Mining actionlet ensemble for action recognition with depth cameras

Wang, Jiang; Liu, Zicheng; Wu, Ying; Yuan, Junsong

2012

Wang, J., Liu, Z., Wu, Y., & Yuan, J. (2012). Mining actionlet ensemble for action recognition with depth cameras. 2012 IEEE Conference on Computer Vision and Pattern Recognition, 1290-1297.

https://hdl.handle.net/10356/100602

https://doi.org/10.1109/CVPR.2012.6247813

Mining Actionlet Ensemble for Action Recognition with Depth Cameras

Jiang Wang[1] Zicheng Liu[2] Ying Wu[1] Junsong Yuan[3]
jwa368@eecs.northwestern.edu zliu@microsoft.com yingwu@northwestern.edu jsyuan@ntu.edu.sg
[1]Northwestern University [2]Microsoft Research [3]Nanyang Technological University

Abstract

Human action recognition is an important yet challenging task. The recently developed commodity depth sensors open up new possibilities of dealing with this problem but also present some unique challenges. The depth maps captured by the depth cameras are very noisy and the 3D positions of the tracked joints may be completely wrong if serious occlusions occur, which increases the intra-class variations in the actions. In this paper, an actionlet ensemble *model is learnt to represent each action and to capture the intra-class variance. In addition, novel features that are suitable for depth data are proposed. They are robust to noise, invariant to translational and temporal misalignments, and capable of characterizing both the human motion and the human-object interactions. The proposed approach is evaluated on two challenging action recognition datasets captured by commodity depth cameras, and another dataset captured by a MoCap system. The experimental evaluations show that the proposed approach achieves superior performance to the state of the art algorithms.*

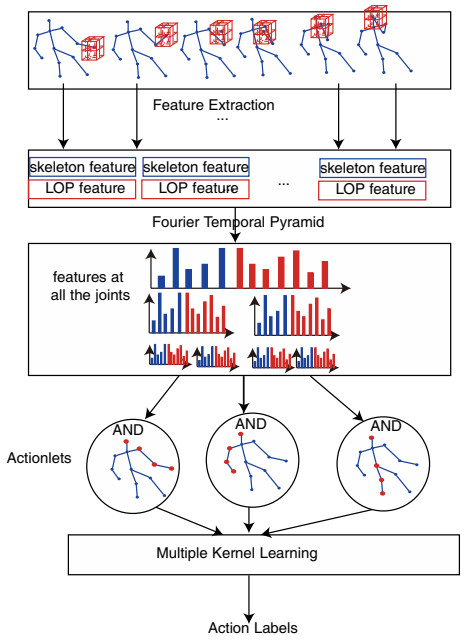

Figure 1. The general framework of the proposed approach.

1. Introduction

Recognizing human actions can have many potential applications including video surveillance, human computer interfaces, sports video analysis and video retrieval. Despite the research efforts in the past decade and many encouraging advances, accurate recognition of the human actions is still a quite challenging task. There are two related major issues for human action recognition. The first one is the sensory input, and the other is the modeling of the human actions that are dynamic, ambiguous and interactive with other objects.

The human motion is articulated, and capturing such highly articulated motion from monocular video sensors is a very difficult task. This difficulty largely limits the performance of video-based human action recognition, as indicated in the studies in the past decade. The recent introduction of the cost-effective depth cameras may change the picture by providing 3D depth data of the scene, which largely eases the task of object segmentation. Moreover, it has facilitated a rather powerful human motion capturing technique [20] that outputs the 3D joint positions of the human skeleton.

Although the depth cameras in general produce better quality 3D motion than those estimated from monocular video sensors, simply using such 3D motion sensory data and the estimated 3D joint positions for human action recognition is not plausible. One reason is that the estimated 3D joint positions are noisy and may have significant errors when there are occlusions such as one leg being in front of the other, a hand touching another body part, two hands crossing, etc. In addition, the 3D skeleton motion alone is not sufficient to distinguish some actions. For example, "drinking" and "eating snacks" give very similar motion for the human skeleton. Extra inputs need to be included and exploited for better recognition.

This paper presents a novel human action recognition ap-

proach using a depth camera. The basic idea is illustrated in Fig. 1. Based on the depth data and the estimated 3D joint positions, we propose a new feature called *local occupancy pattern* or LOP feature. Each 3D joint is associated with a LOP feature, which can be treated as the "depth appearance" of this 3D joint. Translational invariant and highly discriminative, this new feature is also able to capture the relations between the human body parts and the environmental objects in the interaction. In addition, to represent the temporal structure of an individual joint in an action, we propose a new temporal pattern representation called *Fourier Temporal Pyramid*. This representation is insensitive to temporal sequence misalignment and is robust to noise.

