Accepted Manuscript

International Journal of Pattern Recognition and Artificial Intelligence

Article Title:	Contributive Representation based Reconstruction for Online 3D Action Recognition
Author(s):	Mohsen Tabejamaat, Hoda Mohammadzade
DOI:	10.1142/S0218001421500051
Received:	07 June 2019
Accepted:	13 April 2020
To be cited as:	Mohsen Tabejamaat, Hoda Mohammadzade, Contributive Representation based Reconstruction for Online 3D Action Recognition, *International Journal of Pattern Recognition and Artificial Intelligence*, doi: 10.1142/S0218001421500051
Link to final version:	https://doi.org/10.1142/S0218001421500051

World Scientific
www.worldscientific.com

International Journal of Pattern Recognition and Artificial Intelligence

Contributive Representation based Reconstruction for Online 3D Action Recognition

Mohsen Tabejamaat
Department of Electrical Engineering, Sharif University of Technology, Tehran 11155-8639, Iran
m.tabejamaat@sharif.edu

Hoda Mohammadzade*
Department of Electrical Engineering, Sharif University of Technology, Tehran 11155-8639, Iran
hoda@sharif.edu

Recent years have seen an increasing trend in developing 3D action recognition methods. However, despite the advances, existing models still suffer from some major drawbacks including the lack of any provision for recognizing action sequences with some missing frames. This significantly hampers the applicability of these methods for online scenarios where only an initial part of sequences are already provided. In this paper, we introduce a novel sequence-to-sequence representation based algorithm in which a query sample is characterized using a collaborative frame representation of all the training sequences. This way, an optimal classifier is tailored for the existing frames of each query sample, making the model robust to the effect of missing frames in sequences (e.g. in online scenarios). Moreover, due to the collaborative nature of the representation, it implicitly handles the problem of varying styles during the course of activities. Experimental results on three publicly available databases, UTKinect, TST fall, and UTD-MHAD, respectively show 95.48%, 90.91%, and 91.67% accuracy when using the beginning 75% portion of query sequences and 84.42%, 60.98%, and 87.27% accuracy for their initial 50%.

Keywords: Human Machine Interaction; 3D Action Recognition; Contributive Representation based Reconstruction.

1. Introduction

Human Action Recognition (HAR) is one of the fundamental problems in computer vision and automatic surveillance that has been widely used in many applications such as interactive games, smart houses, care robots, automated behavioral monitoring, and video captioning. From the view point of modeling the kinematic topology, HAR shares a great deal of properties with the areas like gait recognition, keystroke dynamics, and signature analysis. However, there is a major difference, a HAR system aims to answer the question "what is happening in a scene?" without

*Corresponding Author.

paying attention to the functor while other methods aims to recognize the identity of actors. Despite the breadth of research carried out in recent years, some problems of HAR systems still remain unsolved. These problems can be divided into three main groups; (1) how to acquire the most suitable type of data, (2) how to localize the interesting parts of sequences, and (3) how to model the kinematic topology of actions.

Traditional methods mostly focused on RGB video cameras. Despite the simplicity and high speed, such streams are dramatically influenced by such factors as scene occlusion, cluttered background, viewpoint variation and non-uniform illumination.

This causes the community's attention to be directed at 3D sensing approaches. The primary models mainely focused on wearable sensors attached on human body. While these methods have been very successful for alleviating the above mentioned shortcomings, also suffer from their own drawbacks which are listed as following; limitations in acquiring data from a distance (due to wires and batteries), the need for cooperation from users (impossibility to be used in people monitoring without their knowledge), and restriction on subject's movements. To overcome these challenges, some researches focused on multi-camera based approaches. These approaches utilize the intrinsic parameters and relative positions of multiple cameras to extract the 3D positions of a scene. Despite the alleviation of wearable sensors' shortcomings, these methods are very time consuming and also suffer from some difficult-to-set parameters. In recent years, the advent of real-time depth sensors (in particular Microsoft Kinect) could overcome all these shortcomings, making a fundamental change in the way of recognizing the actions captured under complex environment. These sensors utilize an infrared launcher to provide a sense of depth in the framework of a 2D-stream. Although, such representation of depth simplifies the process of the recognition (by removing the information on dressing and scene details), it still depends on some disturbing factors like body mass variations and still requires an impressive volume of memory. Inspired by human ability in detecting actions from skeleton trajectories, some researchers were encouraged to develop such an aptitude as a computer vision task. Besides, in 2011, Shotton et al. [1, 2] developed a real-time strategy for body joint estimation from depth-streams which further helped fulfil the realization of real-world action recognition systems.

From the view point of feature representation, HAR algorithms can be mainly disaggregated into three groups; in the first category, the spatio-temporal characteristics of a sequence is represented by a single feature vector. These features can then be used as input to any off-the-shelf dimensionality reduction or classification algorithm. In the second category, actions are modeled as a set of multiple time series. Then, modeling of their temporal characteristics are relegated to a dynamic matching strategy. The methods in the third category represent sequences using a set of initially learned key poses and classify them by histogram based algorithms or Hidden Markov Models (HMMs).

Despite the great success achieved, all the existing action recognition algorithms

suffer from two major drawbacks; In one hand, they have no provisions for dealing with transient changes in the speed (removing some frames from sequences) or style (where some parts of an action is similar to parts of a training sample and some others are alike to parts of another sample in the same class) of performing actions. On the other hand, incorporating attention mechanism is only limited to deep learning strategies, where the performance is remarkably influenced by the shortage of 3D data.

In addition, in spite of the remarkable advances in offline scenario, online activity recognition still remains a challenging task which has been less developed in the literature. This type of recognition refers to identifying an action using a limited number of frames (usually the first part of a sequence) instead of the whole action. It is clear that the main goal of automatic activity recognition is to provide a proper reaction in the same time as or even faster than any human. Therefore, like a human model, it is required for machines to recognize the ongoing actions in the early stages before losing the opportunity for a proper reaction. This has very important implications for many situations like monitoring or protecting humans from harms. As an example of the importance of online recognition, consider the robot-based elderly monitoring, where recognizing a falling action would be valuable only before its completion and not after the fact. Though like humans, machines also do not always require to get all the estimates right while it is not scientifically achievable because many actions have significant initial motion similarities. Note that, it is often sufficient to get the correct prediction among the top estimates, to properly penalize or encourage the subsequent reactions.

