Human object interaction recognition using rate-invariant shape analysis of inter joint distances trajectories

Meng Meng
Telecom Lille, CRIStAL laboratory
UMR CNRS 9189, Lille France
meng@telecom-lille.fr

Hassen Drira
Telecom Lille, CRIStAL laboratory
UMR CNRS 9189, Lille France
drira@telecom-lille.fr

Mohamed Daoudi
Telecom Lille, CRIStAL laboratory
UMR CNRS 9189, Lille France
daoudi@telecom-lille.fr

Jacques Boonaert
Ecole de Mines de Douai
France
jacques.boonaert@mines-douai.fr

Abstract

Human action recognition has emerged as one of the most challenging and active areas of research in the computer vision domain. In addition to pose variation and scale variability, high complexity of human motions and the variability of object interactions represent additional significant challenges. In this paper, we present an approach for human-object interaction modeling and classification. Towards that goal, we adopt relevant frame-level features; the inter-joint distances and joints-object distances. These proposed features are efficiently insensitive to position and pose variation. The evolution of the these distances in time is modeled by trajectories in a high dimension space and a shape analysis framework is used to model and compare the trajectories corresponding to human-object interaction in a Riemannian manifold. The experiments conducted following state-of-the-art settings and results demonstrate the strength of the proposed method. Using only the skeletal information, we achieve state-of-the-art classification results on the benchmark dataset.

1. Introduction

Analysis of human activities and behavior through visual data has attracted a tremendous interest in the computer vision community. Indeed, this represents a task of interest for a wide spectrum of areas due to its huge potential, like human-machine interaction, physical rehabilitation, surveillance security, health care and social assistance, video games, etc[13]. The recent development and widespread use of portable, commodity, high-quality and accurate depth cameras such as Microsoft Kinect[1] has changed the picture by providing 3D depth data of video-based human action recognition. Thus several datasets have been collected to serve as benchmark for researchers algorithms like the MSR dataset [23].

In the literature of activity recognition, most of the previous works have focused on simple human action recognition such as boxing, kicking, walking, etc. However, human activity understanding is a more challenging problem due to the diversity and complexity of human behaviors [2] and accurate human action recognition is still a quite challenging task and is gradually moving towards more structured interpretation of complex human activities involving multiple people and especially interaction with objects.

Actually, during a human object interaction scene, the hands may hold objects and are hardly detected or recognized due to heavy occlusions and appearance variations [22]. A high level of information of the objects is needed to recognize the human-object interaction.

To the best of our knowledge, the majority of action recognition past approaches investigate simple action recognition and less effort have been spent on human object interaction. In this paper, we propose to apply spatio-temporal modeling (STM) and shape analysis framework to perform human-object interaction. The main contributions of this work are the following:

- The use of STM of skeletons and objects in time as trajectories.
- A rate-invariant comparison of these trajectories and compute rate-invariant means of them.
- The proposed method is performed on and gets competitive results with respect to state-of-the-art work on one representative benchmark dataset.

The remainder of the paper is organized as follows. Section 2 briefly describes the related works. In Section 3, we introduce the overview of our method. The spatio-temporal modeling and the shape analysis framework and presented in Section 4 and Section 5 respectively. In Section 6, the classification algorithm is described. The recognition results of the proposed approach on MDRDaily activity dataset which represents the becnchmark of of human activities and comparison with the state-of-the-art algorithms are presented in Section 7. Finally Section 8 summarizes the work, addresses several aspects of the model that can be improved and future research directions.

2. Related Work

The approaches on action recognition can be roughly divided into the following two main categories. The first category includes methods based on static and 2D video. There is emerging interest in exploiting human pose for action recognition. The release of the low-cost RGBD sensor Kinect has brought excitement to the research in computer vision, gaming, gesture-based control, and virtual reality. [7] adopted grouplet encode detailed and structured information from the images to estimate the 2D poses. In [8], it treated object and human pose as the context of each other in human-object interaction activities. [4] [25] developed spatio-temporal AND-OR graph to model the spatio-temporal structure of the poses in an actions. [9] [10] learns a discriminative deformable part model(DPM) that estimates both human poses and object location.

