2014 IEEE International Conference on Robotics & Automation (ICRA)
Hong Kong Convention and Exhibition Center
May 31 - June 7, 2014. Hong Kong, China

Action Recognition Using Ensemble Weighted Multi-Instance Learning

Guang Chen[1], Manuel Giuliani[2], Daniel Clarke[2], Andre Gaschler[2], Alois Knoll[1]

***Abstract*— This paper deals with recognizing human actions in depth video data. Current *state-of-the-art* action recognition methods use hand-designed features, which are difficult to produce and time-consuming to extend to new modalities. In this paper, we propose a novel, 3.5D representation of a depth video for action recognition. A 3.5D graph of the depth video consists of a set of nodes that are the joints of the human body. Each joint is represented by a set of spatio-temporal features, which are computed by an unsupervised learning approach. However, if occlusions occur, the 3D positions of the joints are noisy which increases the intra-class variations in action classes. To address this problem, we propose the Ensemble Weighted Multi-Instance Learning approach (EnwMi) for the action recognition task. It considers the class imbalance and intra-class variations. We formulate the action recognition task with depth videos as a weighted multi-instance problem. We further integrate an ensemble learning method into the weighted multi-instance learning framework. Our approach is evaluated on Microsoft Research Action3D dataset, and the results show that it outperforms *state-of-the-art* methods.**

I. Introduction

Human action recognition has played an important role in a number of real-word applications such as video surveillance, health care, and a variety of systems that involve interactions between persons and computers. Especially in robotics, the ability of a robot to understand the action of its human peers is critical for the robot to collaborate effectively and efficiently with humans in a peer-to-peer human-robot team. With recent developments to low-cost sensors, depth cameras have received a great deal of attention from researchers.

Compared to a visible light camera, depth sensors have several advantages. For example, depth images provide 3D structural information of a scene, which can often be more discriminative than color and texture in many applications including detection, segmentation and action recognition. These advantages have facilitated a rather powerful human motion capturing technique [16] that generates 3D joint positions of the human skeleton.

In action recognition, which is the topic of this paper, two significant questions arise when using depth sequences. First, will RGB-based methods for action recognition perform well when using depth sensors? There is no rich texture in depth data, which hinders the extension of hand-designed features from color-based data to depth data, such as STIP [7], and HOG [2]. Furthermore, the depth images are often contaminated with undefined depth points, which appear in the sequences as large shadows. Second, will the noisy human skeleton data perform well in action recognition? Skeleton data are able to provide additional body part information to differentiate actions. However, the skeleton tracking algorithm proposed in [16] produces inaccurate results or even fails when occlusion occurs.

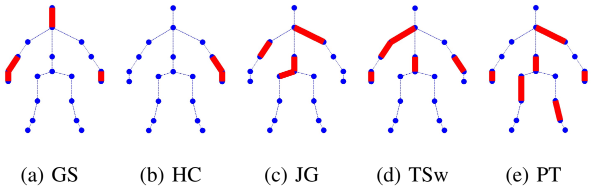


Fig. 1. Examples of the skeleton for different action classes. The discriminative joints discovered by our method are marked as thick and red lines. (a): Golf Swing, (b): Hand Catch, (c): Jogging, (d): Tennis Swing, (e): Pickup Throw. (best viewed in color).

These challenges motivate us to seek for feature representations that are highly discriminative and robust to occlusions. Our work in this paper proceeds along this direction. We propose a novel action recognition approach to address the above two challenges. Specifically, we make two key contributions:

First, we learn 3.5D graph from depth video data using unsupervised learning approache. We provide an unsupervised learning method to learn a 3.5D representation of depth video inspired by [6], [9]. At the heart of our method is the use of the Independent Subspace Analysis (ISA). The ISA algorithm is a well-known algorithm in the field of natural image statics [6]. An advantage of ISA is that it learns features that are robust to local translation while being selective to rotation and velocity. A disadvantage of ISA is that it can be slow to train with high dimensionality data (e.g. video data). In this paper, we extend the ISA algorithm for the use of depth video data (see Fig .2). Instead of training the model with the entire video, we apply the ISA algorithm to local regions of joints to improve the training efficiency. Based on the depth video and the estimated 3D joint positions, we learn spatio-temporal features directly for each joint. The spatio-temporal features can be treated as the resulting descriptors of the local spatio-temporal interest points. These points are densely sampled from a local region around the joints. Each joint is associated with a histogram feature. We call this histogram feature *joint-based ISA feature* or JISA.

