Distances evolution analysis for online and off-line human object interaction recognition

Meng Meng, Hassen Drira, Jacques Boonaert

Distances evolution analysis for online and off-line human object interaction recognition

Meng Meng[a], Hassen Drira[b,c], Jacques Boonaert[b]

[a]*North Carolina Central University, USA*
[b]*IMT Lille Douai, France*
[c]*CRIStAL laboratory university Lille1, UMR CNRS 9189, France*

Abstract

Human action recognition in 3D sequences is one of the most challenging and active areas of research in the computer vision domain. However designing automatic systems that are robust to significant variability due to object combinations and high complexity of human motions are more challenging in addition to the typical requirements such as rotation, translation, and scale invariance is challenging task. In this paper, we propose a spatio-temporal modeling of human-object interaction videos for on-line and off-line recognition. The inter joint distances and the object are considered as low-level features for online classification. For off-line recognition, we propose rate-invariant classification of full video and early recognition. A shape analysis of trajectories of the inter-joint and object-joints distances is proposed for this end. The experiments conducted following state-of-the-art settings using MSR Daily Activity 3D Dataset and Online RGBD Action Dataset and on a new Multi-view dataset for human object interaction demonstrate that the proposed approach is effective and discriminative for human object interaction classification as demonstrated here.

Keywords: Human object interaction, rate invariance, shape analysis, temporal modeling.

*Corresponding author
Email addresses: mmeng@nccu.edu (Meng Meng), drira@imt-lille-douai.fr (Hassen Drira), jacques.boonaert@imt-lille-douai.fr (Jacques Boonaert)

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[1]. Comparing to verbal or vocal communication data, visual data forms one of the most important cues in developing systems for understanding human behavior. The applications range are from tracking daily activities to classifying emotional states, as well as detecting abnormal and suspicious activities.

The recent development and wide-spread use of portable, commodity, high-quality and accurate depth cameras such as Microsoft Kinect[2] has changed the picture by providing 3D depth data for video-based human action recognition. This type of data brings several advantages as it makes the background easy to remove and allows extracting and tracking the human body, thus capturing the human motion at each frame. Additionally, the 3D depth data is independent of the human appearance (texture) providing a more complete human silhouette relative to the silhouette information used in the past[3]. So the emergence of 3D data reduces the challenges to human behavior analysis. In this context, several datasets have been collected such as the MSR 3D-Action dataset, MSR Daily Activity 3D Dataset[4] and the Online RGBD Action Dataset [5].

However, human activity understanding is a more challenging problem due to the diversity and complexity of human behaviors [6] 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. To the best of our knowledge, the majority of action recognition past approaches investigate simple action recognition [7] [8] [9] [10] [11] [12] such as boxing, kicking, walking, etc. and less effort have been spent on human object interaction. There are two different scenarios for human-object interaction recognition. The first one is an on-line

classification that needs low level features and the second is off-line and brings a new challenge which is the difference in rate and execution time.

Even though the depth cameras have generally a better quality of 3D action data than those estimated from monocular video sensors, adopting the 3D joint positions for human-object interaction is not sufficient to classify actions that includes interaction with objects. During a human object interaction scene, the hands may hold objects thus are hardly detected or recognized due to heavy occlusions and appearance variations [13]. A high level information of the objects is needed to recognize the human-object interaction. On the other hand, the use of 3D skeleton joints is not sufficient to distinguish some actions like *drinking* and *picking phone*. Extra inputs need to be included and exploited for more accurate recognition.

In this work, we propose two methods using both skeletal data and local depth information for on-line and off-line human-object interaction videos classification. Several parts are common for the two scenarios such as the extraction of low-level features and the classification task. For the on-line classification, we used [14]'s modeling framework to represent individual skeletons. Additionally, local depth information of objects extracted for building feature vectors provides more accurate human-object interaction recognition. For rate-invariant classification, we use the shape analysis of features trajectories. An action is defined as an evolution of the inter-joint distances and thus actions are represented by trajectories.

1.1. Related work

In this section, we briefly review related work from four streams of research and discuss our contributions compared with the existing work.

In the literature of activity recognition, many previous work in behavior analysis used videos produced by conventional cameras [15], [16], [17], [18]. There are a large amount of existing methods for human-object interaction recognition based on static and 2D videos such as [19] [18] [20] [21] [22] [23] [24] [25]. [26] adopted grouplet encode detailed and structured information from

the images to estimate the 2D poses. In [27], the authors treated object and human pose as the context of each other in human-object interaction activities. [28] inferred the spatial information of objects by modeling the 2D geometric relations between human body and objects. [29] combined spatial and functional constraints between human and objects to recognize actions and objects on static images. [30] [31] developed spatio-temporal AND-OR graph to model the spatio-temporal structure of the poses in an actions. [32] [33] learns a discriminative deformable part model(DPM) that estimates both human poses and object location. These methods define the human-object interactions on 2D image. Such contextual cues are often compromised by the viewpoint changes and occlusions.

Recently, with the development of the commodity depth sensors like Kinect, there has been a lot of interests in human action recognition from depth data such as [9], [8], [10], [11], [12], [34], [35], [36]. Instead of covering all ideas exhaustively, we direct interested readers to some recent surveys [7], [37], [38], [39] that together overview this domain.

Here we focus on the human-object interaction and action recognition approaches using human body descriptors which most closely related to our approach. The approaches on human-object interaction and action recognition can be roughly divided into the following three main categories.

Approaches based on depth information

Human pose estimation has generated a vast literature (surveyed in [40], [41]). The computer vision researchers have been more interested in this area with the availability of depth sensors [42],[43], [44]. Grest et al. [42] use Iterated Closest Point to track a skeleton of a known size and starting position. Anguelov et al. [45] segment puppets in 3D range scan data into head, limbs, torso, and background using spin images and a MRF. Thanks to the work of [46] by using the depth cameras which offers a cost-effective method to track 3D human poses, many approaches in the literature adopted skeleton, RGB and depth features to model human activities (including human-object interaction). In this part, we

mainly discuss the works based on depth sequences. [47] extracted the objects from depth data to perform sub-activity (referred to as action) classification and functional categorization of objects. Their method first detected the sub-activity being performed using the estimated human pose from depth data, and then performed object localization and clustering of the objects into functional categories based on the detected sub-activity. [48] took advantage of pose tracks and depth readings and employed the latent structural SVM to train the model with part-based pose trajectories and object manipulations. [49] proposed a shape analysis tool predicting human pose based on object affordance.

Many works of human activities using depth maps obtained decent performance. [4] employed an action graph to model the dynamics of the actions and sample a bag of 3D points from the depth map to characterize a set of salient postures that correspond to the nodes in the action graph. But there are limitations of this work such as noise and occlusions in the depth maps and sampling scheme is view dependent. In [50], they presented a descriptor histogram of oriented 4d surface normals (HON4D) capturing the distribution of the surface normal orientation in the 4d volume of time, depth and spatial coordinates from depth maps. [51] learned dictionaries of sparse codes of sampled spatial-temporal 3D volumes from depth maps and achieved real-time human action recognition.

In most cases, the works based on depth images adopt the whole depth maps which is difficult to achieve the real-time recognition. Skeleton descriptor can effectively solve this problem. So we take advantage of this property of skeleton information applied in our approaches.

Approaches based on skeleton information

To the best of our knowledge, there are a few works on recognizing human-object interactions only based on skeleton joints. [13] 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. [5] pro-

posed a novel middle level representation called orderlet [52] for recognizing human object interactions. It presented an orderlet mining algorithm to discover the discriminative orderlets from a large pool of candidates.

But in current research field of action recognition, the skeleton information is adopted in many works. [53] performed the Dynamic Time Warping (DTW) on feature vectors defined by 3d joint trajectories. In [54], 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). [55] proposed a general method for online estimation of quality of movement on stairs and used depth sensors on staircases only skeleton data adopted. [56] 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. The depth data is not suitable for online action recognition from unsegmented streams. [57] created a joint space can then be used to predict potential human poses and joint locations from a single image. This joint space modeled he physical interaction between human poses and scene geometry. [58] introduced a human representation by comparing the similarity between human skeletal joint trajectories in a Riemannian manifold [59].

