


# Action recognition from point cloud patches using discrete orthogonal moments

Huaining Cheng<sup>1</sup> · Soon M. Chung<sup>2</sup> 

Received: 9 September 2016 / Revised: 14 February 2017 / Accepted: 12 April 2017 /

Published online: 27 April 2017

© Springer Science+Business Media New York 2017

**Abstract** 3D sensors such as standoff Light Detection and Ranging (LIDAR) generate partial 3D point clouds that resemble patches of irregularly-shaped, coarse groups of points. 3D modeling of this type of data for human action recognition has been rarely studied. Although 2D-based depth image analysis is an option, its effectiveness on this type of low-resolution data hasn't been well answered. This paper investigates a new multi-scale 3D shape descriptor, based on the discrete orthogonal Tchebichef Moments, for the characterization of 3D action pose shapes made of low-resolution point cloud patches. Our shape descriptor consists of low-order 3D Tchebichef moments computed with respect to a new point cloud voxelization scheme that normalizes translation, scale, and resolution. The action recognition is built on the Naïve Bayes classifier using temporal statistics of a 'bag of pose shapes'. For performance evaluation, a synthetic LIDAR pose shape baseline was developed with 62 human subjects performing three actions — digging, jogging, and throwing. Our action classification experiments demonstrated that the 3D Tchebichef moment representation of point clouds achieves excellent action and viewing direction predictions with superb consistency across a large range of scale and viewing angle variations.

**Keywords** LIDAR · Point cloud · Action recognition · Discrete orthogonal moment · Tchebichef moment

---

✉ Soon M. Chung  
soon.chung@wright.edu

Huaining Cheng  
huaining.cheng@us.af.mil

<sup>1</sup> 711th Human Performance Wing, Air Force Research Laboratory, Wright-Patterson AFB, Dayton, OH 45433, USA

<sup>2</sup> Department of Computer Science and Engineering, Wright State University, Dayton, OH 45435, USA

## 1 Introduction

Recently, 3D sensors such as Light Detection and Ranging (LIDAR) started appearing in commercial applications for object detection and recognition. In contrast to 2D imagery, 3D shape captures the true form of human action. It eliminates many of the hard-to-estimate random factors in 2D imagery that are not related to actions, such as projection, lighting, color, texture, etc.; hence, it confines the varying factors to pose, style, scale, and viewing angle. This can lead to simplified probabilistic models for pose learning and action inference as well as better recognition rate at low resolutions.

The objective of this study is to prove conceptually that 1) standoff LIDAR-like 3D sensing data could be exploited for shape-based action recognition through direct modeling of 3D point clouds and 2) the corresponding classifier's performance and consistency are better than those of common 2D depth image analysis methods, under varying viewing angles and small scales. We envision that our 3D sensing and modeling methods could augment existing 2D technologies, especially for the applications of air-to-ground target recognition where human targets are typically much smaller in size than those seen in many ground-level public benchmark datasets created using Microsoft Kinect, due to much longer standoff distance.

A main reason that we promote direct modeling of 3D point clouds over extracting 2D features from depth images is that many existing 2D features are not available or effective in our application, due to the shape degeneracy in standoff LIDAR data. The degeneracy is manifested by irregular gaps and topology among patches of human body parts resulted from self-occlusions and varying viewing angles, as well as point sparsity due to low resolution settings. Consequently, there is a lack of pairwise point relationships and meaningful anatomical reference markers, which makes the registration of points over different image frames difficult and hinders the identification of key points for extracting local features along temporal dimension.

In 3D domain, spatial patterns are typically abstracted into feature vectors called shape descriptors. The degeneracy limits the applicability or efficacy of many existing 3D shape descriptors that are designed to work with smooth, water-tight full body surface models [5]. This calls for degeneracy-tolerant shape representation, which is met by the Tchebichef Moment Shape Descriptor (TMSD) proposed in our recent study [5]. TMSD consists of low-order Tchebichef moments [6, 27] used to characterize 3D point density distribution with respect to a new point cloud voxelization scheme that offers translation, scale, and resolution normalization. Tchebichef moments are discrete orthogonal moments, so they provide orthogonality, completeness, and consistency that are well-suited for multi-scale representation and recognition of discrete spatial patterns. Our previous study indicates that TMSD outperforms other orthogonal transform based 3D descriptors, such as 3D discrete Fourier transform [5].

Unlike our previous focus on 3D shape descriptors only, the current work furthers the investigation into a unique direction of 3D modeling versus 2D modeling because LIDAR images can be easily converted into 2D depth images, and hence people may opt for using popular 2D features. Part of our goal is to demonstrate that, for low resolution LIDAR data, native 3D modeling may be a better alternative than 2D depth image analysis. Moreover, unlike our previous focus on unsupervised single-frame shape search, the current work explores ways to model, learn, and classify spatial-temporal patterns over a sequence of degenerated point cloud patches.

In this study, the types of action we considered are three atomic ones — digging, jogging, and throwing. The raw data of an action are a sequence of point cloud patches that record

temporal changes of partial 3D pose shapes over the length of the action. TMSDs are used to represent individual frames of point cloud patches. To incorporate temporal information, we looked into the popular bag-of-words (BoW) framework, and propose a new bag-of-pose-shapes (BoPS) scheme to accommodate unstructured and uncontrolled use scenarios. Unlike the common local feature based BoW [37], our BoPS constitutes a temporal statistics model of 3D global pose shapes, encoded using a learned vocabulary of pose shape words. The follow-on action classification is performed using the Naïve Bayes classifier. Two posterior distribution models based on word frequency and word appearance, respectively, were investigated.

In order to provide statistically meaningful shape modeling, classifier training, and performance testing, we created a simulated LIDAR pose shape baseline consisting of two subsets of sensing at the ground and from a slant  $45^\circ$  elevation angle, respectively. Each subset has 62 volunteers performing the aforementioned three actions, resulting over 47,000 frames of point clouds when viewed from 12 different azimuth angles spaced  $30^\circ$  apart from  $0^\circ$  to  $360^\circ$ . With the benefit of varying viewing angles in the pose shape baseline, we were able to exploit action label and view angle at the same time and produce some interesting findings. Compared to other datasets, the large numbers of subjects and viewing angles introduce a good level of shape and action style variations, even though the number of action types is only three.

Considering that TMSD is a global feature and LIDAR data can have a wide range of scale and resolution variations, we also thoroughly evaluated our approach's capability to achieve scale and resolution normalization, especially for very low resolution cases that are common in long distance surveillance and target recognition. Finally, we compared the performance of action recognition using 3D TMSD of point clouds with that of using 2D Histogram of Oriented Gradients (HOG) [8] of depth images. This comparison was to support our assertion on the advantage of 3D shape analysis over 2D depth image analysis.

The main contribution of this paper is the introduction of TMSD and BoPS for action recognition from point cloud patches of human pose shapes, under various viewing angles and scales. To our best knowledge, there was no significant study on exploring 3D discrete orthogonal moments for action recognition. Another contribution is the demonstration of the advantage of native 3D shape analysis over 2D depth image analysis in terms of classification performance and consistency for low-resolution point cloud data.

The rest of this paper is organized as follows. Section 2 provides a brief overview of related action recognition models. Section 3 presents the characterization of point cloud patches in the form of TMSD. Section 4 describes a new bag-of-pose-shapes (BoPS) scheme for action representation and inference. Our experimental results are given in section 5, and section 6 includes some conclusions.

## 2 Related work

On the general 3D shape representation and characterization of partial point clouds of LIDAR data, we provided an extensive discussion of different methods in our previous study [5], especially the discrete orthogonal moments. Therefore, in this section, we focus only on feature representations in the context of action recognition from low resolution point clouds.

We categorize action recognition methods into two broad groups: spatial-temporal features/templates and temporal dynamics models. The former could be further divided into global and local features/templates, and the latter could be further divided into temporal state and temporal statistics models. Our discussion here is primarily to provide some background to

our proposed TMSD plus BoPS approach for 3D shape-based action representation. For more comprehensive and broad reviews of action recognition, readers are referred to [33, 49] for 2D imagery and [1, 56] for depth images.

## 2.1 Spatial-temporal features and dynamic templates

In many action recognition studies, the global or local spatial-temporal features and dynamic templates are often considered as 3D features — 2D spatial plus 1D temporal components in the form of  $(x, y, t)$ , which is different from the 3D convention of  $(x, y, z)$  used in this study.

The global spatial-temporal features are typically computed from the derivatives or differences between consecutive frames. The global dynamic templates are made up by stacking up frames over time. Some of the representative 2D imagery examples are spatial-time derivative statistics of optical flow [10], dynamic silhouette templates in the forms of motion energy and history images [4], and space-time shape [11]. Their 3D variations were introduced in [3, 28, 44]. Among them, 3D optical flow may be ill-suited for our application because it is difficult to obtain point registration and stable derivatives from sparse, degenerated point clouds. Occupancy-based dynamic silhouette templates are applicable if they are implemented over a 3D grid. Their potential shortcoming is that they are grid-based 4D representations which may result in large feature vectors involving high computational cost and high dimensionality.

Unlike global features and templates, local feature representations are based on the local spatial information collected along the temporal axis at the points selected by gradient-based maxima detectors. Some of the representative 2D imagery examples are space-time interest points [18] and the spatial-temporal descriptor of Histogram of Gradients (HOG) [16]. Some recent 3D features are: a bag of sampled 3D points [20], the random occupancy pattern using sampled sub-volumes [47], and the depth motion map from depth image projection to multiple orthogonal planes [54]. In general, local feature representations have the advantages of scale invariance and robustness under occlusions.

