3D Action Recognition Using Multi-temporal Depth Motion Maps and Fisher Vector

Chen Chen[1,*], Mengyuan Liu[2,*], Baochang Zhang[3,†], Jungong Han[4], Junjun Jiang[5], Hong Liu[2]

[1] Department of Electrical Engineering, University of Texas at Dallas
[2] Key Laboratory of Machine Perception, Shenzhen Graduate School, Peking University
[3] School of Automation Science and Electrical Engineering, Beihang University
[4] Department of Computer Science and Digital Technologies, Northumbria University
[5] School of Computer Science, China University of Geosciences

Abstract

This paper presents an effective local spatio-temporal descriptor for action recognition from depth video sequences. The unique property of our descriptor is that it takes the shape discrimination and action speed variations into account, intending to solve the problems of distinguishing different pose shapes and identifying the actions with different speeds in one goal. The entire algorithm is carried out in three stages. In the first stage, a depth sequence is divided into temporally overlapping depth segments which are used to generate three depth motion maps (DMMs), capturing the shape and motion cues. To cope with speed variations in actions, multiple frame lengths of depth segments are utilized, leading to a multi-temporal DMMs representation. In the second stage, all the DMMs are first partitioned into dense patches. Then, the local binary patterns (LBP) descriptor is exploited to characterize local rotation invariant texture information in those patches. In the third stage, the Fisher kernel is employed to encode the patch descriptors for a compact feature representation, which is fed into a kernel-based extreme learning machine classifier. Extensive experiments on the public *MSRAction3D*, *MSRGesture3D* and *DHA* datasets show that our proposed method outperforms state-of-the-art approaches for depth-based action recognition.

1 Introduction

Action recognition plays a significant role in a number of computer vision applications such as context-based video retrieval, human-computer interaction and intelligent surveillance systems, e.g., [Chen *et al.*, 2014a; 2014b; Bloom *et al.*, 2012]. Previous works focus on recognizing actions captured by conventional RGB video cameras, e.g., [Wang and Schmid, 2013]. Based on compact local descriptors, state-of-the-art results have been achieved on benchmark RGB action datasets. However, these works suffer from several common problems such as various lighting conditions and cluttered backgrounds, due to the limitations of conventional RGB video cameras.

Recent progresses witness the change of action recognition from conventional RGB cameras to depth cameras. Compared with RGB cameras, depth cameras have several advantages: 1) depth data is more robust to changes in lighting conditions and depth cameras can even work in dark environment; 2) color and texture are ignored in depth images, which makes the tasks of human detection and foreground extraction from cluttered backgrounds easier [Yang and Tian, 2014]; 3) depth cameras provide depth images with appropriate resolution and accuracy, which capture the 3D structure information of subjects/objects in the scene [Ni *et al.*, 2011]; 4) human skeleton information (e.g., 3D joints positions and rotation angles) can be efficiently estimated from depth images providing additional information for action recognition [Shotton *et al.*, 2011].

Since the release of cost-effective depth cameras (in particular Microsoft Kinect), more recent works on action recognition have been conducted using depth images. Various representations of depth sequences have been explored including bag of 3D points [Li *et al.*, 2010], spatio-temporal depth cuboid [Xia and Aggarwal, 2013], depth motion maps (DMMs) [Yang *et al.*, 2012; Chen *et al.*, 2013; 2015], surface normals [Oreifej and Liu, 2013; Yang and Tian, 2014] and skeleton joints [Vemulapalli *et al.*, 2014]. Among those, DMMs-based representations effectively transform the action recognition problem from 3D to 2D and have been successfully applied to depth-based action recognition. Specifically, DMMs [Yang *et al.*, 2012] are obtained by projecting the depth frames onto three orthogonal Cartesian planes and accumulating the difference between projected maps over the entire sequence. They can be used to describe the shape and motion cues of a depth action sequence.

Motivation and contributions However, DMMs based on an entire depth sequence may not be able to capture detailed temporal motion in a subset of depth images. Old motion his-

[*] Equal contribution by Chen Chen (chenchen870713@gmail.com) and Mengyuan Liu (liumengyuan@pku.edu.cn).

[†] Correspondence to Baochang Zhang (bczhang@buaa.edu.cn).

This research was supported by Beijing Municipal Science & Technology Commission, the National Natural Science Foundation of China (Grant No. 61272052 and 61473086), the Program for New Century Excellent Talents of the University of Ministry of Education of China, and the Royal Society Newton Mobility Grant (IE150997).

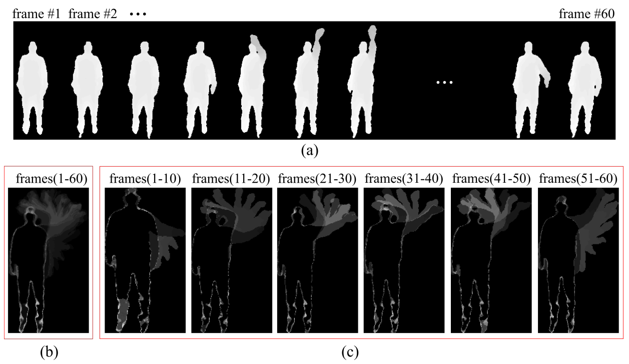


Figure 1: (a) is an example depth action sequence *high wave*. (b) shows the DMM of the front view projection generated using all the depth frames (60 frames) in a depth action sequence (*high wave*). (b) shows 6 DMMs of the front view projection generated using 6 different subsets of depth frames (e.g., frames 1-10, 11-20, 21-30, etc.) in the same action sequence. It can be observed that the detailed motion (e.g., raising hand over head and waving) of a hand waving action can be observed in DMMs generated using subsets of depth frames in a depth action sequence. In other words, the waving motion exhibited in the DMMs generated from subsets of depth frames is more obvious and clear than that in the DMMs generated using the entire depth sequence (all frames).