For fine-grained human object interaction recognition, [60] used the MSR Daily Activity 3D Dataset obtained by Kinect and linked object proposals followed by feature pooling in selected regions. In their work, the proposed method only analyzed 2D video content without depth map. But they added skeleton information to localize useful interaction parts and remove background noise and obtained best result on this dataset compared with state of the art. [3]

proposed to model the evolution of human skeleton shapes as trajectories on Kendalls shape manifolds, and used a parameterization-invariant metric [61] for aligning, comparing, and modeling skeleton joint trajectories, which can deal with noise caused by large variability of execution rates within and across humans. We also performed our approached by using depth and skeleton information on this dataset and details will be shown in the following sections. There are several works [62] and a comprehensive survey [63] of existing space-time representations of people based on 3D skeletal data.

Approaches based on hybrid information

Here we introduce some works proposed hybrid approaches by combining both depth information and skeleton data features in order to improve recognition performances. [64] defined a Markov Random Field MRF over the spatio-temporal sequence where nodes represent objects and sub-activities, and edges represent the relationships between object affordances, their relations with sub-activities, and their evolution over time. This method needs the video to be pre-segmented. And the object detection is independent of the contextual feedback from human actions. [52] used relative skeleton position and local occupancy patterns (LOP) features to model the human-object interaction, and developed Fourier Temporal Pyramid to characterize temporal dynamics.

1.2. Overview of the proposed method

In this paper, we develop an approach for two different scenarios with different challenges. Actually, several applications require human object interaction recognition after the action is done. The main challenges revealed in this scenario are the execution time differing for same interaction and significant spatial variation in the way of performing an action. The pipeline of the proposed approach (denoted off-line scenario) is depicted in Figure 1. The object detection and the spatio-temporal modeling are common steps in both scenarios. The input sequences are modeled as trajectories in $\mathbb{R}^{210}$ via a Spatio-Temporal Modeling (STM). A rate invariant shape analysis of these trajectories is then

performed and this make the comparison of the sequences invariant to the rate. The shape analysis framework includes calculation of intrinsic mean of the trajectories issued from the same interaction for training data. The rate invariant distance between a trajectory issued from testing data and all mean trajectories calculated on training data built the final feature vector that represents the input of Random Forest classifier.

Furthermore, the trade-off between the accuracy and observation size for rapid and real-time recognition is an important topic in a wide spectrum of real applications, this motivates the second scenario we develop in this paper: online human-object interaction recognition. The main challenge here is the accurate and real time recognition thus we propose to use the low-level features used for the first scenario as input to classifier. Actually, the skeleton features (low-level) are easy to extract and track from depth maps thanks to the work of [46], utilizing low level features to describe interactions and the most relevant parts of human poses with respect to object can make it possible to achieve rapid and online recognition of human-object interactions. The pipeline of this scenario is depicted in Figure 1. The LOP algorithm [52] is applied on each frame of input sequence to detect the presence and the position of the object. Then the low-level features are calculated in each frame and the classification will be done using one frame with N previous frames in the memory (N can be 0). The position of the object is necessary to calculate the object-joints distances used as low features. Moreover, the rough shape and the size of the object are estimated to allow for online human-object interaction recognition.

This paper is an extension of a previous conference paper [14]. The main contributions of this work are the following:

- Extending the low-level features (inter-joint distances) used previously in [14] and propose a new object feature describing roughly the size and the shape of the object for online classification.
- Modeling the evolution of the low-level features along videos by trajectories in $\mathbb{R}^{210}$ and perform an elastic shape analysis of these trajectories to

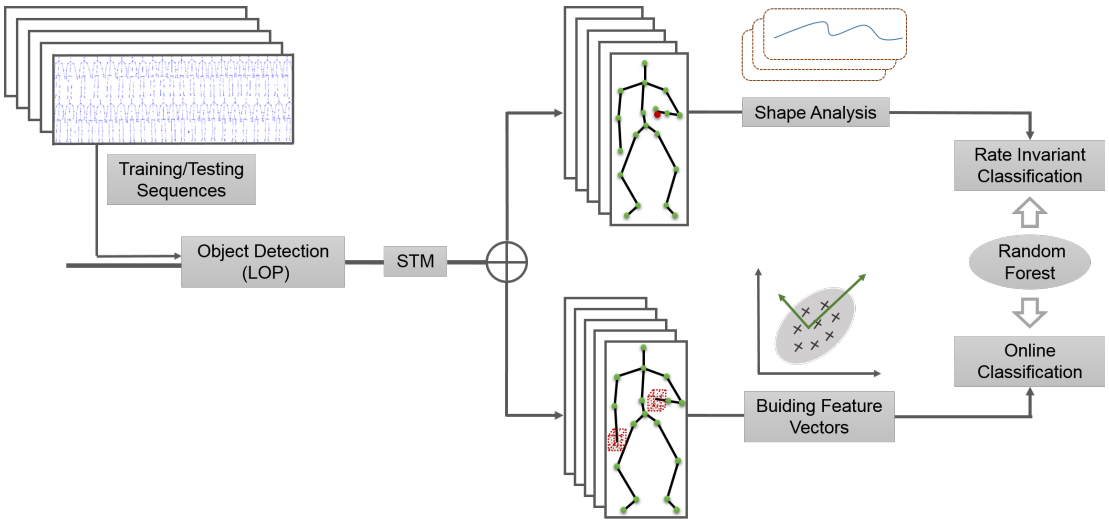

Figure 1: Overview of our method. Four main steps are shown: low-feature extraction from each frame; Buiding feature vector by spatio-temporal modeling; Shape analysis of feature vector for rate invariant classification; For Online classification, Random Forest-based classification.

allow rate-invariant human-object recognition.

- Presenting comparative evaluations on two challenging datasets presenting different challenges.

- Propose a new scenario in Human Object interaction including abnormal gait and multi view. Actually a new multiview dataset is collected for this end (collected by 2 kinects), a new protocol has been designed and results are presented on this new dataset.

The rest of the paper is articulated as follow. Section 2 briefly describes the related works. In Section 3, we introduce the overview of our method. The spatio-temporal modeling of low-level joint feature is introduced in Section 4. The online classification and rate-invariant sequence classification presented in Section 5 and Section 6 respectively. The recognition results of the proposed approach on MSR Daily Activity 3D Dataset and Online RGBD Action Dataset

which represent the benchmark 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. Spatio-temporal representation of actions

The invariance to the translation and rotation of the subject in the scene is a necessary condition of human-object interaction recognition systems: two instances of the same action differing only for the position and orientation of the person with respect to the scanning device should be recognized as belonging to the same action class. This goal can be achieved either by adopting a translation and rotation invariant representation of the action sequence or providing a suitable distance measure that copes with translation and rotation variations. Similarly to [14], we propose to use the inter-joints and the object-joints distances that handles well with the situations discussed above. The object position is detected by the LOP algorithm [52]. 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 are filled using an imputation technique [65]. 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 denoted as J which contains 20 joints from the original data and object joint represented by j_o.

$$J = \{j_1, j_2, ..., j_{20}, j_o\} \tag{1}$$

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

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

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

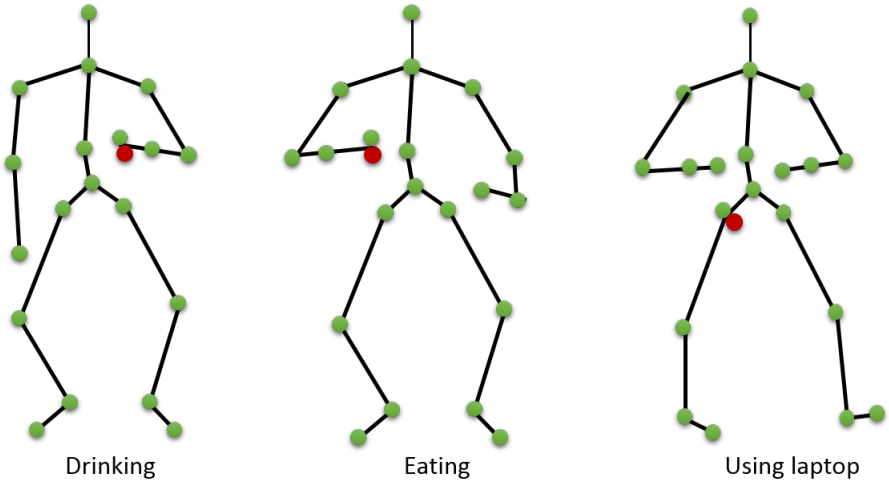

Figure 2: Examples of our pairwise joint distance features on MSR Daily Activity 3D Dataset. The red one refers to object joint for each action.

