hlzhang@hnu.edu.cn

***Abstract*—Recently, mid-level features have shown promising performance in computer vision. Mid-level features learned by incorporating class-level information are potentially more discriminative than traditional low-level local features. In this paper, an effective method is proposed to extract mid-level features from Kinect skeletons for 3D human action recognition. Firstly, the orientations of limbs connected by two skeleton joints are computed and each orientation is encoded into one of the 27 states indicating the spatial relationship of the joints. Secondly, limbs are combined into parts and the limb's states are mapped into part states. Finally, frequent pattern mining is employed to mine the most frequent and relevant (discriminative, representative and non-redundant) states of parts in continuous several frames. These parts are referred to as *Frequent Local Parts* or *FLPs*. The *FLPs* allow us to build powerful *bag-of-FLP*-based action representation. This new representation yields state-of-the-art results on MSR DailyActivity3D and MSR ActionPairs3D.**

I. Introduction

Human action recognition has been an active research topic in computer vision due to its wide range of applications, such as smart surveillance and human-computer interactions. Despite remarkable research efforts and encouraging advances in the past decade, accurate recognition of human actions is still an open problem.

A common and intuitive method to represent human motion is to use a sequence of skeletons. With the development of the cost-effective depth cameras and algorithms for real-time pose estimation [1], skeleton extraction has become more and more robust and skeleton-based action representation is becoming one of the most practical and promising approaches. Up to date, the skeleton-based approach primarily focuses on low-level features and models the dynamics of the skeletons holistically, such as moving pose [2] and trajectories of human joints [3]. The full skeletal description is highly subject to the noise introduced during the extraction of the skeleton and less effective in the cases where some actions involve motion of the whole body and others are preformed using only small number of body parts. A key fact we observed is that during the temporal axis of actions, only a few body parts in several continuous frames are activated during the performance of the actions. These parts are more robust and discriminative to represent an action. In our method we take advantage of this observation to capture mid-level features for action recognition.

Inspired by the mid-level features mining techniques [4] for image classification, we propose a new scheme applying pattern mining to obtain the most relevant combinations of parts in several continuous frames for action recognition rather than to utilize all the joints as most previous works did. In particular, a new descriptor called *bag-of-FLPs* is proposed to describe an action as illustrated in Fig. 1. The overall

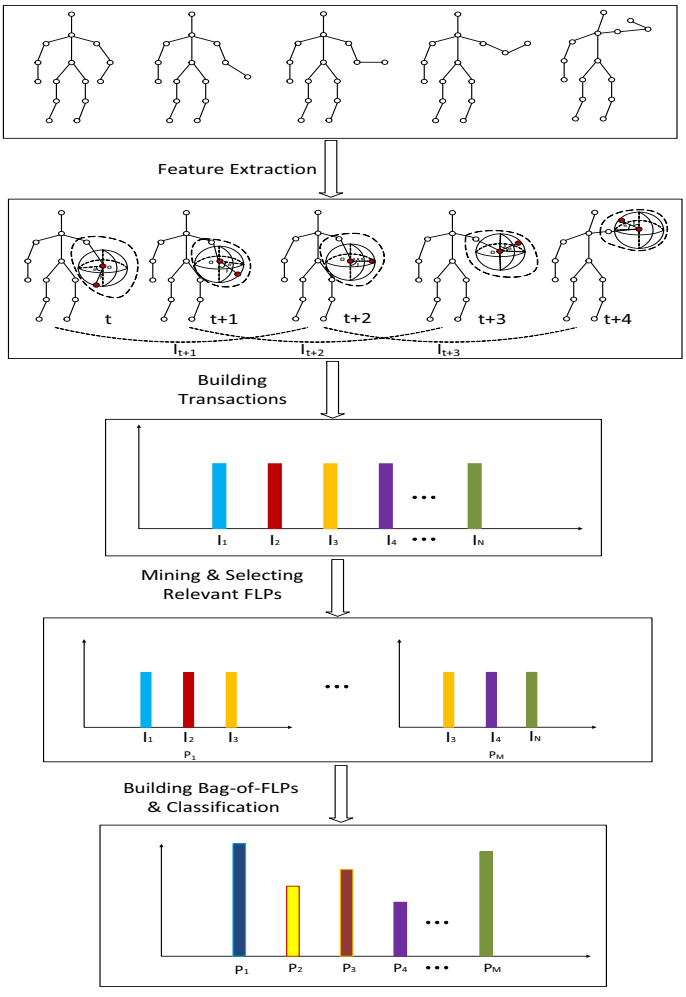

Fig. 1. The general framework of the proposed method.

process of our method can be divided into four steps: feature extraction, building transactions, mining & selecting relevant patterns and building *Bag-of-FLPs* & classification. We first compute the orientations of limbs, i.e. connected joints, and then encode each orientation into one of the 27 states indicating

the spatial relationship of the joints. Limbs are combined into parts and limb's states are mapped to part states. Local temporal information is included by combining part states of several, say, 5, continuous frames into one transaction for mining, with each state as one item. In order to keep motion information after frequent pattern mining, the unique states of parts of the continuous frames are reserved, removing the repeated ones, ensuring the pose information and motion information be included in each transaction. The most relevant patterns, which we referred to *FLPs*, are mined and selected o represent frames and build *bag-of-FLPs* as new representation for a whole action. The new representation is much robust to the errors in the features, because the errors are usually not frequent patterns.