Unfortunately, there is a notable lack of studies on online activity recognition via skeletal information. To the best of our knowledge, there are only five published works that have mostly raised the issue, but addressed it with the traditional idea of localized alignment that has been originally developed for dealing with the issue of temporal ordering in offline recognition tasks. Among these methods, the study in [48] is the most related work that has been specifically developed for use in online scenarios. The work tries to find the canonical frame of a query sequence that has the maximum similarity to a specific class of actions. However, using only one frame to represent an action causes it to lose the valuable temporal information of data, which significantly hampers the applicability of this method in recognizing activities with complex spatiotemporal patterns. Moreover, this mechanism does not support a human interpretable justification because there would be a contradiction in collecting a video stream in one hand, and on the other hand, disregarding its temporal information. In addition, missing the canonical frame causes the algorithm to fail. For the works in [34] and [49], the reason behind the ability to use in online mode is related to the ability of dynamic matching strategies (finite observability matrix of the linear dynamic modeling for [34], and dynamic time warping for [49]) in comparing two signals with unequal lengths. Another issue of DTW (Dynamic Time Warping) is with the role of ending points in the aligning process. Silva et

al. [50] showed that only 6% of different additional prefix for a signal may cause a 70% error in its DTW alignment with respect to the original signal. In comparison, the works in [44] and [45] proposed to use a piece-wise matching strategy to localize the aligning of a truncated query sample with a complete training sequence. Such a scheme can properly deal with the latency issue of sequences, however it is highly sensitive to the suitable choice of window size. In fact, it requires one of the selected windows from the training samples to exactly match the truncated query signal which is a very challenging task.

Accordingly, this paper aims to propose a linear representation based 3D action recognition algorithm called Contributive Representation based Reconstruction (CRR) that highly alleviates the influence of missing frames (transient changes in the speed or using only a partial trajectory of actions) on the recognition process. This way, our method is enabled to be applied for online activity recognition where only an initial part of query sequences are provided. In addition, CRR incorporates all the frames of training sequences into the reconstruction of each query sample, resulting in a robust representation against the style change of subjects over time. For this purpose, each action sequence is first encoded as two set of spatial and temporal time series. To avoid any heterogeneity, temporal attributes are only applied as a set of constraints on the reconstruction coefficients of the spatial series (instead of direct combination which is used in the most of the state-of-the-art algorithms). The existing linear representation algorithms have a common drawback. They are only applicable for the signals of the same size. To address this problem, we propose a sequence-to-sequence representation model that represents a weighted combination of a query sample as a weighted linear combination of all the training sequences. This way, the idea of linear representation can be extended to the problems of time series where sequences are not required to have exactly the same length.

In a nutshell, the main contributions of this paper is as follows:(1) Unlike the traditional HAR algorithms, our method seeks the optimal classifier for the current frames of each query sequence, making it suitable for use in online activity recognition tasks. (2) Our method represents each query sample by using the frame combination of various training sequences, leading to a more robust representation against the style change of subjects over time. (3) To the best of our knowledge, our method is the first attempt to develop a sequence-to-sequence linear representation based algorithms. (4) Unlike deep learning or pose based HAR strategies, our method does not rely on massive training data or a challenging task of determining a set of key poses. (5) our method can be easily extended to a nonlinear version using the idea of kernel trick (Section 3.3). (6) the reconstruction coefficients of query samples introduce an attention mechanism into the recognition process of CRR. As far as we know, it is the first time that such mechanism is introduced in a shallow learning based strategy. A set of extensive experimental results for online activity recognition demonstrates the superiority of our method against the-state-of-the-art

approaches.

The rest of the paper is organized as follows. In Section 2, we propose a brief review on the related works for the skeleton based action recognition. Section 3 describes the sparse and collaborative representations for image based applications. Section 4 presents the proposed CRR approach for skeleton based action recognition. Experiments are conducted in section 5 and section 6 concludes the paper.

2. Related works

This section provides a short review on the current practices of recognizing actions from skeleton data, discusses their limitations and analyzes their improvements. Because of different motivations and ideas, this review is conducted from two different perspectives; (1) the way of representing the sequences, and (2) the strategy of temporal encoding.

2.1. *Representation model*

From this viewpoint, available methods can be disaggregated into three main categories; (i) geometrical representation, (ii) manifold techniques, and (ii) deep learning. -Geometrical representation:

The idea dates back to 1995, when Campbell et al. [3] utilized the joint information of skeletons for recognizing ballet moves. They first represented each movement as a set of points in a phase space. Then, a unique curve, with low-order polynomials, is fitted into a subset of this space. Finally, the maximum correlation between the curve models is used for classification. This way, the model only consider the spatial characteristics of sequences, disregarding the temporal ones, which makes the model much simpler, but at the cost of losing some valuable information. Hussein et al. [4] used the covariance matrix of joint trajectories to characterize some daily living activities. Motivated by the idea of temporal pyramids, they exploited a hierarchy of the matrices to incorporate temporal orders into the recognition process. Xia et al. [5] calculated the histograms of body skeletons by counting the joints falling into a set of predefined spatial bins. The K-means algorithm is used to generate a set of categorical posture vocabularies. The visual words of these vocabularies are modeled by a set of class-specific HMMs which are then voted on to perform classification. In [6], a concatenation of the spatial position, speed and acceleration was utilized to describe a set of pose descriptors. Qiao et al. [7] proposed a local action descriptor called Trajectorylet that characterised the location of joints as well as their velocity and displacement information in some short time intervals. They also introduced a discriminative version of their method by exploiting a set of exemplar-SVMs trained on candidate Trajectorylets. Yang et al. [8] utilized the path theory to learn a signature from the statistical and dynamic characteristics of a sequence. Seddik et al. [9] proposed to integrate multiple descriptors including 3D joint positions, temporal gradients, joint pair-wise distances, Euler joint-rotation

angles, and inter-bone rotation quaternion angles in order to enhance the discriminative ability of feature representation. They further applied PCA on the resultant feature vectors as a remedy for the curse of dimensionality problem. Luvizon et al. [10] proposed to characterize each sequence using the relative positions and displacement vectors of skeleton joints and used a combined strategy of the K-means clustering, dimensionality reduction, and VLAD representation to acquire more reliable features. These feature vectors are finally mapped into a discriminative space and classified by using the KNN classifier. Huang et al. [11] proposed to extract the discriminative parts of skeletons using the Out-of-Bag (OB) error estimation of the Random Forest (RF) classifiers trained on the features of skeleton parts. Finally, only joints with high discrimination power are selected as feature descriptors. Jiang et al. [12] presented an algorithm to select a set of more informative joints and used their relative positions to describe motion trajectories. Ding et al. [13] represented each action using a set of discrete symbol sequences. Then, action is modeled by feeding these sequences to a set of Profile HMMs. In [14], normalized per-limb orientations were calculated as the features of each sequence. They further divided the sequences into several pose and motion segments and represented them using a set of multi-layer codebooks. Finally, Random Forest (RF), SVM, and Nave Bayes Nearest Neighbor (NBNN) classifiers were used for classification. Ohn-Bar et al. [15] characterized each action by pairwise affinities between limb-specific relative angles in the spherical coordinate system. The method in [16], utilized the Gaussian mixture model to encode a set of skeletal quads to a Fisher vector. To incorporate the temporal orders into the encoding process, a hierarchy of FVs were extracted at multilevel splits of sequences. Lin et al. [17] characterized each skeleton using the averages velocity of body parts and utilized a graph model to encode the trajectories. In [18], a concatenation of the pair-wise joints differences, atomic motion property of each joint, and offset feature of skeletons was used to describe each action. Then, PCA was applied for reducing the dimensionality of the feature vectors and Nave Bayes Nearest Neighbor was used for classification. Azis et al. [19] introduced a weighted averaging fusion scheme to merge the skeletal data of two or more camera views and claimed to obtain about 10% improvement over the traditional single viewpoint strategies, but at the expense of a heavy skeleton tracking cost. Ofli et al. [20] focused on the role of informative joints in recognizing specific classes of actions. In [21], the normalized distances between skeleton joints and torso are considered to describe the sequences. The K-means algorithm was then applied to select the main postures of each sequence and multiclass SVM with Gaussian kernel was used as classier. Mokari et al. [22] proposed to use Fisher Linear Discriminant Analysis (LDA) for categorizing the frames of each action into a set of pre-defined body states. The actions were then characterized as a sequence of these states and modeled by using a Hidden Markov Model (HMM). The method in [23] proposed to extend this idea with a windowing strategy to remove the interstitial frames that do not belong to any of the pre-defined body states. Wang et al. [24] offered a framework to construct a combined bag-of-word feature representation of poten-

