Learning Weighted Joint-based Features for Action Recognition using Depth Camera

Guang Chen[1,2], Daniel Clarke[2] and Alois Knoll[1]
[1]*Robotics and Embedded Systems, Fakultät für Informatik, Technische Universität München, München, Germany*
[2]*fortiss GmbH, An-Institut Technische Universität München, München, Germany*

Keywords: Unsupervised Learning, Weighted Joint-based Features, Action Recognition, Depth Video Data.

Abstract: Human action recognition based on joints is a challenging task. The 3D positions of the tracked joints are very noisy if occlusions occur, which increases the intra-class variations in the actions. In this paper, we propose a novel approach to recognize human actions with weighted joint-based features. Previous work has focused on hand-tuned joint-based features, which are difficult and time-consuming to be extended to other modalities. In contrast, we compute the joint-based features using an unsupervised learning approach. To capture the intra-class variance, a multiple kernel learning approach is employed to learn the skeleton structure that combine these joints-base features. We test our algorithm on action application using Microsoft Research Action3D (MSRAction3D) dataset. Experimental evaluation shows that the proposed approach outperforms state-of-the-art action recognition algorithms on depth videos.

1 INTRODUCTION

With the recent advent of the low-cost sensors such as Kinect, depth cameras have received a great deal of attention from researchers. It triggered significant attention to revisit problems such as object detection and action recognition using depth videos.

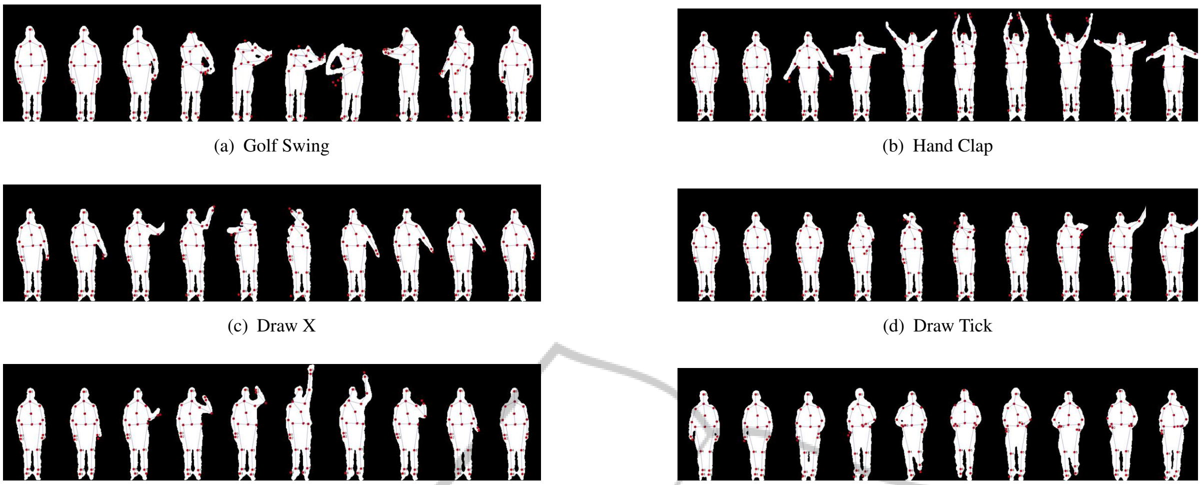
Compared to the visible light camera, depth sensor has several advantages. For example, depth image provides 3D structural information of the scene, which can often be more discriminative than color and texture in many applications including detection, segmentation and action recognition. Moreover, the depth camera can wok in different lighting conditions. These advantages have facilitated a rather powerful human motion capturing technique (Shotton et al., 2011) that generate the 3D joint positions of the human skeleton. However, simply using such 3D depth data and the estimated 3D joint positions for human action recognition is not plausible. One reason is that the estimation of the 3D joints positions may not reliable when the human body is partly in view such as a hand touching another body part, two hands crossing, bending the back, etc.