Our main contributions include the following four aspects. First, an effective and efficient method is proposed to extract skeleton features. Second, a novel method is developed to explore spatial and temporal information in skeleton data, simultaneously. Third, an effective scheme is proposed for applying pattern mining to action recognition by adapting the generic pattern mining tools to the features of skeleton. Our scheme is much robust to noise as most noisy data does not form frequent patterns. In addition, our scheme has achieved the state-of-the-art results on several benchmark datasets.

The rest of the paper is organized as follows. Section II reviews the related work. Section III presents our scheme in detail. Section IV shows experimental results. Conclusion is made in section V.

II. Related work

The process of action recognition can be generally divided into two main steps, action representation and action classification. Action representation consists of feature extraction and feature selection. Features can be extracted from input sources such as depth maps, skeleton and/or RGB images. Regardless of the input source, there are two main approaches, space-time approach and sequential approach [5], [6], [7], to the representation of actions. The space-time approach usually extracts local or holistic features from space-time volume, without explicit modelling of temporal dynamics. By contrast, the sequential approach normally extracts local features from each frame of the input source and models the dynamics explicitly. Action classification is the step of learning a classifier based on action representation and classifying any new observations using the classifier. For space-time approaches, discriminative classifier, such as Support Vector Machine (SVM), is often used for classification. For the sequential approach, generative statistical models, such as Hidden Markov Model (HMM), are commonly used. Our method belongs to the skeleton-based space-time volume approach. In this section, we mainly review the existing work of skeleton-based action representation for action recognition.

For the skeleton-based sequential approach, Xia et al. [8] proposed a feature called Histograms of 3D Joint Locations (HOJ3D) as a representation of postures. The HOJ3D essentially encodes spatial occupancy information relative to the root joint, e.g. hip centre. A modified spherical coordinate system is defined on the root joint and the 3D space is divided into N bins. The HOJ3D is reprojected using Linear Discriminant Analysis (LDA) to reduce dimensionality and then clustered into K posture visual words which represent the prototypical poses of actions. HMMs are adopted to model the visual words and recognize actions. Radial distance is adopted in this spherical coordinate system which makes the method to some extend view-invariant.

Koppula et al. [9] explicitly modelled the motion hierarchy to enable their method to handle simple human-object interactions. The human activities and object affordances are jointly modelled as a Markov Random Field (MRF) where the nodes represent objects and sub-activities, and the edges represent the relationships between object affordances, their relations with sub-activities, and their evolution over time. Feature vectors that represent the object's location and the changing information in the scene are defined by training a Structural Support Vector Machine (SSVM). Similar to this approach, Sung et al. [10] proposed a hierarchical two-layer Maximum Entropy Markov Model (MEMM) to represent an activity. The lower layer nodes represent sub-activities while higher level nodes describe more complex activities, for example, "lifting left hand" and "pouring water" can be described as a sub-activity and a complex activity, respectively. Wang et al. [11] proposed an Local Occupancy Patterns (LOP) feature calculated from the 3D point cloud around a particular joint to discriminate different types of interactions and Fourier Temporal Pyramid (FTP) to represent the temporal structure. Based on above two types of features, a model called Actionlet Ensemble Model (AEM) is proposed which is a combination of the features for a subset of the joints. Due to the numerous actionlets, data mining technique is used to discover discriminative actionlets. Both skeleton and point cloud information are utilized to recognize human-objects interactions.

For the skeleton-based space-time volume approach, Yang et al. [12] proposed a new feature descriptor called EigenJoints features which contain posture features, motion features and offset features. The pair-wise joint differences in current frames and their consecutive frames are used to encode the spatial and temporal information, which are called posture features and motion features, respectively. The difference of a pose with respect to the initial pose is called offset features. The initial pose is generally assumed as a neutral pose. The three channels are normalized and Principal Component Analysis (PCA) is applied to reduce redundancy and noise to obtain the EigenJoints descriptor. A Naive-Bayes-Nearest-Neighbor (NBNN) classifier is adopted to recognize actions. Gowayyed et al. [3] proposed a new descriptor called Histograms of Oriented Displacements (HOD) to recognize actions. The displacement of each joint votes with its length in a histogram of oriented angles. Each 3D trajectory is represented by the HOD of its three 2D projection. In order to reserve temporal information, a temporal pyramid is proposed, where trajectories are considered as a whole, halves and quarters and then all the descriptors in these three levels are concatenated to form the final descriptor. A linear SVM is used to classify actions based on the histograms. Similar to this work, Hussein et al. [13] proposed a descriptor called Covariance of 3D Joints (Cov3DJ) for human action recognition. This descriptor uses covariance matrix to capture the dependence of locations of different joints on one another during an action. In order to capture the order of motion in time, a hierarchy of Cov3DJs is used, similarly to the work in [3].