tial energy, kinetic energy, direction variation, and spatial information for skeleton joints. The main disadvantage of these methods is that, such low-level features are usually engineered for some particular types of actions, while there is no guarantees of success if they are applied for other types of activities. -Manifold techniques:

Recent findings show that better accuracies can be achieved if the geometry of non-Euclidean manifolds is incorporated into the spatiotemporal modeling of action sequences. Accordingly, lots of manifold based techniques have been proposed in the past decades which can be broadly categorized into two groups (because manifold based learning is an off-the-shelf strategy, silhouette and RGB based techniques are also included in this categorization): (a) mapping of frame-set, and (b) mapping of dynamic systems. In the first category, each frame of action is represented as a point on a typical manifold and temporal modeling is performed on the manifold structure. In [25], Kendall's shape theory was used for mapping frames on a spherical manifold. Then, distance between two trajectories is calculated using an innovative geodesic-DTW based technique. They also proposed two parametric AR and ARMA models for the tangent space projections of shape sequences so as to resolve the nonvalidity of these models on non-Euclidean manifolds. In [26], quare-root elastic representation algorithm [27,28] was utilized to represent sequences as closed curves in a shape space. Then, similar to [25], two trajectories on this manifold could be compared using a geodesic-DTW. In addition, a graphical-based HMM is also presented to model the trajectories using the high order statistical characteristics of their variations.

Unlike the frame mapping, methods in the second category aim to produce temporal features before characterising the geometry of manifold. Turaga et al. [29] utilized the ARMA model for characterizing a sequence on a Grassman manifold and then classify the models using the Procrustes Distance Metric (PDM) or kernel density functions. In [30], Harandi et al. used infinite-dimensional Covariance Descriptors (CovDs) in a Hilbert space to map a trajectory on a manifold of Symmetric Positive Definite (SPD) and then utilized the properties of Bregman divergences [31] for comparing each two CovDs. Motivated by the theory of block Hankel, Zhang et al. [32] utilized the Gram matrices to embed sequences on a set of Positive Definite (PD) Riemannian manifolds. Then, four distance-like metrics including Affine Invariant Riemannien, Log-Euclidean Riemannian, Jensen-Bregman Log-det Divergence, KL-Divergence were used for comparing the embedded trajectories on the manifolds. The method in [33] used a modified High Order Singular Value Decomposition (HOSVD) to factorize the third order tensor representation of action sequences on a Grassmann manifold. They also applied the concept of tangent bundles on the Grassmann manifold to calculate the distance between two action sequences. Slama et al. [34] utilized the observability matrix of ARMA model for embedding an action on a Grassmann manifold. Then, motivated by the concepts of class specific tangent space and tangent bundle, two Truncated Wrapped Gaussian (TWG) and Local Tangent Bundle SVM (LBTSVM) models were proposed for classification on this manifold. Cherian et al. [57] used a Rank Pooling

technique on a Riemannian manifold embedded in the reproducing kernel Hilbert space for recognizing actions. The major drawback of these methods, however, is that they need some simplifying assumptions about the geometry of the manifold, which may not reflect the actual distribution of data, particularly for databases with large number of actions and too wiggly distribution. -Deep learning:
Currently, deep learning seems to be the main stream of research on pattern recognition methods. Reviews of the literature reveal that Recurrent Neural Network (RNN) and its cousin Long Short-Term Memory (LSTM) network are the most used architectures in the field of 3D action recognition. This comes from the machine state nature of these networks that can easily capture the dynamic structure of actions in addition to their contextual information. In this context, Du et al. [35] proposed a hierarchical RNN framework to fully characterize the multi-part structure of body skeletons. At the first layer, skeleton was partitioned into five major parts including two hands, two feats, and a trunk. Then, each part was fed to a Bidirectional RNN (BRNN). On next layers, the outputs of these networks were hierarchically stuck and then fed to further bidirectional networks. Finally, classification was performed by fully concatenating the representations of the last layer and feeding the result into a softmax one. Zhu et al [36] proposed a regularized LSTM framework to incorporate the conjunction and discriminative information of joints into the learning process of an LSTM network. Veeriah et al. [37] incorporated spatio-temporal information, derived from the derivative of sequences, in the learning process of an LSTM network. They found that the resultant network learns the salient dynamics of an action much faster than the conventional LSTM. Liu et al. [55] used the contextual information of action sequences to push LSTMs toward learning more informative joints. Liu et al. [56], proposed an alternative framework of spatio-temporal LSTM. For this purpose, two sequences of actions over time and labels are respectively considered as temporal and spatial inputs of a 2D LSTM network. Some researchers also utilized the Convolutional Neural Networks (CNNs) [38-40] to remedy the overfitting problem of RNNs (LSTMs) which is mostly induced by the shortage of training data. However, unlike the manifold and geometrical based techniques, the success of deep neural networks is heavily reliant on massive training data, while the current 3D action databases are too small.

2.2. *Encoding Strategy*

Temporal information of sequences can be encoded in two different ways (i) holistic and (ii) atomistic. In holistic strategies, sequences are treated as a whole so that their pose ordering information could not be retrieved from their encoding models. This causes them to fail to recognize partial trajectories from learned activity models. In contrast, atomistic methods take advantages of an object-oriented ensemble coding strategy which allows for preserving the ordering information of skeletons into a learned model, resulting in an ability to recognize actions from partial trajectories which is referred to as online recognition. It is noteworthy to point out that

this sense of online recognition is quite different from the meaning introduced in the literature [64-66] which refers to recognizing different actions from unsegmented streams of data in a continuous manner.