tory may get overwritten when a more recent action occurs at the same point. We provide an example in Fig. 1 to illustrate this limitation of DMMs in capturing detailed motion cues. In addition, action speed variations may result in large intraclass variations in DMMs. To this end, in this paper we develop a novel local spatio-temporal descriptor which takes the shape discrimination and action speed variations into account. More specifically, we propose to divide a depth sequence into overlapping segments and generate multiple sets of DMMs in order to preserve more detailed motion cues that might be lost in DMMs based on an entire sequence. To cope with speed variations in actions, we employ different temporal lengths of the depth segments, leading to a multi-temporal DMMs representation. A set of local patch descriptors are then extracted by partitioning all the DMMs into dense patches and using the local binary patterns (LBP) [Ojala *et al.*, 2002] descriptor to characterize local rotation invariant texture information in those patches. To build a compact representation, the Fisher kernel [Perronnin *et al.*, 2010] is adopted to encode the patch descriptors. Our proposed approach is validated on several benchmark depth datasets for human action recognition and demonstrates superior performances over other state-of-the-art approaches.

2 Related Work

In this section, we briefly review recent methods on action recognition using depth information, which can be broadly categorized into depth images-based, skeleton-based, and depth and skeleton fusion-based methods. A comprehensive review on action recognition from 3D data is provided in [Agarwal and Lu, 2014].

In the field of 3D object retrieval, surface normal vectors can efficiently reflect local shapes of 3D objects [Tang *et al.*, 2013]. By extending the same concept to temporal dimension, [Oreifej and Liu, 2013] described a depth sequence by a Histogram of Oriented Normal vectors in the 4D space of depth, spatial coordinates and time (HON4D). To increase the descriptive power of HON4D, [Rahmani *et al.*, 2014a] characterized each 3D point by encoding Histogram of Oriented Principal Components (HOPC) within a volume around that point, which is more informative than HON4D as it captures the spread of data in three principal directions. To alleviate the loss of information in quantization part of constructing HON4D, [Kong *et al.*, 2015] adopted the concept of surface normal and proposed kernel descriptors to convert pixel-level 3D gradient into patch-level features. Rather than describing a depth sequence by using surface normal vectors, [Rahmani *et al.*, 2015] divided a depth sequence into equally spatio-temporal cells, which were represented by a Histogram of Oriented 3D Gradients (HOG3D) and encoded by locality-constrained linear coding. [Lu *et al.*, 2014] developed a binary descriptor by conducting τ test to encode relative depth relationships among pairwise 3D points. [Chen *et al.*, 2016] presented a weighted fusion framework of combining 2D and 3D auto-correlation of gradients features from depth images for action recognition. [Zhang and Tian, 2015] proposed an effective descriptor, the Histogram of 3D Facets (H3DF), to explicitly encode the 3D shape and structures of various depth images by coding and pooling 3D Facets from depth images.

Estimating skeleton joints from depth images [Shotton *et al.*, 2011] provides a more intuitive way to perceive human actions. Existing skeleton-based approaches can be broadly grouped into joint-based and body part-based approaches. [Wang *et al.*, 2014] selected an informative subset of joints for one specific action type, and extracted pairwise relative position features to represent each selected joint. Instead of using joint locations as features, [Vemulapalli *et al.*, 2014] represented skeletons as points in the Lie group $SE(3) \times ... \times SE(3)$, which explicitly models the 3D geometric relationships among human body parts.

Obviously, skeleton joints only reflect the state of human bodies, therefore skeleton-based methods gain limited recognition rates in human object interaction scenarios. To improve the recognition performance using skeleton joints, [Wang *et al.*, 2014] proposed an ensemble model which associates local occupancy pattern features from depth images with skeleton joints. [Ohn-Bar and M. Trivedi, 2013] utilized joint angles pairwise similarities to represent skeletons and extracted HOG features involving depth information around joints. These two can be considered as representatives of combining skeleton joints and depth information. Although multimodal fusion methods generally achieve good recognition performance, running a depth descriptor on top of a complicated skeleton tracker makes such algorithms computationally expensive, limiting their use in real-time applications.