In action recognition, which is the topic of this paper, two significant aspects arise when using the depth sequences. First, the performance of adopting conventional color-based methods is unknown. The depth images are often contaminated with undefined depth points, which appear in the sequences as spatially and temporally discontinues black holes. Furthermore, there is no texture in the depth data. Hence, the discrimination of the depth video data is considered doubtful. These hinder the extending of the hand-designed features from color-based data to depth data, such as STIP (Laptev, 2005), HOF (Laptev et al., 2008) and HOG (Dalal and Triggs, 2005).

Second, with the release of depth sensors and the associated SDK, we are able to obtain 3D positions of the joints in real time. This arises one question: will the noisy human skeleton data perform well in action recognition? Skeleton data are able to provide additional body part information to differentiate actions. For instance, it was recently shown in (Yang and Tian, 2012) that for the action recognition, they proposed a type of features by adopting the differences of joints in both temporal and spatial domains to explicitly model the dynamics of individual joints and the configuration of different joints.

Our work in this paper proceeds along this direction. We propose a novel human action recognition approach using a depth camera. Fig. 1 demonstrates the depth images with 20 extracted body joints of each depth map for actions *Golf Swing, Hand Clap, Draw X, Draw Tick, High Throw and Jogging*. The basic

Chen G., Clarke D. and Knoll A..
Learning Weighted Joint-based Features for Action Recognition using Depth Camera.
DOI: 10.5220/0004735705490556
In *Proceedings of the 9th International Conference on Computer Vision Theory and Applications* (VISAPP-2014), pages 549-556
ISBN: 978-989-758-004-8



(a) Golf Swing

(b) Hand Clap

(c) Draw X

(d) Draw Tick

(e) High Throw

(f) Jogging

Figure 1: The sequences of depth maps and skeleton for different action classes. Each depth image includes 20 joints.

idea is illustrated in Fig. 2. We provide an unsupervised learning method to learn the joint-based features inspired by (Hyvrinen et al., 2009; Le et al., 2011). At the heart of our method is the use of the Independent Subspace Analysis (ISA). ISA algorithm is a well-known algorithm in the field of natural image statics (Hyvrinen et al., 2009). Experimental studies have shown that this algorithm can learn powerful features from static image or color-based sequences. 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 when the dimension of the input data is large e.g video data. In this paper, we extend the ISA algorithm to the depth video data. Instead of training the model with the full video, we apply the ISA algorithm to the local region of joints to improve the training efficiency. Based on the depth video and the estimated 3D joint positions, we learn the spatio-temporal features directly for each action. The spatio-temporal features can be treated as the the resulting descriptors of the local spatio-temporal interest points. These interest points are dense sampled from a local region around the joint. We perform the vector quantization by clustering the spatio-temporal feature for each joint. Each 3D joint is associated with a histogram feature. We call this histogram feature *joint-based ISA feature* or *JISA* feature.

More importantly, to deal with the tracking errors of the skeleton data and better characterize the intra-class variations, a multiple kernel learning approach is employed to learn the skeleton structure that combines these discriminative *JISA* features. The articulated human body has a large number of kinematic joints, but a certain action is usually only associated with a subset of them. For example, the joints “head” and “right wrist” are discriminative for action “drinking”. In our paper, each action is represented as a linear combination of joints, and their discriminative weights are learnt via a multiple kernel learning method. This weighted joint-base model is more robust to the errors in the features, and it can better characterize the intra-class variations in the actions.

Our main contributions include the following three aspects. First, we proposes a novel joint-based features for action recognition under an unsupervised learning paradigm. The features are discriminative enough to characterize variations in different joints. Second, we demonstrate how to deal with the noisy skeleton data using the multi kernel learning approach. The weighted joint-base features are capable of characterizing both the human action and their internal variation. Third, We explore how many joints are sufficient for action recognition in our framework We observe that a small subset of joints is sufficient to perform action recognition. Our experimental results show that, despite the simplicity of our proposed method, this method is able to achieve better recognition accuracy than the state-of-the-art methods.