[1] Guang Chen, and Alois Knoll are with Technische Universität München, Garching bei München, Germany, email addresses: {guang, knoll}@in.tum.de.

[2] Manuel Giuliani, Daniel Clarke, and Andre Gaschler are with fortiss GmbH, An-Institut Technische Universität München, Guerickestr. 25, 80805 München, Germany, email addresses: {giuliani, clarke, gaschler}@fortiss.org.

978-1-4799-3685-4/14/$31.00 ©2014 IEEE

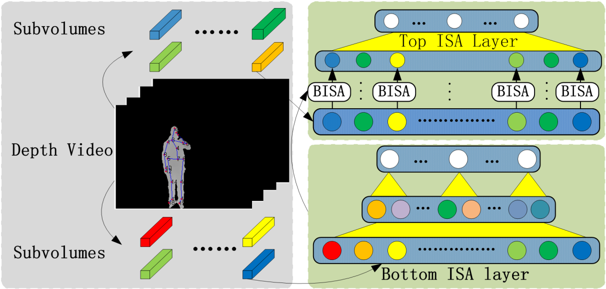


Fig. 2. An overview of our ISA model.

Second, we provide the ensemble weighted multi-instance learning approach. By training and combining multiple classifiers, ensemble methods [22] are state-of-the-art techniques with strong generalization abilities. Considering tracking errors of the skeleton data and to better characterize the intra-class variations, we propose *an ensemble weighted multi-instance learning approach* (EnwMi) for action recognition using depth video. Inspired by [11], this method firstly samples several subsets from a majority class independently, then trains multiple basic classifiers using the subsets and the minority class, and finally combines all classifiers for the final decision. It can deal with the class imbalance and the long training time of an SVM simultaneously. We formulate action recognition task with depth video as a multiple instance problem. We solve the multi-instance problem by a multiple kernel learning (MKL) approach. MKL is able to discover the discriminative JISA features. The basic idea for employing the MKL approach is that a certain action class is usually only associated with a subset of kinematic joints of the articulated human body.

The reminder of this paper is organized as follows: Section 2 reviews related work. Section 3 gives details of learning the 3.5D Graph Representation for depth video data. In Section 4, we present the ensemble weighted multi-instance learning approach. Section 5 provides the experimental results. Finally, Section 6 concludes the paper.

II. Related Work

Research in action recognition focused on analyzing spatio-temporal patterns in traditional 2D videos captured by a single camera. As RGBD sensors become available, action recognition researchers attempted to adopt techniques developed for color sequences to depth sequences. For instance, Li *et al.* [10] proposed a Bag of 3D points model by sampling points from the silhouette of the depth images. Lv and Nevatia [12] employed a hidden markov model (HMM) to represent the transition probability for pre-defined 3D joint positions. Similarly, Han *et al.* [4] used conditional random filed (CRF) to describe the 3D joint positions. However, adopting local interest points-based methods is difficult, because features such as STIP [7] and HOG [2] are not reliable in depth sequences. Until recently, a few spatial-temporal cuboid descriptors for depth videos were proposed. Cheng *et al.* [1] built a comparative coding descriptor to describe the depth cuboid by comparing the depth value of the center point with the nearby 26 points. Zhao *et al.* [21] built local depth patterns which describe the local region of interest points in depth map. Xia *et al.* [20] proposed the depth cuboid similarity feature as descriptor for the spatio-temporal depth cuboid. Oreifej *et al.* [14] presented a new descriptor HON4D using a histogram which captures the distribution of the surface normal orientation in the 4D space of time, depth, and spatial coordinates.

Besides these algorithms, there has been another category of methods for action recognition using depth images: algorithms based on high-level features. It is generally agreed that knowing the 3D joint position of human subject is helpful for action recognition. Wang *et al.* [19] combined joint location features and local occupancy features and employ a Fourier temporal pyramid to represent the temporal dynamics of the actions. Another method for modeling actions is dynamic temporal warping (DTW), Müller *et al.* [13] matched the 3D joint positions to the templates, and action recognition can be done through a nearest-neighbor classification method. However, the 3D joint positions that are generated via skeleton tracking from the depth map sequences are noisy. Moreover, with limited amount of training data, training a complex model is easy to overfit.