3 Proposed Depth Video Representation

3.1 Multi-temporal Depth Motion Maps

According to [Chen *et al.*, 2013], the DMMs of a depth sequence with N frames are computed as follows:

$$DMM_{\{f,s,t\}} = \sum_{i=2}^{N} |map^{i}_{\{f,s,t\}} - map^{i-1}_{\{f,s,t\}}| \qquad (1)$$

where map^i_f, map^i_s and map^i_t indicate three projected maps of the i^{th} depth frame on three orthogonal Cartesian planes corresponding to the front view (f), side view (s) and top view (t). As mentioned before, the DMMs based on the entire depth sequence may not be able to capture the detailed motion cues. Therefore, to overcome this shortcoming, we divide a depth sequence into a set of overlapping 3D depth segments with equal number of frames (i.e., same frame length for each depth segment) and compute three DMMs for each depth segment. Since different people may perform an action in different speeds, we further employ multiple frame lengths to represent multiple temporal resolutions to cope with action speed variations. The proposed multi-temporal DMMs representation framework is shown in Fig. 2. Take this figure as an example, generating DMMs using the entire depth sequence (i.e., all the frames in the sequence) is considered as a default level of temporal resolution (denoted by Level 0 in Fig. 2). In the second level (Level 1 in Fig. 2), the frame length (L_1) of a depth segment is set to 5 (i.e., 5 frames in a depth segment). In the third level (Level 2 in Fig. 2), the frame length (L_2) of a depth segment is set to 10. Note that L_1 and L_2 can be changed. Obviously, the computational complexity increases with the increase of temporal levels. Therefore, we limit the maximum number of levels to be 3 including a default level, i.e., Level 0 which considers all the frames. The frame interval (R, $R < L_1$ and $R < L_2$) in Fig. 2 is the number of frames between the first frames (or the starting frames) respectively in two neighboring depth segments, indicating how much overlapping between the two segments. For simplicity, we use the same R in Level 1 and Level 2.

3.2 Patch-based LBP Features

DMMs can effectively capture the shape and motion cues of a depth sequence. However, DMMs are pixel-level features. To enhance the discriminative power of DMMs, we adopt the patch-based LBP feature extraction approach in [Chen *et al.*, 2015] to characterize the rich texture information (e.g., edges, contours, etc.) in the LBP coded DMMs. Fig. 3 shows the process of patch-based LBP feature extraction. The overlap between two patches is controlled by the pixel shift (ps) illustrated in Fig. 3. Under each projection view, a set of patch-based LBP histogram features are generated to describe the corresponding multi-temporal DMMs. Therefore, three feature matrices H_f, H_s and H_t are generated associated with front view DMMs, side view DMMs and top view DMMs, respectively. Each column of the feature matrix (e.g., H_f) is a histogram feature vector of a local patch.

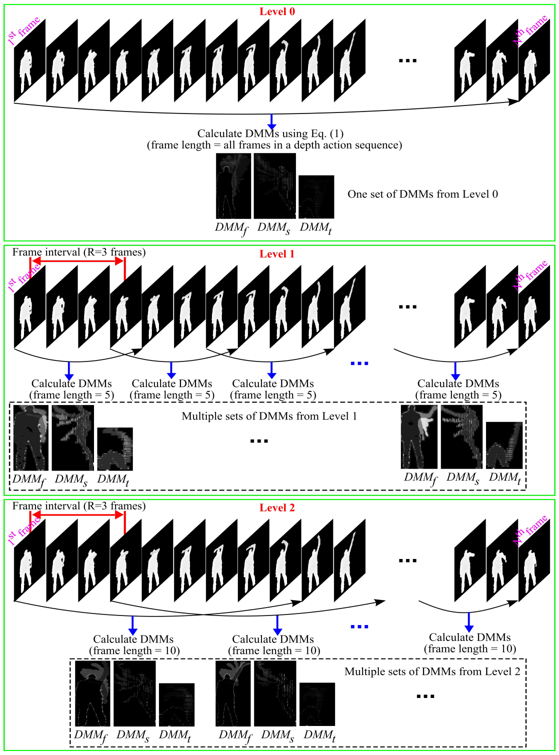

Level 0
1st frame
Nth frame
...
Calculate DMMs using Eq. (1)
(frame length = all frames in a depth action sequence)
One set of DMMs from Level 0
DMM_f DMM_s DMM_t
Frame interval (R=3 frames)
Level 1
Calculate DMMs (frame length = 5)
Multiple sets of DMMs from Level 1
Level 2
Calculate DMMs (frame length = 10)
Multiple sets of DMMs from Level 2

Figure 2: Proposed multi-temporal DMMs representation of a depth sequence.

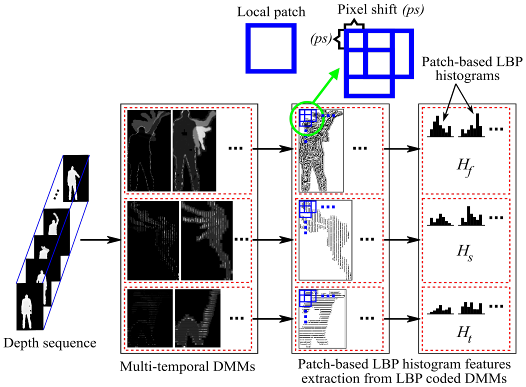

Local patch
Pixel shift (ps)
(ps)
Patch-based LBP histograms
H_f
H_s
H_t
Depth sequence
Multi-temporal DMMs
Patch-based LBP histogram features extraction from LBP coded DMMs

Figure 3: Patch-based LBP feature extraction.

3.3 A Fisher Kernel Representation

Fisher kernel representation [Perronnin *et al.*, 2010] is an effective patch aggregation mechanism to characterize a set of low-level features, which shows superior performance over the popular Bag-of-Visual-Words (BoVW) model. Therefore, we employ the Fisher kernel to build a compact and descriptive representation of the patch-based LBP features.