The reminder of this paper is organized as follows. Section 2 reviews the related work. Section 3 gives details of learning joint-based features. In Section 4, the skeleton structure is learnt using multiple kernel learning approach. The experimental results are presented in Section 5. Finally, Section 6 concludes the paper.

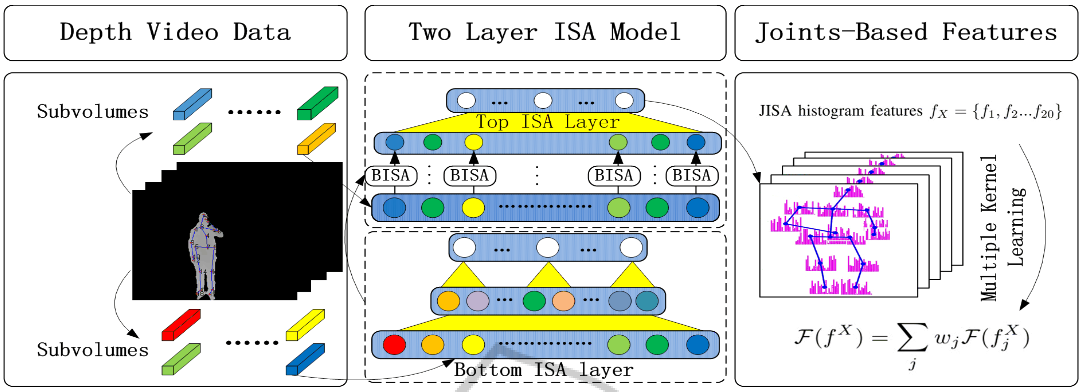


Figure 2: An overview of our method: We randomly sample the subvolumes from local regions of each joint of the depth video data. The subvolumes are given as input to the bottom ISA network. The learned bottom ISA model are copied to the top ISA network. The stacked ISA model learn spatio-temporal features. A multiple kernel learning approach is employed to learn a linear combination of joint-based features and classify the actions. (best viewed in color)

2 RELATED WORK

In traditional 2D videos captured by a single camera, action recognition mainly focused on analyzing spatio-temporal patterns. As RGBD sensors become available, research of action recognition based on the depth sequences attempted to adopt techniques originally developed for color sequences. For instance, Lv and Nevatia in (Lv and Nevatia, 2006) employ a hidden markov model (HMM) to represent the transition probability for pre-define 3D joint positions. (Li et al., 2010) proposed a Bag of 3D points model by sampling points from the silhouette of the depth images. Consequently, a GMM is used to globally model the postures, and an action graph (Li et al., 2008) is used for inference. Similarly, in (Han et al., 2010), the 3D joint position is described using the conditional random filed (CRF).

Local interest point is the most popular method for classification and recognition task in computer vision. Interest point provides a compact representation of image content by describing local parts of the scene thus offer robustness to occlusions, clutter, and intra-class variations. However, as discussed earlier, features such as STIP (Laptev, 2005) and HOG (Dalal and Triggs, 2005) are not reliable in depth sequences, adopting local interest points based methods operate in depth sequences is difficult.

Until recently, a few spatial-temporal cuboid descriptors for depth videos were proposed. (Zhao et al., 2012) build local depth pattern by computing the difference of the average depth values between the cells. (Cheng et al., 2012) build a comparative coding descriptor to describe the depth cuboid by comparing the depth value of the center point with the nearby 26 points. (Xia and Aggarwal, 2013) propose depth cuboid similarity feature as the descriptor for the spatio-temporal depth cuboid that describes the local “appearance” in the depth video. (Oreifej and Liu, 2013) present a new descriptor HON4D using a histogram capturing the distribution of the surface normal orientation in the 4D space of time, depth, and spatial coordinates. (Zhang and Parker, 2011) extract STIPS by calculating a response function fromt both the depth and RGB channels and use the gradients along *x, y, t* directions as the descriptor. Notice that some existing methods still depend on the detectors and descriptors designed for RGB images.