Zanfir et al. [2] proposed a descriptor called moving pose which is formed by the position, velocity and acceleration of skeleton joints within a short time window around the current fame. To learn discriminative pose, a modified k-Nearest Neighbours (kNN) classifier is used that considers both the temporal location of a particular frame within the action sequence as well as the discrimination power of its moving pose descriptor compared to other frames in the training set. Thanh et al. [14] extracted key frames which are the central frames in the short temporal segments of videos and labelled each key frame as a pattern for a unit action. An improved Term Frequency-Inverse Document Frequency (TF-IDF) method is used to learn the discriminative patterns and learned patterns is defined as local features for action recognition. Wang et al. [15] first estimated human joints positions from videos and then grouped the estimated joints into five parts. Each action is represented by computing sets of co-occurring spatial and temporal configurations of body parts. The authors use a bag of words method with the extracted features for classification. Ohn-Bar and Trivedi [16] tracked the joint angles and built a descriptor based on similarities between angle trajectories. This feature is further combined with a double-HOG descriptor that accounts for the spatio-temporal distribution of depth values around the joints. Theodorakopoulos et al. [17] initially processed the skeleton data from sensor coordinate to torso PCA frame in order to gain robust and invariant pose representation. Sparse coding in dissimilarity space is utilized to sparsely represent the actions. Chaaraoui et al. [18] proposed to use an evolutionary algorithm to determine the optimal subset of skeleton joints, taking into account the topological structure of the skeleton.

To fuse depth-based features with skeleton-based features, Althloothi et al. [19] presented two sets of features, features for shape representation extracted from depth data by using a spherical harmonics representation and features for kinematic structure extracted from skeleton data by estimating 3D joint positions. The shape features are used to describe the 3D silhouette structure while the kinematic features are used to describe the movement of the human body. Both sets of features are fused at the kernel level for action recognition by using Multiple Kernel Learning (MKL) technique. Similar to this direction, Chaaraoui et al. [20] proposed a fusion method to combine skeleton and silhouette-based features. The skeletal features are obtained by normalising the 3D position of original skeleton data while the silhouette-based features are generated by extracting contour points of the silhouette. After feature fusion, a model called bag of key poses is employed for action recognition. The key poses are obtained by K-means clustering algorithm and the words are made up of key poses. In recognition stage, unknown video sequences are classified based on sequence matching. Rahmani et al. [21] proposed an algorithm combining the discriminative information from depth maps as well as from 3D joints positions for action recognition. To avoid the suppression of subtle discriminative information, local information integration and normalization are performed. Joint importance is encoded by using joint motion volume. Random Decision Forest (RDF) is trained to select the discriminant features. Because of the low dimensionality of their features, their method turns to be efficient.

In above methods, most of them are based low-level features and need the whole skeletal description which leads to their weak adaptation to noise. In addition, most of them need to explore the spatial and temporal information, separately, and then combine them together. Besides, most of the methods used to explore temporal information are subject to the neural poses, which are shared by all actions. However, in our method, we use a parts-based mid-level feature to represent actions and explore the spatial and temporal information simultaneously. This makes our method more robust.

III. Proposed Method

The overall process of the proposed method is illustrated in Fig. 1. It can be divided into four steps: feature extraction, building transactions, mining & selecting relevant patterns and building *Bag-of-FLPs* & classification.

A. Feature Extraction

In our method, the orientations of human body limbs are considered as low-level features and they can be calculated from the two joints of the limbs. For Kinect skeleton data, 20 joint positions, as shown in Fig. 2, are tracked [1]. The skeleton data is first normalized using Algorithm 1 in [2] to suppress noise in the original skeleton data and to compensate for length variations across different subjects and different body parts. Each joint i has 3 coordinates, denoted as (x_i, y_i, z_i) after normalization.

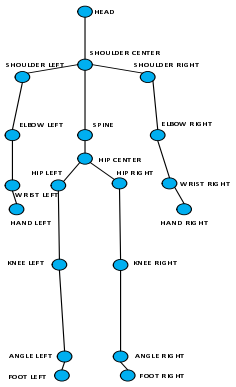

Fig. 2. The human joints tracked with the skeleton tracker [1].

For Kinect skeleton, it is found that the *Hand Left*, *Hand Right*, *Foot Left*, *Foot Right*, and *Spine* joints are often not reliable and, hence they are not used in our method. Thus, there are 15 joints 14 limbs. The joint Head is considered as the origin of the 15 points. For each limb, we compute a unit difference vector between its two joints:

$$(\Delta x_{ij}, \Delta y_{ij}, \Delta z_{ij}) = \frac{(x_i, y_i, z_i) - (x_j, y_j, z_j)}{d_{ij}} \qquad (1)$$

where i and j represent the current joint and reference joint, respectively; d_{ij} is the Euclidean distance between the two joints. For example, as illustrated in Fig. 1, to compute the orientation of the limb between joint Hand Right and Wrist Right (highlighted in red), the Wrist Right joint is regarded as the sphere center and Eq. (1) is used to compute the unit difference vector.