For low-resolution and irregular LIDAR data, a potential problem of local feature representation is the difficulty in identifying any meaningful spatial extremity, maxima of curvature, and inflection point to use as a key point. In addition, some local features require stable and smooth local surface approximations around key points which are difficult to obtain from degenerated and sparse point cloud patches. Therefore, some local features, such as those introduced in [16, 20, 54], are often generated using a sampling or a grid over depth images and then aggregated globally. However, this approach results in semi-global representations which face the similar high computational cost and dimensionality problems of global features, because they rely on large number of local components to achieve a similar spatial granularity as their global counterparts. Even though some data reduction techniques, such as the Principal Component Analysis (PCA), have been used to reduce the dimensionality, the resulted subspaces may not be consistent [13] because typical training datasets are limited in size compared to the feature dimensionality. To avoid this problem, we did not attempt to develop a 4D descriptor composed of TMSD and time; instead, we incorporated the temporal information through our BoPS scheme.

Recently, deep neural networks, particularly Convolutional Neural Networks (CNN), have demonstrated great success in image classification by automatically learning multi-level of complex features through layers of trainable filters and feature pooling operations over large-volume datasets [17, 40]. The framework was extended to action recognition by applying 3D CNN (spatial plus temporal) or multi-stream CNN over short video clips to learn spatial-

temporal features for feature-aggregated classification [12, 14, 36]. For more challenging tasks of video interpretation, recurrent neural networks, such as Long Short Term Memory (LSTM) was introduced on top of CNN to further learn and decode long-term temporal clues [9, 45, 55]. However, the inputs to these CNNs were typically limited to 2D images. The application of 3D CNN to 3D data has not been studied extensively, except for a few studies on 3D object classification [24, 51].

## 2.2 Temporal dynamics models

Among the temporal dynamics models, the subgroup of temporal state models applies a probabilistic graph to model joint (generative) or conditional (discriminative) probability distribution for temporal state transitions. The temporal states are typically encoded using the temporal labels of action context or the part-based body models in the form of joint location and joint angle profiles. Some of the representative graph models are hidden Markov model (HMM) [53], maximum entropy Markov model (MEMM) [25], and conditional random field (CRF) [38]. Theoretically the graph-based temporal state models are capable of modeling the details of a wide range of motions. However, they often encounter tractability issues in learning and inference, thus have to make assumptions that greatly limit their expressiveness. For our application, these models are overly complicated to estimate, so they were not used.

The subgroup of temporal statistics models usually works with a collection of local features using the BoW framework, which was originated from text categorization and extended to 2D image segmentation and categorization [37] as well as 2D image-based action recognition [34]. In the BoW framework, the temporal information is implicitly encoded using a vocabulary of feature words that are learned through an unsupervised quantization of the feature space. The temporal statistics models using BoW ignore the temporal orders, so may not be as discriminative as the temporal state models. However, in real-world scenarios, the exact temporal profile of an action could be affected by many factors, such as sensing rate, detection and tracking performance, action segmentation, and varying action style and speed, etc. Therefore, we may not be able to acquire consistent temporal order information anyway; hence, the BoW framework is a good alternative with more robustness, simplicity, and flexibility.

The basic concept of BoW also works for global features, although this type of usage has been much less than its usage with local features. In [46], global optical flows of human figures are combined with BoW for 2D video-based action recognition. In this case, the BoW representation is frame-based; i.e., each frame is a word. Our framework of TMSD plus BoPS has a similar setup: we used the proposed TMSD for the representation of individual frames of pose shapes and BoPS for the encoding of temporal statistics.

In the broader domain including 3D shape analysis, a few studies adopted BoW for shape retrieval by quantizing scale-invariant feature transform (SIFT)-based local features extracted from multi-view projected images [21, 31]. BoW was also used for non-rigid 3D shape retrieval with local extremity point based features, such as the heat kernel signature (HKS) [32] and local patch surface model [41]. In general, applications of BoW in 3D domain are still mainly with local features from dense surface models. Our BoPS approach is the first to explore the potential of BoW for action recognition from the perspective of sequences of global 3D shapes.

The classification performance of the BoW framework could be enhanced by coupling more sophisticated inference models such as topic models [29, 46]. The topic models allow

modeling complicated, hierarchical joint distributions which fit well to the multi-level semantics of human activities. Since the actions in our pose shape baseline are well segmented atomic actions, more efficient and less complicated Naïve Bayes models were used for action learning and inference.

### 2.3 Orthogonal moments for action recognition

Most of previous applications of orthogonal moments are for 2D image analysis. The primary focus was on 2D radial kernel based (rotation-invariant), continuous orthogonal moments, such as Zernike moment [42], pseudo-Zernike moment [43], and Fourier-Merlin moment [35]. Among them, the best-performing Zernike moment has been introduced as a feature for action classification from 2D images. Using the BoW framework, Sun et al. [39] investigated the classification performance of different combinations of SIFT-based local features and Zernike moments of individual frames. Costantini et al. [7] used Zernike moments to form a multi-scale kernel descriptor of the space-time shape of local patches [18], although the subsequent classification was based on a nearest neighbor search that is more like a shape retrieval process. Lassoued et al. [19] proposed a Zernike moment representation for the global space-time volume of action silhouettes.

Compared to the discrete counterparts, the continuous orthogonal moments have a problem of shape approximation error due to the difference between their continuous radial kernels and the discrete nature of digital images. In the family of discrete orthogonal moments, Tchebichef moment [27] demonstrates superior 2D image reconstruction performance, compared with Zernike moment. It has been used for action recognition in 2D imagery [22] and for modeling motion energy and history images [4]. In general, there has been little interest in exploring orthogonal moments in 3D domain, except for a few studies on 3D shape retrievals using Zernike moments [30] and Krawtchouk moments [23].

## 3 Scale-invariant Tchebichef moment shape descriptor (TMSD)

### 3.1 3D pose shape baseline for human action recognition

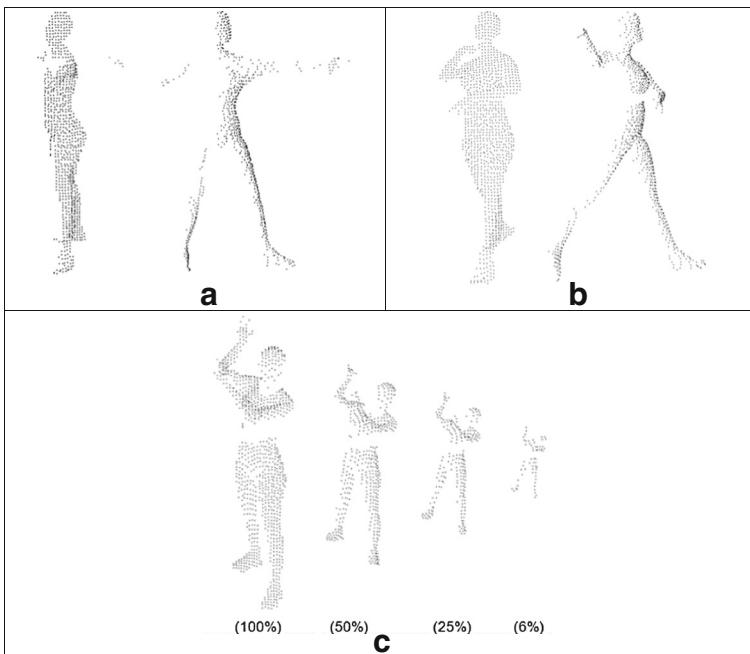
With the release of low-cost range cameras such as Kinect, new depth image datasets were generated for the purpose of human action recognition. Some of publically available ones are MSR action 3D [20], MSR daily activity [48], LIRIS human activity [50], and UT Kinect action [52]. These depth images were acquired in much closer ranges (< 4 m) than the typical operational range of low-grade commercial LIDARs (80 ~ 100 m). Their resolutions are also higher than those offered by typical LIDARs. In addition, they have limited variation in individual anthropometry, action style, and viewing angle.

In light of these limitations, we developed a simulated LIDAR baseline of human action pose shapes using a hybrid experimental/modeling approach. Unlike many common avatar animations produced by artists, action simulation in our baseline is individualized with respect to each human test subject. We first created the digital human model of a subject by rigging his/her full-body scan with the skeleton estimated from his/her anatomical landmarks and motion capture markers. We then reproduced the full-body action by driving the subject's digital model with the joint angle time history derived from his/her motion capture of individual actions. Finally, we applied the ray tracing

using a 100-by-100 detector array to simulate the LIDAR illumination over the human subjects, and captured the corresponding point cloud patches in every other frame at a frame rate of 15 Hz. This process was repeated at 12 evenly-spaced horizontal azimuth angles (every 30° from 0° to 330°) and two vertical elevation angles (where 0° represents ground platforms, and 45° represents aerial platforms). Figure 1 shows some examples of such point cloud patches, magnified in size and rendered using Blender.

The development of this shape baseline was initiated in our previous study for 9 subjects with ground truthing on pose shape class labels, which were conveniently used in this study for BoPS vocabulary learning. In order to train and test the action classifier, the size of the original baseline was increased from 9 to 62 subjects. Currently, the baseline is organized into two subsets according to the two elevation angles. Each subset has the same 62 subjects (25 females and 37 males) with 47,398 point cloud patches of the three types of action. Moreover, to facilitate the research on scale and resolution invariance, 12 subjects (6 males and 6 females) were randomly selected from the 62 subjects to produce simulated, scale-reduced LIDAR captures at 50%, 25%, and 6% of the detector panel area, as shown in Fig. 1c.