In the literature, there has been another category for action recognition using depth images: algorithms based on high-level features. It is generally agreed that knowing the 3D joint position is helpful for action recognition. (Wang et al., 2012b) combine joint location features and local occupancy features and employ a Fourier temporal pyramid to represent the temporal dynamics of the actions. (Xia et al., 2012) take the skeletal joint locations and vote them into 3D spatial bins and build posture words for action recognition. Another method for modeling actions is dynamic temporal warping (DTW), which matches the 3D joint positions to a template (Müller and Röder, 2006), 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. However, the 3D joint position 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.

Different from these approaches, we propose to recognize human actions with weighted joints-based features. The above approach focus on hand-tuned features, which are difficult and time-consuming to be extended to depth images. In contrast, we compute the joint-based features using an unsupervised learning approach. It is easy to extend our approach to other modalities. To deal with the tracking errors of the skeleton, a multiple kernel learning approach is employed to learn the skeleton structure that combine these joint-based features.

3 LEARNING JOINTS-BASED FEATURES

In this section, we first briefly describe the ISA algorithm. Next we give details of how deep learning techniques can be used to obtain the joint-based features.

3.1 Independent Subspace Analysis

ISA is an unsupervised learning algorithm that learns features from unlabeled subvolumes. First, random subvolumes are extracted from the local region of 20 joints. The set of subvolumes is then normalized and whitened. The pre-processed subvolumes are feed to ISA networks as input units. An ISA network (Hyvrinen et al., 2009) 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

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

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

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

Here, $W \in \mathbb{R}^{k\times n}$ is the weights connecting the input units to the first layer units. $V \in \mathbb{R}^{m\times k}$ is the weights 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 is to ensure the features are diverse.

The model so far has been unsupervised. The bottom ISA model learns spatio-temporal features that detect a moving edge in time. As is common in neural network, we stack another ISA layer with PCA on top of 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 (Bengio et al., 2007; Hinton et al., 2006; Le et al., 2011).

3.2 Learning Joint-based Features

One of the disadvantages in training the ISA model is that it could be slow 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 comparing to the whole image, we could train the model efficiently. Additionally, it is possible to dense sample the local region of the joint to capture more discriminative information. Moreover, the features are discriminative enough to characterize variations in different joints.

For a human subject, 20 joint positions are tracked by the skeleton tracker (Shotton et al., 2011). For each joint i at frame t, its local region LR_t^i is of size (V_x, V_y) pixels. Let T denote the temporal dimension of the depth video. Given a sequence of depth images $\{I_1, I_2...I_T\}$ containing a human subject performing an activity, the depth video is represented as the set of joints volumes $\{JV_1, JV_2...JV_{20}\}$. Each joints volume can be considered as a sequence of local region $JV_i = \{LR_1^i, LR_2^i...LR_T^i\}$. The size of JV_i is $V_x \times V_y \times T$. Based on the above ISA model, we compute the spatio-temporal features directly from JV_i for each joint. The spatio-temporal features can be treated as the the resulting descriptors of the local spatio-temporal interest points. Each interest point is represented by a subvolume, which is of size $Vs_x \times Vs_y \times Vs_t$. The interest points are dense sampled from JV_i. We perform the vector quantization by clustering the spatio-temporal feature for each joint. Each 3D joint is associated with a histogram feature $JISA_i$.

In order to capture the 3D position to fully model an 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 (Lazebnik et al., 2006), we group the adjacent joints together as a joints pair to capture the spatial structure of the action. Therefore, for a human subject, we have 19 joints pairs. Each joints pair is represented as a histogram feature $JISAp_{ij} = [JISA_i, JISA_j]$.

Table 1: The three action subsets used in our experiments.