In the second category, there are still two sub-categories because of feature types. Some works adopted all of or two of skeleton, RGB and depth information and others only used skeleton-based algorithms. Recently, the development of depth cameras offers a cost-effective method to track 3D human poses [18]. In [24], an approach for human action recognition with histograms of 3D joint locations (HOJ3D) as a compact representation of postures is proposed. The HOJ3D computed from the action depth sequences are re-projected using LDA and then clustered into several posture visual words, which represent the prototypical poses of actions. The temporal evolutions of those visual words are modeled by discrete hidden Markov models (HMMs). [20] represented a human skeleton as a point in the Lie group which is curved manifold, by explicitly modeling the 3D geometric relationships between various body parts using rotations and translations. Using the proposed skeletal representation, it modeled human actions as curves in this Lie group then mapped all the curves to its Lie algebra, which is a vector space, and performed temporal modeling and classification in the Lie algebra. [14] used a spatio-temporal modeling of skeleton joint position in a Riemannian manifold. There are several works [11] [12] [15] [17] [23] relyed on skeleton information and developed features based on depth images for human object interaction recognition.

To the best of our knowledge, there are a few works on recognizing human-object interactions only based on skeleton joints. [22] presented a 4D human-object interaction model for joint event recognition through joint inference from RGBD videos. The 4DHOI model represents the geometric, temporal and semantic relations in daily events involving human object interactions. [26] proposed a novel middle level representation called orderlet [21] for recognizing human object interactions. It presented an orderlet mining algorithm to discover the discriminative orderlets from a large pool of candidates.

3. Overview of our method

An overview of the proposed approach is given in Figure 1. The human-object interaction videos are modeled as trajectories in $\mathbb{R}^{210*n}$ via a Spatio-Temporal Modeling (STM), then a rate invariant shape analysis of these trajectories is performed and this make the comparison of the videos invariant to the rate.

First, STM is applied on each video of training and testing data to get trajectories of dimension $\mathbb{R}^{210*n}$ (where n is the number of frames for each video).

Then, the rate-invariant mean shape μ_i of each action $a_i, i = 1..k$ is calculated. The feature vector for a given trajectory is then built by concatenating the distances d_S between this trajectory and all of the mean trajectories. Lastly, Random Forest-based classification is performed.

4. Spatio-temporal modeling

The 3-D humanoid skeleton can be extracted from depth images (via RGB-D cameras, such the Microsoft Kinect) in real-time thanks to the work of [18]. This skeleton contains the 3-D position of a certain number of joints representing different parts of the human body and provides strong cues to recognize human-object interaction.

Similarly to [16] we propose to use the inter-joints and the object-joints distances. The object position is detected by the LOP algorithm [21] unlike [16] where the authors manually detect the position of the object by associating it to the hand holding it. For each frame, all pairwise distances of 20 skeleton joints and object one are calculated. When the action does not have object, the corresponding entries in the distance matrix are blank and we fill them using an imputation technique [5]. In our experiments we employed the mean imputation method, which consists of replacing the missing values by the means of values already calculated in presence of the object from the training set. The skeleton information is donated as S which contains 20 joints from the original data and object joint represented by j_o.