Let $H = \{\boldsymbol{h}_i \in \mathbb{R}^D, 1 \leq i \leq M\}$ be a set of M D-dimensional patch-based LBP feature vectors extracted from the multi-temporal DMMs of a particular projection view (e.g., front veiw) for a depth sequence. By assuming statistical independence, H can be modeled by a K-component Gaussian mixture model (GMM):

$$p(H|\theta) = \prod_{i=1}^{M} \sum_{k=1}^{K} \omega_k \mathcal{N}(\boldsymbol{h}_i|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k), \qquad (2)$$

where $\theta = \{\omega_k, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\}, k = 1, ..., K$ is the parameter set with mixing parameters ω_k, means $\boldsymbol{\mu}_k$ and diagonal covariance matrices $\boldsymbol{\Sigma}_k$ with the variance vector $\boldsymbol{\sigma}_k^2$. These GMM parameters can be estimated by using the Expectation-Maximization (EM) algorithm based on a training dataset (or feature set).

Two D-dimensional gradients with respect to the mean vector $\boldsymbol{\mu}_k$ and standard deviation $\boldsymbol{\sigma}_k$ of the k^{th} Gaussian component are defined as

$$\begin{aligned} \boldsymbol{\rho}_k &= \frac{1}{M\sqrt{\pi_k}} \sum_{i=1}^{M} \gamma_{k,i} \frac{\boldsymbol{h}_i - \boldsymbol{\mu}_k}{\boldsymbol{\sigma}_k}, \\ \boldsymbol{\tau}_k &= \frac{1}{M\sqrt{2\pi_k}} \sum_{i=1}^{M} \gamma_{k,i} \left(\left(\frac{\boldsymbol{h}_i - \boldsymbol{\mu}_k}{\boldsymbol{\sigma}_k} \right)^2 - 1 \right), \end{aligned} \tag{3}$$

where $\gamma_{k,i}$ is the posterior probability that $\boldsymbol{q}_i$ belongs to the kth Gaussian component. The Fisher vector (FV) of H is represented as $\Phi(H) = (\boldsymbol{\rho}_1^T, \boldsymbol{\tau}_1^T, ..., \boldsymbol{\rho}_K^T, \boldsymbol{\tau}_K^T)^T$. The dimensionality of the FV is $2KD$.

A power-normalisation step introduced in [Perronnin *et al.*, 2010], i.e., signed square rooting (SSR) and ℓ_2 normalization, is applied to eliminate the sparseness of the FV as follows:

$$T(\Phi(H)) = sgn(\Phi(H)) * |\Phi(H)|^\alpha, \quad 0 < \alpha \leq 1. \tag{4}$$

Let H_f, H_s and H_t denote three sets of patch-based LBP feature vectors from three projection views, each depth sequence is then represented by concatenating three FVs $[\Phi(H_f); \Phi(H_s); \Phi(H_t)]$ as the final feature representation.

4 Experiments

In this section we extensively evaluate our proposed method on three public benchmark datasets: *MSRAction3D* [Li *et al.*, 2010], *MSRGesture3D* [Wang *et al.*, 2012] and *DHA* [Lin *et al.*, 2012]. We employ kernel-based extreme learning machine (KELM) [Huang *et al.*, 2006] with a radial basis function (RBF) kernel as the classifier due to its general good classification performance and efficient computation.

4.1 *Datasets*

MSRAction3D dataset [Li *et al.*, 2010] is one of the most popular depth datasets for action recognition as reported in the literature. It contains 20 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", "sideboxing", "bend", "forward kick", "side kick", "jogging", "tennis swing", "tennis serve", "golf swing", "pick up & throw". Each action is performed 2 or 3 times by 10 subjects facing the depth camera. It is a challenging dataset due to similarity of actions and large speed variations in actions.

MSRGesture3D dataset [Wang *et al.*, 2012] is a benchmark dataset for depth-based hand gesture recognition. It consists of 12 gestures defined by American Sign Language: "bathroom", "blue", "finish", "green", "hungry", "milk", "past", "pig", "store", "where", "j", "z". Each action is performed 2 or 3 times by each subject, resulting in 336 depth sequences.

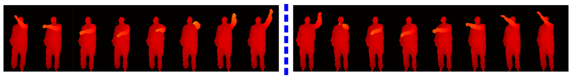


Figure 4: Actions "drawTick" (left) and "drawX" (right) in the *MSRAction3D* dataset.

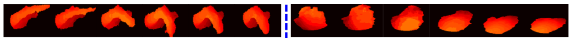


Figure 5: Actions "milk" (left) and "hungry" (right) in the *MSRGesture3D* dataset.

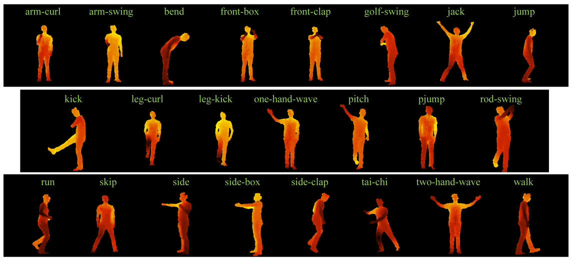


Figure 6: Action snaps in the *DHA* dataset.