III. Learning 3.5D Graph Representations

In this section, we first briefly describe how to implement the ISA algorithm to depth video data. Next, we discuss details of the 3.5D graph representations of action images.

A. Independent Subspace Analysis

ISA is an unsupervised learning algorithm that learns features from unlabeled subvolumes (see Fig. 2). First, we extract random subvolumes from the local region of 20 joints of depth video data. We then normalize and whiten the set of subvolumes. We feed the pre-processed subvolumes to ISA networks as input units. An ISA network [6] is described as a two-layer neural network, with square and square-root nonlinearities in the first and second layers respectively.

We start with any input unit $x^t \in \mathbb{R}^n$ for each random sampled subvolume. We split each subvolume into a sequence of image patches and flatten them into a vector x^t with the dimension n. The activation of each second layer unit is

$$p_i(x^t; W, V) = \sqrt{\sum_{k=1}^{m} V_{ik} \left(\sum_{j=1}^{n} W_{kj} x_j^t\right)^2} \quad (1)$$

ISA learns parameters W through finding sparse feature representations in the second layer by solving

$$\begin{aligned} &\min_W \sum_{t=1}^{T} \sum_{i=1}^{m} p_i(x^t; W, V) \\ &s.t. WW^T = \mathbf{I} \end{aligned} \quad (2)$$

Here, $W \in \mathbb{R}^{k \times n}$ is the weight connecting the input units to the first layer units. $V \in \mathbb{R}^{m \times k}$ is the weight connecting the first layer units to the second layer units; n, k, m are the input dimension number of the first layer units and second layer units respectively. The orthonormal constraint ensures feature diversity.

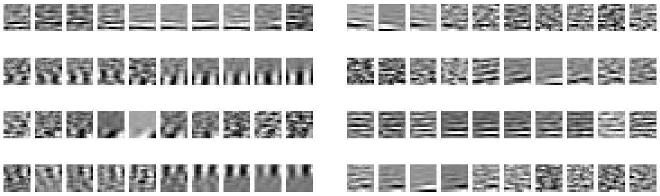


Fig. 3. Visualization of 10 ISA filters learned from the MSRAction3D dataset. These filters capture a moving edge in time.

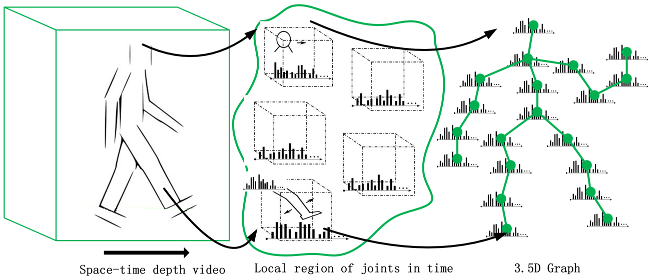


Fig. 4. Instead of treating an action class as a space-time pattern entire depth video (left), we propose to define an action as a collection of local regions of joints in time (middle). EnwMi is used to learn the 3.5D Graph of the depth video (right).

The model so far has been unsupervised. The bottom ISA model learns spatio-temporal features that detect a moving edge in time as shown in Fig. 3. It shows that the learned feature (each row in Fig. 3) is able to assign similar features in a group thereby achieving spatial invariance. The features have sharper edges like Gabor filters.As is common in neural networks, we stack another ISA layer with PCA on top of the bottom ISA. We use PCA to whiten the data and reduce the dimensions of the input unit. The model is trained greedily layerwise in the same manner as other algorithms described in [5], [9].

B. The 3.5D Graph Representation

We borrow the term, *3.5D graph*, from stereoscopic vision [15]. It refers to the outcome of reconstructing 4D information from spatio-temporal features and 3D joints positions. Fig. 4 shows a graphical illustration of our 3.5D representation of action videos. It combines the 3D configuration of human skeletons and 3D appearance features of each joint.

A 3.5D graph $\mathcal{G}^{\mathcal{X}}$ representing a depth video $\mathcal{X}$ consists of *V* nodes connected by *E* edges. The nodes correspond to a set of key points (joints) of the human body, as shown in Fig. 4. A node *v* is represented by the 3D position of this node p_v and the histogram features $f_v^{\mathcal{X}}$ extracted in a local image region surrounding this node in time. An edge *e* is a histogram feature $f_e^{\mathcal{X}} = [f_v^{\mathcal{X}}, f_{v'}^{\mathcal{X}}]$, where node *v* and node *v'* are connected by *e*.