Each element of the unit difference vector is quantized into three states: -1, 0 and 1. If $|\Delta x_{ij}| \leq threshold$ then $q(\Delta x_{ij}) = 0$; if $\Delta x_{ij} > threshold$ then $q(\Delta x_{ij}) = 1$; else $q(\Delta x_{ij}) = -1$. Thus, there are 27 possible states for each unit difference vector, and each state is encoded as one element of a feature vector, so the dimension of the feature vector for each pose is 14×27 = 378 after concatenating all feature vectors of the 14 limbs. For each element of the feature vector, if the corresponding orientation between two joints is bid to one state, then the relative position is labelled to 1, otherwise, it is 0. Therefore, the feature vectors are very sparse, only 14 positions in each feature vector are 1 (not zeros). The threshold is an empirical value which is dependent on the noise characteristics of the skeleton data.

For each frame of skeleton, a quantized 378 dimensional feature vector is calculated as described above. This feature vector is reduced to a 14 dimensional feature vector with each element being the index to a non-zero element of the 378-dimensional feature vector.

To extract mid-level features for action representation, the 14 limbs are combined into 7 body parts. As illustrated in Fig. 1, the dotted line contains joints Hand Right, Wrist Right and Elbow Right, and these three limbs form one part. In this way, seven body parts are formed, namely, Head-Shoulder Center, Should Center-Shoulder Left-Elbow Left-Wrist Left, Shoulder Center-Shoulder Right-Elbow Right-Wrist Right, Shoulder Center-Hip Center-Hip Left, Hip Left-Knee Left-Angle Left,Shoulder Center-Hip Center-Hip Right and Hip-Right-Knee Right-Angle Right. According to the Degree of Freedom (DoF) of joints [22], each body part is encoded with different number of states and the total number of states is denoted as NDF, which is currently an empirical parameter. It should be adjusted according to the complexity of the actions to be recognized and noise level of the dataset.

To explore temporal information and keep motion information at the same time after frequent data mining (generally, frequent data mining can only mine the most frequent pattens which can not be guaranteed as discriminative patterns), a novel way is proposed. Seven states for each frame will be obtained after combination, and the unique states of continuous C frames, as illustrated in Fig. 1, where $C = 3$, are counted and form a new mid-level feature vector, denoted as $\{f_i|i = 1, ..., n_A\}$. This new feature vector contains both pose information of the current frame and the motion information in the continuous C frames, because the repeated states in the continuous frames can be regarded as static pose information and the different ones with other frames can capture the motion information. This feature vector is used to build transactions described in the next section. The pattens after mining can be the combinations of several body parts in different frames, thus the temporal order information can be easily maintained.

B. Building Transactions

Each instance of action A is represented by a set of above mid-level features $\{f_i|i = 1, ..., n_A\}$ and a class label c, $c \in \{1...C\}$. The set of features for all the action samples is denoted by Ω. The dimensionality of the feature vector is denoted as W and in our case $|\mathrm{W}| \geq 7$.

1) Items, Transactions and Frequencies: Each element in a feature vector for continuous C poses is defined as an item, and an item is denoted as ω, where $\omega \in (0, NDF]$ and $\omega \in \mathbb{N}$.

The set of *transactions* X from the set Ω is created next. For each $\mathbf{x} \in \Omega$ there is one transaction x (i.e. a set of items). This transaction x contains all the items ω_j. A *local pattern* is an itemset $t \subseteq \Gamma$, where Γ represents the set of all possible items. For a local pattern t, the set of transactions that include the pattern t is defined as: $X(t) = \{x \in X | t \subseteq x\}$. The *frequency* of t is $|X(t)|$, also known as the *support* of the pattern t or $supp(t)$.

2) Frequent Local Part: For a given constant T, also known as the minimum support threshold, a local pattern t is *frequent* if $supp(t) \geq T$. A pattern t is said to be *closed* if there exists no pattern $t^{'}$ that $t \subset t^{'}$ and $supp(t) = supp(t^{'})$. The set of frequent closed patterns is a compact representation of the frequent patterns, and such a frequent and closed local part pattern is referred to as *Frequent Local Part* of *FLP*.

C. Mining & Selecting Relevant FLPs

1) FLPs Mining: Given the set of transaction X, any existing frequent mining algorithm can be used to find the set of *FLPs* Υ. In our work, the optimised *LCM* algorithm [23] is used as in [4]. *LCM* uses a *prefix preserving closure extension* to completely enumerate closed itemsets.

2) Encoding a New Action with FLPs: Given a new action, the features can be extracted according to the section A and each feature vector can be converted into a transaction x and for each *FLP* pattern $t \in \Upsilon$ it can be checked whether $t \subseteq x$. If $t \subseteq x$ is true, then x is an *instance* of the *FLP* pattern t. The frequency of a pattern t in a given action A_j (i.e. the number of instances of t in A_j) is denoted as $F(t|A_j)$.

3) Selecting the Best FLPs for Action Recognition: The *FLPs* set Υ is considered as a candidate set of mid-level features to represent an action. Therefore, the most useful *FLP* patterns from Υ is needed to be selected because $i)$ the number of generated *FLP* patterns is huge and $ii)$ not all discovered *FLP* patterns are equally important to the action recognition task. Usually, relevant patterns are those *discriminative* and *non-redundant*. On top of that, a new criterion, *representativity* is also used. As a result, some patterns may be frequent and appear to be discriminative but they may occur in very few actions (e.g. noise pose). Such features are not representative and therefore not the best choice for action recognition. A good *FLP* pattern should be at the same time discriminative, representative and non-redundant. In this section, how to select such patterns is discussed.