More importantly, we propose a new model for human actions, called the *Actionlet Ensemble Model*. The articulated human body has a large number of kinematic joints, but a certain action is usually only associated with and characterized by the interactions and combinations of a subset of them. For example, the joints "right wrist" and "head" are discriminative for the action "drinking". Therefore, we introduce the concept of *actionlet*. An *actionlet* is a particular conjunction of the features for a subset of the joints, indicating a structure of the features. As there are an enormous number of possible *actionlets*, we propose a novel data mining solution to discover *discriminative actionlets*. Then an action is represented as an *Actionlet Ensemble*, which is a linear combination of the *actionlets*, and their discriminative weights are learnt via a multiple kernel learning method. This new action model is more robust to the errors in the features, and it can better characterize the intra-class variations in the actions. For example, for the action "call cellphone", some people use their right hands while others use their left hands. This variation can be characterized by the proposed *actionlet ensemble* model.

Our main contributions include the following three aspects. First, this paper proposes the *actionlet ensemble* model as a new way of characterizing and recognizing human actions. Second, our extensive experiments have shown that the proposed features are well suitable for the depth data-based action recognition task. Third, the proposed Fourier temporal pyramid is a new representation of temporal patterns, and it is shown to be robust to temporal misalignment and noise.

The proposed features and models are evaluated on three benchmark datasets: CMU MoCap dataset [1], MSR-Action3D dataset [14] and DailyActivity3D dataset. The first dataset contains 3D joint positions captured by a multi-camera motion capturing system, and the other two datasets are captured with commodity depth cameras. Our extensive experimental results show that the proposed method is able to achieve significantly better recognition accuracy than the state-of-the-art methods.

2. Related Work

Actions are spatio-temporal patterns. There are two important issues in action recognition: the representation of suitable spatio-temporal features, and the modeling of dynamical patterns.

Features can be sensor-dependent. In video-based methods, it is a common practice to locate spatio-temporal interest points like STIP [10], and use the distributions of the local features like HOF [11] or HOG [7] to represent such local spatio-temporal pattern. When we want to use depth data, however, because there is no texture in the depth map, these local features are not discriminative enough for classification.

It is generally agreed that knowing the 3D joint position is helpful for action recognition. Multi-camera motion capture (MoCap) systems [3] have been used for human action recognition, but such special equipment is marker-based and expensive. It is still a challenging problem for marker-free motion capturing via regular video sensors. Cost-effective depth cameras have been used for motion capturing, and produced reasonable results, despite of the noise when occlusion occurs. Because of the different quality of the motion data, the action recognition methods designed for MoCap data may not be suitable for depth camera.

In the literature, there have been many different approaches for temporal modeling. One way to model the human actions is to employ generative models, such as a Hidden Markov model (HMM) for a number of pre-defined relative position features from 3D joint positions [15], or a conditional random field (CRF) for the 3D joint positions [9]. Similar approaches are also proposed to model human actions in normal videos [18, 5]. However, the 3D joint positions that are generated via skeleton tracking from the depth map sequences are generally more noisy than that of the MoCap data. When the difference between the actions is subtle, it is usually difficult to determine the accurate states from the observation without careful selection of the features, which undermines the performance of such generative models. Moreover, with limited amount of training data, training a complex generative model is easy to overfit.

Another generative approach to dynamical patterns can also be modeled by linear dynamical systems, and the states of the system can be used for MoCap action categorization [9]. In addition, the complex and nonlinear dynamics can also be characterized by a Recurrent Neural Network [16]. Although these two approaches are good models for time series data and are robust to temporal misalignment, it is generally difficult to learn these models from limited amount of training data.

Another method for modeling actions is dynamic temporal warping (DTW), which matches the 3D joint positions to a template [17], and action recognition can be done through

a nearest-neighbor classification method. Its performance heavily depends on a good metric to measure the similarity of frames. Moreover, for periodic actions (such as "waving"), DTW is likely to produce large temporal misalignment which may ruin action classification [13].

Different from these approaches, we propose to employ Fourier Temporal Pyramid to represent the temporal patterns. The Fourier temporal pyramid is a descriptive model. It does not involve complicated learning as in the generative models (e.g., HMM, CRF and dynamical systems), and it is much more robust than DTW to noise and temporal misalignment.

For the action of a complex articulated structure, the motion of the individual parts are correlated. The relationship among these parts (or high-order features) may be more discriminative than the individual ones. Such combinatorial features can be represented by stochastic AND/OR structures. This idea has been pursued for face detection [6], human body parsing [24], object recognition [23], and human object interaction recognition [22]. This paper presents an initial attempt of using the AND/OR ensemble approach to action recognition. We propose a novel data mining solution to discover the discriminative conjunction rules, and apply multiple kernel learning to learn the ensemble.

3. Spatio-Temporal Features

This section gives a detailed description of two types of features that we utilize to represent the actions: the 3D joint position feature and the Local Occupancy Pattern (LOP). These features can characterize the human motions as well as the interactions between the objects and the human. In addition, the Fourier Temporal Pyramid is proposed to represent the temporal dynamics. The proposed features are invariant to the translation of the human and robust to noise and temporal misalignment.

3.1. Invariant Features for 3D Joint Positions

The 3D joint positions are employed to shape the motion of the human body. Our key observation is that representing the human movement as the pairwise relative positions of the joints results in more discriminative features.