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 a trajectory in $\mathbb{R}^{210}$.

We start by outlining a mathematical framework for helping in analyzing the temporal evolution of human object interactions when viewed as trajectories on shape space of distance trajectories. This framework respects the underlying geometry of the shape space of the trajectories and helps maintain desired invariance. Then calculate the distance between the mean trajectory of each action to the trajectories from testing set based on square-root velocity function (SRVF[66]. So that the trajectories from different action classes can be fairly compared in a another shape space.

2.1. Shape analysis of distance trajectories

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. The shape of trajectories of the

3D joints have been studied previously [3], [58]. In this work we analyze the shape of the trajectory engendered by the evolution of the distances between the joints in time. To analyze the shape of β, we shall represent it mathematically using the *square-root velocity function* (SRVF) [66], 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

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

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 [66], $\mathcal{S}$ 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:

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

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^*.

2.2. Statistics of the trajectories

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 [59] 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:

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

The Karcher mean is then defined by:

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

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. The karcher mean has been previously used in biometrics [67], [68].

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.

2.3. Rate-invariant Action classification

The classification step is performed by Random Forest classifier based on an Euclidean signature for each sequence (trajectory) -training or test. The euclidean signature calculation process is illustrated in Figure 3 Given n sequences $\{v^1, ..., v^n\}$ in the training set, we first calculate the corresponding trajectories $\{\beta^1, ..., \beta^n\} \in \mathbb{R}^{210*n}$ using the spatio-temporal modeling. Each resulting trajectory β^i corresponds to an action class $label_i \in \{a_1, ..., a_k\}$. Using algorithm 1, the mean μ_i for each class is first calculated on the training set. Next, the geodesic distance d_S between any query trajectory -training or test- and the mean curves μ_i is calculated and denoted d_i.

A concatenation of the k individual Euclidean representations results in a large Euclidean vector that represents the final Euclidean signature of any a given trajectory.

The classification is performed using the multi-class version of Random Forest algorithm. The Random Forest algorithm was proposed by Leo Breiman in [69] and defined as a meta-learner comprised of many individual trees. It

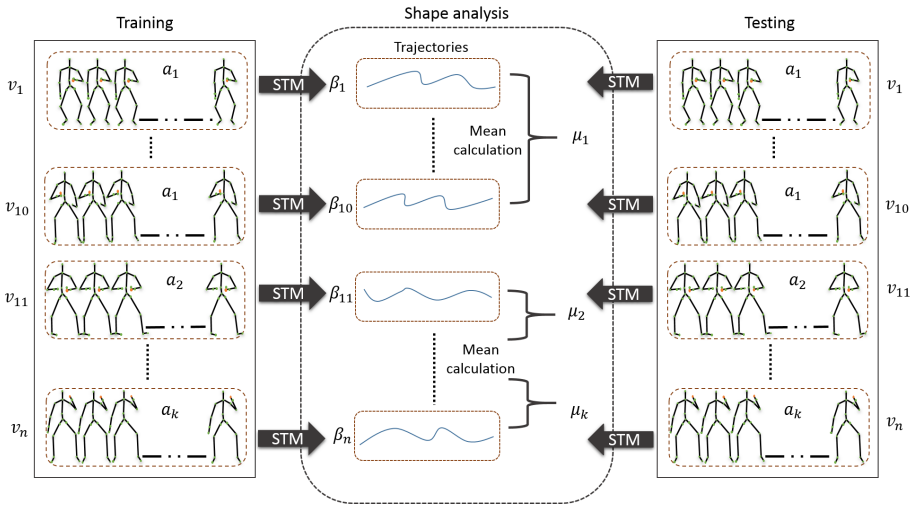

Figure 3: Overview of off-line classification. Note that the both training and testing data are built by spatio-temporal modeling and the red point is the object position we assumed. First, Spatio-Temporal Modeling (STM) is applied on each video of training and testing data to get trajectories of dimension $\mathbb{R}^{210*n}$ (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.

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 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. A study of the effect of the number of the trees is reported later in the experimental part.

3. Online recognition

Based on the inter-joints and object-joints distances presented previously, we propose in this section an algorithm for on-line recognition. The classification is based on one frame with N previous ones in the memory (N can be zero). When N is not null, the N-frames sliding window is considered for the on-line classification. The first step in the on-line recognition system we propose is the object feature. This object feature will be fused later with the low-level extracted features to built the final features vector which will be classified on-line using random forest classifier.

3.1. Object Feature

A specific object description can be helpful to characterize the human object interaction. But this is a difficult and time consuming way to realize online classification. As we discussed in the previous part, it is insufficient to only use the 3D joint positions to fully model an action, especially when the action includes the interactions between the subject and other objects such as *drinking* and *picking phone*. The extra input like depth information need be adopted in order to have more precise classification.

Motivated by properties of objects, we try to utilize the size and shape information of objects which is more efficient and convenient way for online

Figure 4: The illustration of object cube size. In the first row, the green rectangular and red rectangular represent small and large cube respectively. The yellow rectangular in the second row represent the appropriate size of object cube.

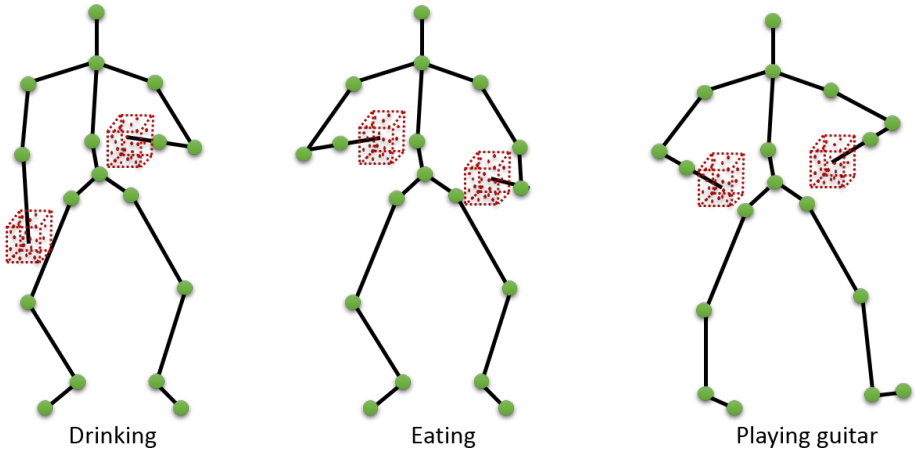

Figure 5: Examples of our object features on Online RGBD Action Dataset (ORGBD) dataset. The red cube refers to object cube for each action.

human-object interaction recognition. When performing an interaction, human usually hold objects by two hands. Moreover, the depth points located around the skeleton joints of two hands contain a lot of messages about the size and shape of objects.

The object is assumed to be present around one hand, thus similarly to the LOP algorithm [52] that counts the number of points inside a given cube around given point (hand for example) and decides the presence of an object given a threshold, we extend this algorithm to exploit the number of the points inside the cube and the 3D coordinates of these points to built the object feature. The number of depth points refers to the rough size of objects and the coordinates of these points refer to the rough shape of objects. The PCA algorithm is applied on the coordinates in order to determine the principal directions of the object inside the cube. These directions are concatenated with the number of the points to build the object feature. The feature vector calculation depends on the size of chosen cubes to detect these points. If the cube size is too small like the situation shown at the top left in Fig. 4, the green rectangular is too small to show the features of different object. So the resulting feature will not be discriminative for interaction classification. If the cube size is too big like the situation shown at the top right in Fig. 4, the red rectangular is so big that contains a lot of context from background and other parts of body. So we have to detect objects in a appropriate size as shown in the second row in Fig. 4. In the experiment, the retained size of cubes is 50. A trade of the size of this cube and the results will be discussed later in experimental section. We report in Fig. 2 examples of different human action interactions. The object is reported in red and the joints are reported in green. The proposed features are the pairwise distances between all these points in 3D.