$$S = \{j_1, j_2, ..., j_{20}, j_o\} \quad (1)$$

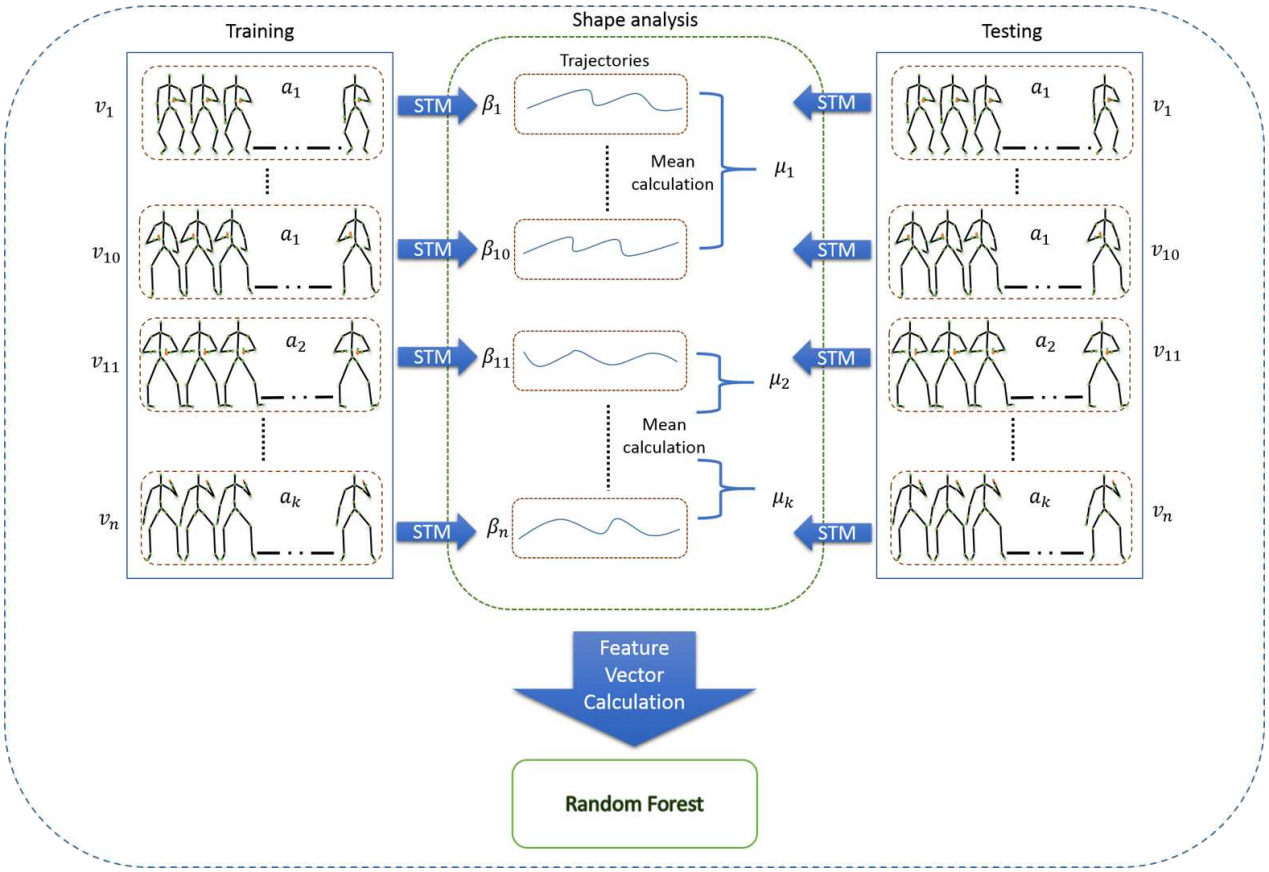

Figure 1. Overview of our method. Four main steps are shown: low-feature extraction from each frame; Buiding feature vector by spatio-temporal modeling; Mean calculation of feature vector; Random Forest-based classification. Note that the both training and testing data are built by spatio-temporal modeling and the red point is the object position we assumed.

D refers to the set of the pairwise distances between the joint a and joint b from S.

$$D = \{d(a,b)\}, a \in S\, b \in S \tag{2}$$

Thus the low-level feature vector is composed by the all pairwise distances between the joints and the distances between the object and the joints. The size of this vector is equal to $m \times (m-1)/2$, with $m = 21$: the 20 joints and the object joint. The concatenation of this feature vector along frames gives rise to a trajectory in $\mathbb{R}^{210}$. The shape of the resulting trajectories will be investigated in next section for human object classification.