The large number of subjects and viewing angles in the baseline provides some randomness in pose shapes, resulted from varying anthropometry, viewing angle, scale, and action style. For example, the initial throwing poses of two subjects shown in Fig. 1a and b are very different because different people have different action styles.



**Fig. 1** Partial point cloud examples of some initial poses of two female subjects at 0° azimuth angle: (a) subject 1057 throwing, 0° elevation angle, (b) subject 1075 throwing, 0° elevation angle, and (c) subject 1057 digging, 45° elevation angle, varying scales. In (a) and (b), the left drawing is the view from the simulated sensor and the right one is a 90° rotation of the left for better illustration purpose. In (c), the drawing is rotated upward to expose the occlusions caused by the left arms

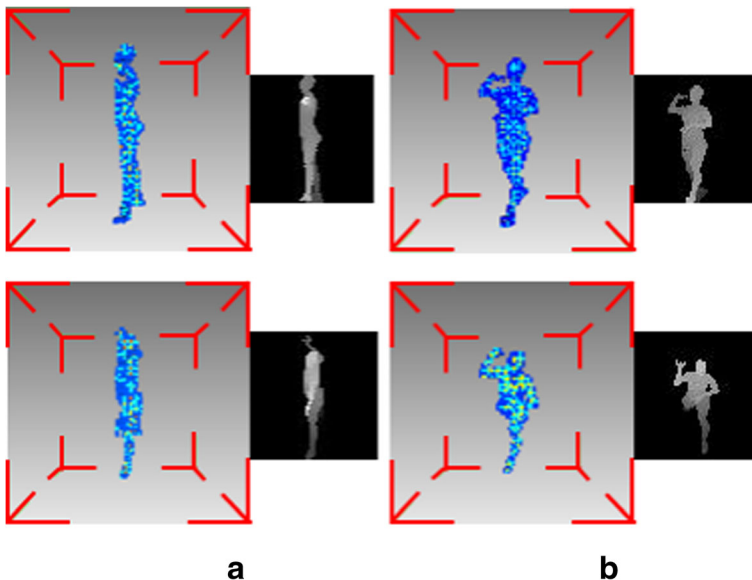


### 3.2 Scale-invariant voxelization and normalization of point clouds

The shapes of raw point clouds are not translation and scale normalized. Even for the full-scale captures in the baseline, there are variations resulted from the initial uncalibrated rough positioning of the simulated detector array during the data capturing process as well as the body size difference among human subjects. In addition, there is a resolution variation in the form of varying global point density among different sets of point cloud captures, because of different sensor resolutions in real-world 3D applications or different mesh refinements in the case of simulated digital models. The scale and resolution variations are typically intertwined.

To normalize these variations, we employed our recently-developed voxelization scheme, called Proportional Grid Voxelization and Normalization (PGVN) [5], to approximate a point cloud with a discrete volumetric point counting (shape) function  $f(x, y, z)$ . It consists of the voxelization with a one-side bounding box originated at the center of mass of a point cloud and the normalization of the total point cloud mass to a fixed value. The size of this bounding box is determined by the longest semi-axis from the center of mass to the boundary of a point cloud, hence the body is only one-side bounded.

Figure 2 presents renderings of some PGVN examples of the initial throwing pose shapes of two female subjects. The voxelization grid size is  $N = 64$ , which is a common voxelization setting of  $64 \times 64 \times 64$ , used in 3D shape analysis [15, 23] and reasonable for the coarse nature of LIDAR data. Our previous experiments [5] also indicate that the grid size does not make any observable changes in shape retrieval performance. The proportional bounding boxes are the red boxes around the picture edges. The one-side bounding can be seen more clearly in the depth images corresponding to the voxelization. Note that some of the voxelization density unevenness shown in the figure is due to the difficulty in achieving and maintaining a strictly uniformed mesh during the capture of simulated data, which actually makes the simulated data closer to the real-



**Fig. 2** Examples of PGVN 3D shapes of initial throwing poses: (a) subject 1057 and (b) subject 1075, viewed from  $0^\circ$  azimuth and  $0^\circ$  (1st row) or  $45^\circ$  (2nd row) elevation angles. The corresponding depth images are intended to show one-side bounding



world LIDAR signal with random noise. The moments computed with respect to  $f(x, y, z)$  under PGVN grid reference should be translation, scale, and resolution invariant.

### 3.3 Tchebichef moment shape descriptor (TMSD)

In [5], we proposed TMSD to approximate the global pattern of point cloud patches in an embedded subspace using low-order Tchebichef moments. Moments in general can be defined as an inner product projection of a real function  $f$ , such as the aforementioned shape function, to a set of basis (kernel) functions. For Tchebichef moments, its basis function set consists of the family of discrete orthogonal Tchebichef polynomials [6] which support the completeness and orthogonality. The completeness and orthonormality means a unique decomposition of  $f$  with respect to the basis set and also a least-squared reconstruction of  $f$  from the corresponding set of moments.

The  $n$ -th order discrete Tchebichef polynomial,  $t_n(x)$ , can be expressed in the form of a generalized hypergeometric function  ${}_3F_2(\cdot)$  [6] as:

$$\begin{aligned}
 t_n(x) &= (1-N)_n {}_3F_2(-n, -x, 1+n; 1, 1-N; 1) \\
 &= (1-N)_n \sum_{k=0}^n \frac{(-n)_k (-x)_k (1+n)_k}{(k!)^2 (1-N)_k},
 \end{aligned}
 \tag{1}$$

where  $n, x = 0, 1, \dots, N-1$ , and  $(a)_k$  is the Pochhammer symbol. In our case,  $N$  is the size of a 3D  $(N \times N \times N)$  voxelization grid, and  $x$  corresponds to one of their coordinate variables. Correspondingly, the orthogonality is defined as:

$$\sum_{x=0}^{N-1} t_n(x)t_m(x) = \rho(n, N)\delta_{nm},
 \tag{2}$$

where  $m, n = 0, 1, \dots, N-1$ ;  $\delta_{nm}$  is the kronecker symbol, and  $\rho(n, N)$  is a normalization function given as follows [6]:

$$\begin{aligned}
 \rho(n, N) &= (2n+1)^{-1}N(N^2-1)(N^2-2^2)\cdots(N^2-n^2) \\
 &= (2n)!\binom{N+n}{2n+1}.
 \end{aligned}
 \tag{3}$$

The corresponding order-scale normalized Tchebichef polynomials is [27]:

$$\tilde{t}_n(x) = \frac{t_n(x)}{\sqrt{\rho(n, N)}}.
 \tag{4}$$

Taking  $\{\tilde{t}_n(x)\}$  as the basis function set, an individual 3D Tchebichef moment of order  $(n + m + l)$  for the voxel mass distribution  $f(x, y, z)$ , over an  $N \times N \times N$  grid, can be defined as the inner product functional between  $\{\tilde{t}_n(x)\}$  and  $f$ :

$$T_{nml} = \sum_{x=0}^{N-1} \sum_{y=0}^{N-1} \sum_{z=0}^{N-1} \tilde{t}_n(x)\tilde{t}_m(y)\tilde{t}_l(z)f(x, y, z),
 \tag{5}$$

where  $0 \leq n, m, l \leq N-1$ . There are total  $N^3$  number of  $T_{nml}$ s with the maximum order of  $3 \times (N-1)$ . Among them, a small subset consisting of the first  $R$ -th order moments,  $R \ll N^3$ , is used to form the 3D Tchebichef moment shape descriptor (TMSD):

$$\text{TMSD} = [T_{001}, T_{010}, T_{100}, \dots, T_{nml}, \dots, T_{R00}]^T,
 \tag{6}$$

where  $0 < n + m + l \leq R$ .

Excluding the constant zero-order term, if  $R < N$ , the dimension of TMSD is  $\frac{1}{6}(R+1)(R+2)(R+3)-1$ . We can reconstruct shapes in varying details using TMSD up to different orders [5]. Using the full set of Tchebichef moments (maximum order  $R=128$ ), we can achieve an exact duplicate of the voxelization (PGVN) of a point cloud. Using the subset of low-order moments, we can approximate the general pattern of a pose shape. In other words, TMSD provides a multi-scale shape representation for the point cloud patches.

Our previous experiments found that the optimal TMSD order for pose shape representation is  $R=16$  [5], which reduces the problem's dimension from the original voxel model's 262,144 voxels in a grid size of  $N=64$  to mere 968 moment components in TMSD. This compact TMSD representation captures the intrinsic dimensions of a pose shape and is particularly suited for aerial mobile platforms. Utilizing the sparsity in the low-resolution point cloud data [5], together with the recurrence and symmetry relationships of Tchebichef polynomials [6, 27], TMSD could be computed efficiently and stably.