3.2. Online action recognition

The feature vector is the concatenation of the pairwise distances of elements from J in equation 1 and the object feature that contains the number of points inside the cube around the hand holding the object and the main directions describing the rough shape of the object given by PCA algorithm. The proposed approach handles also the human action with no object interaction. In this case, the LOP algorithm detects the absence of objects around the hands and an imputation technique is used to fill the missing informations. In our exper-

iments 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. For the classification task we used the multi-class version of Random Forest algorithm. The Random Forest algorithm was proposed by Leo Breiman in [69] 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 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 100 trees. A study of the effect of the number of the trees is reported later in the experimental part.

4. Multi-view Human Object Interaction Recognition

4.1. Dataset description

To evaluate our method, we built a large-scale 3D event dataset with abnormal and normal human activities involved human object interactions. It is captured by two stationary Kinect sensors from different viewpoints simultaneously. It includes 8 event categories: *press button with injured arm or with injured leg, pick phone with injured arm or with injure leg, use remote and take it back with injured arm or with injure leg, fetch water from dispenser with injured arm, walk around holding cane with injured leg, walk around holding umbrella with injured leg, remove chair with injured leg, walk with plate and put it back on the table with injured arm or with injure leg* and 3 modalities include normal, injured arm and injured leg. All these activities were performed by 10 different subjects each two times in normal and abnormal way. Each event category

includes about 30 video sequence instances. For each frame, the Kinect V2 provides 25 skeleton joints which is different from Kinect V1 which provides 20 skeleton joints.

Figure 6: The setting up of dataset collection

Figure 7: The setting up of dataset collection

Here, Fig. 6 and Fig. 7 are the photographs of the system we set up. We have two Kinects (one on the left and one on the right), mounted on tripods so that we get a big enough common fields (see next forwarded mail to see the trace on the floor (dashed lines) that corresponds to the area where both Kinects provide a good detection of the skeleton. This represent a surface of about 3 x 3 meters starting at (about, again) 1,5 meters from the Kinects' lenses).

There are several characteristics which make the new multi-view dataset challenging. In the first place, we use two Kinect sensors to capture the video. As the various types of subjects' action, the synchronization from different view for rate invariant recognition is a big issue to address. In the second place, we not only capture the normal persons holding different objects but also abnormal persons executing activities with objects. At last, there are two abnormal modalities which means our new dataset has large variety when each subject

performing an event.

4.2. Multi view data synchronization and fusion

For the synchronized view experiments, we propose two steps for synchronizing the same action from different views. First, we use the resampling algorithm and the function 4 of shape analysis framework for the alignment of trajectories from each frame. Second, after alignment, we adopt the proposed fusion framework to achieve a fused trajectories by selecting the best distance attributes from each trajectory. For the fusion algorithm, we will detail it later.

As shown in Fig. 8, we apply the same framework like we did in the temporal modeling. The β_1 and β_2 refer to the feature matrix obtained from the low-level feature based on the same action from different views. $\beta_1(t_0)$ is the trajectory of first frame of β_1 and $\beta_1(t_1)$ is the trajectory of second frame of β_1. So the $\beta_1(t)$ is the one of trajectories of the frames of β_1 and β_2 is the same. So after resampling, we align them frame by frame.

$$\beta_2 * \gamma^*$$

is the aligned trajectory after applying the function γ^* (4). Then, we fuse the β_1 and $\beta_2 * \gamma^*$ as following for the same action.

As explained above, we need fuse the trajectories for each frame of the videos from K1 and K2 which refer to these two Kinect sensors. In Fig. 9, we can notice that there are three colors to represent different distance attributes. Here, we calculate the distance between $\beta_1(t)$ and $\beta_2(t)$: if they are not close, we choose the mean of the corresponding distances (red); if they are close, we choose the smoother one by comparing their own curvature 7.

$$k = |dt/ds| \tag{7}$$

t is the velocity vector which is also the difference between the distance attributes on $\beta_1(t)$ and $\beta_2(t)$.

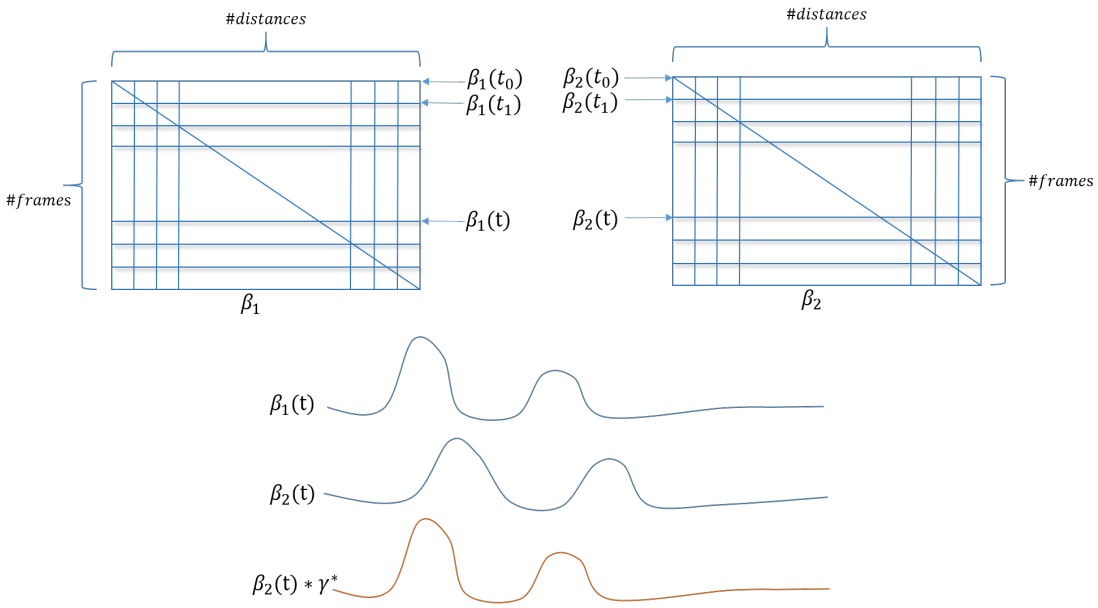

Figure 8: The illustration of trajectories synchronization

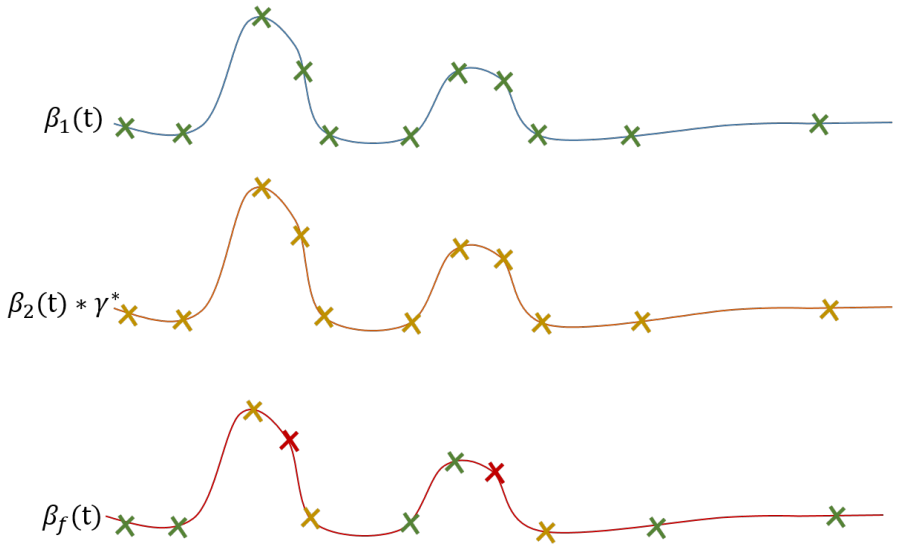

Figure 9: The illustration of trajectories fusion. The green, yellow and red points refer to the distance on $\beta(t)$ of the videos from K1, K2 after synchronization and mean respectively.

5. Experimental evaluation

5.1. Datasets

1. MSRDaily Activity 3D dataset is a daily activity dataset captured by Kinect [2] 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. We propose to evaluate the MSR Daily Activity 3D dataset[52] for both of online and rate-invariant classifications (off-line scenario).