5. Shape analysis of distance vector

5.1. SRVF calculation

Let $\beta : I \to \mathbb{R}^{210}$, where $I = [0,1]$, represents a parameterized curve encoding the trajectory of pairwise distances along a video. For each frame t, $\beta(t) = D_t$ encodes the pairwise distances at this frame.

To analyze the shape of β, we shall represent it mathematically using the *square-root velocity function* (SRVF) [19], denoted by $q(t)$, according to: $q(t) = \frac{\dot{\beta}(t)}{\sqrt{\|\dot{\beta}(t)\|}}$; $q(t)$ is a special function of β that simplifies computations under elastic metric.

Actually, under $\mathbb{L}^2$-metric, the re-parametrization group acts by isometries on the manifold of q functions, which is not the case for the original curve β. To elaborate on the last point, let q be the SRVF of a curve β. Then, the SRVF of a re-parameterized curve $\beta \circ \gamma$ is given by $\sqrt{\dot{\gamma}}(q \circ \gamma)$. Here $\gamma : I \to I$ is a re-parameterization function and let Γ be the set of all such functions.

Define the preshape space of such curves: $\mathcal{C} = \{q : I \to \mathbb{R}^{210} | \|q\| = 1\} \subset \mathbb{L}^2(I, \mathbb{R}^{210})$, where $\|\cdot\|$ implies the $\mathbb{L}^2$ norm. With the $\mathbb{L}^2$ metric on its tangent spaces, $\mathcal{C}$ becomes a Riemannian manifold. Also, since the elements of $\mathcal{C}$ have a unit $\mathbb{L}^2$ norm, $\mathcal{C}$ is a hypersphere in the Hilbert space $\mathbb{L}^2(I, \mathbb{R}^{210})$. The geodesic path between any two points $q_1, q_2 \in \mathcal{C}$ is given by the great circle, $\psi : [0,1] \to \mathcal{C}$,

where

$$\psi(\tau) = \frac{1}{\sin(\theta)} \left(\sin((1-\tau)\theta)q_1 + \sin(\theta\tau)q_2\right), \quad (3)$$

and the geodesic length is $\theta = d_c(q_1, q_2) = cos^{-1}(\langle q_1, q_2 \rangle)$.

In order to study *shapes* of curves, one identifies all re-parameterizations of a curve as an equivalence class.

Note that the parameterization of a trajectory during an action corresponds to the rate of the action. Thus comparison of equivalent classes rather than trajectories themselves is rate invariant differentiation which reduces the difference in rate between actions and facilitates the action recognition.

Let's define the equivalent class of q as: $[q] = \{\sqrt{\dot{\gamma}(t)}q(\gamma(t)), \quad \gamma \in \Gamma\}$. The set of such equivalence classes, denoted by $\mathcal{S} \doteq \{[q] | q \in \mathcal{C}\}$ is called the *shape space* of open curves in $\mathbb{R}^{210}$. As described in [19], $\mathcal{S}$ inherits a Riemannian metric from the larger space $\mathcal{C}$ due to the quotient structure. To obtain geodesics and geodesic distances between elements of $\mathcal{S}$, one needs to solve the optimization problem:

$$\gamma^* = argmin_{\gamma \in \Gamma} d_c(q_1, \sqrt{\dot{\gamma}}(q_2 \circ \gamma)). \quad (4)$$

The optimization over Γ is done using the dynamic programming algorithm. Let $q_2^*(t) = \sqrt{\dot{\gamma^*}(t)}q_2(\gamma^*(t)))$ be the optimal element of $[q_2]$, associated with the optimal re-parameterization γ^* of the second trajectory, then the geodesic distance between $[q_1]$ and $[q_2]$ in $\mathcal{S}$ is $d_s([q_1], [q_2]) \doteq d_c(q_1, q_2^*)$ and the geodesic is given by Eqn. 3, with q_2 replaced by q_2^*.