Cross Subject Subset 1 (CS1)	Cross Subject Subset2 (CS2)	Cross Subject Subset 3 (CS3)
Horizontal Wave (HoW)	High Wave (HiW)	High Throw (HT)
Hammer (HM)	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)
Tennis Serve (TSr)	Forward Kick (FK)	Golf Swing (GS)
Pickup Throw (PT)	Side Boxing (SB)	Pickup Throw (PT)

4 LEARNING SKELETON STRUCTURE

Although the proposed feature is robust to noise, to deal with the tracking errors of the skeleton data and better characterize the intra-class variations, a multiple kernel learning approach is employed to learn the skeleton structures that combines these discriminative joint-based features.

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

The function $\mathcal{F}$ is learnt, along with the optimal combination of histogram features f^X, by using the Multiple Kernel Learning techniques proposed in (Vishwanathan et al., 2010). The function $\mathcal{F}(f^X)$ is the discriminant function of a Support Vector Machine, and is expressed as

$$\mathcal{F}(f^X) = \sum_{i=1}^{M} y_i \alpha_i K(f^X, f^i) + b \tag{3}$$

Here, f^i, $i = 1, ..., M$ denote 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

$$K(f^X, f^i) = \sum_j w_j K(f_j^X, f_j^i) \tag{4}$$

MKL learns both the coefficient α_i and the kernels 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 .3 can be rewrittten as

$$\mathcal{F}(f^X) = \sum_j w_j \mathcal{F}(f_j^X) \tag{5}$$

where

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

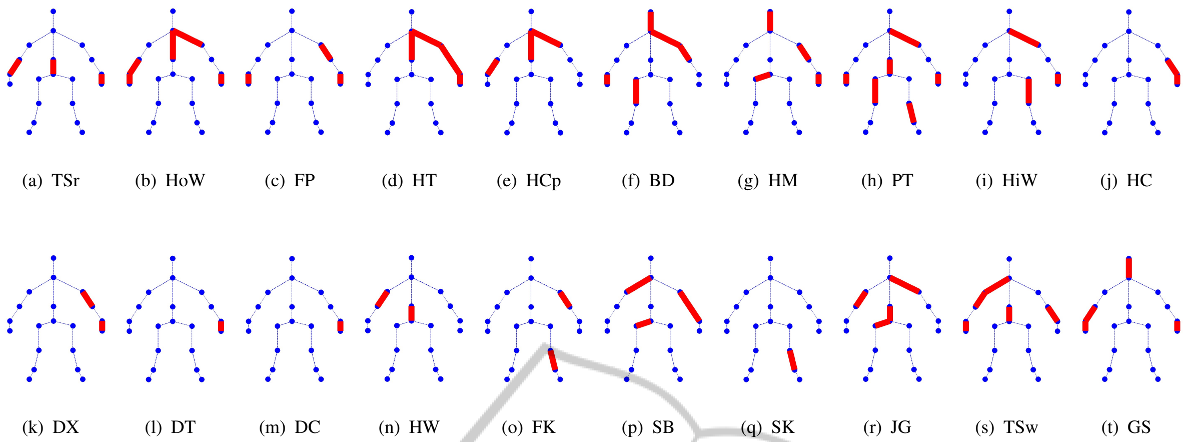
With each kernel corresponding to each feature, there are 20 weights w_j to be learned for the linear combination for IJSA features, and 19 weights w_j to be learned for *JISA*p features. Weights can therefore emphasis more discriminative joints for an action and we can even ignore joints that are not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

5 EXPERIMENTS

We evaluate our proposed method on the MSRAction3D dataset (Li et al., 2010). In this section, we first describe the dataset and experimental setup. Next, we give the details of the model. Then, we show the leaned structure of skeleton for each action. We study how many joints in a depth video are sufficient to perform action detection and recognition. Finally we present the performance of our approach. We compare our algorithm with state-of-the-art methods on action recognitions from depth videos. Experimental results show that our algorithm gives significantly better recognition accuracy than algorithm based on low-level hand-designed features and high-level joint-based features.