2. ORGBD dataset contains seven types of actions which all those actions are human-object interactions: drinking, eating, using laptop, picking up phone, reading phone (sending SMS), reading book, and using remote. The bounding box of the object in each frames is manually labeled. In our approach, we use the object labels to locate our object feature. All of the videos are captured by Kinect. Each action was performed by 16 subjects for two times. We compare our approach with state-of-the-art methods on the cross-subject test setting, where half of the subjects are used as training data and the rest of the subjects are used as test data. This dataset was proposed for online recognition [5]. Thus we evaluate the performance of our online algorithm on this dataset.

3. Multi-view Human Object Interaction Recognition Lille Douai dataset To the best of our knowledge, this dataset represents the first multi-view human object interaction dataset that include normal and abnormal gaits. We collect it to test the performance of the proposed method within this different scenarios for human object interaction. The dataset description was presented in the previous section.

5.2. Comparative results on ORGBD dataset

We compare our online approach with results achieved by the state-of-the-art methods on ORGBD dataset and compare our results to [5] which is, to our best knowledge, the only work on this database for real time classification. We used the same protocol as [5], where half of the subjects are used as training data and the rest of the subjects are used as test data. As shown in Table 1, we reveal different experimental results based on the different features we used in our approach. As illustrated in the second row of Table 1, based on the low-level features (object to joints and inter joints distances), the recognition rate is 75.58% compared to 71.4% in [5]. By adding the rough shape of the object, using PCA on the points inside the cube, we reach 77.26% as illustrated in the third row of the table. Based of the low-level features and the rough size of the object, the recognition rate is 77.22% as presented in the fourth row of the table. Finally, the fusion of the all the features (the low-level features (all the distances), the rough size and the rough shape of the objects) gives 78.4% The size of the sliding window used here is 50 frames. More results with different sizes of sliding windows will be presented later.

Table 1: Comparison of Online Classification on ORGBD Dataset with state of the art results

Method	Accuracy
Discriminative Orderlet Mining [5]	71.4%
Proposed approach: low-level feature	75.58%
Proposed approach: low-level features and the rough shape of the object	77.26%
Proposed approach: low-level features and the rough size of the object	77.22%
Proposed approach: low-level features, the rough shape and the rough size of the object	78.4%

Relevant Features

We reveal the relevant features for human object interaction recognition. The distances between the object and the joints are selected ones in general. In order to better understand the behavior of the proposed approach, we perform binary classification of each interaction. For action 1 (drinking), we label the

data from action 1 as the first class, the second class includes all the remaining actions. The best features to classify action 1 (drinking) are revealed. We repeat this experiments for all the remaining actions separately. Fig. 13 shows the results of this experiment. The pairwise distances between the the yellow and red joints are the best features to recognize each human object interaction.

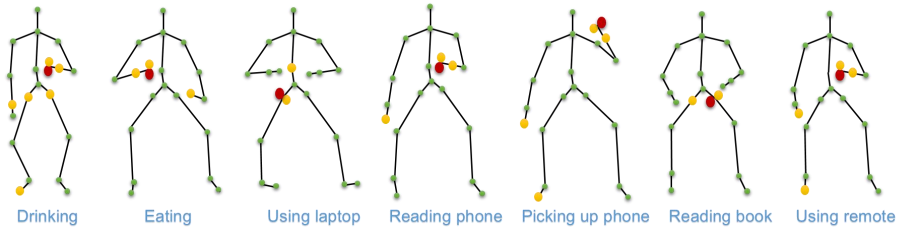

Figure 10: Selected features for each interaction, the best features are the distances between the yellow and red joints.

For example, the best features for drinking (action 1) are the pairwise distances between the object joint and the skeleton joints which are on the right hands, on both sides of the crotch, on the left hand and on the left feet. Another example, for eating (action 2), the best features are the pairwise distances between the object joint and the skeleton joints which are on the left hand and on the right hand. There is another situation, for using laptop (action 3), the best features are the pairwise distances between the object joint and the skeleton joints which are on the crotch and on the spinal part. Based on the attributed distances, we know which joint on the skeleton data for each action is more meaningful for recognizing different human object interaction.

Effect of Temporal Size of the Sliding Window

We have conducted additional experiments when varying the temporal size of the sliding window used to define the sub-sequences. We test different sizes of sliding window on the two kinds of feature discussed above and report results on these two datasets. The recognition rate is reported in Fig. 11 for different window sizes (from 10 to 80). As these results show the fusion of the rough

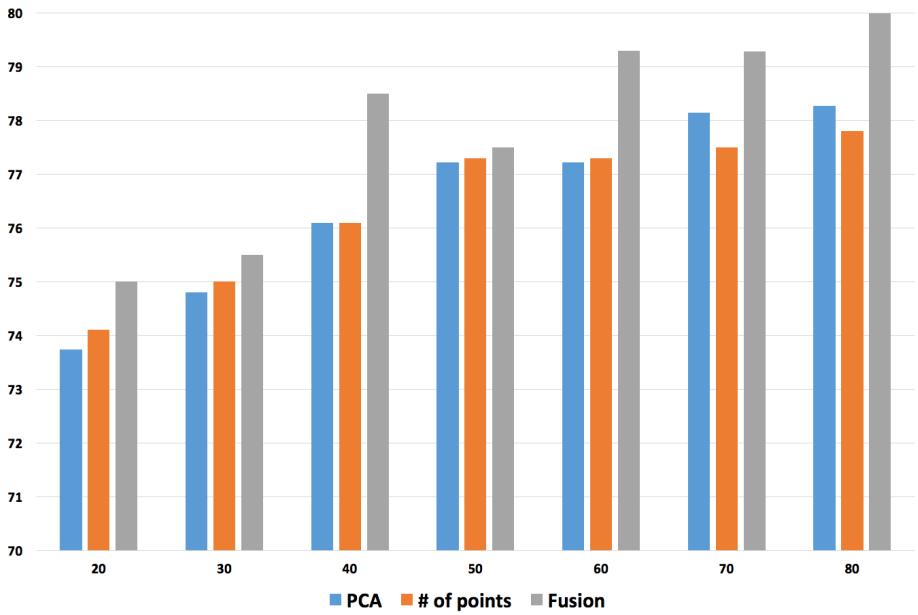

Figure 11: Effect of the temporal size of the sliding window on ORGBD Dataset

shape (PCA) and the rough size (number of points) increases the accuracy of the recognition for the different sizes of window. Moreover, the accuracy increases with the size of the window, the recognition rate is about 75% for a window size of 20 and reaches about 80% when the window size is 80. Actually the bigger the window is, a bigger memory is used for the recognition.

5.3. Comparative results on MSR Daily Activity 3D dataset

Most current methods worked on MSR Daily Activity 3D Dataset only based on the whole sequences, not online classification. We used the same protocol as [5], where we use the videos from half of the subjects for training and the other half for testing. For fair comparison, we show our results comparing with online methods in Table 2. Here we also adopt four different features as introduced in previous subsection with the same memory set to compare with state of the art results. We can see from Table 2, we achieves the classification accuracy

of 76.35% based on the low-level features (all the distances) and 76.9% based on the rough shape. The accuracy reaches 77.2% based on the rough size of the object and 77.5% based on the fusion of all the features (the distances, the rough size and the rough shape). The accuracy revealed in this experience is much better than the frame level accuracy of discriminative orderlet mining [5].

Table 2: Comparison of Online Classification on MSR Daily Activity 3D Dataset with state of the art results

Method	Accuracy
Discriminative Orderlet Mining on Continuous Recognition [5]	60.1%
Proposed approach: low-level feature	76.35%
Proposed approach: low-level feature and the rough size of the object	76.9%
Proposed approach: low-level feature and the rough shape of the object	77.2%
Proposed approach: low-level feature and the rough shape and size of the object	77.5%

In Fig. 12, the recognition rate is reported in for different window sizes (from 10 to 80). As these results show the fusion of the rough shape (PCA) and the rough size (number of points) increases the accuracy of the recognition for the different sizes of window. Moreover, the accuracy increases with the size of the window, the recognition rate is about 72% for a window size of 20 and reaches about 76.5% when the window size is 80. It clearly emerges that the action recognition rate increases when increasing the temporal length of the window. This reveals the importance of the temporal dynamics and shows that the spatio-temporal analysis outperforms a spatial analysis of the frames.

As the property of rate-invariant classification which is not online method, we just evaluated this method on MSR Daily Activity 3D Dataset.