For a human subject, 20 joint positions are tracked (the Motion Capture system captures 30 joints) by the skeleton tracker [20] and each joint i has 3 coordinates $\boldsymbol{p}_i(t) = (x_i(t), y_i(t), z_i(t))$ at a frame t. The coordinates are normalized so that the motion is invariant to the absolute body position, the initial body orientation and the body size.

For each joint i, we extract the pairwise relative position features by taking the difference between the position of joint i and that of each other joint j:

$$\boldsymbol{p}_{ij} = \boldsymbol{p}_i - \boldsymbol{p}_j, \tag{1}$$

The 3D joint feature for joint i is defined as:

$$\boldsymbol{p}_i = \{\boldsymbol{p}_{ij} | i \neq j\}.$$

Although enumerating all the joint pairs introduces some information that is irrelevant to our classification task, our approach is able to select the joints that are most relevant to our recognition task. The selection will be handled by the *Actionlet* mining as discussed in Section 4.

Representing the human motion as the relative joint positions results in more discriminative and intuitive features. For example, the action "waving" is generally interpreted as "arms above the shoulder and move left and right". This can be better characterized through the pairwise relative positions.

3.2. Local Occupancy Patterns

It is insufficient to only use the 3D joint positions to fully model an action, especially when the action includes the interactions between the subject and other objects. Therefore, it is necessary to design a feature to describe the local "depth appearance" for the joints. In this paper, the interaction between the human subject and the environmental objects is characterized by the *Local Occupancy Patterns* or LOP at each joint. For example, suppose a person is drinking a cup of water. When the person fetches the cup, the space around his/her hand is occupied by the cup. Afterwards, when the person lifts the cup to his/her mouth, the space around both the hand and the head is occupied. This information can be useful to characterize this interaction and to differentiate the drinking action from other actions.

In each frame, as described below, an LOP feature computes the local occupancy information based on the 3D point cloud around a particular joint, so that the temporal dynamics of all such occupancy patterns can roughly discriminate different types of interactions.

At frame t, we have the point cloud generated from the depth map of this frame. For each joint j, its local region is partitioned into $N_x \times N_y \times N_z$ spatial grid. Each bin of the grid is of size (S_x, S_y, S_z) pixels. For example, if $(N_x, N_y, N_z) = (12, 12, 4)$ and $(S_x, S_y, S_z) = (6, 6, 80)$, the local $(96, 96, 320)$ region around a joint is partitioned into $12 \times 12 \times 4$ bins, and the size of each bin is (6,6,80).

The number of points at the current frame that fall into each bin b_{xyz} of the grid is counted, and a sigmoid normalization function is applied to obtain the feature o_{xyz} for this bin. In this way, the local occupancy information of this bin is:

$$o_{xyz} = \delta\Big(\sum_{q \in \text{bin}_{xyz}} I_q\Big) \tag{2}$$

where $I_q = 1$ if the point cloud has a point in the location q and $I_q = 0$ otherwise. $\delta(.)$ is a sigmoid normalization

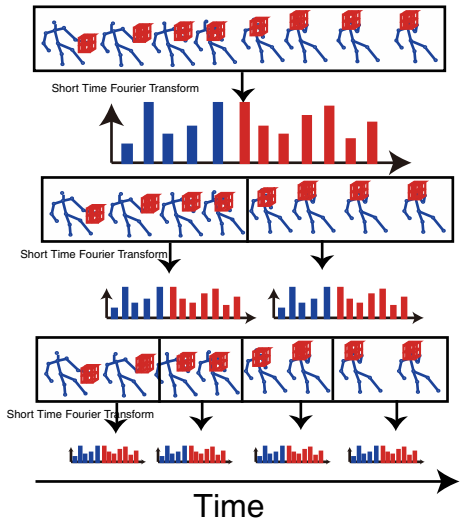

Figure 2. A Illustration of the Fourier Temporal Pyramid.

function: $\delta(x) = \frac{1}{1+e^{-\beta x}}$. The LOP feature of a joint i is a vector consisting of the feature o_{xyz} of all the bins in the spatial grid around the joint, denoted by $\boldsymbol{o}_i$.

3.3. Fourier Temporal Pyramid

Two types of features are extracted from each frame t : the 3D joint position features $\boldsymbol{p}_i[t]$, and the LOP features $\boldsymbol{o}_i[t]$. In this subsection, we propose the Fourier temporal pyramid to represent the temporal dynamics of these frame-level features.

When using the current cost-effective depth camera, we always experience noisy depth data and temporal misalignment. We aim to design temporal representations that are robust to both the data noise and the temporal misalignment. We also want such temporal features to be a good representation of the temporal structure of the actions. For example, one action may contain two consecutive sub-actions: "bend the body" and "pick up". The proposed Fourier Temporal Pyramid is a descriptive representation that satisfies these properties.

Inspired by the Spatial Pyramid approach [12], in order to capture the temporal structure of the action, in addition to the global Fourier coefficients, we recursively partition the action into a pyramid, and use the short time Fourier transform for all the segments, as illustrated in Fig. 2. The final feature is the concatenation of the Fourier coefficients from all the segments.