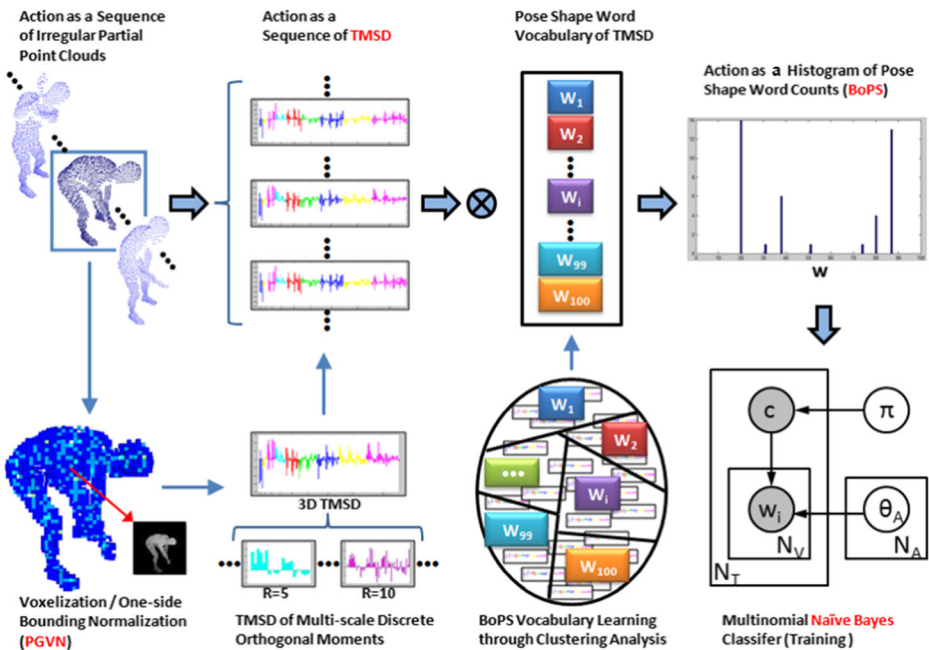
More importantly, this approximation decouples and aligns the spatially correlated point distributions into low-order 'modes' determined solely by the basis function  $\{\tilde{t}_n(x)\}$ . That means, unlike the aforementioned dimension reduction techniques such as the PCA, TMSD's spanning of the embedded subspace is consistent and not dependent on datasets. This is particularly valuable for the task of shape clustering analysis in our BoPS approach. Moreover, the orthonormality of TMSD guarantees no false dismissal of qualified nearest neighbors in its subspace query returns [5], which is a necessary condition for clustering analysis and nearest neighbor search on TMSDs. The former is used for learning BoPS vocabulary and the latter is needed for assigning pose shape words.

## 4 Action representation and inference using a bag of pose shapes (BoPS)

Our action recognition consists of four steps in the order of PGVN, TMSD, BoPS, and Naïve Bayes, corresponding to shape normalization, feature representation, temporal modeling, and action classification, respectively, as illustrated in Fig. 3. The block arrows in the figure indicate the solution path for action recognition from raw point clouds, whereas the line arrows are associated with the intermediate supporting functions. Starting from the left side of Fig. 3, the PGVN process shown in the first column normalizes each frame of raw point cloud patches in the action sequence into a grid-based voxelization. This is followed by the computation of Tchebichef moments to form the corresponding TMSD, shown in the second column as a bar plot. This procedure is repeated for each frame and results in a collection of TMSDs representing the pose shapes forming the action. These TMSDs are then mapped to pose shape words  $\{w_j\}$  through nearest neighbor search against the pose shape vocabulary, shown in the third column. Finally in the right-most column, a histogram counting the pose shape words is used to classify the action type. The steps in the last two columns are further explained in sections 4.1 and 4.2.

### 4.1 Action as a bag of pose shapes

For a general bag-of-words (BoW) representation, the quantization of feature space into a fixed vocabulary of words is performed typically through an unsupervised clustering analysis. Using the vocabulary, individual features of an input spatial-temporal sample are replaced by their closest words according to some distance measure. A histogram of word counts is formed to



**Fig. 3** Action recognition pipeline of PGVN + TMSD + BoPS + Naïve Bayes. ⊗ stands for nearest neighbor search

provide a compact representation of the spatial-temporal pattern in the input sample. The histogram can then be used for pattern inference or classification. A clear benefit of the BoW scheme is the mapping of a large and varying number of high-dimensional feature vectors into a fixed low-dimensional context space. Thus, it offers natural flexibility and scalability in handling widely different and unknown inputs.

We extended the general BoW concept to our BoPS representation of action by mapping the sequences of pose shape point clouds to a pose shape vocabulary. More specifically, suppose that a clip of point cloud patches of an atomic action, called an action clip, can be characterized by a sequence of pose shapes  $\mathcal{S} = \{s_1, s_2, \dots, s_f, \dots\}$  where  $s_f$  is either our Tchebichef moment shape descriptor (TMSD) for a point cloud or the comparative HOG-based shape descriptor (HSD, see section 4.3 for details) for a depth image at frame  $f$ , then the BoPS representation of the action clip can be defined in the context of pose shape quantization as follows. If the pose shape descriptor space is quantized into a vocabulary of  $V = \{w_1, w_2, \dots, w_{N_V}\}$ , where  $w_j$ , the virtual pose shape word, is the index to the  $j$ -th cluster of the  $N_V$  clusters produced by the quantization, the pose shape  $s_f$  can be mapped to its corresponding pose shape word  $w_j$  as:

$$s_f \mapsto w_j = \underset{w \in V}{\operatorname{argmin}} d(s_f, s_w) \tag{7}$$

where  $d(s_f, s_w)$  is a proximity function measuring the Manhattan distance between  $s_f$  and the mean descriptor  $s_w$  of cluster  $w$ . Subsequently, an action clip  $x$  can be represented in the form of BoPS by collecting its visual pose shape words into a histogram:

$$\mathcal{S} \mapsto \mathbf{x} = \{m_j = |w_j|, j = 1, 2, \dots, N_V\} \tag{8}$$

where  $|w_j|$  denotes the cardinality of word  $w_j$ . The mapping in Eq. (7) is achieved through a nearest neighbor search on TMSD or HSD of pose shape against the learned pose shape vocabulary.

For our application, each elevation angle has its own vocabulary. Using the  $k$ -means clustering algorithm, a vocabulary was learned from 9 subjects (5 males and 4 females) randomly selected from the 62 subjects in the pose shape baseline. These 9 subjects were not used in the later classifier learning, cross-validation, and testing. The size of the vocabulary may affect the classification performance. Too few words may decrease the discriminative power of the BoPS representation, whereas too many words may cause over-fitting. For the  $k$ -means clustering, we tested the  $k$  values of 100 and 400. The overall classification accuracy difference between the two is less than 2% improvement with 400-word vocabulary, which is not a significant benefit, considering the much higher computational cost for word matching and potential generalization issue. Thus, the 100-word vocabulary is used for later experiments.

### 4.2 Naïve Bayes classifier for action recognition

An action classifier takes an input vector  $\mathbf{x}$  that encodes an action clip and outputs a scalar  $y$  representing the action category (label). It is a hypothesis function  $\mathcal{H}$  whose parameters are learned through a set of training observation pairs  $\{(\mathbf{x}_i, y_i) \mid i = 1, \dots, N_T\}$ , by minimizing an empirical error  $\sum_{i=1}^{N_T} err(\mathcal{H}(\mathbf{x}_i), y_i) / N_T$ .

In this study, the hypothesis function is the generative Naïve Bayes model. The main consideration for this choice was: 1) the semantics and assumption of BoPS correspond to a Naïve Bayes model where the action label is the hidden node and the pose shape words are the observable nodes, and 2) Naïve Bayes is very efficient to implement for our BoPS-based multiclass classification problem, in which the dimensionality is still relative high even after the quantization of the original feature space, and the size of the vocabulary would grow if more actions are added later. Even though the class independence assumption for pose shape words may not be true in the context of an individual’s specific action, it holds better across the population pool, considering the style variation among people.

Specifically for our application, the input feature vectors are action clips in the form of Eq. (8). Assuming that there are  $\mathbf{C} = \{c_k \mid k = 1, 2, \dots, N_A\}$  action class labels and the lengths of action clips are independent of action labels, the Naïve Bayes assumption could lead to several models of factorizing  $P(\mathbf{x} \mid c_k)$  into the products of  $P(w_j \mid c_k)$  [26]. Among them, the multinomial distribution model demonstrates better performance because it models the word frequency as:

$$P(\mathbf{x} \mid c_k) = P(|\mathbf{S}|) |\mathbf{S}|! \prod_{j=1}^{N_V} \frac{P(w_j \mid c_k)^{m_j}}{m_j!}. \tag{9}$$

Consequently, the classification can be accomplished through the maximum a posteriori rule:

$$y = \underset{c_k}{\operatorname{argmax}} P(c_k) \prod_{j=1}^{N_V} P(w_j \mid c_k)^{m_j}. \tag{10}$$

In this study, we assumed uniform prior probabilities; hence the end result is a maximum likelihood solution. We also limited the pose shapes in an action clip to have the same azimuth and elevation viewing angle.

Using a training set  $\{(\mathbf{x}_i, y_i) \mid i = 1, \dots, N_T\}$  where  $i$  indexes individual training action clips,  $P(w_j | c_k)$  can be estimated with Laplace smoothing as:

$$P(w_j | c_k) = \frac{1 + N_{kj}}{N_V + N_k}, \quad (11)$$

where  $N_{kj} = \sum_{i=1}^{N_T} m_{ij}[y_i = c_k]$  is the number of  $w_j$  in the action clips belonging to action class  $c_k$ , and  $N_k = \sum_{s=1}^{N_V} \sum_{i=1}^{N_T} m_{is}[y_i = c_k]$  is the total number of words in action class  $c_k$ .  $[\bullet]$  is the Iverson bracket.