The methods used in [4] are followed to find the most suitable pattern subset χ, where $\chi \subset \Upsilon$, for action recognition. To do this the *gain* of a pattern t is denoted by $G(t)$ (s.t. $t \notin \chi$ and $t \in \Upsilon$) and defined as follows:

$$G(t) = S(t) - \max_{s \in \chi} \{R(s,t) \cdot \min(S(t), S(s))\} \qquad (2)$$

where $S(t)$ is the overall relevance of a pattern t and $R(s,t)$ is the redundancy between two patterns s, t. In Eq. (2), a pattern t has a higher gain $G(t)$ if it has a higher relevance $S(t)$ (i.e. it is discriminative and representative) and if the pattern t is non

redundant with any pattern s in set χ (i.e. $R(s,t)$ is small). $S(t)$ is defined as:

$$S(t) = D(t) \times O(t), \tag{3}$$

and $R(s,t)$ is defined as:

$$\begin{aligned} R(s,t) = &\exp\{-[p(t) \cdot D_{KL}(p(A|t)||p(A|\{t,s\})) \\ &+ p(s) \cdot D_{KL}(p(A|s)||p(A|\{t,s\}))]\}. \end{aligned} \tag{4}$$

Following a similar approach in [24] to find affinity between patterns, two patterns t and $s \in \Upsilon$ are redundant if they follow similar document distributions, i.e. if $p(A|t) \approx p(A|s) \approx p(A|\{t,s\})$ where $p(A|\{t,s\})$ is the document distribution given both patterns $\{t,s\}$.

In Eq. (3), $D(t)$ is the *discriminability score*. Following the entropy-based approach in [25], and a high value of $D(t)$ implies that the pattern t occurs only in very few actions; $O(t)$ is the *representativity score* for a pattern t and it considers the divergence between the optimal distribution for class c $p(A|t_c^*)$ and the distribution for pattern t $p(A|t)$, and then takes the best match over all classes. The optimal distribution is such that $i)$ the pattern occurs only in actions of class c, i.e. $p(c|t_c^*) = 1$ (giving also a discriminability score of 1), and $ii)$ the pattern instances are equally distributed among all the actions of class c, i.e. $\forall A_j, A_k$ in class c, $p(A_j|t_c^*) = p(A_k|t_c^*) = (1/N_c)$ where N_c is the number of samples of class c. An optimal pattern, denoted by t_c^* for class c, is a pattern which has above two properties.

The *discriminability score* and *representativity score* are defined as:

$$D(t) = 1 + \frac{\sum_c p(c|t) \cdot \log p(c|t)}{\log C}, \tag{5}$$

$$O(t) = \max_c(\exp\{-[D_{KL}(p(A|t_c^*)||p(A|t))]\}) \tag{6}$$

where $p(c|t)$ is the probability of class c given the pattern t, computed as follows:

$$p(c|t) = \frac{\sum_{j=1}^{N} F(t|A_j) \cdot p(c|A_j)}{\sum_{j=1}^{N} F(t|A_j)}; \tag{7}$$

$D_{KL}(.||.)$ is the Kullback-Leibler divergence between two distributions; $p(A|t)$ is computed empirically from the frequencies $F(t|A_j)$ of the pattern t:

$$p(A|t) = \frac{F(t|A)}{\sum_j F(t|A_j)} \tag{8}$$

Here, A_j is the j^{th} action and N is the total number of actions in the dataset. $p(c|A) = 1$ if the class label of A_j is c and 0 otherwise; $p(c|t_c^*)$ is the optimal distribution with respect to a class c.
In Eq. (4), $p(t)$ is the probability of pattern t and it is defined as:

$$p(t) = \frac{\sum_{A_j} F(t|A_j)}{\sum_{t_j \in \Upsilon} \sum_{A_j} F(t_j|A_j)} \tag{9}$$

while $p(A|\{t,s\})$ is the document distribution given both patterns $\{t,s\}$ and it is defined as:

$$p(A|\{t,s\}) = \frac{F(t|A) + F(s|A)}{\sum_j F(t|A_j) + F(s|A_j)} \tag{10}$$

To find the best K patterns the following greedy process is used. First the most relevant pattern is added to the relevant pattern set χ. Then the pattern with the highest gain (non redundant but relevant) is searched out and this pattern is added into the set χ until K patterns are added (or until no more relevant patterns can be found). For more detailed discussions, [4] is recommended to refer to.

D. Building Bag-of-FLPs & Classification

After computing the K most relevant and non-redundant *FLPs*, each action can be represented by a new representation called *bag-of-FLPs* by counting the occurrences of such *FLPs* in the action. Let L be such a *bag-of-FLPs* for action A_L and M be the *bag-of-FLPs* for action A_M.

An SVM [26] is trained to classify the actions. The SVM uses the following kernel to calculate the similarities between the *bag-of-FLPs* of L and M.

$$K(L,M) = \sum_i \min(\sqrt{L(i)}, \sqrt{M(i)}) \tag{11}$$

Here $L(i)$ is the frequency of the i^{th} selected pattern in histogram L. It is a standard histogram intersection kernel with non-linear weighting. This reduces the importance of highly frequent patterns and is necessary since there is a large variability in pattern frequencies.