5.1 Dataset and Experimental Setup

The MSRAction3D dataset (Li et al., 2010) is a public dataset that provides sequences of depth maps and skeletons captured by a RGBD camera. It includes 20 actions performed by 10 subjects facing the camera during performance. Each subject performed each action 2 or 3 times. As shown in Fig 1, actions in this dataset reasonably capture a variety of motions related to arms, legs, torso, and their combinations.



(a) TSr (b) HoW (c) FP (d) HT (e) HCp (f) BD (g) HM (h) PT (i) HiW (j) HC

(k) DX (l) DT (m) DC (n) HW (o) FK (p) SB (q) SK (r) JG (s) TSw (t) GS

Figure 3: The skeleton structure for 20 action classes in MSRAction3D dataset. Our method can learn the discriminative joint pairs for each action class. The weight w_j is used to describe the discrimination of of the *joint pairs*. The *joint pairs* with the w_j>0 are marked as thick and red lines. All abbreviations of action classes are written out in Table 1. (best viewed in color).

In order to facilitate a fair comparison, we follow the same experimental settings as (Li et al., 2010; Oreifej and Liu, 2013; Xia and Aggarwal, 2013) to split 20 actions into three subsets as listed in Table 1. In each subset, 1/2 subjects are used as training and the rest ones used as testing.

5.2 Model Details

We focus on the problem of human action recognition from depth video data. We train ISA model on the MSR Action3D training sets. The input units to the bottom layer of ISA model are of size $12 \times 12 \times 10$, 12, 12, 10 means 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. The model parameters for different joints are the same. We performs vector quantizatoin by K-means on the learned spatio-temporal features for each joint. The codebook size k is 700. We choose χ^2 as the histogram kernel for multi class SVM classifier. Finally, each depth video is represented by 20 *JISA* features or 19 *JISA*p features.

5.3 Skeleton Structure

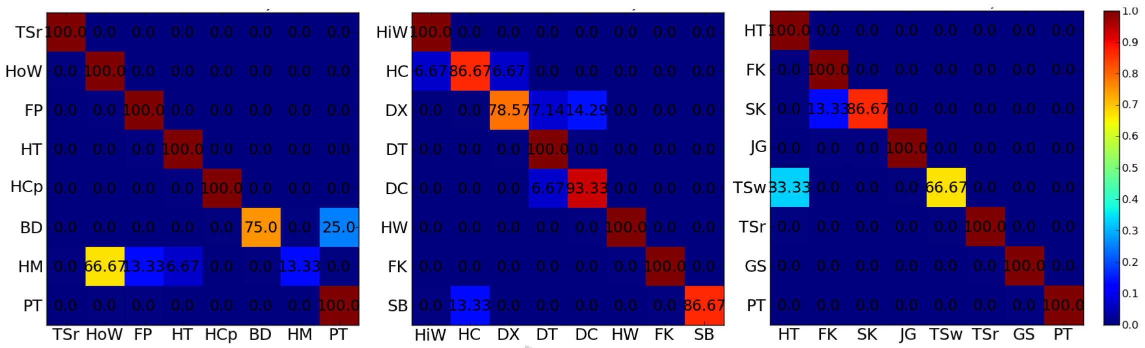
Our method is able to deal with the tracking errors of the skeleton data and better characterize the intra-class variations. We start from the intuition that, although the human body has a large number of kinematic joints, a certain action usually only associated with a subset of them.

In our experiments, each action is represented as a linear combination of joints-based features. We learned their weight via a multiple kernel learning method. Fig 3 illustrates the skeleton with the weight of *joint pairs* discovered by our method. The *joint pairs* with the weight w_j>0 are marked as thick and red lines. The average number of *joint pairs* for 20 action class in the MSRAction3D dataset is 4. 7 of 20 action classes have less than 2 discriminative *joint pairs*. 3 of 20 action classes have only 1 discriminative *joint pair*. The maximum number of discriminative *joint pairs* is 6 for action class *Pickup and Throw*. As the action class *Pickup and Throw* consist of two sub-action, it is more complex than other action classes in MSRAction3D dataset.