For our experiments, we set aside the data of 41 subjects in the pose shape baseline as the training set for three classes of actions — digging, jogging, and throwing. In addition, we also looked into the scenario of inferring both action and azimuth viewing angle at the same time by changing the class label to a tuple of <action, azimuth angle>, in which azimuth angle indexes one of the 30° azimuth intervals in the pose shape baseline. This results in 3 actions  $\times$  12 azimuth angles = 36 classes for each elevation angle. Note that this is a simplified model, compared to a more formal representation of two separate hidden nodes of action and viewing angle, respectively. However, since each action clip has the same viewing angle and our baseline is well-balanced in both action and viewing angle distributions, this model is semantically and statistically sound. Probably it is also valid for real-world scenarios because these atomic actions are typically executed within a second or so, during which the sensor platform may look stationary at a distance.

Other study [26] found that multinomial distribution with Boolean values for  $w_j$ , which is equivalent to modeling word appearance only, outperforms the word frequency counterpart in many text categorization cases. This is due to the fact that  $P(w_j | c_k)$  has a Poisson distribution under multinomial model, which may not fit well to the burstiness of the same word in many situations, including action recognition. Thus, we evaluated both word appearance and word frequency based multinomial models. Note that multinomial Boolean value model is different from another common binary model of word appearance — multivariate Bernoulli model, which tends to perform worse because of the counts of negative events, i.e., the absent pose shape words.

### 4.3 2D histogram of oriented gradients (HOG) based representation of depth images

To test our hypothesis that TMSD performs better than typical 2D depth image analysis in action recognition, we converted each pose shape point cloud into a depth image. These depth images were then subject to a similar four-step action recognition pipeline aforementioned, but with different 2D-based normalization and features. Note that this conversion is done before the voxelization of point clouds, thus these depth images are not limited by the grid size of  $N = 64$ . They can have up to 100 pixels along each dimension.

We chose the Histogram of Oriented Gradients (HOG) [8] as the features for these depth images and named the representation as 2D HOG-based shape descriptor (2D HSD). For a full-scale depth image captured by our 100-by-100 simulated LIDAR detector, its 2D HSD has a size of 5184. Readers are referred to [8] for configuration details of HOG. Like the scale normalization for computing TMSD, we also need to normalize the scale of depth images to the full scale of 100-by-100 pixels before computing HSDs, especially for those with reduced scales. To that end, a depth image is first padded symmetrically to a square image and then isometrically enlarged to

100-by-100 pixels. This normalization provides scale invariance that allows us to train the classifier once through the full-scale baseline and then use it in varying scales.

## 5 Experimental results

### 5.1 Experiment setup and performance measures

Our experiments presented here were designed to test various comparative hypotheses regarding the performance on action recognition and azimuth viewing angle identification. They were conducted separately for two elevation angles of  $EL = 0^\circ$  and  $EL = 45^\circ$ . We divided the pose shape baseline of 62 subjects into the following three groups for each elevation angle:

- (1) 9 subjects (5 males and 4 females) consisting of 5890 point clouds were used for learning the vocabulary of pose shape words.
- (2) 41 subjects (26 males and 14 females) consisting of 32,088 point clouds in 1476 action clips were used for classifier training and cross validation.
- (3) 12 subjects (6 males and 6 females) consisting of 9420 point clouds in 432 action clips were used for independent testing on scale invariance.

Three groups of experiments were conducted to evaluate/compare the followings: 1) word frequency model vs. word appearance model, 2) 3D TMSD classifier vs. 2D HSD classifier, and 3) scale invariance of 3D TMSD classifier vs. that of 2D HSD classifier. The Manhattan distance ( $L_1$  norm) is used in both pose shape quantization and action clip (histogram) formation because it behaves better than the Euclidean distance under high dimensionality [2], which was confirmed by our previous experiments on shape retrieval [5].

The classification performance is quantified through cross validation using the confusion matrix (aka contingency table) as well as the classification accuracy rates (i.e., percentages of correctly classified), denoted by  $ACR_a$  and  $ACR_{av}$  for action only and for action plus azimuth viewing angle, respectively. For azimuth viewing angle identification, since the 30-degree interval is rather arbitrary, we also looked into an expanded interval of 90-degree azimuth angle that consists of a 30-degree interval and its two adjacent intervals bordering each side of the 30-degree interval. This leads to the quadrant azimuth angle accuracy, denoted by  $ACR_{av}^{90}$ , which treats the confusion matrix's elements in the band of the diagonal and two off-diagonals as the correct assignments. In real-world applications, this quadrant azimuth angle may represent general viewing direction.

The cross validation employs a random 10-fold split over the 41-subject training set. We observed that the performance differences between different 10-fold splits were less than 1%. So, we did not do any averaging over multiple 10-fold cross validation runs.

### 5.2 Word frequency vs. word appearance for the action and viewing angle recognition

Table 1 presents the cross validation results of the word frequency and word appearance models with the 100-word vocabulary. The word appearance model performs slightly better in action recognition, but the word frequency model performs slightly better for the azimuth angle recognition. For action recognition, this trend could be further seen in the confusion

**Table 1** ACR results of 3D TMSD classifiers from cross validations between word frequency and word appearance models with 100-word vocabulary

Models	EL	ACR <sub>a</sub>	ACR <sub>av</sub>	ACR <sub>av</sub> <sup>90</sup>
Word Frequency	0°	96.6%	73.3%	95.3%
	45°	97.0%	74.3%	95.5%
Word Appearance	0°	97.2%	72.0%	94.5%
	45°	97.5%	72.4%	94.0%

matrices shown in Table 2, in which each action has total 492 training clips. The majority of the misclassifications are throwing assigned falsely to jogging or, to a lesser degree, to digging. This may be caused by the facts that 1) some throwing pose shapes may resemble a few jogging or digging pose shapes at certain viewing angles, and 2) throwing tends to have fewer pose shape words. Consequently, when a frame of throwing point cloud is mapped to a wrong word, there is a greater chance that adjacent frames and, hence, a larger portion of the action clip may also be mapped to the same wrong word. The second fact may partially reveal the problem of the burstiness of the same word in the frequency model.

Even though the performance enhancement is within the variation of cross validations, we found that it is consistent for action recognition across different vocabularies and multiple cross validation runs. Since our primary concern is action recognition, we selected the word appearance model as our default classification model in later experiments.

Table 1 also indicates that the quadrant azimuth angle accuracy for azimuth angle recognition,  $ACR_{av}^{90}$ , is significantly higher than the corresponding accuracy of 30-degree interval recognition. This makes us believe that our 30-degree interval may be too refined for inter-class separation of pose shapes, and a larger 90-degree interval is a better choice.

Overall, our experimental results demonstrate very consistent prediction capability with respect to the varying elevation angles up to 45°, which is an important merit for aerial applications. Regarding the varying azimuth angle, we can achieve good performance with the quadrant azimuth angle, which is useful in real-world applications where we do not know the actual orientation of a target with respect to the sensor platform. Considering the simplicity of the Naïve Bayes classifier and degeneracy of the point clouds, these superb performance and consistency prove the power of 3D TMSD and BoPS approach in characterizing dynamic patterns of human actions.

**Table 2** Confusion matrices of TMSD classifiers from cross validations (percentages) of word frequency and word appearance models using 100-word vocabulary

EL	Word Frequency					Word Appearance				
0°										
	Predicted	Dig	Dig	Jog	Throw	Predicted	Dig	Dig	Jog	Throw
	Jog	0	100	10.2		Jog	0	99.6	5.5	
	Throw	0	0	88.2		Throw	0.6	0.4	92.5	
45°										
	Predicted	Dig	Dig	Jog	Throw	Predicted	Dig	Dig	Jog	Throw
	Jog	99.8	0	1.6		Jog	99.6	0	1.0	
	Throw	0	100	7.5		Throw	0	98.8	4.9	
		0.2	0	90.9			0.4	1.2	94.1	



**Table 3** ACRs of 3D TMSD and 2D HSD classifiers from cross validations using word appearance model with 100-word vocabulary

Features	EL	ACR <sub>a</sub>	ACR <sub>av</sub>	ACR <sub>av</sub> <sup>90</sup>
3D TMSD	0°	97.2%	72.0%	94.5%
	45°	97.5%	72.4%	94.0%
2D HSD	0°	96.7%	53.0%	80.0%
	45°	94.4%	69.7%	91.7%