Different from these works, our proposed method aims to focus on recognizing actions with some missing frames in their sequences. A novel sequence-to-sequence collaborative learning strategy is introduced to extend the idea of the Linear Representation (LR) to higher dimensional video processing. Furthermore, unlike the previous online strategies, it does not simply fit a learning model for each temporal segment, but instead concurrently uses the contributive representations of all the training sequences to describe an unknown query sample and therefore provide a robust representation against the style variations a subject over the time of performing an activity while has much lower computational cost.

3. Consensus Representation based Algorithms

Sparse Representation based Classification (SRC) [53] and Collaborative Representation based Classification (CRC) [54] are two most related works to our method which are briefly reviled in this section. Sparse representation is a fundamental theory of compress sensing upon which each signal can be passably reconstructed with as few as possible training samples. Let $\mathbf{X} = [X_1, X_2, ..., X_C] \in R^{\mathfrak{d} \times n}$ be an over-complete dictionary composed of the sub-dictionaries $X_i = [x_{i1}, x_{i2}, ..., x_{in_i}]$; $i = 1, ..., C$ belonging to C difference classes, where $n = \sum_{i=1}^{C} n_i$ is the total number of training samples, $\mathfrak{d}$ is the dimension of feature space, $x_{ij} \in R^{\mathfrak{d}}$ is the j-th sample of the i-th class, and n_i denotes the number of training samples belonging to the i-th class.

Given a query sample y, SRC aims to encode it over the over-complete dictionary $\mathbf{X}$ so that the following equation is approximately satisfied:

$$\begin{aligned} & y = \mathbf{X}\alpha = x_{11}a_{11} + x_{12}a_{12} + ... + x_{Cn_C}a_{Cn_C} \\ & s.t. \quad \tilde{\alpha} = argmin\|\alpha\|_0 \end{aligned} \tag{1}$$

where $\alpha = (a_{11}, a_{12}, ..., a_{Cn_C})^T$ denotes the sparse representation coefficient vector and $\|.\|$ stands for l_0-norm which refers to the number of nonzero values in the vector. Due to the noise, the equation $y = \mathbf{X}\alpha$ is rarely held in the real world applications and hence is mostly revised into the form of $\|y - \mathbf{X}\alpha\|_2 < \epsilon$ to allow for some bounded representation noise. On the other hand, as such a l_0 form of minimization problems is nonconvex and difficult to solve [52], it is often approximated by an l_1-term which results in a convex problem while almost satisfies the condition of sparsity. Considering these issues, equation (1) can be rewritten as follows;

$$\tilde{\alpha} = argmin\|\alpha\|_1 \quad s.t. \quad \|y - \mathbf{X}\alpha\|_2^2 < \epsilon \tag{2}$$

According to the Lagrange multiplier theorem, this problem can be reformulated

as the following unconstrained minimization form;

$$L(\alpha, \lambda) = argmin\|y - \mathbf{X}\alpha\|_2^2 + \lambda\|\alpha\|_1 \tag{3}$$

where λ is a scalar value. This problem has an analytical solution that is calculated in an iterative manner [52]. Then, classification is performed by computing the deviation of the linear combination for the training samples of a specific class from the query sample:

$$identity(y) = argmin(\|y - X_i\alpha_i\|_2^2) \tag{4}$$

where α_i is the coefficient vector associated to the i-th class.
Though this method imputes the discriminative ability of the representation to the sparseness achieved by the l_1-norm minimization problem, Zhang et al. [54] demonstrated that it will be almost preserved while using an l_2-norm that alternatively emphasizes on the collaborative use of samples. Such representation scheme is referred to as Collaborative Representation (CR) and formulated as follows;

$$\tilde{\alpha} = argmin\|\alpha\|_2 \quad s.t. \quad \|y - \mathbf{X}\alpha\|_2^2 < \epsilon \tag{5}$$

Unlike the sparse representation, collaborative manner has a closed-form solution that is achieved by differentiating the equation with respect to α and equating the derivative to zero:

$$\alpha = (X^T X + \lambda I)^{-1} Xy \tag{6}$$

Despite the success of SRC/CRC in signal (image) processing, they do not provide any provisions for such video-stream based applications as action recognition. As a solution one may compress them into one dimensional vectors so that to be applied to SRC/CRC algorithm. However, such a coding scheme does not incorporate the dynamic information of streams into the recognition process. On the other hand, due to the different intrinsic natures, integrating spatial information and motion dynamics into a single feature vector does not seem to be an elegant trick. In addition, it neglects the role of single frames and therefore can not deal with the issue of composite videos made of the frames belonging to different subjects.

4. Contributive Representation based Reconstruction

The central goal is to derive a set of discriminative geometrical features from single skeletons and then to classify the sequences based on the best frame-to-frame similarity between the training and testing sequences. As geometrical features, we use relative angles between the skeleton joints in which the angles reference points are specifically designed for each limb, resulting in a more discriminative ability compared to the algorithms that utilize some fixed reference angles [5,15,67]. As the classification scheme, we try to reconstruct a linear expression of frames of a test sample as a linear expression of whole frames of the training samples. According to

our minimization scheme, those training and test frames with maximum similarities will be assigned to largest reconstruction coefficients. In this way, it brings three major advantages: (1) For test sample, these coefficients determine the relevancy of each frame in performing action. This way, we can simply ignore those assigned almost negligible reconstruction coefficients. (2) The training sample whose frames got the most reconstruction values is determined as the most similar specimen to the test sequence. (3) similar frames in other training samples mitigate the influence of the noisy frames in the most similar one leading to a more complete reconstruction of the test sample.

4.1. *Feature extraction*

Given the 3D coordinate of skeleton joints, our method aims to characterize the sequence using two distinct set of features including the spatial information of poses (relative angles) and their corresponding motion dynamics (spatial displacement). Then, the spatial set is directly used as the basis vectors of a linear representation model while motion dynamics are applied as their corresponding coefficients (or vise versa). This way, such distinct-nature attributes could be integrated in a more efficient manner than the direct combination used in the state-of-the-art algorithms. To reduce the influence of disturbing factors like off-centric movements and varying camera angles, each skeleton is preprocessed before extracting its main characterising features. Accordingly, the hip joint of a skeleton is anchored to the origin and its configuration is rotated so as to be parallel to the x-axis. A further preprocessing is also performed to mitigate the issue of repetitive frames. To do so, we utilize the concept of the kinetic energy in micro-motion patterns. Let $p^{k_f}_{j,i} = \{x_{j,i}, y_{j,i}, z_{j,i}\}; ^{i=1,...,n}_{j=1,...,m}, p^1 = x, p^2 = y, p^3 = z$ be the coordinate of the j-th joint at the i-th frame of an action sequence (in the aligned coordinate system), where m denotes the number of tracked joints and n stands for the length of the sequence. First, the energy of micro-motions is calculated over the consecutive frames;

$$KE(i) = \sum_{j=1}^{m} \sum_{k_f=1}^{3} |p^{kf}_{j,i+1} - p^{kf}_{j,i}| \tag{7}$$

A frame is considered as repetitive if the kinetic energy of its transition is less than a threshold value. However, KE is a content-dependent measure which needs to be normalized so that applying a predefined threshold works with all the sequences;

$$NKE(i) = \frac{KE(i) - min(KE)}{max(KE) - min(KE)};\ i = 1, ..., n-1 \tag{8}$$

Finally, the frames with the normalized energy less than the threshold value (0.15 in this work) are removed from the sequence.