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 [52], [13] and [5]. [3] only used skeleton information that is the same as our work. We used the same experimental setting as [5] and performed on the 2-fold cross-validation which is using the samples of half of the subjects as training

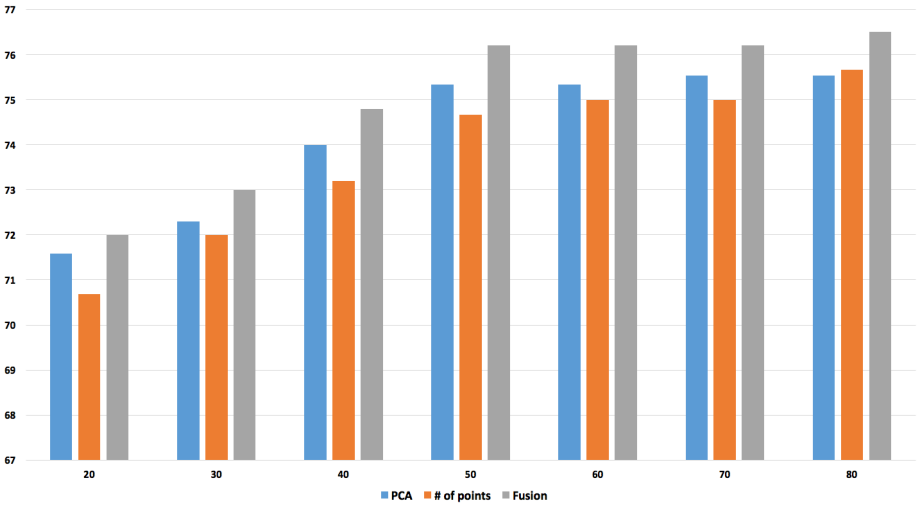

Figure 12: Effect of the temporal size of the sliding window on MSR Daily Activity 3D Dataset

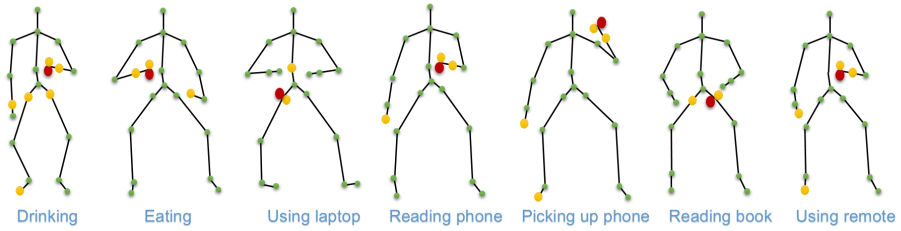

Figure 13: Selected features for each interaction, the best features are the distances between the yellow and red joints.

data, and the samples of the rest half as testing data. The comparison of the performance is shown in Table 3. We can notice in Table 3 that we obtained better accuracy than other works. The accuracy of our approach is 77.05%.

Effect of the number of trees in Random Forest algorithm on off-line Classification

The performance of Random Forest classifier varies with the number of trees. Thus, we perform the experiments with different numbers of trees; the results of this experimentation is shown in Fig. As illustrated in this figure, the recognition

Table 3: Comparison of Rate-invariant Classification on MSR Daily Activity 3D Dataset with state of the art results

Method	Accuracy
Skeleton in [52]	68.0%
4DHOI model [13]	70.0%
Skeletal shape trajectories [3]	70.0%
Discriminative Orderlet Mining on Batch Recognition [5]	73.8%
Proposed approach	77.05%

rate raises with the increasing number of trees until 60, when the recognition rate reaches 72.5%, and then becomes quite stable. Thus, in the following we consider 50 trees and we report detailed results with this number of trees in Fig.14.

To fully evaluate our method, we perform 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 14. 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. A similar behavior of this effect is revealed in the online scenario.

Latency Analysis

The recognition here is still off-line, however, the earlier the decision can be provided the better it is in several applications like abnormal activities detection. For this end, we provide the recognition rate using the first $k \times 10\%$ of the data, with $k = 1, 2...10$. A given test sequence is first modeled as a trajectory in $\mathbb{R}^{210}$ and then the corresponding part of the trajectory is compared to the corresponding part of karcher mean trajectories.

As illustrated in Figure 15, the recognition rate increases slowly using 10 to 40% of the data. Then the slope becomes greater until 70% of the data where the performance reaches about 67%. The improvement further is slower using more data.

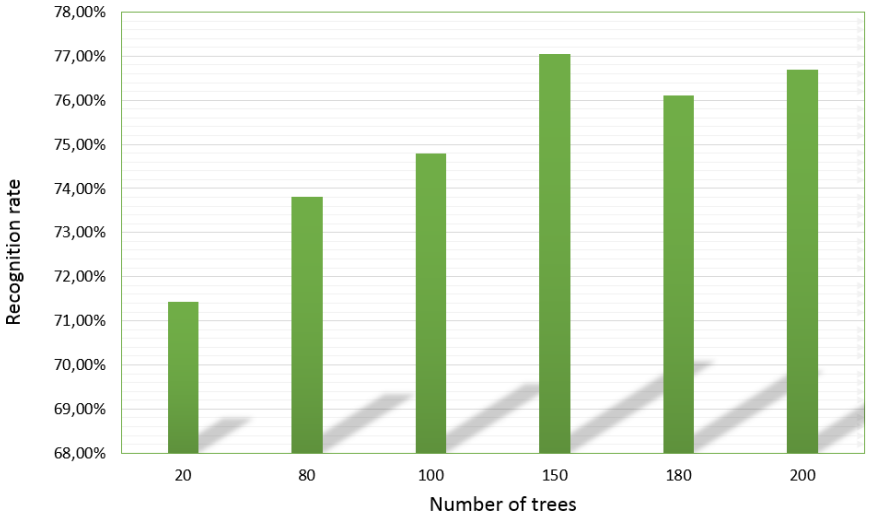

Figure 14: Human-Object interaction recognition results using a Random Forest classifier when varying the number of trees.

5.4. Experiments on Multi-view Human Object Interaction Recognition Lille Douai dataset

Experimental protocol

Due to the new multi-view data, we test our on-line framework on the new dataset in two main scenarios: different views and synchronized view. In the scenario of synchronized view, we divide it into two different experiments based on person independent or not. For each scenario, there are three different protocols according to the properties of the new dataset. In the first protocol, there are two classes which are normal and abnormal. As there are two types of abnormal modalities, we make three classes: normal, injured arm and injured leg in the second protocol. In the third protocol, we class them by different types of activities.

So in the experiments of different views, we use all of the videos from the one of Kinect sensors as the training set and all of the videos from the other. In the experiments of synchronized view on person dependent, due to each

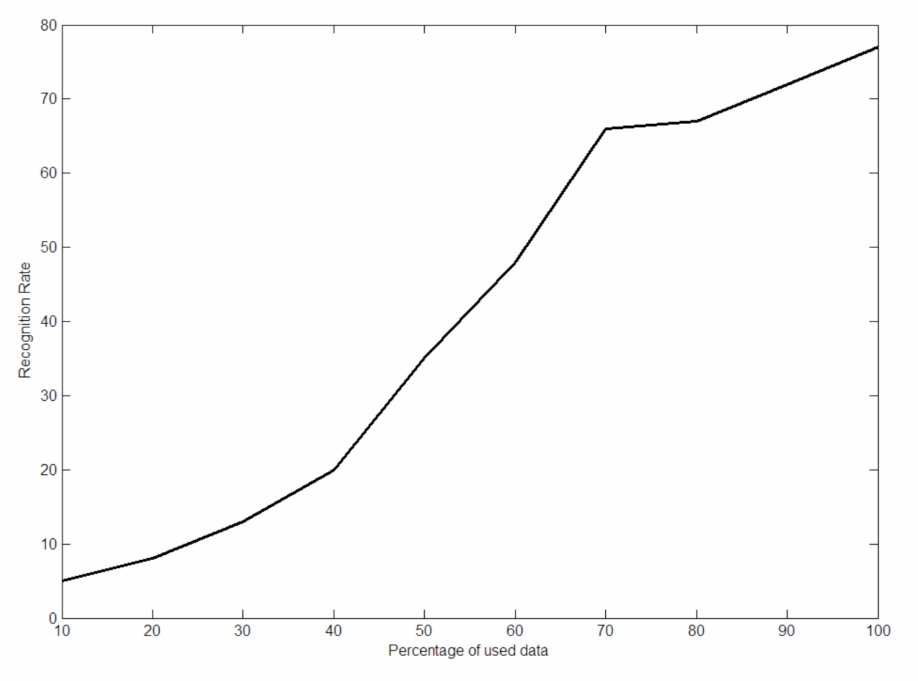

Figure 15: Early detection: trade off between the recognition rate and the percentage of used data.

action performing twice, we use all the first iteration videos as training set and the second time as testing set. In addition, for synchronized view on person independent, we choose all the videos from half actors as training set and the other half of actors' videos as testing data. All these experiments were based on the 2-fold cross-validation.