IV. Experimental Results

Two benchmark datasets, MSR-DailyActivity3D [27] and MSR-ActionPairs3D [28], were used to evaluate the proposed method and the results are compared with those reported in other papers on the same datasets and under the same training and testing configuration.

A. Experimental Setup

In our method, there are several parameters that need to be tuned, the *threshold* T, the number of states NDF, the number of relevant patterns K, the continuous frames C, minimum support S and maximum support U. For different datasets, different sets of parameters were learned through cross-validation to optimise the performance. Specifically, two-third of the entire training dataset was used as training and the rest one-third was used for validation to tune the parameters. The ranges of the parameters are empirical. In general, the threshold T is dependent on the noise level of the dataset. The higher the noise the larger its value. This is an important parameter because it affects the states of limbs computed from the skeleton data. However, such sensitivity can be reduced by setting a large number, NDF (i.e. over 600) of states. The number of relevant patterns K is dependent on the complexity of the actions to be recognized, the more actions in the dataset, the larger number it should be. The number of continuous frames C is affected by the complexity of required temporal information to encode the actions. If the dataset has pair actions, for example, two actions of each pair are similar in motion (have similar trajectories) and shape (have similar objects), the value of C should be large. However, a large C leads to high memory and post-processing requirement. The values of the minimum support S and maximum support U effect the number of generated patterns before pattern

selection. We observed that if S is large, U should also be large; If S is small, U should also be small. Generally, S and U are set to reduce the computational time for post-processing. In fact, there are many combinations of these two parameters to get the best results. In the other words, the performance of the proposed method is not much sensitive to the choice of S and U.

B. MSR DailyActivity3D

The MSR DailyActivity3D dataset consists of 10 subjects and 16 activities: *drink, eat, read book, call cellphone, write on a paper, use laptop, use vacuum cleaner, cheer up, sit still, toss paper, play game, lay down on sofa, walk, play guitar, stand up, sit down*. Fig. 3 shows some sample frames for the activities. Each subject performed each activity twice, once in

Fig. 3. Sample frames of the MSR DailyActivity3D dataset.

standing position and once in sitting position. In total, there are 320 samples. This dataset has large intra-class variations and involves human-object interactions, which is challenging for recognition only by 3D joints. Experiments were performed based on cross-subject test setting described in [2], i.e. five subjects (1, 2, 3, 4, 5) were used for training and the rest 5 subjects were used for testing. Table I shows the results of our methods compared with other published results. For

TABLE I. COMPARISON ON MSR-DAILYACTIVITY DATASET

Methods	Accuracy (%)
Dynamic Temporal Warping [29]	54.0
Moving Pose [2]	73.8
Actionlet Ensemble on Joint Features [11]	74.0
Proposed Method	78.8

this dataset, $T = 0.15, NDF = 600, K = 30000, C = 3, S = 15, U = 180$. As seen, although this dataset is quite challenging, our method obtained promising results based only on skeleton data. The confusion matrix is illustrated in Fig. 4. From the confusion matrix, it can be seen that activities such as "Drink", "Cheer Up", "Sit Still", "Toss Paper" are relatively easy to recognise, while "Eat" and "Use laptop" are relatively difficult to recognise. The reason for the difficulties is that for these human-object interactions, object information was not

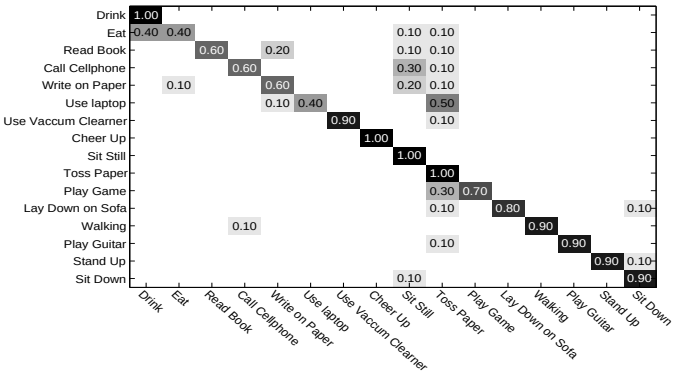

Fig. 4. The confusion matrix of our proposed method for MSR-DailyActivity3D.

available from skeleton data which makes these interactions are almost the same in terms of motion reflected in the skeleton data.

C. MSR ActionPairs3D

The MSR ActionPairs3D dataset [28] is a paired-activity dataset captured by a Kinect camera. This dataset contains 12 activities (i.e. six pairs) of 10 subjects with each subject performing each activity 3 times. The pair actions are: Pick up a box/Put down a box, Lift a box/Place a box, Push a chair/Pull a chair, Wear a hat/Take off hat, Put on a backpack/Take off a backpack, Stick a poster/Remove a poster. Some sample frames for the activities of this dataset are shown in Fig. 5. This dataset is collected to investigate how the temporal