For each joint i, let $\boldsymbol{g}_i = (\boldsymbol{p}_i, \boldsymbol{o}_i)$ denote its overall feature vector where $\boldsymbol{p}_i$ is its 3D pairwise position vector and $\boldsymbol{o}_i$ is its LOP vector. Let N_i denote the dimension of $\boldsymbol{g}_i$, i.e., $\boldsymbol{g}_i = (g_1, \ldots, g_{N_i})$. Note that each element g_j is a function of time and we can write it as $g_j[t]$. For each time segment at each pyramid level, we apply Short Fourier Transform [19] to element $g_j[t]$ and obtain its Fourier coefficients, and we utilize its low-frequency coefficients as features. The Fourier Temporal Pyramid feature at joint i is defined as the low-frequency coefficients at all levels of the pyramid, and is denoted as $\boldsymbol{G}_i$.

The proposed Fourier Temporal Pyramid feature has several benefits. First, by discarding the high-frequency Fourier coefficients, the proposed feature is robust to noise. Second, this feature is insensitive to temporal misalignment, because time series with temporal translation have the same Fourier coefficient magnitude. Finally, the temporal structure of the actions can be characterized by the pyramid structure.

4. Actionlet Ensemble

Although the proposed feature is robust to noise, to deal with the errors of the skeleton tracking and better characterize the intra-class variations, an *actionlet ensemble* approach is proposed in this section as a representation of the actions.

An *actionlet* is defined as a conjunctive (or AND) structure on the base features. One base feature is defined as a Fourier Pyramid feature of one joint. A *discriminative actionlet* should be highly representative of one action and highly discriminative compared to other actions. A novel data mining algorithm is proposed to discover the discriminative actionlets.

Once we have mined a set of discriminative actionlets, a multiple kernel learning [4] approach is employed to learn an actionlet ensemble structure that combines these discriminative actionlets.

4.1. Mining Discriminative Actionlets

An actionlet is denoted as a subset of joints $S \subseteq \{1, 2, \cdots, N_j\}$, where N_j is the total number of joints.

Suppose we have training pairs $(\boldsymbol{x}^{(j)}, t^{(j)})$. In order to determine how discriminative each individual joint is, a SVM model is trained on feature $\boldsymbol{G}_i$ of each joint i. For each training sample $\boldsymbol{x}^{(j)}$ and the SVM model on the joint i, the probability that its classification label $y^{(j)}$ is equal to an action class c is denoted as $P_i(y^{(j)} = c|\boldsymbol{x}^{(j)})$, which can be estimated from the pairwise probabilities by using pairwise coupling approach[21].

Since an actionlet takes a conjunctive operation, it predicts $y^{(j)} = c$ if and only if every joint $i \in S$ predicts $y^{(j)} = c$. Thus, assuming the joints are independent, the probability that the predicted label $y^{(j)}$ is equal to an action class c given an example $\boldsymbol{x}^{(j)}$ for an actionlet S can be computed as:

$$P_S(y^{(j)} = c|\boldsymbol{x}^{(j)}) = \prod_{i \in S} P_i(y^{(j)} = c|\boldsymbol{x}^{(j)}) \qquad (3)$$

Define $\mathcal{X}_c$ as $\{j : t^{(j)} = c\}$. For an actionlet to be discriminative, the probability $P_S(y^{(j)} = c|\boldsymbol{x}^{(j)})$ should be large for some data in $\mathcal{X}_c$, and be small for all the data that does not belong to $\mathcal{X}_c$. Define the confidence for actionlet S as

$$\text{Conf}_S = \max_{j \in \mathcal{X}_c} \log P_S(y^{(j)} = c|\boldsymbol{x}^{(j)}) \tag{4}$$

and the ambiguity for actionlet S as

$$\text{Amb}_S = \sum_{j \notin \mathcal{X}_c} \log P_S(y^{(j)} = c|\boldsymbol{x}^{(j)}) \tag{5}$$

We would like a discriminative actionlet to have large confidence Conf_S and small ambiguity Amb_S. An actionlet S is called an l-actionlet if its cardinality $|S| = l$. One important property is that if we add a joint $i \notin S$ to an $(l-1)$-actionlet S to generate an l-actionlet $S \cup \{i\}$, we have $\text{Conf}_{S\cup\{i\}} \leq \text{Conf}_S$, i.e., adding a new joint into one actionlet will always reduce the confidence. As a result, the Aprior mining process [2] can be applied to select the actionlets with large Conf_S and small Amb_S. If Conf_S is less than the threshold, we do not need to consider any S' with $S' \supset S$. The outline of the mining process is shown in Alg. 1. For each class c, the mining algorithm outputs a discriminative actionlet pool P_c which contains the actionlets that meet our criteria: $\text{Amb}_S \leq T_{\text{amb}}$ and $\text{Conf}_S \geq T_{\text{conf}}$.

```
Take the set of joints, the feature G_i on each joint i,
the number of the classes C, thresholds T_conf and T_amb.
Train the base classifier on the features G_i of each
joint i.
for Class c = 1 to C do
    Set P_c, the discriminative actionlet pool for class c
    to be empty : P_c = {}. Set l = 1.
    repeat
        Generate the l-actionlets by adding one joint
        into each (l-1)-actionlet in the
        discriminative actionlet pool P_c.
        Add the l-actionlets whose confidences are
        larger than T_conf to the pool P_c.
        l = l + 1
    until no discriminative actionlet is added to P_c in
    this iteration;
    remove the actionlets whose ambiguities are larger
    than T_amb in the pool P_c.
end
return the discriminative actionlet pool for all the
classes
```