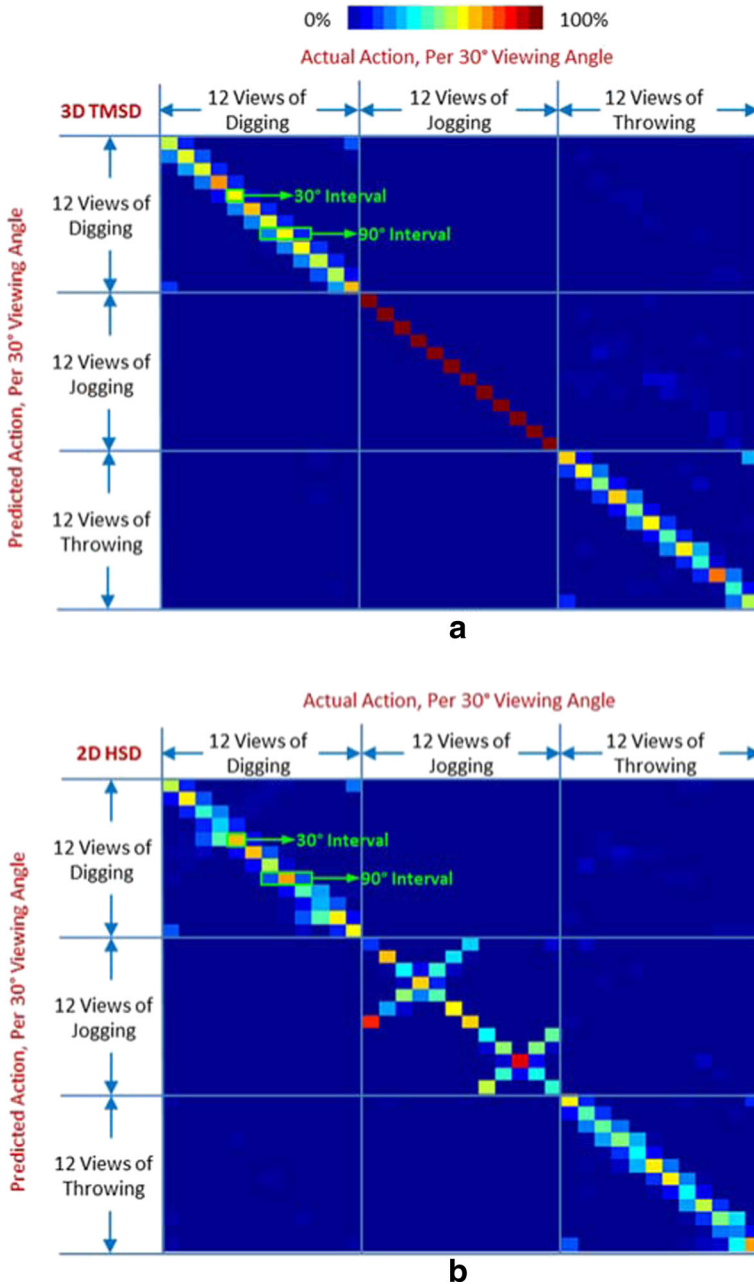
### 5.3 Performance comparison between 3D TMSD and 2D HOG-based shape descriptor (2D HSD)

This experiment was designed to provide a comparison in classification performance between our 3D based method and the traditional 2D feature based depth image analysis. Table 3 presents the cross-validation on the classification accuracies of 3D TMSD classifier against 2D HSD classifier for action and viewing angle recognitions using the word appearance model. On the accuracy of action only recognition —  $ACR_a$ , 2D HSD underperforms 3D TMSD by 1–2%. Although the difference is small, the trend is consistent across many cross validation runs. The similar conclusion can also be drawn from the confusion matrices presented in Table 4. There are more misclassifications between digging and throwing as well as between jogging and throwing for 2D HSD. Some of the cause is revealed in the following discussion where viewing angle results are included. The only exception is the case of telling throwing from jogging at the zero elevation in which 3D TMSD is less accurate than 2D HSD. We suspect that this may be caused by the similarity between a few persons' jogging poses and their beginning and ending poses of throwing.

On the classification of action plus azimuth viewing angle —  $ACR_{av}$  for 30° interval and  $ACR_{av}^{90}$  for quadrant azimuth angle — 2D HSD performs significantly worse than 3D TMSD, indicated by the much lower values of  $ACR_{av}$  and  $ACR_{av}^{90}$  for 2D HSD against those for 3D TMSD in Table 3. This is mainly because 2D HSD has difficulty in maintaining performance consistency for jogging over the viewing angle variation, evidenced by the larger spread of misclassification on jogging in the <action, azimuth angle> confusion matrix of 2D HSD presented in the middle panel of the 9-panel confusion matrix shown in Fig. 4b.

**Table 4** Confusion matrices of 2D HSD and 3D TMSD classifiers from cross validations (percentages) using word appearance model with 100-word vocabulary

EL	2D HSD					3D TMSD				
0°										
	Predicted	Dig	98.2	0	3.1	Predicted	Dig	99.4	0	2.0
	Jog	0	95.9	0.8		Jog	0	99.6	5.5	
	Throw	1.8	4.1	96.1		Throw	0.6	0.4	92.5	
45°										
	Predicted	Dig	92.9	0	4.5	Predicted	Dig	99.6	0	1.0
	Jog	0	97.6	2.6		Jog	0	98.8	4.9	
	Throw	7.1	2.4	92.9		Throw	0.4	1.2	94.1	



**Fig. 4** Confusion matrices of cross validations of actions plus viewing angles at 0° elevation angle using word appearance model with 100-word vocabulary. Each small cell represents a 30° azimuth interval, ordered from 0°–30° to 330°–360° along the diagonal for each action: (a) 3D TMSD classification on point clouds, (b) 2D HSD classification on depth images

Compared to other types of actions such as digging and throwing, jogging actually tends to vary much less on pose shapes across different people. Therefore, when aligned and modeled

**Table 5** Independent test of scale invariance using word appearance model with 100-word vocabulary

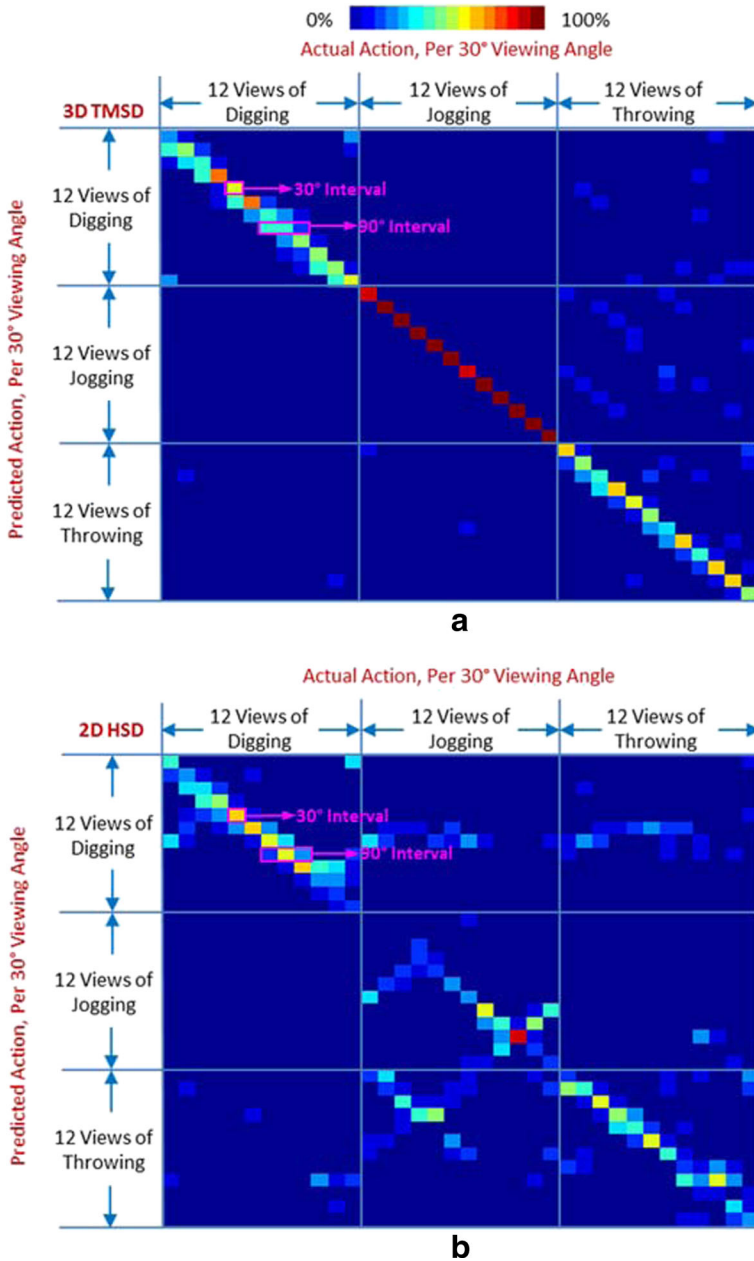
Descriptor	Scale	EL	ACR <sub>a</sub>	ACR <sub>av</sub>	ACR <sub>av</sub> <sup>90</sup>
3D TMSD	100%	0°	95.8%	69.2%	88.7%
		45°	94.7%	71.3%	88.0%
	50%	0°	95.4%	69.4%	89.1%
		45°	94.2%	71.3%	88.7%
	25%	0°	96.8%	69.2%	88.9%
		45°	94.9%	70.1%	88.4%
2D HSD	100%	0°	94.9%	68.3%	88.2%
		45°	93.5%	65.7%	87.3%
	50%	0°	96.0%	54.4%	76.8%
		45°	94.0%	63.7%	85.9%
	25%	0°	92.6%	52.1%	75.5%
		45°	91.9%	63.4%	84.0%
6%	0°	90.3%	48.4%	74.1%	
	45°	86.1%	59.3%	83.3%	
		0°	73.4%	34.7%	58.3%
		45°	63.4%	36.6%	58.3%

properly by capturing the true 3D spatial relationship, it should have better directional classification results than others. This is confirmed by the best and most consistent classification rate on jogging with respect to different viewing angles when modeled by 3D TMSD — close to 100% correctness even at a smaller 30-degree interval of azimuth angle, represented by the middle panel of Fig. 4a. This highlights the benefit and power of our 3D representation using TMSD.

On the other hand, the misclassification pattern in the corresponding middle panel of 2D HSD confusion matrix in Fig. 4b clearly reveals the problem of motion ambiguity encountered by 2D shape based features such as HSD; i.e., the classifier has difficulty to tell a person is jogging towards or away from the sensor. In other words, 2D HSD is good at capturing the prominent shape silhouette in a 2D depth image, but not the subtle variation of depth inside the silhouette. Even though shape silhouette is an important feature for action recognition, discerning of viewing angles may require capturing more subtle depth changes. In 3D TMSD, the depth dimension receives the equal treatment as the height and width dimensions, which alleviates this problem.