To encode the spatial features, each skeleton is represented with the relative

angles of its projected joints on the three Cartesian planes xy, xz, and yz. First, we utilize the position of head, torso, left- and right- shoulders, and also left- and right- knees to define five reference points:

$$\begin{aligned} O_{1,i}^{k_\mathfrak{f}} &= \frac{p_{head,i}^{k_\mathfrak{f}} + p_{L.shoulder,i}^{k_\mathfrak{f}}}{2}, O_{2,i}^{k_\mathfrak{f}} = \frac{p_{head,i}^{k_\mathfrak{f}} + p_{R.shoulder,i}^{k_\mathfrak{f}}}{2} \\ O_{3,i}^{k_\mathfrak{f}} &= \frac{p_{torso,i}^{k_\mathfrak{f}} + p_{L.shoulder,i}^{k_\mathfrak{f}}}{2}, O_{4,i}^{k_\mathfrak{f}} = \frac{p_{torso,i}^{k_\mathfrak{f}} + p_{R.shoulder,i}^{k_\mathfrak{f}}}{2} \\ O_{5,i}^{k_\mathfrak{f}} &= \frac{p_{L.knee,i}^{k_\mathfrak{f}} + p_{R.knee,i}^{k_\mathfrak{f}}}{2} \end{aligned} \tag{9}$$

where $\mathfrak{f} \in \{1,2,3\},\ k_1 = x, k_2 = y; k_3 = z$. As the coordinate of these points varies with the posture of skeletons, they are referred to as 'dynamic reference points'. Then, the pairwise cosine distance of joints with respect to these anchors are considered as the spatial features of skeletons;

$$\begin{aligned} f_{j_1 j_2,i}^{\{k_1,k_2\},v} &= \frac{< \xi_{j_1,v,i}^{k_1,k_2}, \varsigma_{j_2,v,i}^{k_1,k_2} >}{\|\xi_{j_1,v,i}^{k_1,k_2}\| \|\varsigma_{j_2,v,i}^{k_1,k_2}\|} \\ \xi_{j_1,v,i}^{k_1,k_2} &= \left(p_{j_2,i}^{k_1} - O_{v,i}^{k_1}\ p_{j_1,i}^{k_2} - O_{v,i}^{k_2}\right)^T \\ \varsigma_{j_2,v,i}^{k_1,k_2} &= \left(p_{j_2,i}^{k_1} - O_{v,i}^{k_1}\ p_{j_1,i}^{k_2} - O_{v,i}^{k_2}\right)^T \end{aligned} \tag{10}$$

where $< .,. >$ stands for the inner product, $j_1, j_2 \in \{1,...,m\}$, and $i \in \{1,...,n\}$. Finally, all the relative angles of the three Cartesian planes are concatenated to form the spatial feature vector F_s;

$$F_s^i = \left(f_{11,i}^{\{x,y\},O_1}, f_{12,i}^{\{x,y\},O_1}, ..., f_{(m-1)m,i}^{\{y,z\},O_5}\right)^T \tag{11}$$

Moreover, a dynamic feature vector is also created using the displacement of joints in consecutive frames;

$$F_d^i = \left(p_{1,i+1}^x - p_{1,i}^x, ..., p_{j,i+1}^x - p_{j,i}^x, ..., p_{m,i+1}^z - p_{m,i}^z\right)^T \tag{12}$$

The pseudo code of the feature extraction procedure is listed in Algorithm 1.

4.2. *Coding Scheme*

This section proposes a novel Contributive Representation based Reconstruction (CRR) algorithm for recognizing action sequences from skeletal information.

Let $D = [d_1, ..., d_\mathcal{N}] = [D_{11}, ..., D_{n_1,1}, ..., D_{1,C}, ..., D_{n_C,C}]$ be a dictionary including $\mathfrak{N} = \sum_{\mathfrak{e}=1}^{C} n_\mathfrak{e}$ training sequences belonging to C different classes, where $D_{\iota,\kappa} = [F_1, ..., F_{n_\mathfrak{r}}]$ is the concatenation of $n_\mathfrak{r}$ spatial feature vectors (extracted by the strategy described in section 4.1) corresponding to the skeleton poses of the ι-th action in the κ-th class, ($n_\mathfrak{r}$ is the number of frames for this typical sequence), and $\mathcal{N}$ is the total number of the frames in the dictionary. Given a query sequence

Algorithm 1 Spatiotemporal feature representation of an action sequence

Input: Position of skeleton joints $p_{j,i}^{k_i} = \{x_{j,i}, y_{j,i}, z_{j,i}\}_{j=1,...,m}^{i=1,...,n}$
Output: Spatial and temporal feature vectors of the sequence
for i=1:n **do**
  # Calculate the position of the reference points using equation (9)
  for j_1=1:m **do**
    for j_2=1:m **do**
      for $\{k_1, k_2\} \in \{x, y\}\{x, z\}\{y, z\}$, $j_1, j_2 \in \{1, ..., m\}$ **do**
        # Calculate the joint angles of each skeleton using equation (10)
        # Calculate the spatial feature vector by concatenating all the joint angles during the course of the action (equation (11))
        # Calculate the temporal feature vector using equation (12)
      end for
    end for
  end for
end for

$Y = [y_1, ..., y_{n_Y}]$ with n_Y frames, CRR aims to minimize the deviation between a linear combination of its frames ($\sum \mathfrak{p}_i y_i$) and a linear combination of all the frames in the dictionary ($\sum \mathfrak{q}_i d_i$);

$$min(\| \sum_{i=1}^{n_Y} \mathfrak{p}_i y_i - \sum_{j=1}^{\mathcal{N}} \mathfrak{q}_j d_j \|) \tag{13}$$

This way, any linear combination of skeletons is treated as a valid representation of an action. Therefore, the reconstruction coefficients are allowed to be arbitrary shared among different classes which is not appropriate for such classification tasks as action recognition. So, we need some constraints to be applied on the reconstruction coefficients to push them towards learning a class-specific solution. For this, we introduce three constraints on the minimization problem of equation (13) which incorporates the structural properties of classes into the linear representation of a sequence;