Experimental results

During these experiments, we built our feature vectors based on the online framework. Due to the different characters of two scenarios, different views experiments use the low-level feature for the task of abnormal and normal human object interaction recognition. In the synchronized view experiments, we build the feature vectors based on fusing the low-level feature of the same action from different view by shape analysis framework.

Table 4: The results of different scenarios for the task of multi-view human object interaction recognition

Protocols	Accuracy (%)
Different views	
Protocol1	74.32
Protocol2	60.61
Protocol3	77.05
Synchronized view on person dependent	
Protocol1	67.12
Protocol2	55.24
Protocol3	70.17
Synchronized view on person independent	
Protocol1	62.21
Protocol3	51.41
Protocol3	65.27

By analyzing Table. 4, it can be noticed that the results of the two scenarios based on the three protocols show the success of the proposed method.

6. Conclusion and future work

In this paper we have presented an effective human object interaction approach. We have also presented results on human object interaction designed to handle variations of pose and rate. This method has several properties that make it appropriate for non-cooperative recognition (rate different, pose variation) and on-line recognition. Firstly, to handle pose variation, we have proposed the inter-joints and objects to joints distances as low-level features. Secondly, to handle rate variations, we have presented a Riemannian analysis of the distance trajectories. This framework offers the possibility to calculate an intrinsic trajectory means which represent the mean variations in several sequences belonging to the same class to build the feature vector that will be fed to SVM classifier. Thirdly, to handle on-line recognition, we have proposed, in addition to the low-level features, an object feature describing roughly the shape and the size of the object. Finally, we have collected a new Multi-view dataset (IMT Lille Douai dataset) with normal and abnormal human-object interaction sequences using two kinects. The experimental results on the MSR Daily Activity 3D, ORGBD and IMT Lille Douai datasets demonstrate the effectiveness of the proposed approach. As future work, we plan to integrate in our framework other descriptors based on both depth and skeleton information. We also expect widespread applicability in domains such as physical therapy and rehabilitation.

References

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

[2] Microsoft, Microsoft kinect, url = http://www.microsoft.com/en-us/kinectfor-windows/.,.

[3] B. B. Amor, J. Su, A. Srivastava, Action recognition using rate-invariant analysis of skeletal shape trajectories, IEEE transactions on pattern analysis and machine intelligence 38 (1) (2016) 1–13.

[4] W. Li, Z. Zhang, Z. Liu, Action recognition based on a bag of 3d points, in: Computer Vision and Pattern Recognition Workshops (CVPRW), 2010 IEEE Computer Society Conference on, IEEE, 2010, pp. 9–14.

[5] G. Yu, Z. Liu, J. Yuan, Discriminative orderlet mining for real-time recognition of human-object interaction, in: Asian Conference on Computer Vision, Springer, 2014, pp. 50–65.

[6] M. Ye, Q. Zhang, L. Wang, J. Zhu, R. Yang, J. Gall, A survey on human motion analysis from depth data, in: Time-of-Flight and Depth Imaging. Sensors, Algorithms, and Applications, Springer, 2013, pp. 149–187.

[7] J. K. Aggarwal, M. S. Ryoo, Human activity analysis: A review, ACM Computing Surveys (CSUR) 43 (3) (2011) 16.

[8] P. Turaga, A. Veeraraghavan, A. Srivastava, R. Chellappa, Statistical computations on grassmann and stiefel manifolds for image and video-based recognition, IEEE Transactions on Pattern Analysis and Machine Intelligence 33 (11) (2011) 2273–2286.

[9] X. Yang, Y. Tian, Super normal vector for activity recognition using depth sequences, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2014, pp. 804–811.

[10] R. Slama, H. Wannous, M. Daoudi, A. Srivastava, Accurate 3d action recognition using learning on the grassmann manifold, Pattern Recognition 48 (2) (2015) 556–567.

[11] R. Anirudh, P. Turaga, J. Su, A. Srivastava, Elastic functional coding of human actions: From vector-fields to latent variables, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 3147–3155.

[12] J. Xiang, R. Liang, Motion recognition and synthesis based on 3d sparse representation, Signal Processing 110 (2015) 82–93.

[13] P. Wei, Y. Zhao, N. Zheng, S.-C. Zhu, Modeling 4d human-object interactions for joint event segmentation, recognition, and object localization, IEEE transactions on pattern analysis and machine intelligence 39 (6) (2017) 1165–1179.

[14] M. Meng, H. Drira, M. Daoudi, J. Boonaert, Human-object interaction recognition by learning the distances between the object and the skeleton joints, in: Automatic Face and Gesture Recognition (FG), 2015 11th IEEE International Conference and Workshops on, Vol. 7, IEEE, 2015, pp. 1–6.

[15] A. Veeraraghavan, A. Srivastava, A. K. Roy-Chowdhury, R. Chellappa, Rate-invariant recognition of humans and their activities, IEEE Transactions on Image Processing 18 (6) (2009) 1326–1339.

[16] A. Srivastava, P. Turaga, S. Kurtek, On advances in differential-geometric approaches for 2d and 3d shape analyses and activity recognition, Image and Vision Computing 30 (6) (2012) 398–416.

[17] M. F. Abdelkader, W. Abd-Almageed, A. Srivastava, R. Chellappa, Silhouette-based gesture and action recognition via modeling trajectories on riemannian shape manifolds, Computer Vision and Image Understanding 115 (3) (2011) 439–455.

[18] M. Marszalek, I. Laptev, C. Schmid, Actions in context, in: Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on, IEEE, 2009, pp. 2929–2936.

[19] A. Prest, V. Ferrari, C. Schmid, Explicit modeling of human-object interactions in realistic videos, IEEE transactions on pattern analysis and machine intelligence 35 (4) (2013) 835–848.

[20] V. Delaitre, D. F. Fouhey, I. Laptev, J. Sivic, A. Gupta, A. A. Efros, Scene semantics from long-term observation of people, in: European conference on computer vision, Springer, 2012, pp. 284–298.

[21] J. Wu, A. Osuntogun, T. Choudhury, M. Philipose, J. M. Rehg, A scalable approach to activity recognition based on object use, in: Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on, IEEE, 2007, pp. 1–8.

[22] H. Kjellström, J. Romero, D. Kragić, Visual object-action recognition: Inferring object affordances from human demonstration, Computer Vision and Image Understanding 115 (1) (2011) 81–90.

[23] E. E. Aksoy, A. Abramov, J. Dörr, K. Ning, B. Dellen, F. Wörgötter, Learning the semantics of object–action relations by observation, The International Journal of Robotics Research 30 (10) (2011) 1229–1249.

[24] D. J. Moore, I. A. Essa, M. H. Hayes, Exploiting human actions and object context for recognition tasks, in: Computer Vision, 1999. The Proceedings of the Seventh IEEE International Conference on, Vol. 1, IEEE, 1999, pp. 80–86.

[25] Y. Zhu, A. Fathi, L. Fei-Fei, Reasoning about object affordances in a knowledge base representation, in: European conference on computer vision, Springer, 2014, pp. 408–424.

[26] B. Yao, L. Fei-Fei, Grouplet: A structured image representation for recognizing human and object interactions, in: Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on, IEEE, 2010, pp. 9–16.

[27] B. Yao, L. Fei-Fei, Recognizing human-object interactions in still images by modeling the mutual context of objects and human poses, IEEE Transactions on Pattern Analysis and Machine Intelligence 34 (9) (2012) 1691–1703.