C. Implementation Details

For a human subject in a depth video $\mathcal{X}$, the skeleton tracker tracks 20 joint positions[16], which correspond to 20 nodes of a 3.5D graph $\mathcal{G}^{\mathcal{X}}$. For each joint i at frame t, its local region S_t^i is of size (v_x, v_y) pixels. Let $\mathcal{T}$ denote the temporal dimension of the depth video $\mathcal{X}$. The depth video $\mathcal{X}$ is represented as the set of joint volumes $\{JV_1, JV_2...JV_{20}\}$. Each joint volume can be considered as a sequence of local regions $JV_i = \{S_1^i, S_2^i...S_t^i\}$. The size of JV_i is $v_x \times v_y \times \mathcal{T}$.

One of the disadvantages in training the ISA model is that it could be time-consuming when the dimension of the input data is large. In this paper, we apply the ISA algorithm to the local region of joints. As the local region of each joint is small compared to the whole image, we reduce the dimensionality and greatly improve efficiency. Additionally, it is possible to densely sample the local region of the joint to capture more discriminative information. Moreover, the features are discriminative enough to characterize variations in different joints. Based on the above ISA model, we compute the spatio-temporal features directly from JV_i for each joint (see Fig. 4). We treat the spatio-temporal features as the resulting descriptors of the local spatio-temporal interest points. Each interest point is represented by a subvolume, which is of size $s_x \times s_y \times s_t$. We densely sample the interest points from JV_i. We perform the vector quantization by clustering the spatio-temporal feature for each joint. Hence each 3D joint is associated with a histogram feature $JISA_i$, which corresponds to the feature $f_v^{\mathcal{X}}$ of a node *v* in $\mathcal{G}^{\mathcal{X}}$.

In order to capture the 3D position to fully model the joint, it is necessary to integrate the position information of joint i into the final feature $JISA_i$. For each joint i at frame t, we extract the pairwise relative position features P_i^t by taking the difference between the 3D position p_i of joints i and that of each other joint j: $P_i^t = \{p_i - p_j | i \neq j\}$.

Inspired by the Spatial Pyramid approach [8], we group the adjacent joints together as a *joint pair* to capture the spatial structure of the action. Therefore, for a human subject, we have 19 *joint pairs*. Each *joint pair* is represented as a histogram feature $JISAp_{ij} = [JISA_i, JISA_j]$, which corresponds to the feature $f_e^{\mathcal{X}}$ of en edge *e* in graph $\mathcal{G}^{\mathcal{X}}$.

IV. Ensemble Weighted Multi-instance Learning

To better characterize the intra-class variations and be robust to the errors of the skeleton tracker [16], we propose an ensemble weighted multi-instance learning algorithm (EnwMi) for action recognition using depth videos. We first describe the basic approach. Next, we give the details of the kernel design.

A. Basic Approach

The properties of training datasets such as size, distribution and number of attributes significantly contribute to the generalization error of a learning machine. In most action recognition tasks, there are serious class imbalances and not-well-distributed samples.In addition, different subjects perform actions with considerable variations. These problems are prone to lead to a partial over-fitting model.

To deal with these problems, under-sampling is an efficient method. It uses a subset of majority class samples to train a classifier. Although the training set becomes balanced and the training process becomes faster, standard under-sampling often suffers from the loss of helpful information concealed in the ignored majority class samples. Inspired by [11], our EnwMi method considers the distributions of different samples in the training dataset. Rather than randomly sampling subsets of the majority class samples, we try to select the samples which are hardest to be trained, and remove the samples which already have been learned well. Similar to other ensemble learning approaches, AdaBoost algorithm [3] is used in EnwMi to train a number of weighted component classifiers. For each iteration of the AdaBoost algorithm, a subset of top-weighted majority class samples are selected as negative samples. An ensemble of all component classifiers together creates the final classifier. A detailed presentation of the EnwMi method is given in Algorithm 1.