5.2. Mean calculation

One advantage of a shape analysis framework of the trajectories is that one has the actual deformations in addition to distances. In particular, we have a geodesic path in $\mathcal{S}$ between the two trajectories β^1 and β^2 in $\mathbb{R}^{210}$. This geodesic corresponds to the optimal elastic deformations of two trajectories. The Riemannian structure defined on the manifold of shape of the trajectories in $\mathcal{S}$ enables us to perform such statistical analysis for computing curves (trajectories) mean and variance. The Karcher mean utilizes the intrinsic geometry of the manifold to define and compute a mean on that manifold. It is defined as follows: Let $d_s(\beta^i, \beta^j)$ denote the length of the geodesic from β^i to β^j in $\mathcal{S}$.

To calculate the Karcher mean of trajectories $\{\beta^1, ..., \beta^n\}$ in $\mathcal{S}$, define the variance function:

$$\mathcal{V} : \mathcal{S} \to \mathbb{R}, \mathcal{V}(N) = \sum_{i=1}^{n} d_s(SRVF(\beta^i), SRVF(\beta^j))^2 \quad (5)$$

The Karcher mean is then defined by:

$$\overline{\beta} = \arg\min_{\mu \in \mathcal{S}} \mathcal{V}(\mu) \quad (6)$$

The intrinsic mean may not be unique, i.e. there may be a set of points in $\mathcal{S}$ for which the minimizer of $\mathcal{V}$ is obtained. To interpret geometrically, $\overline{\beta}$ is an element of $\mathcal{S}$, that has the smallest total deformation from all given trajectories.

Algorithm 1 Karcher mean algorithm

Set k = 0. Choose some time increment $\epsilon \leq \frac{1}{n}$. Choose a point $\mu_0 \in \mathcal{S}$ as an initial guess of the mean. (For example, one could just take $\mu_0 = \beta^1$.)

1- For each $i = 1, ..., n$ choose the tangent vector $t_i \in T_{\mu_k}(\mathcal{S})$ which is tangent to the geodesic from μ_k to β^i. The vector $g = \sum_{i=1}^{i=n} t_i$ is proportional to the gradient at μ_k of the function $\mathcal{V}$.
2- Flow for time ϵ along the geodesic which starts at μ_k and has velocity vector g. Call the point where you end up μ_{k+1}.
3- Set $k = k + 1$ and go to step 1.

The mean are calculated on trajectories belonging to the same action in order to get mean of the trajectory for each action. These means will be used in the classification of the actions. Moreover, the mean trajectory is invariant to the rate of execution of given videos due to the elastic metric used in the calculation of the mean.

6. Classification

6.1. Feature vector

The feature vector is built by using the distances to the means of the actions calculated on train data. Given train set $T = \{\beta^1, ..., \beta^n\} \in \mathbb{R}^{210*n}$, each trajectory corresponds to an action class $label_i \in \{a_1, ..., a_k\}$. We first calculate, using algorithm 1, the mean μ_i for each class. Next, we calculate the geodesic distance d_S between a given curve β and the mean curves. Thus a vector of distance of size k is provided as feature vector to classify the curve β. For example, this is a feature vector size k of one video sequence:

$$d_S = \left\{d(\beta^1, \mu_1), d(\beta^1, \mu_2), ..., d(\beta^1, \mu_k)\right\}$$

6.2. Random Forest

For the classification task we used the Multi-class version of Random Forest algorithm. The Random Forest algorithm was proposed by Leo Breiman in [6] and defined as a meta-learner comprised of many individual trees. It was designed to operate quickly over large datasets and more importantly to be diverse by using random samples to build each tree in the forest. Diversity is obtained by randomly choosing attributes at each node of the tree and then using the attribute that provides the highest level of learning. Once trained, Random Forest classify a new action from an input feature vector by putting it down each of the trees in

Method	Accuracy
Skeleton in [21]	68.0%
4DHOI model [22]	70.0%
Skeletal shape trajectories [3]	70.0%
Discriminative Orderlet Mining [26]	73.8%
Proposed approach	77.05%

Table 1. Reported results comparison to state of the art

the forest. Each tree gives a classification decision by voting for that class. Then, the forest chooses the classification having the most votes (over all the trees in the forest). In our experiments we used Weka Multi-class implementation of Random Forest algorithm by considering 150 trees. A study of the effect of the number of the trees is reported later in the experimental part.