DHA dataset is proposed in [Lin *et al.*, 2012], whose action types are extended from the Weizmann dataset [Gorelick *et al.*, 2007] which is widely used in action recognition from RGB sequences. It contains 23 action categories:"arm-curl", "arm-swing", "bend", "front-box", "front-clap", "golf-swing", "jack", "jump", "kick", "leg-curl", "leg-kick", "one-hand-wave", "pitch", "pjump", "rod-swing", "run", "skip", "side", "side-box", "side-clap", "tai-chi", "two-hand-wave", "walk". Each action is performed by 21 subjects (12 males and 9 females), resulting in 483 depth sequences.

4.2 *Experimental Settings*

Several action snaps from the three datasets are shown in Figs. 4-6, where inter-similarity among different types of actions are observed. In the *MSRAction3D* dataset, actions such as "drawX" and "drawTick" are similar except for a slight difference in the movement of one hand. In the *MSRGesture3D* dataset, actions such as "milk" and "hungry" are alike, since both actions involve the motion of bending palm. What's more, self-occlusion is also a challenge for this dataset. In the *DHA* dataset, "golf-swing" and "rod-swing" actions share similar motions by moving hands from one side up to the other side. More similar pairs can be found in "leg-curl" and "leg-kick", "run" and "walk", etc.

In order to keep the reported results consistent with other works, we follow the same evaluation protocols in [Wang *et al.*, 2014], [Wang *et al.*, 2012] and [Lin *et al.*, 2012] respectively for the three datasets.

We adopt the same parameter values in [Chen *et al.*, 2015] for the patch sizes and parameters for the LBP operator in our method. The other parameters are determined empirically. The overall accuracies on three datasets with different parameters are shown in Figure 7, where frame length L_1, frame length L_2, frame interval R, pixel shift ps and the number of Gaussians (K) respectively change from 3 to 11, 10 to 18, 1

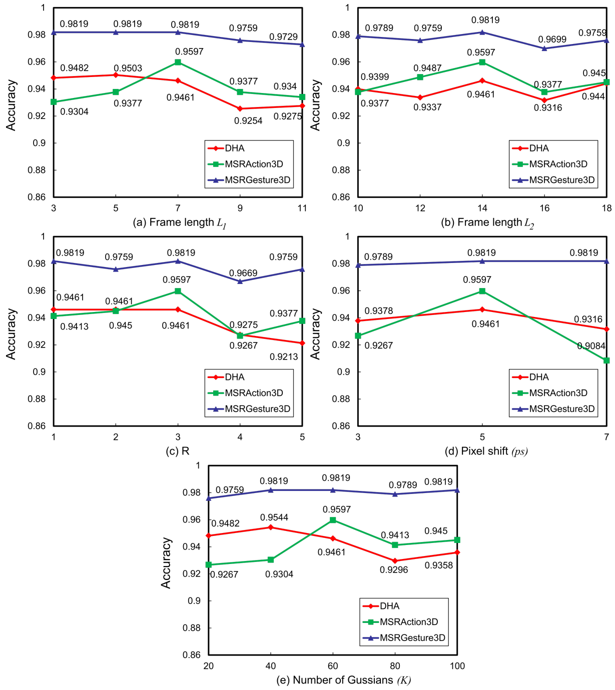


Figure 7: Recognition accuracies with changing parameters.

Table 1: Recognition accuracy and average feature computation time of our method with different numbers of temporal levels on the *MSRAction3D* dataset.

Temporal levels	Accuracy	Time/sequence (s)
1 level (Level 0)	89.95%	0.35
2 levels (Levels 0, 1)	93.34%	2.51
3 levels (Levels 0, 1, 2)	95.97%	4.49

to 5, 3 to 7 and 20 to 100 at equal intervals. Experiments are conducted with one parameter changes and the other parameters in default values: $L_1 = 7$, $L_2 = 14$, $R = 3$, $ps = 5$ and $K = 60$. From Figure 7, we can see that more than 90% accuracies are achieved with different parameters, which reflect the robustness of our method to parameter settings. Since default parameters work well for all three datasets, the following experiments are conducted with these values in default.

In our method, we use three levels for the multi-temporal DMMs representation. We test the algorithm on the *MSRAction3D* dataset using different numbers of temporal levels. The recognition accuracy and average feature computation time are reported in Table 1. It is worth mentioning that our algorithm is implemented in MATLAB and executed on CPU platform with an Intel(R)Core(TM)i7 CPU @2.60GHz and 8GB of RAM. It can easily gain the efficiency by converting the code to C++ and running the multi-temporal DMMs representation in parallel.

4.3 *Results on the MSRAction3D Dataset*

Table 2: Recognition accuracy comparison on the *MSRAction3D* dataset.