**Table 6** Confusion matrices (percentages) of independent test at 6% scale using word appearance model with 100-word vocabulary

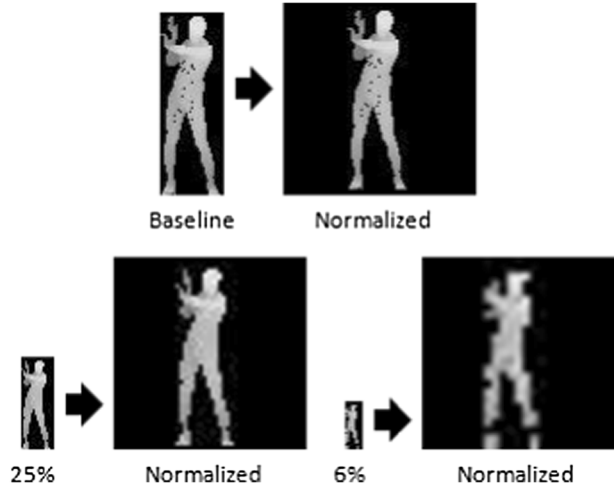
EL	3D TMSD				2D HSD				
0°	Actual				Actual				
	Predicted	Dig	Jog	Throw	Dig	Jog	Throw	Throw	
		Dig	96.5	0	4.1	Dig	93.7	9.0	11.8
		Jog	0.7	97.2	4.9	Jog	0	39.6	1.4
Throw	2.8	2.8	91.0	Throw	6.3	51.4	86.8		
45°	Actual				Actual				
	Predicted	Dig	Jog	Throw	Dig	Jog	Throw	Throw	
		Dig	96.5	0	3.5	Dig	91.7	21.5	43.8
		Jog	2.8	96.5	9.0	Jog	1.4	54.2	11.8
Throw	0.7	3.5	87.5	Throw	6.9	24.3	44.4		



**Fig. 5** Confusion matrices of actions plus viewing angle intervals for the 12-subject independent test at 6% scale and 0° elevation angle using word appearance model with 100-word vocabulary: (a) 3D TMSD classification on point clouds, (b) 2D HSD classification on depth images

Another advantage of 3D TMSD over 2D HSD is its compact size. At the order of  $R = 16$ , 3D TMSD has 968 components, regardless how the volume is voxelized. On the other hand, HSD size varies with the image size and implementation. In our cases, HSD has 5184 components, based on the typical HOG configuration for 100-by-100 pixel depth images.

**Fig. 6** Examples of depth image normalization for 2D HSD from various scales (elevation angle = 0°)



Almost five times smaller in size could reduce the computation cost considerably for TMSD during the BoPS mapping of Eq. (7) and Eq. (8). This would be a great value for real-time mobile applications. Finally, the two confusion matrices in Fig. 4 further support the conclusion that quadrant azimuth angles could be better discerned than the small 30-degree intervals.

In summary, 3D TMSD classifier has better performance, consistency, and efficiency than 2D HSD classifier. This supports our assertion that native 3D characterization of point clouds is superior to 2D characterization of depth images for analyzing low-resolution LIDAR-type data.

#### 5.4 Scale invariance comparison of 3D TMSD and 2D HOG-based shape descriptor (HSD)

This experiment compares the classification performance between 3D TMSD and 2D HSD classifiers, using the independent 12-subject test subsets of 4 different scales of 100%, 50%, 25%, and 6% (see Fig. 1c). There are total 144 action clips for each type of action in a test subset. The classifiers were trained using only the group of 41-subject training dataset in the baseline which is near full scale (see section 3.1). Thus, the classifiers do not have any clue on the varying scales of the independent test datasets. This arrangement allows us to compare our pipeline of PGVN + TMSD + BoPS + Naïve Bayes with the pipeline of depth image normalization + HSD + BoPS + Naïve Bayes, in terms of scale invariance.

Table 5 presents their classification accuracies at various scales. 3D TMSD demonstrates solid performance consistency, down to the scale of 6%. At this extremely small scale, a pose shape is hard for human eyes to discern, because the point cloud has no more than 90 points and the corresponding depth image is roughly equivalent to a body height of less than 20 pixels. In contrast, 2D HSD starts showing performance deterioration at 50% scale. At 6% scale, it has great difficulty in predicting jogging and throwing actions correctly, as shown in the confusion matrices in Table 6.

Figure 5 shows the confusion matrices for action prediction per viewing angle at 6% scale. It is safe to say that 3D TMSD classifier can still roughly detect viewing angles at this small scale, whereas 2D HSD classifier performs poorly. This result may not be quite conclusive due to the small number of action clips at each viewing angle interval (12 for each action type).

The significant performance difference between TMSD and HSD in terms of scale invariance also shows the advantage of native 3D spatial modeling of point clouds, compared to converting them into 2D depth images. Using our PGVN scheme, a reduced-scale point cloud is able to retain its local density and pairwise spatial relationships well in all three dimensions. By combining PGVN with the capability of Tchebichef moment in characterizing discrete density distribution, we can achieve consistent 3D representations for low-quality point clouds. In contrast, during the depth image normalization (see section 4.3), the pixel-based enlargement of a reduce-scale depth image may introduce artifacts that alter the edge pattern and intensity distribution. We can see this in Fig. 6 by comparing the normalizations from different scales. The local gradient and 2D nature of HOG aggravate this problem.

## 6 Conclusions

This study investigated new feature representation methods for recognizing low-resolution, degenerated, and sparse point cloud patches of human body shapes, often seen in the outputs of standoff 3D sensors such as LIDARs. We have leveraged our recently proposed Tchebichef Moment Shape Descriptor (TMSD) to achieve effective and compact shape characterization by approximating patterns of point clouds through low-order discrete orthogonal Tchebichef moments. Our action recognition uses a new bag-of-pose-shapes (BoPS) scheme to model temporal statistics and the Naïve Bayes model to infer the action label and viewing direction. We found that a small word vocabulary is sufficient to encode each action's pose shape sequences using our solution pipeline composed of Proportional Grid Voxelization and Normalization (PGVN), TMSD, and BoPS.

The cross validations and independent tests indicate that our 3D TMSD-based action classifier can achieve and maintain accurate predictions across a large range of scales and viewing angles. Moreover, our experiments demonstrated that native 3D characterization of point clouds by TMSD outperforms some representative 2D-based methods such as the depth image analysis based on the Histogram of Oriented Gradients (HOG) in terms of classification performance and consistency, especially at small sensing sizes. These advantages are significant benefits for mobile and aerial sensor platforms.

We plan to add other types of actions and more extreme elevation angles into the simulated pose shape baseline and develop more sophisticated inference models. Our datasets will be made public once cleared by the US Air Force.

**Acknowledgements** The authors would like to thank Isiah Davenport, Max Grattan, and Jeanne Smith for their indispensable help in the creation of biofidelic pose shape baseline.

## References

1. Aggarwal JK, Xia L (2014) Human activity recognition from 3D data: a review. *Pattern Recogn Lett* 48:70–80
2. Aggarwal CC, Hinneburg A, Keim DA (2001) On the surprising behavior of distance metrics in high dimensional space. In: *Proc. Int. Conf. database theory*, pp 420–434
3. Ballin G, Munaro M, Menegatti E (2012) Human action recognition from RGB-D frames based on real-time 3d optical flow estimation. *Biologically Inspired Cognitive Architectures*, Springer-Verlag, pp 65–74