7. Experiments

7.1. Dataset

MSRDailyActivity3D dataset [21] is a daily activity dataset captured by Kinect [1] device, to cover human daily activities in the living room. There are 16 action classes: *drink, eat, read book, call cellphone, write on a paper, use lap- top, use vacuum cleaner, cheer up, sit still, toss paper, play game, lay down on sofa, walk, play guitar, stand up, sit down* each of which was performed twice by 10 subjects. For each video, it provides 3 kinds of data: RGB, depth image and joint and 320 samples in total. Additionally, the activities includes human-object interactions and human motion that is the most important reason we choose this dataset.

7.2. Results

As our feature vectors built only based on skeleton joint information, this dataset is very challenging if the depth information is not used. To make it fair for comparison, we mainly compared with the algorithms on skeleton feature [21], [22] and [26]. [3] only used skeleton information that is the same as our work. We used the same experimental setting as [26] and performed on the 2-fold cross-validation which is using the samples of half of the subjects as training data, and the samples of the rest half as testing data. The comparison of the performance is shown in Table 1. We can notice in Table 1 that we obtained better accuracy than other works. The accuracy of our approach is 77.05%.

To fully evaluate our method, we performed the experiments with different numbers of trees. So we can see clearly that the performance of Random Forest classifier varies with the number of trees from Figure 2. As illustrated in this figure, the recognition rate raises with the increasing number of trees until 150; the recognition rate reaches the peak 77.05% and then becomes quite stable.

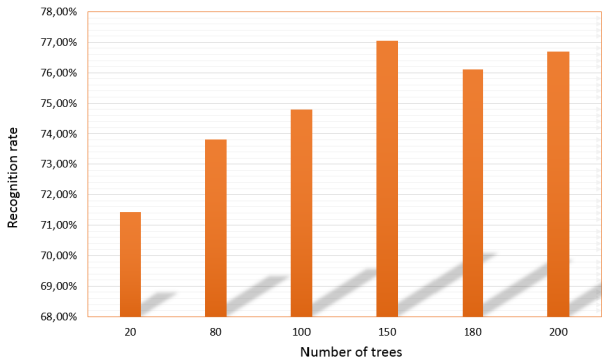

Figure 2. Human-Object interaction recognition results using a Random Forest classifier when varying the number of trees.

8. Conclusion and future work

This paper proposed an human-object interaction approach that use STM to model the pairwise distances of skeleton joints and object joints in each video as a trajectory. Then we compute the mean shape of trajectories corresponding to each action in a rate-invariant way. Human-object interaction classification is solved using Random Forest algorithm applied the feature vector calculated based on the distances to the means of actions. Experiments performed on MSRDaily Activity dataset testing on human motion and human-object interaction have demonstrated that our proposed approach gives comparative results with respect to state-of-the-art work. As the object assumed as one of skeleton joints in this paper, we will focus on object itself such as its shape in future.

Acknowledgment

This work was partially supported by the FUI project MAGNUM 2 and the Programme dInvestissements dAvenir (PIA) and Agence Nationale pour la Recherche (grant ANR-11-EQPX-0023), and European Founds for the Regional Development (Grant FEDER- Presage 41779).