Method	Accuracy
Bag of 3D Points [Li *et al.*, 2010]	74.70%
Random Occupancy Pattern [Wang *et al.*, 2012]	86.50%
Actionlet Ensemble [Wang *et al.*, 2014]	88.20%
Depth Motion Maps [Yang *et al.*, 2012]	88.73%
HON4D [Oreifej and Liu, 2013]	88.89%
DSTIP [Xia and Aggarwal, 2013]	89.30%
H3DF [Zhang and Tian, 2015]	89.45%
Skeletons Lie group [Vemulapalli *et al.*, 2014]	89.48%
Skeletal Quads [Evangelidis *et al.*, 2014]	89.86%
HOG3D+LLC [Rahmani *et al.*, 2015]	90.90%
Moving Pose [Zanfir *et al.*, 2013]	91.70%
Hierarchical 3D Kernel [Kong *et al.*, 2015]	92.73%
DMM-LBP-DF [Chen *et al.*, 2015]	93.00%
Super Normal Vector [Yang and Tian, 2014]	93.09%
Hierarchical RNN [Du *et al.*, 2015]	94.49%
Range-Sample [Lu *et al.*, 2014]	95.62%
Our Method	**95.97%**

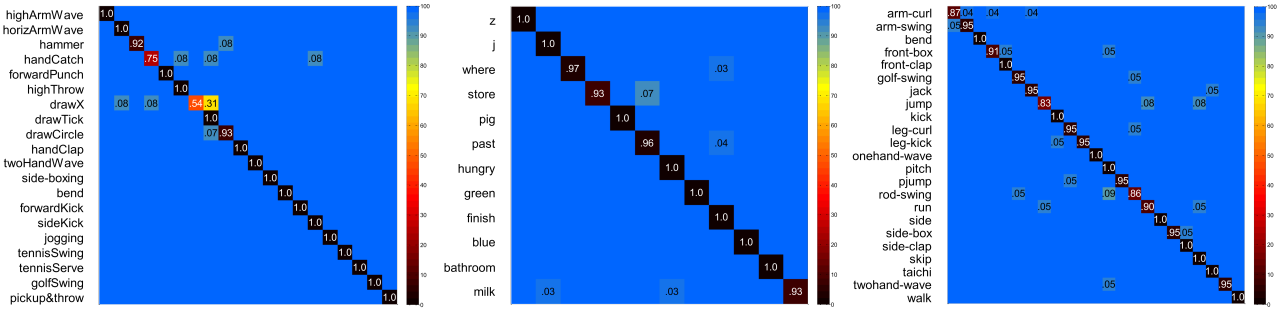
In Figure 8 (a), we show the confusion matrix of the *MSRAction3D* dataset with the accuracy of 95.97%. It is observed that large ambiguities exist between similar action pairs, for example "handCatch" and "highThrow", and "drawX" and "drawTick", due to the similarities of their DMMs. We also compare our method with the state-of-the-art methods in Table 2. "Moving Pose" [Zanfir *et al.*, 2013], "Skeletons in a Lie group" [Vemulapalli *et al.*, 2014] and "Skeletal Quads" [Evangelidis *et al.*, 2014] belong to skeleton-based features, and "Actionlet Ensemble" [Wang *et al.*, 2014] belongs to skeleton+depth based features. Our method outperforms these methods for two reasons: first, skeleton joints used by these methods contain a lot of noises, which bring ambiguities to distinguish similar actions; second, our method directly using DMMs, which provide more affluent motion information. Our result is also better than the recent depth-based features such as Super Normal Vector [Yang and Tian, 2014] and Range-Sample [Lu *et al.*, 2014], which demonstrates the superior discriminatory power of our multi-temporal DMMs representation.

Table 3: Recognition accuracy comparison on the *MSRGesture3D* dataset.

Method	Accuracy
Random Occupancy Pattern [Wang *et al.*, 2012]	88.50%
HON4D [Oreifej and Liu, 2013]	92.45%
HOG3D+LLC [Rahmani *et al.*, 2015]	94.10%
DMM-LBP-DF [Chen *et al.*, 2015]	94.60%
Super Normal Vector [Yang and Tian, 2014]	94.74%
H3DF [Zhang and Tian, 2015]	95.00%
Depth Gradients+RDF [Rahmani *et al.*, 2014b]	95.29%
Hierarchical 3D Kernel [Kong *et al.*, 2015]	95.66%
HOPC [Rahmani *et al.*, 2014a]	96.23%
Our Method	**98.19%**

4.4 *Results on the MSRGesture3D Dataset*

In Figure 8 (b), we show the confusion matrix of the *MSRGesture3D* dataset with the accuracy of 98.19%. It is observed



(a) Overall accuracy = 95.97%

(b) Overall accuracy = 98.19%

(c) Overall accuracy = 95.44%

Figure 8: Confusion matrices of our method on the (a) *MSRAction3D*, (b) *MSRGesture3D* and (c) *DHA* datasets.

Table 4: Recognition accuracy comparison on the *DHA* dataset.

Method	Accuracy
D-STV/ASM [Lin *et al.*, 2012]	86.80%
SDM-BSM [Liu *et al.*, 2015]	89.50%
D-DMHI-PHOG [Gao *et al.*, 2015]	92.40%
DMPP-PHOG [Gao *et al.*, 2015]	95.00%
Our Method	**95.44%**

that similar action pairs like “milk” and “hungry” can be distinguished with high accuracy. We compare our method with several existing methods in Table 3. As can be seen from this table, our method outperforms Histogram of Oriented Principal Components (HOPC) [Rahmani *et al.*, 2014a] by 1.96%, leading to a new state-of-the-art result.