Algorithm 1 EnwMi

```
Input:
  For the training set of each action class, select all positive samples $\mathcal{P}$, and all negative samples $\mathcal{N}$, $|\mathcal{P}| < |\mathcal{N}|$, $y^i \in \{+1, -1\}$ are their class labels. Define T the number of iterations to train an AdaBoost ensemble $\mathcal{C}$.

  Weights initialization for each sample: $r_\tau^i = 1/(|\mathcal{P}|+|\mathcal{N}|)$, $i = 1, ..., |\mathcal{P}| + |\mathcal{N}|, \tau = 1$, $mode = top$
  while $\tau \leq T$ do
    Weights normalization: $\bar{r}_\tau^i = r_\tau^i / \sum_i r_\tau^j, \ \forall i$
    if $mode == top$ then
      Select top weighted samples: a subset $\mathcal{N}_\tau$ from $\mathcal{N}$
    end if
    Training an MKLSVM component classifier, $\mathcal{F}_\tau$ on $\mathcal{P}$ and $\mathcal{N}_\tau$
    Compute the performance of $\mathcal{F}_\tau$ over $\mathcal{P}$ and $\mathcal{N}$:
        $p_\tau = \sum_i r_\tau^i g_\tau^i (1 - abs(sgn(\mathcal{F}_\tau^i) - y^i))$    (3)
    where
        $g_\tau^i = ((1 - sgn(\mathcal{F}_\tau^i))/2 + \ pro(\mathcal{F}_\tau^i) sgn(\mathcal{F}_\tau^i))$
        pro() means the probability output of $\mathcal{F}_\tau^i$
    Choose $\alpha_\tau = -\frac{1}{2} log(\frac{1-p_\tau}{p_\tau})$
    if $\alpha_\tau > \theta$ then
      $mode = top$
      $\tau = \tau + 1$
      Update the weights:
  $r_\tau^{i+1} = \bar{r}_\tau^i e^{(-2|g_\tau^i|+\alpha_\tau)(1-abs(sgn(\mathcal{F}_\tau^i)-y^i))}$    $\forall i$    (4)
    else
      $mode = random$
      Select a random subset $\mathcal{N}_\tau$ from $\mathcal{N}$
      continue
    end if
  end while
Output:
        $\mathcal{C} = \frac{\sum_{\tau=1}^T \alpha_\tau pro(\mathcal{F}_\tau)}{\sum_{\tau=1}^T \alpha_\tau}$    (5)
```

B. Kernel Design of Component Classifiers

Our aim is to learn a component classifier where rather than using a pre-specified kernel, the kernel is learnt to be a linear combination of given base kernels. Suppose that the bags of the depth video $\mathcal{X}$ are represented as $f_\mathcal{X} = \{f_1, f_2, ..., f_{t-1}, f_t\}$, where t is the number of the features for each depth video. The classifier defines a function $\mathcal{F}(f^\mathcal{X})$ that is used to rank the depth video $\mathcal{X}$ by the likelihood of containing an action of interest.

The function $\mathcal{F}$ is learnt, along with the optimal combination of histogram features $f^\mathcal{X}$, by using the Multiple Kernel Learning techniques proposed in [17]. The function $\mathcal{F}(f^\mathcal{X})$ is the discriminant function of a Support Vector Machine, and is expressed as

$$\mathcal{F}(f^\mathcal{X}) = \sum_{i=1}^{M} y_i \alpha_i K(f^x, f^i) + b \tag{6}$$

Here, f^i, $i = 1, ..., M$ denotes the feature histograms of M training depth video data, selected as representative by the SVM, $y^i \in \{+1, -1\}$ are their class labels, and K is a positive definite kernel, obtained as a linear combination of base kernels

$$K(f^\mathcal{X}, f^i) = \sum_j w_j K(f_j^\mathcal{X}, f_j^i) \tag{7}$$

MKL learns both the coefficient α_i and the kernel combination weight w_j. For a multi class problem, a different set of weights $\{w_j\}$ are learnt for each class. We choose one-against-rest to decompose a multi-class problem.

Because of linearity, Eq .6 can be rewrittten as

$$\mathcal{F}(f^\mathcal{X}) = \sum_j w_j \mathcal{F}(f_j^\mathcal{X}) \tag{8}$$

where

$$\mathcal{F}(f_j^\mathcal{X}) = \sum_{i=1}^{M} y_i \alpha_i K(f_j^x, f_j^i) + b \tag{9}$$

With each kernel corresponding to each feature, there are 20 weights w_j to be learned for the linear combination for IJSA features, and 19 weights w_j to be learned for JISAp features. Weights can therefore highlight more discriminative joints for an action and we can even ignore joints that are not discriminative by setting w_j to zero.