References

[1] Microsoft kinect. *http://www.microsoft.com/en-us/ kinectforwindows/*, 2013. 1, 5

[2] A. Alpher, J. P. N. Fotheringham-Smythe, and G. Gamow. A survey on human motion analysis from depth data. *Time-of-Flight and Depth Imaging*, pages 149–187, 2013. 1

[3] B. B. Amor, J. Su, and A. Srivastava. Action recognition using rate-invariant analysis of skeletal shape trajectories. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 38(1):1–13, 2016. 5

[4] Z. L. B. Yao, X. Nie and S. Zhu. Animated pose templates for modelling and detecting human actions. *PAMI*, 36(3):436–452, 2014. 2

[5] G. Batista and M. C. Monard. An analysis of four missing data treatment methods for supervised learning. *Applied Artificial Intelligence*, 17:519533, 2003. 2

[6] L. Breiman. Machine learning. 45:5–32, 2001. 4

[7] B.Yao and F. Li. Grouplet: A structured image representation for recognizing human and object interactions. *CVPR*, pages 9–16, 2010. 2

[8] B.Yao and F. Li. Recognizing human-object interactions in still images by modeling the mutual context of objects and human poses. *Pattern Analysis and Machine Intelligence*, 34(9):1691–1703, 2012. 2

[9] C. Desai and D. Ramanan. Detecting actions, poses, and objects with relational phraselets. *ECCV*, 2012. 2

[10] C. Desai, D. Ramanan, and C. Fowlkes. Discriminative models for static human-object interactions. *CVPRW*, 95(1):1–12, 2011. 2

[11] G.Dian and G.Medioni. Dynamic manifold warping for view invariant action recognition. *ICCV*, pages 571–578, 2011. 2

[12] J.Qi and Z.Yang. Learning dictionaries of sparse codes of 3d movements of body joints for real-time human activity understanding. *PloS one*, 9(12), 2014. 2

[13] A. Kleinsmith and N. Bianchi-Berthouze. Affective body expression perception and recognition: A survey. *IEEE Transactions on Affective Computing*, 4(1):15–33, 2013. 1

[14] M.Devanne, H.Wannous, S. Berretti, P. Pala, M.Daoudi, and A. Bimbo. 3d human action recognition by shape analysis of motion trajectories on riemannian manifold. *Cybernetics*, 45(7):1340–1352, 2015. 2

[15] M.Devanne, H.Wannous, S. Berretti, P. Pala, M.Daoudi, and A. Bimbo. Combined shape analysis of human poses and motion units for action segmentation and recognition. *FGW*, 7:1–6, 2015. 2

[16] M.Meng, H.Drira, M.Daoudi, and J. Boonaert. Human-object interaction recognition by learning the distances between the object and the skeleton joints. *FGW*, 7:1–6, 2015. 2

[17] O. Omar and Z. Liu. Hon4d: Histogram of oriented 4d normals for activity recognition from depth sequences. *CVPR*, 14(1):234–778, 2013. 2

[18] 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. *Communications of the ACM*, pages 116–124, 2013. 2

[19] A. Srivastava, E. Klassen, S. H. Joshi, and I. H. Jermyn. Shape analysis of elastic curves in euclidean spaces. *PAMI*, 33(7):1415–1428, 2011. 3, 4

[20] R. Vemulapalli, F. Arrate, and R. Chellapan. Human action recognition by representing 3d skeletons as points in a lie group. *CVPR*, pages 588–595, 2014. 2

[21] J. Wang, Z. Liu, Y. Wu, and J. Yuan. Mining actionlet ensemble for action recognition with depth cameras. *CVPR*, pages 1290–1297, 2012. 2, 5

[22] P. Wei, Y. Zhao, N. Zheng, and S. Zhu. Modeling 4d human-object interactions for joint event segmentation, recognition, and object localization. *ICCV*, pages 3272–3279, 2013. 1, 2, 5

[23] W.Li, Z.Zhang, and Z. Liu. Action recognition based on a bag of 3d points. *CVPRW*, pages 9–14, 2010. 1, 2

[24] L. Xia, C. Chen, and J. Aggarwal. View invariant human action recognition using histograms of 3d joints. *CVPRW*, pages 20–27, 2012. 2

[25] B. Yao and S. Zhu. Learning deformable action templates from cluttered videos. *ICCV*, pages 1507–1514, 2009. 2

[26] G. Yu, Z. Liu, and J. Yuan. Discriminative orderlet mining for real-time recognition of human-object interaction. *ACCV*, pages 50–65, 2014. 2, 5