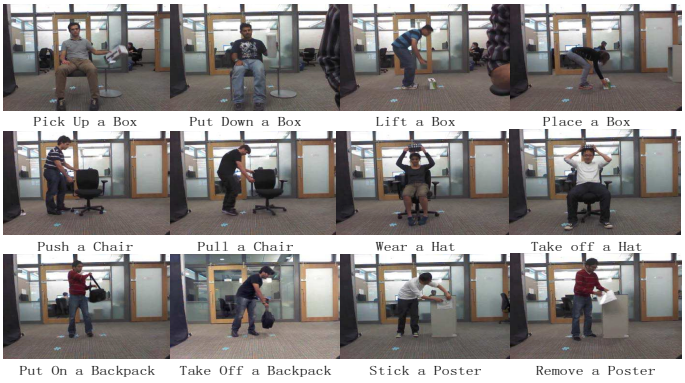

Fig. 5. Sample frames of the MSR ActionPairs3D dataset.

order affects activity recognition. Experiments were set to the

TABLE II. COMPARISON ON MSR-ACTIONPAIRS DATASET

Methods	Accuracy (%)
Skeleton + LOP [27]	63.33
Depth Motion Maps [30]	66.11
Proposed Method	75.56

same configuration as [28], namely, the first five actors are used for testing, and the rest for training. For this dataset, $T = 0.11, NDF = 1000, K = 10000, C = 4, S = 3, U = 100$. We compare our performance in this dataset with two

methods whose results were reported in [28]. Table II shows the comparisons with other methods tested on this dataset.

The confusion matrix is shown in Fig. 6. From the con-

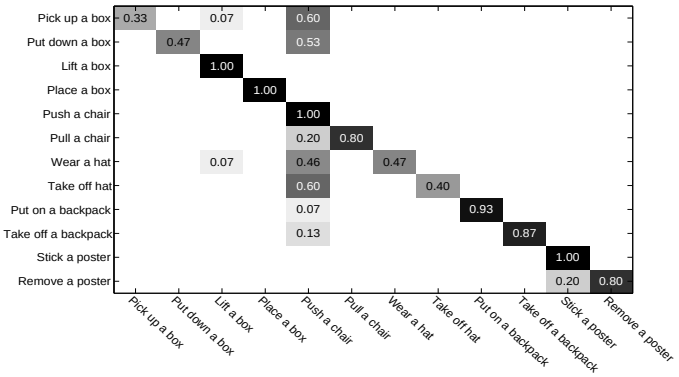

Fig. 6. The confusion matrix of our proposed method for MSR-ActionPairs3D.

fusion matrix, it can be seen that activities such as "Lift a box", "Place a Box", "Push a Chair", "Stick a Poster" are easy for our method to recognise, while "Pich up a Box" and "Take off Hat" are relatively difficult to recognise. The results have verified that our method can distinguish temporal orders in actions, however, it still can be confused with other actions which were not paired. One possible reason for causing the confusion between some actions, for instance, "Pick up a Box" and "Push a Chair", is the 3-state quantization of the unit different vectors. This issue can be addressed by quantizing the vector into more states.

V. Conclusion

In this paper, a new representation is proposed and effective data mining method is adopted to mine the mid-level patterns (different compositions of body parts) for action recognition. A novel method to explore temporal information and mine the different combinations of different body parts in different frames is proposed. The strength of the proposed method has been demonstrated through the state-of-the-art results obtained on the recent and challenging benchmark datasets for activity and action recognition. However, the proposed method can be further improved by combining depth or RGB data to explore the human-object interactions. With the increasing popularity of Kinect-based action recognition and data mining methods in computer vision, the proposed method has promising potentialities in practical applications.

References

[1] J. Shotton, A. Fitzgibbon, M. Cook, T. Sharp, M. Finocchio, R. Moore, A. Kipman, and A. Blake, "Real-time human pose recognition in parts from single depth images," in *Proceedings of the 2011 IEEE Conference on Computer Vision and Pattern Recognition*, ser. CVPR '11. IEEE Computer Society, 2011, pp. 1297–1304.

[2] M. Zanfir, M. Leordeanu, and C. Sminchisescu, "The moving pose: An efficient 3d kinematics descriptor for low-latency action recognition and detection," in *Computer Vision (ICCV), 2013 IEEE International Conference on*, Dec 2013, pp. 2752–2759.

[3] M. A. Gowayyed, M. Torki, M. E. Hussein, and M. El-Saban, "Histogram of oriented displacements (hod): Describing trajectories of human joints for action recognition," in *Proceedings of the Twenty-Third International Joint Conference on Artificial Intelligence*, ser. IJCAI'13. AAAI Press, 2013, pp. 1351–1357.

[4] B. Fernando, E. Fromont, and T. Tuytelaars, "Mining mid-level features for image classification," *International Journal of Computer Vision*, vol. 108, no. 3, pp. 186–203, 2014.

[5] W. Li, Z. Zhang, and Z. Liu, "Expandable data-driven graphical modeling of human actions based on salient postures," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 18, no. 11, pp. 1499–1510, 2008.

[6] W. Li, Z. Zhang, and Z.Liu, "Action recognition based on a bag of 3D points," in *IEEE International Workshop for Human Communicative Behavior Analysis (in conjunction with CVPR2010)*, 2010.

[7] 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*. Springer, 2013, pp. 149–187.

[8] L. Xia, C.-C. Chen, and J. Aggarwal, "View invariant human action recognition using histograms of 3d joints," in *Computer Vision and Pattern Recognition Workshops (CVPRW), 2012 IEEE Computer Society Conference on*. IEEE, 2012, pp. 20–27.