In addition, we can observe that our method can learn the structure of the skeleton very well. Fig. 3(t) shows that *Golf Swing* is represented by the combination of joints *left hand, left wrist, right hand, right wrist and right elbow*. For action classes *Draw X (see Fig. 3(k)), Draw Tick (see Fig. 3(l)) and Draw Circle (see Fig. 3(m))*, they have one common discriminative *joint pair*: *left wrist and left hand*. Fig. 3(j) shows that *Hand Catch* is represented by the combination of joints *left elbow, left wrist and left hand*. One interesting observation in Fig. 3 is: none of the action classes in MSRAction3D dataset has discriminative foot joints like *left foot or right foot*. Normally, action classes like *Jogging, Forward Kick and Side Kick* are related to foot joints. However, for the MSRAction3D dataset, the tracking positions of the foot joints are full of noise, which shows that our method is robust to the tracking errors of the joint 3D positions.

5.4 Experimental Results

We compare our method with the state-of-the-art methods on the MSRAction3D dataset. We report



(a) Cross Subject Subset 1 (CS1)

(b) Cross Subject Subset 2 (CS2)

(c) Cross Subject Subset 3 (CS3)

Figure 4: The confusion matrices for our method with *JISA*p 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 3. (best viewed in color).

Table 2: Comparison of recognition accuracy on MSRAction3D dataset.

Method	Accuracy
EigenJoints (Yang and Tian, 2012)	0.823
Random Occupancy Pattern (Wang et al., 2012a)	0.865
Mining Actionlet Ensemble (Wang et al., 2012b)	0.882
Histogram of Oriented 4D Normals (Oreifej and Liu, 2013)	0.889
Spatio-Temporal Depth Cuboid Similarity Feature (Xia and Aggarwal, 2013)	0.893
Our Method with *JISA* features	0.895
Our Method with *JISA*p features	0.912

Table 3: The performance of our method on three test sets.

Method	CS1	CS2	CS3
Our Method with *JISA*	0.870	0.873	0.942
Our Method with *JISA*p	0.860	0.932	0.942

the performance of our method in Table 2. Our method achieves superior performance compared to state-of-the-art results in the literature. There is an increase in performance between our method (91.2%) 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. Additionally, it is interesting to note that in our method the obtained accuracy using *JISA*p features is 91.2%, which is better than using *JISA* feature. This prove the advantage of 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 are illustrated in Fig. 4. We report the average accuracy of three test sets in Table 3. While the performance in CS2 and CS3 is promising, the accuracy in CS1 is relatively low. This is probably because action in CS1 are with similar movements. For example, in CS1 *Hammer* tends to be confused with *Forward Punch* and *Horizontal Wave*, and *Pickup Throw* consists of *Bend* and *High Throw*. Although our method obtains an accuracy of 100% in 16 out of 24 actions, the accuracy of the *Hammer* in CS1 is only 13.33%. This is probably due to the small number of subjects and also the significant variations of the *Hammer* action performed by different subjects. The performance can be improved by adding more subjects.

6 CONCLUSIONS

We presented a novel, simple and easily implementable approach for action recognition from depth images. The stacked ISA network learns the spatio-temporal features in an unsupervised way. This architecture could leverage the plethora of the unlabeled data and adapt easily to new sensors. The multiple kernel learning learns the optimal combination of joint-based features, which can deal with the tracking errors of the skeleton data and better characterize the intra-class variations. The experiments show that the proposed method outperforms a range of previous approaches on MSRAction3D dataset.

ACKNOWLEDGEMENTS

This work was carried out in the context of the multi-sensor data fusion group at the research institute fortiss.