Algorithm 1: Discriminative Actionlet Mining

4.2. Learning Actionlet Ensemble

For each actionlet S_k in the discriminative actionlet pool, an SVM model on it defines a joint feature map $\Phi_k(\boldsymbol{x}, y)$ on data $\mathcal{X}$ and labels $\mathcal{Y}$ as a linear output function $f_k(\boldsymbol{x}, y) = \langle \boldsymbol{w}_k, \Phi_k(\boldsymbol{x}, y)\rangle + b_k$, parameterized with the hyperplane normal $\boldsymbol{w}_k$ and bias b_k. The predicted class y for $\boldsymbol{x}$ is chosen to maximize the output $f_k(\boldsymbol{x}, y)$.

Multiclass-MKL considers a convex combination of p kernels, $K(\boldsymbol{x}_i, \boldsymbol{x}_j) = \sum_{k=1}^{p} \beta_k K_k(\boldsymbol{x}_i, \boldsymbol{x}_j)$, where each kernel corresponds to an actionlet. Equivalently, we consider the following output function:

$$f_{\text{final}}(\boldsymbol{x}, y) = \sum_{k=1}^{p} [\beta_k \langle \boldsymbol{w}_k, \Phi_k(\boldsymbol{x}, y)\rangle + b_k] \tag{6}$$

We aim at choosing $\boldsymbol{w} = (\boldsymbol{w}_k), \boldsymbol{b} = (\boldsymbol{b}_k), \boldsymbol{\beta} = (\beta_k), k = 1, \ldots, p$, such that given any training data pair $(\boldsymbol{x}^{(i)}, y^{(i)})$, $f_{\text{final}}(\boldsymbol{x}^{(i)}, y^{(i)}) \geq f_{\text{final}}(\boldsymbol{x}^{(i)}, u)$ for all $u \in \mathcal{Y} - \{y^{(i)}\}$. The resulting optimization problem becomes:

$$\begin{aligned} &\min_{\boldsymbol{\beta}, \boldsymbol{w}, \boldsymbol{b}, \xi} \frac{1}{2}\Omega(\boldsymbol{\beta}) + C \sum_{i=1}^{n} \xi_i \\ &\text{s.t. } \forall i : \xi_i = \max_{u \neq y_i} l(f_{\text{final}}(\boldsymbol{x}^{(i)}, y^{(i)}) - f_{\text{final}}(\boldsymbol{x}^{(i)}, u)) \end{aligned} \tag{7}$$

where C is the regularization parameter and l is a convex loss function, and $\Omega(\boldsymbol{\beta})$ is a regularization parameter on the $\boldsymbol{\beta}$. Following the approach in [8], we choose $\Omega(\boldsymbol{\beta}) = \|\boldsymbol{\beta}\|_1^2$ to encourage a sparse β, so that an ensemble of a small number of actionlets is learned.

This problem can be solved by iteratively optimizing $\boldsymbol{\beta}$ with fixed $\boldsymbol{w}$ and $\boldsymbol{b}$ through linear programming, and optimizing $\boldsymbol{w}$ and $\boldsymbol{b}$ with fixed $\boldsymbol{\beta}$ through a generic SVM solver such as LIBSVM.

5. Experimental Results

We choose CMU MoCap dataset [1], MSR-Action3D dataset [14] and MSRDailyActivity3D dataset to evaluate the proposed action recognition approach. In all the experiments, we use three-level Fourier temporal pyramid, with 1/4 length of each segment as low-frequency coefficients. The empirical results show that the proposed framework outperforms the state of the art methods.

5.1. MSR-Action3D Dataset

MSR-Action3D dataset [14] is an action dataset of depth sequences captured by a depth camera. This dataset contains twenty actions: *high arm wave*, *horizontal arm wave*, *hammer*, *hand catch*, *forward punch*, *high throw*, *draw x*, *draw tick*, *draw circle*, *hand clap*, *two hand wave*, *side-boxing*, *bend*, *forward kick*, *side kick*, *jogging*, *tennis swing*, *tennis serve*, *golf swing*, *pick up & throw*. Each action was performed by ten subjects for three times. The frame rate is 15 frames per second and resolution 640×480. Altogether, the dataset has 23797 frames of depth map for 402 action

Figure 3. Sample frames of the MSR-Action3D dataset.