(1) $\sum \mathfrak{p}_i = 1$ and $\sum \mathfrak{q}_j = 1$ to align the lower bounds of the linear combinations
(2) $min\|\mathfrak{P}\|_2^2$ and $min\|\mathfrak{Q}\|_2^2$ to allow correlated variables to enter the model and avoid outliers to be incorporated into the reconstruction precess. Here $\mathfrak{P} = (\mathfrak{p}_1, \mathfrak{p}_2, ..., \mathfrak{p}_{n_Y})$, and $\mathfrak{Q} = (\mathfrak{q}_1, \mathfrak{q}_2, ..., \mathfrak{q}_{\mathcal{N}})$.
(3) $min \sum \mathfrak{q}_j^2 \|H - U_\varpi\|_2^2$ to incorporate the dynamic structure of actions into the reconstruction process. Here, j-th index belongs to the ϖ-th sequence, H and U_ϖ are the down-sampled [a] frames of the temporal feature vec-

[a]The sampling rate has to be determined empirically and in this paper is set to ensure the sequence

tors respectively for the query and ϖ-th training sequences extracted by equation (12): $H \leftarrow F_d^{query}$ and $U_\varpi \leftarrow F_d^\varpi$. This allows the dynamic information of each sequence to be considered as a whole. To better explain the role of this constraint, consider two sequences consisting of similar frames but with different temporal orders, e.g., sequences of actions picking up an object and placing down an object. Without this constraint, these two sequences obtain the same reconstruction coefficients, whereas by incorporating this constraint in the optimization, different reconstruction coefficients are yielded which are appropriate for classification purposes.

Therefore, the optimization problem of CRR can be rewritten as follows;

$$\begin{aligned} \min_{\mathfrak{p},\mathfrak{q}} \quad & (\| \sum_{i=1}^{n_Y} \mathfrak{p}_i y_i - \sum_{j=1}^{\mathcal{N}} \mathfrak{q}_j d_j \|) \\ & min\|\mathfrak{P}\|_2^2 \quad min\|\mathfrak{Q}\|_2^2 \\ \text{s.t.} \quad & \sum \mathfrak{p}_i = 1 \quad \sum \mathfrak{q}_j = 1 \\ & min \sum \mathfrak{q}_j^2 \|H - U_\varpi\|_2^2. \end{aligned} \tag{14}$$

According to the Lagrange multiplier theorem, this constrained problem can be reformulated in an unconstrained form;

$$\begin{aligned} \min_{\mathfrak{P},\mathfrak{Q}} \quad & \|Y\mathfrak{P} - D\mathfrak{Q}\| + \gamma_1\|\mathfrak{P}\|_2 + \gamma_2\|\mathfrak{Q}\|_2 + \gamma_3(1 - \sum_{i=1}^{n_Y} \mathfrak{p}_i) + \\ & \gamma_4(1 - \sum_{j=1}^{\mathcal{N}} \mathfrak{q}_i) + \gamma_5 \sum \mathfrak{q}_j^2 \|H - U_\varpi\|_2^2 \end{aligned} \tag{15}$$

Solving the above equation, we get the optimal coefficients of $\mathfrak{P}$ and $\mathfrak{Q}$ which determine the importance of each skeleton (frame) in the recognition of the query sequence. This way, the model learns which parts of sequences should be paid more attention to. Rewriting this equation in a matrix form, we have:

$$\begin{aligned} \min_{\mathfrak{P},\mathfrak{Q}} \quad & \| \left(Y - D\right) \begin{pmatrix} \mathfrak{P} \\ \mathfrak{Q} \end{pmatrix} \| + \gamma_1 \mathfrak{P}^T \mathfrak{P} + \gamma_2 \mathfrak{Q}^T \mathfrak{Q} \\ & + \gamma_3(1 - e\mathfrak{P}) + \gamma_4(1 - t\mathfrak{Q}) + \gamma_5 \sum \mathfrak{q}_j^2 \|H - U_\varpi\|_2^2 \end{aligned} \tag{16}$$

where $e = \left(1, 1, ..., 1\right) \in R^{1\times n_Y}$ and $t = \left(1, 1, ..., 1\right) \in R^{1\times \mathcal{N}}$ are two horizontal all-ones vectors. After some algebraic simplifications, we obtain a hyper-variable

length is 5.

optimization problem as;

$$\begin{aligned}\min_{\mathfrak{P},\mathfrak{Q}} \ & \| \big(Y - D\big) \begin{pmatrix}\mathfrak{P}\\ \mathfrak{Q}\end{pmatrix} \| + \big(\mathfrak{P}^T \ \mathfrak{Q}^T\big) \begin{pmatrix}\gamma_1 I & 0\\ 0 & \gamma_2 I\end{pmatrix} \big(\mathfrak{P} \ \mathfrak{Q}\big) \\ & \gamma_3 (1 - \big(e \ \varrho\big)) \begin{pmatrix}\mathfrak{P}\\ \mathfrak{Q}\end{pmatrix} + \gamma_4 (1 - \big(\chi \ t\big)) \begin{pmatrix}\mathfrak{P}\\ \mathfrak{Q}\end{pmatrix} \\ & + \gamma_5 \sum \mathfrak{q}_j^2 \|H - U_\varpi\|_2^2 \end{aligned} \tag{17}$$

Where $\varrho = \big(0, 0, ..., 0\big) \in R^{1\times\mathcal{N}}$ and $\chi = \big(0, 0, ..., 0\big) \in R^{1\times n_Y}$ are all-zeros horizontal vectors. Let $A = \big(Y - D\big)$, $v = \begin{pmatrix}\mathfrak{P}\\ \mathfrak{Q}\end{pmatrix}$, $B = \begin{pmatrix}\gamma_1 I & 0\\ 0 & \gamma_2 I\end{pmatrix}$, $l = \big(e \ \varrho\big)$, $h = \big(\chi \ t\big)$. Differentiating this equation and setting the result equal to zero, we get;

$$\begin{aligned} & A^T A v + v^T B v + \gamma_3 (1 - lv) + \gamma_4 (1 - hv) \\ & + \gamma_5 diag(\big(\varrho \ , \big(\|H - U_1\|, \|H - U_2\|, ..., \|H - U_n\|\big)\big)) v = 0 \end{aligned} \tag{18}$$

As all the square matrices are invertible, this problem has a closed-form solution:

$$\begin{aligned} & v = (A^T A + B + \\ & \gamma_5 diag(\big(\varrho \ , \big(\|H - U_1\|, ..., \|H - U_n\|\big)\big))))^{-1} (\gamma_3 l + \gamma_4 h) \end{aligned} \tag{19}$$