4. Bobick AF, Davis JW (2001) The recognition of human movement using temporal templates. *IEEE Trans Pattern Anal Mach Intell* 23(3):257–267
5. Cheng H, Chung SM (2016) Orthogonal moment-based descriptors for pose shape query on 3D point cloud patches. *Pattern Recognition* 52, Elsevier Science:397–406
6. Chihara TS (1978) An introduction to orthogonal polynomials, Gordon and Breach
7. Costantini L, Seidenari L, Serra G, Capodiferro L, Bimbo AD (2011) Space-time Zernike moments and pyramid kernel descriptors for action classification. In: *Proc. Int. Conf. Image Anal. Processing*
8. Dalal N, Triggs B, Schmid C (2006) Human detection using oriented histograms of flow and appearance. *Proc Eur Conf Comput Vis. Lect Notes Comput Sci* 3952:428–441
9. Donahue J, Hendricks LA, Guadarrama S, Rohrbach M, Venugopalan S, Saenko K, Darrell T (2015) Long-term recurrent convolutional networks for visual recognition and description. *IEEE Conf Comput Vis Pattern Recogn* 2625–2634
10. Efros AA, Berg A, Mori G, Malik J (2003) Recognizing action at a distance. *Proc Int Conf Comput Vis* 2:726–733
11. Gorelick L, Blank M, Shechtman E, Irani M, Basri R (2007) Actions as space-time shapes. *IEEE Trans Pattern Anal Mach Intell* 29(12):2247–2253
12. Ji S, Xu W, Yang M, Yu K (2013) 3D convolutional neural networks for human action recognition. *IEEE Trans Pattern Anal Mach Intell* 35(1):221–231
13. Johnstone IM, Lu AY (2009) On consistency and sparsity for principal components analysis in high dimensions. *J. Am. Stat. Assoc* 104:682–693
14. Karpathy A, Toderici G, Shetty S, Leung T, Sukthankar R, Fei-Fei L (2014) Large-scale video classification with convolutional neural networks. *IEEE Conf Comput Vis. Pattern Recogn* 1725–1732
15. Kazhdan M, Funkhouser T, Rusinkiewicz S (2003) Rotation invariant spherical harmonic representation of 3D shape descriptors. In: *Proc. Eurographics Symp. Geometry Processing*, pp 156–164
16. Kläser A, Marszałek M, Schmid C (2008) A spatial-temporal descriptor based on 3D gradients. In: *Proc. British Mach. Vis. Conf*
17. Krizhevsky A, Sutskever I, and Hinton GE (2012) Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems* 25 (NIPS 2012), pp 1097–1105
18. Laptev I, Lindeberg T (2003) Space–time interest points. *Proc Int Conf Comput Vis* 2:432–439
19. Lassoued I, Zagrouba E, Chahir Y (2011) An efficient approach for video action classification based on 3D Zernike moments. In: *Proc. Int. Conf. Future Inf. Tech., Part II*, pp 196–205
20. Li W, Zhang Z, Liu Z (2010) Action recognition based on a bag of 3D points. In: *Proc. IEEE. Conf. Comput. Vis. Pattern Recogn. Workshops*, pp 9–14
21. Lian Z, Godil A, Sun X (2010) Visual similarity based 3D shape retrieval using bag-of-features. *Int Conf Shape Model Appl* 25–36
22. Lu Y, Li Y, Shen Y, Ding F, Wang X, Hu J, Ding S (2012) A human action recognition method based on Tchebichef moment invariants and temporal templates. In: *Proc. Int. Conf. Intelligent Human-Machine Sys. and Cybernetics*, vol. II, pp 76–79
23. Mademlis A, Axenopoulos A, Daras P, Tzovaras D, Srinatzis MG (2006) 3D content-based search based on 3D Krawtchouk moments. In: *Proc. Int. Symp. 3D data processing, visualization, and transmission*, pp 743–749
24. Maturana D, Scherer S (2015) Voxnet: a 3D convolutional neural network for real-time object recognition. *IEEE/RSJ Int Conf Intell Robots Sys* 922–928
25. McCallum A, Freitag D, Pereira F (2000) Maximum entropy Markov models for information extraction and segmentation. *Int Conf Mach Learning* 591–598
26. Metsis V, Androustopoulos I, Paliouras G (2006) Spam filtering with naive Bayes — which naive Bayes? In: *Proc. Conf. Email and anti-spam*, pp 27–28
27. Mukundan R, Ong SH, Lee PA (2001) Image analysis by Tchebichef moments. *IEEE Trans Image Process* 10(9):1357–1364
28. Ni B, Wang G, Moulin P (2011) RGBD-HuDaAct: a color-depth video database for human daily activity recognition. In: *Proc. IEEE. Int. Conf. Comput. Vis. Workshops*, pp 1147–1153
29. Niebles J, Wang H, Wang H, Fei-Fei L (2008) Unsupervised learning of human action categories using spatial-temporal words. *Int J Comput Vis* 79(3):299–318
30. Novotni M, Klein R (2004) Shape retrieval using 3D Zernike descriptors. *Comput Aided Des* 36(11):1047–1062



31. Ohbuchi R, Osada K, Furuya T, Banno T (2008) Salient local visual features for shape-based 3D model retrieval. *IEEE Int Conf Shape Model Appl* 93–102
32. Ovsjanikov M, Bronstein AM, Bronstein MM, Guibas L (2009) Shape google: a computer vision approach to isometry invariant shape retrieval. In: *Proc. workshop on non-rigid shape analysis and deformable image alignment (NORDIA'09)*
33. Poppe R (2010) A survey on vision-based human action recognition. *Image Vis Comput* 28(6):976–990
34. Schüldt C, Laptev I, Caputo B (2004) Recognizing human actions: a local SVM approach *Int Conf Pattern Recogn* 32–36
35. Sheng Y, Shen L (1994) Orthogonal Fourier-Mellin moments for invariant pattern recognition. *J Opt Soc Am* 11(6):1748–1757
36. Simonyan K, Zisserman A (2014) Two-stream convolutional networks for action recognition in videos. *Advances in neural information processing systems 27 (NIPS 2014)*, pp. 568–576
37. Sivic J, Zisserman A (2003) Video Google: A text retrieval approach to object matching in videos. *Proc Int Conf Comput Vis* 2:1470–1477
38. Sminchisescu C, Kanaujia A, Li Z, Metaxas D (2006) Conditional models for contextual human motion recognition. *Comput Vis Image Underst* 104:210–220
39. Sun X, Cheng M, Hauptmann A (2009) Action recognition via local descriptors and holistic features. In: *Proc. IEEE Conf. Comput. Vis. Pattern Recogn. Workshops*, pp 58–65
40. Szegedy C, Liu W, Jia Y, Sermanet P, Reed S, Anguelov D, Erhan D, Vanhoucke V, Rabinovich A (2015) Going deeper with convolutions. *Proc. IEEE Conf Comput Vis Pattern Recogn*
41. Tabia H, Daoudi M, Vandeborre J-P, Colot O (2011) Deformable shape retrieval using bag-of-features techniques. In: *Proc. SPIE-IS&T Electronic Imaging, SPIE*, vol 7864
42. Teague MR (1980) Image analysis via the general theory of moments. *J Opt Soc Am* 70(8):920–930
43. Teh CH, Chin RT (1988) On image analysis by the methods of moments. *IEEE Trans Pattern Anal Mach Intell* 10(4):496–513
44. Vieira A, Nascimento E, Oliveira G, Liu Z, Campos M (2012) STOP: Space-time occupancy patterns for 3D action recognition from depth map sequences. *Progress in Pattern Recognition, Image Analysis, Computer Vision and Application. Lect Notes Comput Sci* 7441:252–259
45. Vinyals O, Toshev A, Bengio S, Erhan D (2015) Show and tell: a neural image caption generator. *IEEE Conf Comput Vis Pattern Recogn* 3156–3164
46. Wang Y, Mori G (2009) Human action recognition by Semilattent topic models. *IEEE Trans Pattern Anal Mach Intell* 31(10):1762–1774
47. Wang J, Liu Z, Chorowski J, Chen Z, Wu Y (2012) Robust 3D action recognition with random occupancy patterns. *European Conf Comput Vis* 872–885
48. Wang J, Liu Z, Wu Y, Yuan J (2014) Learning actionlet ensemble for 3D human action recognition. *IEEE Trans Pattern Anal Mach Intell* 36(5):914–927
49. Weinland D, Ronfard R, Boyer E (2011) A survey of vision-based methods for action representation, segmentation, and recognition. *Comput Vis Image Underst* 115(2):224–241
50. Wolf C, Mille J, Lombardi E, Celiktutan O, Jiu MB, Dellandrea E, Bichot C, Garcia C, Sankur B (2012) The LIRIS human activities dataset and the ICPR 2012 human activities recognition and localization competition. Technical report RR-LIRIS-2012-004, LIRIS Laboratory
51. Wu Z, Song S, Khosla A, Yu F, Zhang L, Tang X, Xiao J (2015) 3D ShapeNets: a deep representation for volumetric shapes. *IEEE Conf Comput Vis Pattern Recogn* 1912–1920
52. Xia L, Chen C.-C., and Aggarwal JK (2012) View invariant human action recognition using histograms of 3D joints. In: *Proc. IEEE Conf. Comput. Vis. Pattern Recogn. Workshops*, pp 20–27
53. Yamato J, Ohya J, Ishii K (1992) Recognizing human action in time-sequential images using hidden Markov model. *IEEE Conf Comput Vis Pattern Recogn* 379–385
54. Yang X, Zhang C, Tian Y (2012) Recognizing actions using depth motion maps based histograms of oriented gradients. *ACM Int Conf Multimed* 1057–1060
55. Yao L, Torabi A, Cho K, Ballas N, Pal C, Larochelle H, Courville A (2015) Describing videos by exploiting temporal structure. *IEEE Conf Comput Vis* 4507–4515
56. Ye M, Zhang Q, Wang L, Zhu J, Yang R, Gail J (2013) A survey on human motion analysis from depth data. *Time-of-Flight and Depth Imaging, Sensors, Algorithms, and Applications. Lect Notes Comput Sci* 8200:149–187



**Huaining Cheng** is a research computer scientist at the 711th Human Performance Wing, USA Air Force Research Laboratory (AFRL). Prior to joining AFRL, he was a mechanical engineer at the General Dynamics Corporation. He holds an M.S. degree in Flight Dynamics from Beijing University of Aeronautics and Astronautics and two M.S. degrees in Mechanical Engineering and Computer Science from Wright State University. He also holds a Ph.D. degree in Computer Science and Engineering from Wright State University. His current research interests include 3D shape analysis and recognition, biosignature data mining, multimedia database and information system, and modeling of human biomechanics systems.



**Soon M. Chung** received a B.S. degree in Electronic Engineering from Seoul National University, Korea, in 1979, an M.S. degree in Electrical Engineering from Korea Advanced Institute of Science and Technology (KAIST), South Korea, in 1981, and a Ph.D. degree in Computer Engineering from Syracuse University, Syracuse, New York, in 1990. He is currently a professor in the department of Computer Science and Engineering at Wright State University, Dayton, Ohio. His current research interests include database, data mining, multimedia, information security, Grid computing, text mining, and parallel and distributed processing.