Method	Accuracy
Recurrent Neural Network [16]	0.425
Dynamic Temporal Warping [17]	0.54
Hidden Markov Model [15]	0.63
Action Graph on Bag of 3D Points [14]	0.747
Proposed Method	**0.882**

Table 1. Recognition Accuracy Comparison for MSR-Action3D dataset.

samples. Some examples of the depth sequences are shown in Fig. 3.

Those actions were chosen to cover various movement of arms, legs, torso and their combinations, and the subjects were advised to use their right arm or leg if an action is performed by a single arm or leg. Although the background of this dataset is clean, this dataset is challenging because many of the actions in the dataset are highly similar to each other.

The 3D joint positions are extracted from the depth sequence by using the real time skeleton tracking algorithm proposed in [20]. Since there is no human-object interaction in this dataset, we only extract the 3D joint position features.

We compare our method with the state-of-the-art methods on the cross-subject test setting [14], where the samples of half of the subjects are used as training data, and the rest of the samples are used as test data. The recognition accuracy of the dynamic temporal warping is only 54%, because some of actions in the dataset are very similar to each other, and there are typical large temporal misalignment in the dataset. The accuracy of recurrent neural network is 42.5%. The accuracy of Hidden Markov Model is 63%. The proposed method achieves an accuracy of 88.2%. This is a very good performance considering that the skeleton tracker sometimes fails and the tracked joint positions are quite noisy. The confusion matrix is illustrated in Fig. 4. For most of the actions, our method works very well. The classification errors occur if two actions are too similar to each other, such as "hand catch" and "high throw", or if the occlusion is so large that the skeleton tracker fails frequently, such as the action "pick up and throw".

The comparison between the robustness of the Fourier

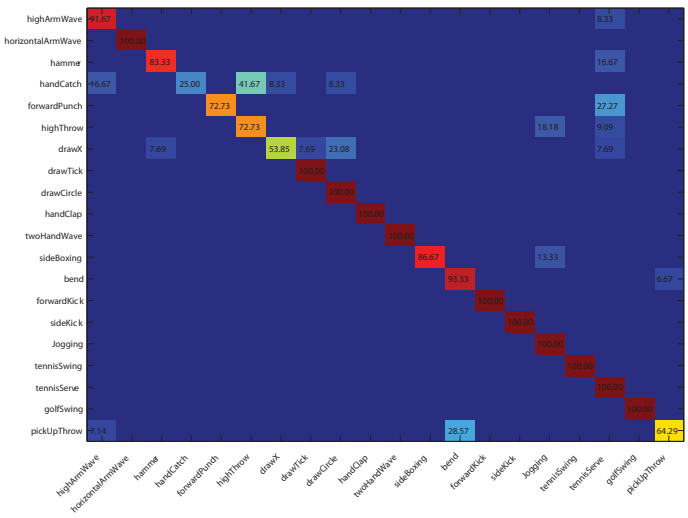

Figure 4. The confusion matrix for MSR-Action3D dataset.

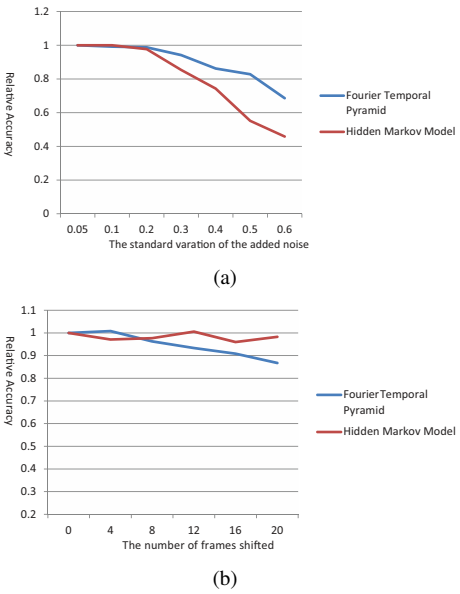

Figure 5. The relationship between the relative accuracy and the noise or temporal shift.

Temporal Pyramid features and that of Hidden Markov Model is shown in Fig. 5(a). In this experiment, we add white Gaussian noise to the 3D joint positions of all samples, and compare the relative accuracies of the two methods. For each method, its relative accuracy is defined as the accuracy under the noisy environment divided by the accuracy under the environment without noise. We can see that the proposed Fourier Temporal Pyramid feature is much more robust to noise than the Hidden Markov Model.

The robustness of the proposed method and the Hidden Markov model to temporal shift is also compared. In

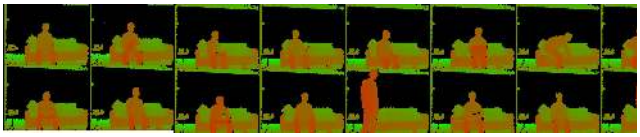

Figure 6. Sample frames of the DailyActivity3D dataset.

this experiment, we circularly shift all the training data, and keep the test data unchanged. The relative accuracy is shown in Fig. 5(b). It can be seen that both methods are robust to the temporal shift of the depth sequences, though the Fourier Temporal Pyramid is slightly more sensitive to temporal shift than the Hidden Markov Model.