[28] A. Prest, C. Schmid, V. Ferrari, Weakly supervised learning of interactions between humans and objects, IEEE Transactions on Pattern Analysis and Machine Intelligence 34 (3) (2012) 601–614.

[29] A. Gupta, A. Kembhavi, L. S. Davis, Observing human-object interactions: Using spatial and functional compatibility for recognition, IEEE Transactions on Pattern Analysis and Machine Intelligence 31 (10) (2009) 1775–1789.

[30] B. Yao, S.-C. Zhu, Learning deformable action templates from cluttered videos, in: Computer Vision, 2009 IEEE 12th International Conference on, IEEE, 2009, pp. 1507–1514.

[31] B. Z. Yao, B. X. Nie, Z. Liu, S.-C. Zhu, Animated pose templates for modeling and detecting human actions, IEEE transactions on pattern analysis and machine intelligence 36 (3) (2014) 436–452.

[32] C. Desai, D. Ramanan, C. Fowlkes, Discriminative models for static human-object interactions, in: Computer vision and pattern recognition workshops (CVPRW), 2010 IEEE computer society conference on, IEEE, 2010, pp. 9–16.

[33] C. Desai, D. Ramanan, Detecting actions, poses, and objects with relational phraselets, Computer Vision–ECCV 2012 (2012) 158–172.

[34] G. Chen, D. Clarke, M. Giuliani, A. Gaschler, A. Knoll, Combining unsupervised learning and discrimination for 3d action recognition, Signal Processing 110 (2015) 67–81.

[35] M. Jiang, J. Kong, G. Bebis, H. Huo, Informative joints based human action recognition using skeleton contexts, Signal Processing: Image Communication 33 (2015) 29–40.

[36] H. Pazhoumand-Dar, C.-P. Lam, M. Masek, Joint movement similarities for robust 3d action recognition using skeletal data, Journal of Visual Communication and Image Representation 30 (2015) 10–21.

[37] J. M. Chaquet, E. J. Carmona, A. Fernández-Caballero, A survey of video datasets for human action and activity recognition, Computer Vision and Image Understanding 117 (6) (2013) 633–659.

[38] S.-R. Ke, H. L. U. Thuc, Y.-J. Lee, J.-N. Hwang, J.-H. Yoo, K.-H. Choi, A review on video-based human activity recognition, Computers 2 (2) (2013) 88–131.

[39] M. Ziaeefard, R. Bergevin, Semantic human activity recognition: a literature review, Pattern Recognition 48 (8) (2015) 2329–2345.

[40] T. B. Moeslund, A. Hilton, V. Krüger, A survey of advances in vision-based human motion capture and analysis, Computer vision and image understanding 104 (2) (2006) 90–126.

[41] R. Poppe, Vision-based human motion analysis: An overview, Computer vision and image understanding 108 (1) (2007) 4–18.

[42] D. Grest, J. Woetzel, R. Koch, Nonlinear body pose estimation from depth images, in: DAGM-Symposium, Vol. 5, Springer, 2005, pp. 285–292.

[43] C. Plagemann, V. Ganapathi, D. Koller, S. Thrun, Real-time identification and localization of body parts from depth images, in: Robotics and Automation (ICRA), 2010 IEEE International Conference on, IEEE, 2010, pp. 3108–3113.

[44] S. Knoop, S. Vacek, R. Dillmann, Sensor fusion for 3d human body tracking with an articulated 3d body model, in: Robotics and Automation, 2006. ICRA 2006. Proceedings 2006 IEEE International Conference on, IEEE, 2006, pp. 1686–1691.

[45] D. Anguelov, B. Taskarf, V. Chatalbashev, D. Koller, D. Gupta, G. Heitz, A. Ng, Discriminative learning of markov random fields for segmentation of 3d scan data, in: Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on, Vol. 2, IEEE, 2005, pp. 169–176.

[46] J. Shotton, T. Sharp, A. Kipman, A. Fitzgibbon, M. Finocchio, A. Blake, M. Cook, R. Moore, Real-time human pose recognition in parts from single depth images, Communications of the ACM 56 (1) (2013) 116–124.

[47] J. Gall, A. Fossati, L. Van Gool, Functional categorization of objects using real-time markerless motion capture, in: Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on, IEEE, 2011, pp. 1969–1976.

[48] B. Packer, K. Saenko, D. Koller, A combined pose, object, and feature model for action understanding., in: CVPR, 2012, pp. 1378–1385.

[49] V. G. Kim, S. Chaudhuri, L. Guibas, T. Funkhouser, Shape2pose: Human-centric shape analysis, ACM Transactions on Graphics (TOG) 33 (4) (2014) 120.

[50] O. Oreifej, Z. Liu, Hon4d: Histogram of oriented 4d normals for activity recognition from depth sequences, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2013, pp. 716–723.

[51] J. Qi, Z. Yang, Learning dictionaries of sparse codes of 3d movements of body joints for real-time human activity understanding, PloS one 9 (12) (2014) e114147.

[52] J. Wang, Z. Liu, Y. Wu, J. Yuan, Mining actionlet ensemble for action recognition with depth cameras, in: Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on, IEEE, 2012, pp. 1290–1297.

[53] D. Gong, G. Medioni, Dynamic manifold warping for view invariant action recognition, in: Computer Vision (ICCV), 2011 IEEE International Conference on, IEEE, 2011, pp. 571–578.

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

[55] A. Paiement, L. Tao, S. Hannuna, M. Camplani, D. Damen, M. Mirmehdi, Online quality assessment of human movement from skeleton data, in: British Machine Vision Conference, BMVA press, 2014, pp. 153–166.

[56] R. Vemulapalli, F. Arrate, R. Chellappa, Human action recognition by representing 3d skeletons as points in a lie group, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2014, pp. 588–595.

[57] A. Gupta, S. Satkin, A. A. Efros, M. Hebert, From 3d scene geometry to human workspace, in: Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on, IEEE, 2011, pp. 1961–1968.

[58] M. Devanne, H. Wannous, S. Berretti, P. Pala, M. Daoudi, A. Del Bimbo, 3-d human action recognition by shape analysis of motion trajectories on riemannian manifold, IEEE transactions on cybernetics 45 (7) (2015) 1340–1352.

[59] H. Karcher, Riemannian center of mass and mollifier smoothing, Communications on pure and applied mathematics 30 (5) (1977) 509–541.

[60] Y. Zhou, B. Ni, R. Hong, M. Wang, Q. Tian, Interaction part mining: A mid-level approach for fine-grained action recognition, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2015, pp. 3323–3331.

[61] J. Su, A. Srivastava, F. D. de Souza, S. Sarkar, Rate-invariant analysis of trajectories on riemannian manifolds with application in visual speech recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2014, pp. 620–627.

[62] M. Devanne, H. Wannous, P. Pala, S. Berretti, M. Daoudi, A. Del Bimbo, Combined shape analysis of human poses and motion units for action segmentation and recognition, in: Automatic Face and Gesture Recognition

(FG), 2015 11th IEEE International Conference and Workshops on, Vol. 7, IEEE, 2015, pp. 1–6.

[63] F. Han, B. Reily, W. Hoff, H. Zhang, Space-time representation of people based on 3d skeletal data: A review, Computer Vision and Image Understanding 158 (2017) 85–105.

[64] H. S. Koppula, R. Gupta, A. Saxena, Learning human activities and object affordances from rgb-d videos, The International Journal of Robotics Research 32 (8) (2013) 951–970.

[65] G. E. Batista, M. C. Monard, An analysis of four missing data treatment methods for supervised learning, Applied artificial intelligence 17 (5-6) (2003) 519–533.

[66] A. Srivastava, E. Klassen, S. H. Joshi, I. H. Jermyn, Shape analysis of elastic curves in euclidean spaces, IEEE Transactions on Pattern Analysis and Machine Intelligence 33 (7) (2011) 1415–1428.

[67] R. Mokni, H. Drira, M. Kherallah, Combining shape analysis and texture pattern for palmprint recognition, Multimedia Tools and Applications Journal.

[68] H. Drira, B. B. Amor, A. Srivastava, M. Daoudi, A riemannian analysis of 3d nose shapes for partial human biometrics, in: Computer Vision, 2009 IEEE 12th International Conference on, IEEE, 2009, pp. 2050–2057.

[69] L. Breiman, Random forests, Machine learning 45 (1) (2001) 5–32.