V. Experiments

To evaluate our method, we conducted experiments on the MSRAction3D dataset [10]. We compared our algorithm with state-of-the-art methods on action recognition using depth videos. Experimental results show that our algorithm gives significantly better recognition accuracy than algorithms based on low-level hand-designed features and high-level joint-based features. In addition, we investigate the discriminative joints for each action class.

TABLE I

THE THREE ACTION SUBSETS USED IN OUR EXPERIMENTS

Cross Subset 1(CS1)	Cross Subset 2(CS2)	Cross Subset 3(CS3)
Tennis Serve(TSr)	High Wave(HiW)	High Throw(HT)
Horizontal Wave(HoW)	Hand Catch(HC)	Forward Kick(FK)
Forward Punch(FP)	Draw X(DX)	Side Kick(SK)
High Throw(HT)	Draw Tick(DT)	Jogging(JG)
Hand Cap(HCp)	Draw Circle(DC)	Tennis Swing(TSw)
Bend(BD)	Hands Wave(HW)	Tennis Serve(TSr)
Hammer(HM)	Forward Kick(FK)	Golf Swing(GS)
Pickup Throw(PT)	Side Boxing(SB)	Pickup Throw(PT)

TABLE II

COMPARISON OF RESULTS ON MSRACTION3D DATASET

Method	Accuracy
Action Graph On Bag of 3D Points [10]	0.747
Random Occupancy Pattern [18]	0.865
Mining Actionlet Ensemble [19]	0.882
Histogram of Oriented 4D Normals [14]	0.889
Spatio-Temporal Depth Cuboid Similarity Feature [20]	0.893
EnwMi-s + JISA features	0.895
EnwMi-s + JISAp features	0.912
EnwMi + JISA features	0.903
EnwMi + JISAp features	**0.920**

A. Experimental Setup

The MSRAction3D dataset [10] is a public dataset that provides sequences of depth maps and skeletons captured by a depth camera. In order to facilitate a fair comparison, we follow the same experimental settings as [10], [14], [20] to split 20 actions into three subsets as listed in Table I, each having 8 action classes. In each subset, half ot the subjects are used for training and the other half for testing.

B. Model Details

We train the ISA model on the MSRAction3D training sets. The input units to the bottom layer of ISA model are of size $12 \times 12 \times 10$, which are the dimensions of the spatial and temporal size of the subvolumes. The subvolumes to the top layer of the ISA model are the same size with the bottom layer.

We perform vector quantizatoin by K-means on the learned spatio-temporal features for each joint. The densely sampling step of the local regions of each joint is 2 pixels. The codebook size k is 700. The model parameters for different joints are the same. Therefore, each depth video is represented by 20 JISA features or 19 JISAp features. We choose χ^2 as the histogram kernel for multi class SVM classifier. For EnwMi, we set the number of subesets $|\mathcal{N}_\tau| = 3|\mathcal{P}|$, and the rounds of the AdaBoost $T = 20$. The threshold for a good component classifier is set to 1.45. All the parameters across three subsets are the same. Note that when we set the number of the samples in subsets $|\mathcal{N}_\tau| = |\mathcal{N}|$, and the rounds of the AdaBoost $T = 1$, EnwMi is cast into an muti-instance problem. We call this special case EnwMi-s.

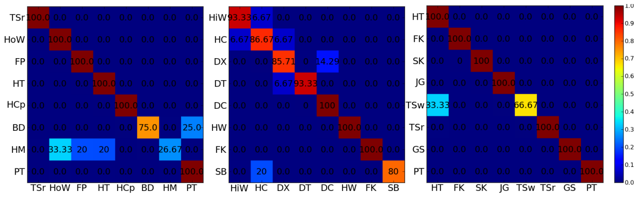
C. Experimental Results

A comparison of our method against best published results for the MSRAction3D dataset is reported in Table II. As can be seen from the table, our approach outperforms a wide range of methods. There is an increase in performance between our method (92.0%) and the closet competitive method (89.3%). This is a very good performance considering that the skeleton tracker sometimes fails and the tracked joint positions are quite noisy.

TABLE III

THE PERFORMANCE OF OUR METHOD ON THREE TEST SETS. CS1, CS2 CS3 ARE THE ABBREVIATIONS OF CROSS SUBSET 1, CROSS SUBSET 2, CROSS SUBSET3 (SEE TABLE I).