Algorithm 2 Contributive Representation base Reconstruction

Input: Spatial training set D, spatial query set Y, dynamic training set F_d^{ϖ}, dynamic query set F_d^{query}
Output: Optimal coefficient vector
Initialization $e = \big(1, 1, ..., 1\big)$, $\varrho = \big(0, 0, ..., 0\big)$, $t = \big(1, 1, ..., 1\big)$, $\chi = \big(0, 0, ..., 0\big)$
Down-sample the dynamic training set of each class and query sample $U_\varpi \leftarrow F_d^{\varpi}$, $H \leftarrow F_d^{query}$
$A \leftarrow \big(Y - D\big)$
$B \leftarrow \begin{pmatrix}\gamma_1 I & 0\\ 0 & \gamma_2 I\end{pmatrix}$
$l \leftarrow \big(e \ \varrho\big)$
$h = \big(\chi \ t\big)$
Code Y over D by equation (19)
Compute the reconstruction coefficients for each class
$\forall \nu \in 1, ..., C, \quad AC_\nu = \frac{1}{n_{\triangle_o}} \sum_{o \in \nu th \ class} \triangle_o; \triangle_o \triangleq (\mathfrak{Q}_o | \mathfrak{Q}_o > \zeta \max(\mathfrak{Q}))$, where $n_{\triangle_o}$ is the number of elements in the sequence $\triangle_o$ and ζ is a regularization parameter
Assign Y to $\mathfrak{g}$-class if $\mathfrak{g} = argmax_\nu \big(Av_\nu\big)$

It is clear that the main cost is to solve the inverse term of $A^T A$. Therefore, the computational complexity could be almost estimated as $O(\mathcal{N}^3)$. Our codes were written in Matlab and run on an Intel(R) Core(TM) i5 (3.4GHz) PC.

4.3. *Kernel CRR*

Sometimes, data points are hard to linearly separate, especially when some of the classes have the same direction distribution. In this case, it is better to nonlinearly map the data into a higher dimensional space and then separate it in a linear manner. Nevertheless, optimizing a separation model in an unknown high-dimensional space is not an easy task. In contrast, kernel trick enables such models to be applied in an implicit high dimensional space without the need for computing the coordinates of the data in that space, and thus has been widely applied to many LR algorithms like PCA, LDA, SVM, LSDA, SRC, and CRC. However, unlike these methods, CRR aims to classify the trajectories (and not single vectors) and therefore requires their points (frames) to be individually transformed into the Hilbert space, leading the dimensionality of this space to be equal to the total number of the frames in the database ($\mathcal{N}$).

Let ϕ be a nonlinear mapping from the space R^m to a high-dimensional space $R^{\mathcal{N}}$; $\phi : R^m \longrightarrow R^{\mathcal{N}}$, so that

$$\phi(y_i) = (\phi_1(y_i), \phi_2(y_i), ..., \phi_{\mathcal{N}}(y_i)) \tag{20}$$

where $\phi(y_i)$ is a point of the trajectory φ in the new space which is defied as follows;

$$\varphi = (\phi(y_1), \phi(y_2), ..., \phi(y_{n_Y})) \tag{21}$$

Similarly, we project each frame of the dictionary D into the new space so that we obtain;

$$\Phi = (\phi(d_1), \phi(d_2), ..., \phi(d_{\mathcal{N}})) \tag{22}$$

As each Now, defining the optimization problem of CRR on the new space, we get the following equation;

$$\begin{aligned} \min_{\mathfrak{P},\mathfrak{Q}} \quad & \|\varphi\mathfrak{P} - \Phi\mathfrak{Q}\| \\ \text{s.t.} \quad & \sum \mathfrak{p}_i = 1 \quad \sum \mathfrak{q}_j = 1 \\ & min\|\mathfrak{P}\|_2^2 \quad min\|\mathfrak{Q}\|_2^2 \\ & min \sum \mathfrak{q}_j^2 \|H - U_{\varpi}\|_2^2. \end{aligned} \tag{23}$$

However, due to the high dimensionality, this equation is much harder to solve than equation (14). To alleviate this problem, we define a coefficient matrix so that to be a linear combination of the dictionary elements in the new space;

$$R = \Phi\mathfrak{Z} = (\phi(d_1), \phi(d_2), ..., \phi(d_{\mathcal{N}}))(z_1, z_2, ..., z_{\mathcal{N}}) \tag{24}$$

where $\mathfrak{Z} = (z_1, z_2, ..., z_{\mathcal{N}})$ is called the pesodu-transformation matrix. Multiplying R to the first row of equation (23), we obtain;

$$\min_{\mathfrak{P},\mathfrak{Q}} \quad \|R^T\varphi\mathfrak{P} - R^T\Phi\mathfrak{Q}\| = \|\mathfrak{Z}^T\Phi^T\varphi\mathfrak{P} - \mathfrak{Z}^T\Phi^T\Phi\mathfrak{Q}\| \tag{25}$$

As $K(D,Y) = \Phi^T\varphi$ and $K(D,D) = \Phi^T\Phi$, equation (25) can be reformulated as follows;

$$\min_{\mathfrak{P},\mathfrak{Q}} \quad \|\mathfrak{Z}^T K(D,Y)\mathfrak{P} - \mathfrak{Z}^T K(D,D)\mathfrak{Q}\| \tag{26}$$

Removing $\mathfrak{Z}^T$ from both sides of the differential equation, the optimization problem of CRR (in the reproducing kernel Hilbert space) can be finally written as follows;

$$\begin{aligned} \min_{a,b} \quad & \|K(D,Y)\mathfrak{P} - K(D,D)\mathfrak{Q}\| \\ \text{s.t.} \quad & \sum \mathfrak{p}_i = 1 \quad \sum \mathfrak{q}_j = 1 \\ & min\|\mathfrak{P}\|_2^2 \quad min\|\mathfrak{Q}\|_2^2 \\ & min \sum \mathfrak{q}_j^2 \|H - U_{\varpi}\|_2^2. \end{aligned} \tag{27}$$

For classification, we calculate the average reconstruction coefficients for all the action classes using $AC_\nu = \frac{1}{n_{\triangle_o}} \sum_{o \in \nu th\ class} \triangle_o; \triangle_o \triangleq (\mathfrak{Q}_o | \mathfrak{Q}_o > \zeta \max(\mathfrak{Q}))$, where $n_{\triangle_o}$ is the number of elements in the sequence $\triangle_o$ and ζ is a regularization parameter that allows suppressing the noisy contributions in the reconstruction process. Then, we choose the class with the maximum average contribution as the prediction label of the query sample Y. The update procedure of the method is listed in Algorithm 2.

5. Experimental Results

This section provides the experimental frameworks for evaluating the performance of our method on three publicly available databases; UTKinect [6] and TST [58], and UTD-MHAD [59].

The UTKinect database includes 199 action sequences acquired from 10 subjects where each action is performed twice. The actions include walk, sit down, stand up, pick up, carry, throw, push, pull, wave and clap hands. All the sequences were collected indoor using a Kinect depth camera at a distance of 4 to 11 ft with varying lengths from 5 to 120 frames. The actions were carried out in a continuous manner and then separated using manual segmentation which results in some latency in movements. This issue along with the varying styles of subjects, and significant variations of camera angle are the most challenges of this database.

The TST database was originally established for detecting falling action, yet it includes a variety of daily living activities, making it also suitable for evaluating action recognition tasks. TST was collected in 2015 by using a low noise Kinect V2 sensor in an indoor environment and contains 264 sequences from 11 subjects, performing 8 different actions. The actions include sit, grasp, walk, and lying down, falling front, back, side and falling backward while ends up sitting. Compared to UTKinect, TST provides four additional skeleton joints for hand fingers which are

ignored in this work. The challenge of this database is more related to the falling actions that are heavily influenced by the varying styles of subjects.