REFERENCES

Bengio, Y., Lamblin, P., Popovici, D., and Larochelle, H. (2007). Greedy layer-wise training of deep networks. pages 153–160.

Cheng, Z., Qin, L., Ye, Y., Huang, Q., and Tian, Q. (2012). 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, Berlin, Heidelberg. Springer-Verlag.

Dalal, N. and Triggs, B. (2005). Histograms of oriented gradients for human detection. In *In CVPR*, pages 886–893.

Han, L., Wu, X., Liang, W., Hou, G., and Jia, Y. (2010). Discriminative human action recognition in the learned hierarchical manifold space. *Image Vision Comput.*, 28(5):836–849.

Hinton, G. E., Osindero, S., and Teh, Y. W. (2006). A fast learning algorithm for deep belief nets. *Neural Computation*, 18(7):1527–1554.

Hyvrinen, A., Hurri, J., and Hoyer, P. O. (2009). *Natural Image Statistics: A Probabilistic Approach to Early Computational Vision*. Springer Publishing Company, Incorporated, 1st edition.

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

Laptev, I., Marszałek, M., Schmid, C., and Rozenfeld, B. (2008). Learning realistic human actions from movies. In *Conference on Computer Vision & Pattern Recognition*.

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

Le, Q., Zou, W., Yeung, S., and Ng, A. (2011). 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.

Li, W., Zhang, Z., and Liu, Z. (2008). Expandable data-driven graphical modeling of human actions based on salient postures. *IEEE Trans. Cir. and Sys. for Video Technol.*, 18(11):1499–1510.

Li, W., Zhang, Z., and Liu, Z. (2010). Action recognition based on a bag of 3d points.

Lv, F. and Nevatia, R. (2006). Recognition and segmentation of 3-d human action using hmm and multi-class adaboost. In Leonardis, A., Bischof, H., and Pinz, A., editors, *Computer Vision ECCV 2006*, volume 3954 of *Lecture Notes in Computer Science*, pages 359–372. Springer Berlin Heidelberg.

Müller, M. and Röder, T. (2006). 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, Aire-la-Ville, Switzerland, Switzerland. Eurographics Association.

Oreifej, O. and Liu, Z. (2013). Hon4d: Histogram of oriented 4d normals for activity recognition from depth sequences. In *Computer Vision and Pattern Recognition (CVPR)*.

Shotton, J., Fitzgibbon, A., Cook, M., Sharp, T., Finocchio, M., Moore, R., Kipman, A., and Blake, A. (2011). Real-time human pose recognition in parts from single depth images. In *Proceedings of the 2011 IEEE Conference on Computer Vision and Pattern Recognition*, CVPR '11, pages 1297–1304, Washington, DC, USA. IEEE Computer Society.

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

Wang, J., Liu, Z., Chorowski, J., Chen, Z., and Wu, Y. (2012a). Robust 3d action recognition with random occupancy patterns. In *Proceedings of the 12th European conference on Computer Vision - Volume Part II*, ECCV'12, pages 872–885, Berlin, Heidelberg. Springer-Verlag.

Wang, J., Liu, Z., Wu, Y., and Yuan, J. (2012b). Mining actionlet ensemble for action recognition with depth cameras. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 1290–1297.

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

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

Yang, X. and Tian, Y. (2012). Eigenjoints-based action recognition using nave-bayes-nearest-neighbor. In *CVPR Workshops*, pages 14–19. IEEE.

Zhang, H. and Parker, L. (2011). 4-dimensional local spatio-temporal features for human activity recognition. In *Intelligent Robots and Systems (IROS), 2011 IEEE/RSJ International Conference on*, pages 2044–2049.

Zhao, Y., Liu, Z., Yang, L., and Cheng, H. (2012). Combing rgb and depth map features for human activity recognition. In *Signal Information Processing Association Annual Summit and Conference (APSIPA ASC), 2012 Asia-Pacific*, pages 1–4.