Method	CS1	CS2	CS3
EnwMi-s + JISA features	0.870	0.873	0.942
EnwMi-s + JISAp features	0.860	**0.932**	0.942
EnwMi + JISA features	0.860	0.882	**0.967**
EnwMi + JISAp features	**0.877**	0.924	0.958



(a) CS1 (b) CS2 (c) CS3

Fig. 5. The confusion matrices for our method *EnwMi + JISAp features* on three subsets of the MSRAction3D dataset. Rows represent the actual classes, and columns represent predicted classes. All abbreviations of action classes are written out in Table I. (best viewed in color).

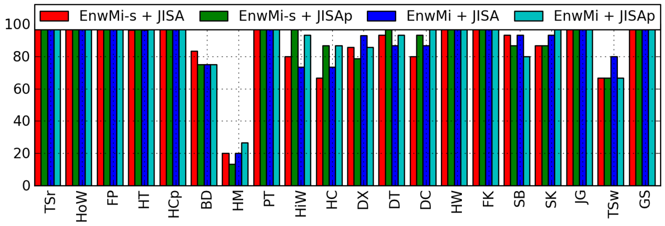


Fig. 6. The accuracies of 20 action classes of MSRAction3D dataset. We compared EnwMi with EnwMi-s using JISA and JISAp features. All abbreviations of action classes are written out in Table I. (best viewed in color).

Compared to EnwMi-s, the improvement of EnwMi is about 1%, which shows that the ensemble learning approach 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. Additionally, it is interesting to note that in our method the obtained accuracies using JISAp features is 92.0% (EnwMi) and 91.2% (EnwMi-s), which are better than using JISA feature 90.3% (EnwMi) and 89.5% (EnwMi-s). This proves the advantage of the spatial pyramid approach, though we just group the adjacent joints together as a *joint pair* to capture the spatial structure of the skeleton.

The confusion tables for three test sets, *Cross Subset 1 (CS1), Cross Subset 2 (CS2), Cross Subset 3 (CS3)*, are illustrated in Fig. 5. We report the average accuracy of three test sets in Table III, and the average accuracy of each action class in Fig. 6. While the performance in CS2 and CS3 is promising, the accuracy in CS1 is relatively low. This is probably because actions in CS1 are done with similar movements. Although our method obtains an accuracy of 100% in 12 out of 20 actions, the accuracy of the *Hammer* in CS1 is only 26.67%. This is probably due to the significant variations of the action *Hammer* performed by different subjects. The performance can be improved by adding more subjects.

D. Mining discriminative joints

It is generally agreed that although the human body has a large number of kinematic joints, a certain action usually only associates with a subset of them. Additionally, feature extraction in action recognition is usually computationally expensive. This encourages us to investigate the discriminative joints for different action classes. In EnwMi-s, each action is represented as a linear combination of joint-based features (JISA features or JISAp features). We learned their weight via a multiple kernel learning method to discover the discriminative joints.

Fig. 1 illustrates the skeleton with the joints weight discovered by our method. The *joint pair*s with the weight >0 are marked as thick and red lines. EnwMi-s is able to discover the discriminative joints and better characterize the intra-class variations. Fig. 1c shows that *Jogging* is represented by the combination of joints *left shoulder, center shoulder, right elbow, spine, center hip and right hip*. Normally, *Jogging* is related to the foot joints like *right/left foot, and right/left ankle*. However, for the MSRAction3D dataset, the tracking positions of the joints, *right/left foot, and right/left ankle*, are full of noise. Therefore, these joints are not discriminative for action class *Jogging*, which is consistent with Fig. 1c. This shows that our method is robust to the tracking errors of the skeleton data.

VI. Conclusion

We presented a novel, simple and easily implementable *ensemble weighted multi-instance learning approach* (EnwMi) method for action recognition from depth video data. We learn the spatio-temporal features using independent subspace analysis in an unsupervised way. This architecture could leverage the plethora of the unlabeled data and adapt easily to new sensors. Furthermore, the ensemble weighted multi-instance learning approach is able to deal with the tracking errors of the skeleton data and better characterize the intra-class variations. Experimental results show that our method outperforms all previous approaches on the MSRAction3D dataset. It also suggests that learning spatio-temporal features directly from depth video data is an important research direction, and the ensemble learning approach can further improve the performance of these features.