4.5 *Results on the DHA Dataset*

In Figure 8 (c), we present the confusion matrix of our method on the *DHA* dataset with the accuracy of 95.44%. The *DHA* dataset is originally collected by [Lin *et al.*, 2012], which only contains 17 action categories. We use an extended version of the *DHA* dataset where extra 6 action categories are involved. [Lin *et al.*, 2012] splited depth sequences into space-time volume and constructed 3bit binary patterns as depth features, which achieved an accuracy of 86.80% on the original dataset. By incorporating multi-temporal information to the DMMs, our proposed method achieves higher accuracy even on the extended *DHA* dataset. In Table 4, we observe that our method outperforms D-DMHI-PHOG [Gao *et al.*, 2015] by 3.04% and outperforms DMPP-PHOG [Gao *et al.*, 2015] by 0.44%. These improvements show that operating LBP on multi-temporal DMMs can produce more informative features than operating PHOG on depth difference motion history image (D-MHI).

4.6 *Execution Rate and Frame Rate Invariance*

Regarding to the execution rate invariance, we have calculated the statistics for the *MSRAction3D* dataset, which contains the actions executed by different subjects with different execution rates. To be more precise, there are 20 actions, each being executed by 10 subjects for 2 or 3 times. The standard derivation of the sequence lengths (numbers of frames) across the actions is 9.21 frames (max: 13.30 frames; min: 4.86 frames), which means that execution rate difference is actually quite large. In view of the achieved 95.97% recognition rate, we would say that our algorithm is resistant to the execution rate.

To test the effect by frame rate difference, we carry out an experiment using the *MSRAction3D* dataset. Specifically, we use the sequences performed by subjects 1, 3, 5, 7, 9, (the original action samples) as training data. We select half number of frames (odd numbers of frames, e.g., 1, 3, 5 ...) of the sequences performed by subjects 2, 4, 6, 8, 10 to form a set of new testing samples with 1/2 of the original frame rate. The achieved recognition result of our method is 93.27%. Therefore, our proposed algorithm is capable of dealing with frame rate changes considering the fact that 1/2 frame rate reduction is actually unrealistic.

5 Conclusion

This paper presents an effective feature representation for action recognition from depth sequences. A multi-temporal DMMs representation is proposed to capture more temporal motion information in depth sequences for better distinguishing similar actions. Multiple temporal resolutions in the proposed representation can also cope with the speed variations in actions. Patch-based LBP features are extracted from dense patches in the DMMs and the Fisher kernel representation is utilized to aggregate local patch features into a compact and discriminative representation. The proposed method is extensively evaluated on three benchmark datasets. Experimental results show that our method outperforms the state-of-the-art methods in all datasets.

References

[Aggarwal and Lu, 2014] J. K. Aggarwal and Xia Lu. Human activity recognition from 3d data: A review. *PRL*, 48(1):70–80, 2014.

[Bloom *et al.*, 2012] Victoria Bloom, Dimitrios Makris, and Vasileios Argyriou. G3d: A gaming action dataset and real time action recognition evaluation framework. In *CVPRW*, pages 7–12, 2012.

[Chen *et al.*, 2013] Chen Chen, Kui Liu, and Nasser Kehtarnavaz. Real-time human action recognition based on depth motion maps. *Journal of Real-Time Image Processing*, pages 1–9, 2013.

[Chen *et al.*, 2014a] Chen Chen, Nasser Kehtarnavaz, and Roozbeh Jafari. A medication adherence monitoring system for pill bottles based on a wearable inertial sensor. In *EMBC*, pages 4983–4986, 2014.

[Chen *et al.*, 2014b] Chen Chen, Kui Liu, Roozbeh Jafari, and Nasser Kehtarnavaz. Home-based senior fitness test measurement system using collaborative inertial and depth sensors. In *EMBC*, pages 4135–4138, 2014.

[Chen *et al.*, 2015] Chen Chen, R. Jafari, and N. Kehtarnavaz. Action recognition from depth sequences using depth motion maps-based local binary patterns. In *WACV*, pages 1092–1099, 2015.

[Chen *et al.*, 2016] Chen Chen, Baochang Zhang, Zhenjie Hou, Junjun Jiang, Mengyuan Liu, and Yun Yang. Action recognition from depth sequences using weighted fusion of 2d and 3d auto-correlation of gradients features. *Multimedia Tools and Applications*, pages 1–19, 2016.

[Du *et al.*, 2015] Yong Du, Wei Wang, and Liang Wang. Hierarchical recurrent neural network for skeleton based action recognition. In *CVPR*, pages 1110–1118, 2015.

[Evangelidis *et al.*, 2014] Georgios Evangelidis, Gurkirt Singh, and Radu Horaud. Skeletal quads:human action recognition using joint quadruples. In *ICPR*, pages 4513–4518, 2014.

[Gao *et al.*, 2015] Z. Gao, H. Zhang, G. P. Xu, and Y. B. Xue. Multi-perspective and multi-modality joint representation and recognition model for 3d action recognition. *Neurocomputing*, 151:554–564, 2015.

[Gorelick *et al.*, 2007] Lena Gorelick, Moshe Blank, Eli Shechtman, Michal Irani, and Ronen Basri. Actions as space-time shapes. *TPMAI*, 29(12):2247–2253, 2007.

[Huang *et al.*, 2006] Guang-Bin Huang, Qin-Yu Zhu, and Chee-Kheong Siew. Extreme learning machine: theory and applications. *Neurocomputing*, 70(1):489–501, 2006.