[9] H. S. Koppula, R. Gupta, and A. Saxena, "Learning human activities and object affordances from RGB-D videos," in *International Journal of Robotics Research (IJRR), 32(8): 951-970, July 2013.*, vol. 32, no. 8, July 2013, pp. 951–970.

[10] J. Sung, C. Ponce, B. Selman, and A. Saxena, "Unstructured human activity detection from rgbd images," in *Robotics and Automation (ICRA), 2012 IEEE International Conference on*. IEEE, 2012, pp. 842–849.

[11] J. Wang, Z. Liu, Y. Wu, and J. Yuan, "Learning actionlet ensemble for 3d human action recognition," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 36, no. 5, pp. 914–927, May 2014.

[12] X. Yang and Y. Tian, "Eigenjoints-based action recognition using naive-bayes-nearest-neighbor," in *International Workshop on Human Activity Understanding from 3D Data (HAU3D) in conjunction with CVPR*, June 2012, pp. 14–19.

[13] M. E. Hussein, M. Torki, M. A. Gowayyed, and M. El-Saban, "Human action recognition using a temporal hierarchy of covariance descriptors on 3d joint locations," in *Proceedings of the Twenty-Third International Joint Conference on Artificial Intelligence*, ser. IJCAI'13. AAAI Press, 2013, pp. 2466–2472.

[14] T. T. Thanh, F. Chen, and K. Kotani, "Extraction of discriminative patterns from skeleton sequences for accurate action recognition," *Fundamenta Informaticae*, pp. 1–15, 2014, in Press.

[15] C. Wang, Y. Wang, and A. L. Yuille, "An approach to pose-based action recognition," in *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*. IEEE, 2013, pp. 915–922.

[16] E. Ohn-Bar and M. Trivedi, "Joint angles similarities and hog2 for action recognition," in *Computer Vision and Pattern Recognition Workshops (CVPRW), 2013 IEEE Conference on*, June 2013, pp. 465–470.

[17] I. Theodorakopoulos, D. Kastaniotis, G. Economou, and S. Fotopoulos, "Pose-based human action recognition via sparse representation in dissimilarity space," *Journal of Visual Communication and Image Representation*, vol. 25, no. 1, pp. 12 – 23, 2014, visual Understanding and Applications with RGB-D Cameras.

[18] A. A. Chaaraoui, J. R. Padilla-Lpez, P. Climent-Prez, and F. Flrez-Revuelta, "Evolutionary joint selection to improve human action recognition with rgb-d devices," *Expert Systems with Applications*, vol. 41, no. 3, pp. 786 – 794, 2014.

[19] S. Althloothi, M. H. Mahoor, X. Zhang, and R. M. Voyles, "Human activity recognition using multi-features and multiple kernel learning," *Pattern Recognition.*, vol. 47, no. 5, pp. 1800–1812, May 2014.

[20] A. Chaaraoui, J. Padilla-Lopez, and F. Florez-Revuelta, "Fusion of skeletal and silhouette-based features for human action recognition with rgb-d devices," in *Computer Vision Workshops (ICCVW), 2013 IEEE International Conference on*, Dec 2013, pp. 91–97.

[21] H. Rahmani, A. Mahmood, A. Mian, and D. Huynh, "Real time action recognition using histograms of depth gradients and random decision

forests," in *IEEE Winter Applications of Computer Vision Conference (WACV)*, March 2014, pp. 14–19.

[22] V. M. Zatsiorsky, *Kinematics of human motion*. Human Kinetics, 1998.

[23] T. Uno, T. Asai, Y. Uchida, and H. Arimura, "Lcm: An efficient algorithm for enumerating frequent closed item sets," in *In Proceedings of Workshop on Frequent itemset Mining Implementations (FIMI03*, 2003.

[24] X. Yan, H. Cheng, J. Han, and D. Xin, "Summarizing itemset patterns: a profile-based approach," in *Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining*. ACM, 2005, pp. 314–323.

[25] H. Cheng, X. Yan, J. Han, and C.-W. Hsu, "Discriminative frequent pattern analysis for effective classification," in *Data Engineering, 2007. ICDE 2007. IEEE 23rd International Conference on*. IEEE, 2007, pp. 716–725.

[26] C.-C. Chang and C.-J. Lin, "LIBSVM: A library for support vector machines," *ACM Transactions on Intelligent Systems and Technology*, vol. 2, pp. 27:1–27:27, 2011, software available at http://www.csie.ntu.edu.tw/∼cjlin/libsvm.

[27] J. Wang, Z. Liu, Y. Wu, and J. Yuan, "Mining actionlet ensemble for action recognition with depth cameras," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012.

[28] O. Oreifej and Z. Liu, "Hon4d: Histogram of oriented 4d normals for activity recognition from depth sequences," in *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*, June 2013, pp. 716–723.

[29] M. Müller and T. Röder, "Motion templates for automatic classification and retrieval of motion capture data," in *Proceedings of the 2006 ACM SIGGRAPH/Eurographics symposium on Computer animation*. Eurographics Association, 2006, pp. 137–146.

[30] 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*. ACM, 2012, pp. 1057–1060.