References

[1] Zhongwei Cheng, Lei Qin, Yituo Ye, Qingming Huang, and Qi Tian. Human daily action analysis with multi-view and color-depth data. In *Proceedings of the 12th international conference on Computer Vision - Volume 2*, ECCV'12, pages 52–61, 2012.

[2] Navneet Dalal and Bill Triggs. Histograms of oriented gradients for human detection. In *In CVPR*, pages 886–893, 2005.

[3] Yoav Freund and Robert E. Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *J. Comput. Syst. Sci.*, 55(1):119–139, August 1997.

[4] Lei Han, Xinxiao Wu, Wei Liang, Guangming Hou, and Yunde Jia. Discriminative human action recognition in the learned hierarchical manifold space. *Image Vision Comput.*, 28(5):836–849, May 2010.

[5] Geoffrey E. Hinton, Simon Osindero, and Yee Whye Teh. A fast learning algorithm for deep belief nets. *Neural Computation*, 18(7):1527–1554, 2006.

[6] Aapo Hyvrinen, Jarmo Hurri, and Patrick O. Hoyer. *Natural Image Statistics: A Probabilistic Approach to Early Computational Vision*. Springer Publishing Company, Incorporated, 1st edition, 2009.

[7] Ivan Laptev. On space-time interest points. *Int. J. Comput. Vision*, 64(2-3):107–123, September 2005.

[8] S. Lazebnik, C. Schmid, and J. Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *Computer Vision and Pattern Recognition (CVPR), 2006 IEEE Conference on*, volume 2, pages 2169–2178, 2006.

[9] Q.V. Le, W.Y. Zou, S.Y. Yeung, and A.Y. Ng. Learning hierarchical invariant spatio-temporal features for action recognition with independent subspace analysis. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pages 3361–3368, 2011.

[10] Wanqing Li, Zhengyou Zhang, and Zicheng Liu. Action recognition based on a bag of 3d points. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2010.

[11] Xu-Ying Liu, Jianxin Wu, and Zhi-Hua Zhou. Exploratory undersampling for class-imbalance learning. *Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on*, 39(2):539–550, 2009.

[12] Fengjun Lv and Ramakant Nevatia. Recognition and segmentation of 3-d human action using hmm and multi-class adaboost. In Ale Leonardis, Horst Bischof, and Axel Pinz, editors, *Computer Vision ECCV 2006*, volume 3954 of *Lecture Notes in Computer Science*, pages 359–372. 2006.

[13] Meinard Müller and Tido Röder. Motion templates for automatic classification and retrieval of motion capture data. In *Proceedings of the 2006 ACM SIGGRAPH/Eurographics symposium on Computer animation*, SCA '06, pages 137–146, 2006.

[14] Omar Oreifej and Zicheng Liu. Hon4d: Histogram of oriented 4d normals for activity recognition from depth sequences. In *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*, 2013.

[15] J.C.A. Read, G.P. Phillipson, I. Serrano-Pedraza, A.D. Milner, and A.J. Parker. Stereoscopic vision in the absence of the lateral occipital cortex. *PLoS One*, 5(9):e12608, 2010.

[16] 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 *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pages 1297–1304, 2011.

[17] S. V. N. Vishwanathan, Z. Sun, N. Theera-Ampornpunt, and M. Varma. Multiple kernel learning and the SMO algorithm. In *Advances in Neural Information Processing Systems*, December 2010.

[18] Jiang Wang, Zicheng Liu, Jan Chorowski, Zhuoyuan Chen, and Ying Wu. Robust 3d action recognition with random occupancy patterns. In *Proceedings of the 12th European conference on Computer Vision - Volume Part II*, pages 872–885, 2012.

[19] Jiang Wang, Zicheng Liu, Ying Wu, and Junsong Yuan. Mining actionlet ensemble for action recognition with depth cameras. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 1290–1297, 2012.

[20] L. Xia and J.K. Aggarwal. Spatio-temporal depth cuboid similarity feature for activity recognition using depth camera. In *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*, 2013.

[21] Yang Zhao, Zicheng Liu, Lu Yang, and Hong Cheng. Combing rgb and depth map features for human activity recognition. In *Signal Information Processing Association Annual Summit and Conference (APSIPA ASC)*, pages 1–4, 2012.

[22] Zhi-Hua Zhou. Ensemble learning. In Stan Z. Li and Anil K. Jain, editors, *Encyclopedia of Biometrics*, pages 270–273. 2009.