5.2. MSRDailyActivity3D Dataset

DailyActivity3D dataset[1] is a daily activity dataset captured by a Kinect device. There are 16 activity types: *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*. If possible, each subject performs an activity in two different poses: "sitting on sofa" and "standing". The total number of the activity samples is 320. Some example activities are shown in Fig. 6.

This dataset is designed to cover human's daily activities in the living room. When the performer stands close to the sofa or sits on the sofa, the 3D joint positions extracted by the skeleton tracker are very noisy. Moreover, most of the activities involve the humans-object interactions. Thus this dataset is more challenging.

Table 2 shows the accuracies of different methods. By employing an actionlet ensemble model, we obtain a recognition accuracy of 85.75%. This is a decent result considering the difficulties in this dataset. If we directly train a SVM on the Fourier Temporal Pyramid features, the accuracy is 78%. When only the LOP feature is employed, the recognition accuracy drops to 42.5%. If we only use 3D joint position features without using LOP, the recognition accuracy is 68%.

Fig. 7 shows the confusion matrix of the proposed method. Fig. 8 compares the accuracy of the actionlet ensemble method and that of the support vector machine on the Fourier Temporal Pyramid features. We can observe that for the activities where the hand gets too close to the body, the proposed actionlet ensemble method can significantly improve the accuracy. Fig. 9 illustrates the actionlets with high weights discovered by our mining algorithm.

5.3. CMU MoCap Dataset

We also evaluate the proposed method on the 3D joint positions extracted by a motion capture system. The dataset we use is the CMU Motion Capture (MoCap) dataset.

Five subtle actions are chosen from CMU MoCap datasets following the configuration in [9]. The five actions differ from each other only in the motion of one or two limbs. The actions in this dataset include: *walking*, *marching*, *dribbling*, *walking with stiff arms*, *walking with wild legs*. The 3D joint positions in CMU MoCap dataset are relatively clean because they are captured with high-

[1] http://research.microsoft.com/~zliu/ActionRecoRsrc

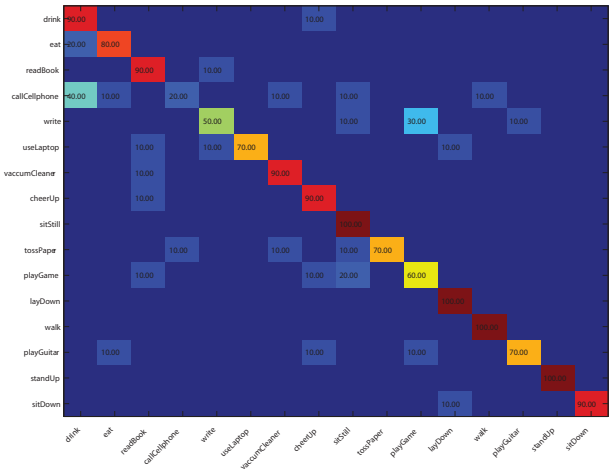

Figure 7. The confusion matrix of the proposed method on DailyActivity3D dataset.

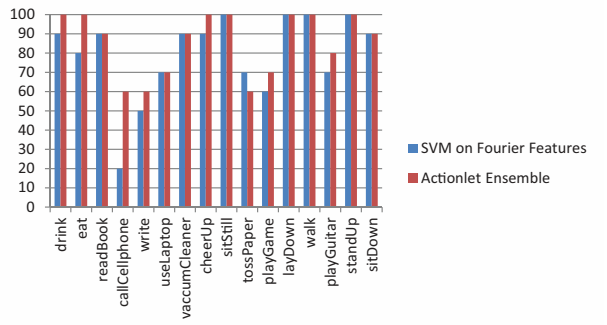

Figure 8. The comparison between the accuracy of the proposed actionlet ensemble method and that of the support vector machine on the Fourier Temporal Pyramid features.

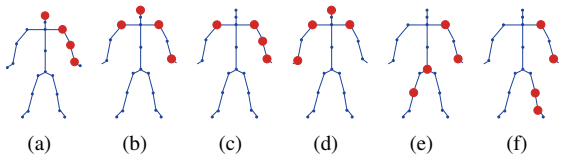

Figure 9. Examples of the mined actionlets. The joints contained in each actionlet are marked as red. (a), (b) are actionlets for "drink" (c), (d) are actionlets for "call". (e), (f) are actionlets for "walk".

Method	Accuracy
Dynamic Temporal Warping [17]	0.54
Only LOP features	0.425
Only Joint Position features	0.68
SVM on Fourier Temporal Pyramid Features	0.78
Actionlet Ensemble	**0.8575**

Table 2. Recognition Accuracy Comparison for DailyActivity3D dataset.

Method	Accuracy
CRF with learned manifold space [9]	0.9827
Proposed Method	**0.9813**

Table 3. Recognition Accuracy Comparison for CMU MoCap dataset.

precision camera array and markers. This dataset is employed to evaluate the performance of the proposed 3D joint position-based features on 3D joint positions captured by Motion Capture system.