[Kong *et al.*, 2015] Yu Kong, B. Satarboroujeni, and Yun Fu. Hierarchical 3d kernel descriptors for action recognition using depth sequences. In *FG*, pages 1–6, 2015.

[Li *et al.*, 2010] Wanqing Li, Zhengyou Zhang, and Zicheng Liu. Action recognition based on a bag of 3d points. In *CVPRW*, pages 9–14, 2010.

[Lin *et al.*, 2012] Yan-Ching Lin, Min-Chun Hu, Wen-Huang Cheng, Yung-Huan Hsieh, and Hong-Ming Chen. Human action recognition and retrieval using sole depth information. In *ACM MM*, pages 1053–1056, 2012.

[Liu *et al.*, 2015] Hong Liu, Lu Tian, Mengyuan Liu, and Hao Tang. Sdm-bsm: A fusing depth scheme for human action recognition. In *ICIP*, pages 4674–4678, 2015.

[Lu *et al.*, 2014] Cewu Lu, Jiaya Jia, and Chi Keung Tang. Range-sample depth feature for action recognition. In *CVPR*, pages 772–779, 2014.

[Ni *et al.*, 2011] Bingbing Ni, Gang Wang, and P. Moulin. Rgbd-hudaact: A color-depth video database for human daily activity recognition. In *ICCVW*, pages 1147–1153, 2011.

[Ohn-Bar and M. Trivedi, 2013] Eshed Ohn-Bar and Mohan M. Trivedi. Joint angles similiarities and hog2 for action recognition. In *CVPRW*, pages 465–470, 2013.

[Ojala *et al.*, 2002] Timo Ojala, Matti Pietikäinen, and Topi Mäenpää. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *TPMAI*, 24(7):971–987, 2002.

[Oreifej and Liu, 2013] O. Oreifej and Zicheng Liu. Hon4d: Histogram of oriented 4d normals for activity recognition from depth sequences. In *CVPR*, pages 716–723, 2013.

[Perronnin *et al.*, 2010] Florent Perronnin, Jorge Sanchez, and Thomas Mensink. Improving the fisher kernel for large-scale image classification. In *ECCV*, pages 143–156, 2010.

[Rahmani *et al.*, 2014a] Hossein Rahmani, Arif Mahmood, Q Huynh Du, and Ajmal Mian. *HOPC: Histogram of Oriented Principal Components of 3D Pointclouds for Action Recognition*. Springer International Publishing, 2014.

[Rahmani *et al.*, 2014b] Hossein Rahmani, Arif Mahmood, Du Q Huynh, and Ajmal Mian. Real time action recognition using histograms of depth gradients and random decision forests. In *WACV*, pages 626–633, 2014.

[Rahmani *et al.*, 2015] Hossein Rahmani, Q. Huynh Du, Arif Mahmood, and Ajmal Mian. Discriminative human action classification using locality-constrained linear coding. *PRL*, 2015.

[Shotton *et al.*, 2011] 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*, pages 1297–1304, 2011.

[Tang *et al.*, 2013] Shuai Tang, Xiaoyu Wang, Xutao Lv, Tony X. Han, James Keller, Zhihai He, Marjorie Skubic, and Shihong Lao. *Histogram of Oriented Normal Vectors for Object Recognition with a Depth Sensor*. Springer Berlin Heidelberg, 2013.

[Vemulapalli *et al.*, 2014] Raviteja Vemulapalli, Felipe Arrate, and Rama Chellappa. Human action recognition by representing 3d human skeletons as points in a lie group. In *CVPR*, pages 588–595, 2014.

[Wang and Schmid, 2013] Heng Wang and Cordelia Schmid. Action recognition with improved trajectories. In *ICCV*, pages 3551–3558, 2013.

[Wang *et al.*, 2012] Jiang Wang, Zicheng Liu, Jan Chorowski, Zhuoyuan Chen, and Ying Wu. Robust 3d action recognition with random occupancy patterns. In *ECCV*, pages 872–885, 2012.

[Wang *et al.*, 2014] Jiang Wang, Zicheng Liu, Ying Wu, and Junsong Yuan. Learning actionlet ensemble for 3D human action recognition. *TPAMI*, 36(5):914–927, 2014.

[Xia and Aggarwal, 2013] Lu Xia and J. K. Aggarwal. Spatio-temporal depth cuboid similarity feature for activity recognition using depth camera. In *CVPR*, pages 2834–2841, 2013.

[Yang and Tian, 2014] Xiaodong Yang and YingLi Tian. Super normal vector for activity recognition using depth sequences. In *CVPR*, pages 804–811, 2014.

[Yang *et al.*, 2012] Xiaodong Yang, Chenyang Zhang, and Ying Li Tian. Recognizing actions using depth motion maps-based histograms of oriented gradients. *ACM MM*, pages 1057–1060, 2012.

[Zanfir *et al.*, 2013] Mihai Zanfir, Marius Leordeanu, and Cristian Sminchisescu. The moving pose: An efficient 3d kinematics descriptor for low-latency action recognition and detection. In *ICCV*, pages 2752–2759, 2013.

[Zhang and Tian, 2015] Chenyang Zhang and Yingli Tian. Histogram of 3D facets: A depth descriptor for human action and hand gesture recognition. *CVIU*, 139, 2015.