The comparison of the performance is shown in Table 3. Since only the 3D joint positions are available, the propose method only utilizes the 3D joint position features. It can be seen that the proposed method achieves comparable results with the state of the art methods on the MoCap dataset.

6. Conclusion

We have proposed novel features and an actionlet ensemble model for human action recognition with depth cameras. The proposed features are discriminative enough to classify human actions with subtle differences as well as human-object interactions and robust to noise and temporal misalignment. The actionlet ensemble model is capable of better capturing the intra-class variations and is more robust to the noises and errors in the depth maps and joint positions. The experiments demonstrated the superior performance of the proposed approach to the state of the art methods. In the future, we aim to exploit the effectiveness of the proposed technique for the understanding of more complex activities.

7. Acknowledgements

This work was supported in part by National Science Foundation grant IIS-0347877, IIS-0916607, US Army Research Laboratory and the US Army Research Office under grant ARO W911NF-08-1-0504, and DARPA Award FA 8650-11-1-7149. This work is partially supported by Microsoft Research.

References

[1] CMU Graphics Lab Motion Capture Database. 2, 5

[2] R. Agrawal and R. Srikant. Fast algorithms for mining association rules. In *VLDB*, volume 1215, pages 487–499. Citeseer, 1994. 5

[3] L. Campbell and A. Bobick. Recognition of human body motion using phase space constraints. In *ICCV*, number 309. Published by the IEEE Computer Society, 1995. 2

[4] O. Chapelle, V. Vapnik, O. Bousquet, and S. Mukherjee. Choosing multiple parameters for support vector machines. *Machine Learning*, 46(1):131–159, 2002. 4

[5] H. Chen, H. Chen, Y. Chen, and S. Lee. Human action recognition using star skeleton. In *Proceedings of the 4th ACM international workshop on Video surveillance and sensor networks*, pages 171–178, New York, New York, USA, 2006. ACM. 2

[6] S. Dai, M. Yang, Y. Wu, and A. Katsaggelos. Detector ensemble. In *CVPR*, pages 1–8. Ieee, June 2007. 3

[7] N. Dalal and B. Triggs. Histograms of Oriented Gradients for Human Detection. In *CVPR*, pages 886–893. Ieee, 2005. 2

[8] J. H. Friedman and B. E. Popescu. Predictive learning via rule ensembles. *The Annals of Applied Statistics*, 2(3):916–954, Sept. 2008. 5

[9] L. Han, X. Wu, W. Liang, G. Hou, and Y. Jia. Discriminative human action recognition in the learned hierarchical manifold space. *Image and Vision Computing*, 28(5):836–849, May 2010. 2, 7, 8

[10] I. Laptev. On Space-Time Interest Points. *IJCV*, 64(2-3):107–123, Sept. 2005. 2

[11] I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld. Learning realistic human actions from movies. In *CVPR*, volume 1, pages 1–8, 2008. 2

[12] S. Lazebnik, C. Schmid, and J. Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *CVPR*, volume 2. IEEE, 2006. 4

[13] L. Li and B. Prakash. Time Series Clustering: Complex is Simpler! In *ICML*, 2011. 3

[14] W. Li, Z. Zhang, and Z. Liu. Action recognition based on a bag of 3d points. In *Human Communicative Behavior Analysis Workshop (in conjunction with CVPR)*, 2010. 2, 5, 6

[15] F. Lv and R. Nevatia. Recognition and Segmentation of 3-D Human Action Using HMM and Multi-class AdaBoost. In *ECCV*, pages 359–372, 2006. 2, 6

[16] J. Martens and I. Sutskever. Learning Recurrent Neural Networks with Hessian-Free Optimization. In *ICML*, 2011. 2, 6

[17] M. Muller 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*, pages 137–146. Eurographics Association, 2006. 2, 6, 8

[18] H. Ning, W. Xu, Y. Gong, and T. Huang. Latent Pose Estimator for Continuous Action. In *ECCV*, pages 419–433, 2008. 2

[19] A. V. Oppenheim, R. W. Schafer, and J. R. Buck. *Discrete Time Signal Processing*. Prentice Hall Signal Processing Series. Prentice Hall, 1999. 4

[20] 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 *CVPR*, 2011. 1, 3, 6

[21] T.-f. Wu, C.-j. Lin, and R. C. Weng. Probability Estimates for Multi-class Classification by Pairwise Coupling. *Journal of Machine Learning Research*, 5:975–1005, 2004. 4

[22] B. Yao and L. Fei-Fei. Grouplet: a structured image representation for recognizing human and object interactions. In *CVPR*, 2010. 3

[23] J. Yuan, M. Yang, and Y. Wu. Mining Discriminative Co-occurrence Patterns for Visual Recognition. In *CVPR*, 2011. 3

[24] L. L. Zhu, Y. Chen, Y. Lu, C. Lin, and A. Yuille. Max Margin AND/OR Graph learning for parsing the human body. In *CVPR*, pages 1–8. Ieee, June 2008. 3