UTD-MHAD [59] includes 861 multimodal sequences (RGB, depth map, skeletal, and inertial sensor signals) from 8 different subjects (4 females and 4 males), performing 27 actions in an indoor environment. All the depth maps have been collected using a Microsoft Kinect camera at a frame rate of 30 fps. The database contains actions from 3 different groups: (i) daily living activities (right arm swipe to the left, right arm swipe to the right, right hand wave, two hand front clap, right arm throw, cross arms in the chest, right hand knock on door, right hand catch an object, right hand pick up and throw, walking in place, sit to stand, stand to sit), (ii) hand gesture (right hand draw x, right hand draw circle (clockwise), right hand draw circle (counter clockwise), draw triangle), and (iii) sport-training exercises (basketball shoot, bowling (right hand), front boxing, baseball swing from right, tennis right hand forehand swing, arm curl (2 arms), tennis serve, two hand push, jogging in place, forward lunge (left foot forward), squat (2 arms stretch out)).

5.1. *Experimental results in online mode*

This section analyzes the efficiency of our proposed algorithm for online activity recognition. For this purpose, we compare the performance of our method with two state of the art algorithms proposed in [44] and [45]. The superior performance of these methods compared to other works in [34], [48] and [49] have been already demonstrated in the original papers. Regarding the work in [34], the offline recognition rate provided by this method (88.5%) on the UTKinect database is much less than the online recognition rate of our method achieved when using the beginning 75% portion of the frames (94.47%). Note that, online recognition rates reported in the literature are always less than their corresponding offline accuracies. For other rivals, in order to avoid any re-implementation uncertainty, the results are directly reported from the original papers.

Following the strategy of Hayes et al. [45] for online recognition, we form our training dictionary using the complete training sequences and then evaluate the system by using the begining 25%, 50%, and 75% portions of query sequences. For both UTKinect and TST fall databases, we conduct k-fold cross subject validation strategy, where in each fold the samples belonging to one subject are excluded for testing while the remaining ones are used to form the dictionary. In contrast, to make a fair comparison with other studies, we apply two-fold cross validation protocol on the UTD-MHAD database where the sequences of odd subjects (only the first trial of each subject) are used for training and all the trials of even subjects are used for test. Note that, UTD-MHAD incudes some very subtle movements in which temporal information are more discriminative than the spatial ones. Therefore, it

Table 1: Our method vs. the state-of-the-arts on UTKinect database in online mode

% of sequence	Our method	SDSR	RAPTOR
25	67.84%	70%<	**79.4**%
50	**84.42**%	72%<	83.1%
75	**95.48**%	85%<	84.7%

Table 2: Our method on TST database in online mode

% of sequence	Our method
25	**20.08**%
50	**60.98**%
75	**90.91**%

Table 3: Our method on UTD-MHAD database in online mode

% of sequence	Our method
25	**55.79**%
50	**87.27**%
75	**91.67**%

is more useful to reformulate equation (14) in the following form:

$$\begin{aligned} \min_{\mathfrak{p},\mathfrak{q}} \quad & (\| \sum_{i=1}^{n_Y} \mathfrak{p}_i h_i - \sum_{j=1}^{\mathcal{N}} \mathfrak{q}_j u_j \|) \\ \text{s.t.} \quad & \sum \mathfrak{p}_i = 1 \quad \sum \mathfrak{q}_j = 1 \\ & min \|\mathfrak{P}\|_2^2 \quad min \|\mathfrak{Q}\|_2^2 \\ & min \sum \mathfrak{q}_j^2 \|Y - D_\varpi\|_2^2. \end{aligned} \tag{28}$$

For all the databases, based on empirical parameter estimates, we set the regularization parameters similarly as $\lambda_1 = 10^{-2}$, $\lambda_2 = 10^{-2}$, $\lambda_3 = 10^{-1}$, and $\lambda_4 = 10^{-1}$. But differently we use $\lambda_5 = 2 \times 10^{-3}$ for UTKinect and TST fall and $\lambda_5 = 2 \times 10^{-5}$ for UTD-MHAD database. The classification parameter ζ, is set to 0 for UTKinect and UTD-MHAD and 0.6 for TST fall. To avoid any overfitting, we reduce the dimensionality of the dictionary and query sequences using PCA and LPP before applying the main phase of CRR. Tables 1-3 show the online recognition rates of our method compared with other state-of-the-art algorithms. As can be seen, our method achieves satisfactory results and outperforms other rivals, if any, especially when using a moderate number of frames.

The confusion matrices of our method in different online scenarios are separately illustrated in Fig. 1. The thing to notice is about the confusion matrix of the TST using only the first 25% of the sequences where most of the actions are confused with walk and grasp. That is because none of the actions in this database are stationary and mostly start with a movement similar to walking. Note that the detailed description of grasp action is "walking and grasping an object from the floor".

An important issue is that, although part of the missclassifications occur due to the alignment error between the complete content and a piece of a signal, but another part is related to the loss of discriminative information in the truncated part. Accordingly, we conduct another experiment to discriminate between the role of alignment precision and the role of discriminative information in early recognition tasks. For this purpose, we truncate the training sequences according to the truncation procedure of query samples. i.e. if we use the 25% of the initial part of a query sample, the dictionary is also created by using the beginning 25% portion of the training sequences. The recognition rates of this experiment are listed in Table 4. The confusion matrices are also shown in Fig. 2. One can compare the results with Tables 1-3 and infer that a piece wise matching strategy would be more helpful for the more initial parts of sequences (that meets the results presented by the works in [44] and [45]), but as the Fig. 2(c), 2(f), and 2(i) indicate, it is not very suitable for larger portions of sequences (the truncated sample does not exactly fit the selected window of the training sample).

Table 4: Performance of our method in online mode using similar truncated training samples

Database	25	50	75
UTKinect	80.90%	86.43%	92.96%
TST	46.21%	71.97%	89.77%
UTD-MHAD	59.95%	89.58%	92.13%

In addition, we also evaluate the performance of our method in an offline scenario where the entire frames of sequences are considered to be available. Splitting of databases into the training and query sequences is also performed in the similar way as performed in the online scenario. Tables 5-7 list the offline recognition rates of our method compared with a set of the state-of-the-art methods respectively on UTKinect, TST fall, and UTD-MHAD databases.

Abbreviations are as follows:

RR: Recognition rate, Ntr: Nature, Ex. Pr.: Experiment protocol, Alg. Pr.: alignment protocol, RP: Rank Pooling, RM: Recurrent Memory, FP: Fourier Pyramid, OBARMA: Observability Matrix of ARMA model, PWM: Piece-Wise Matching. LOSubO: Leave One Subject Out where in each fold, one Subject is excluded for

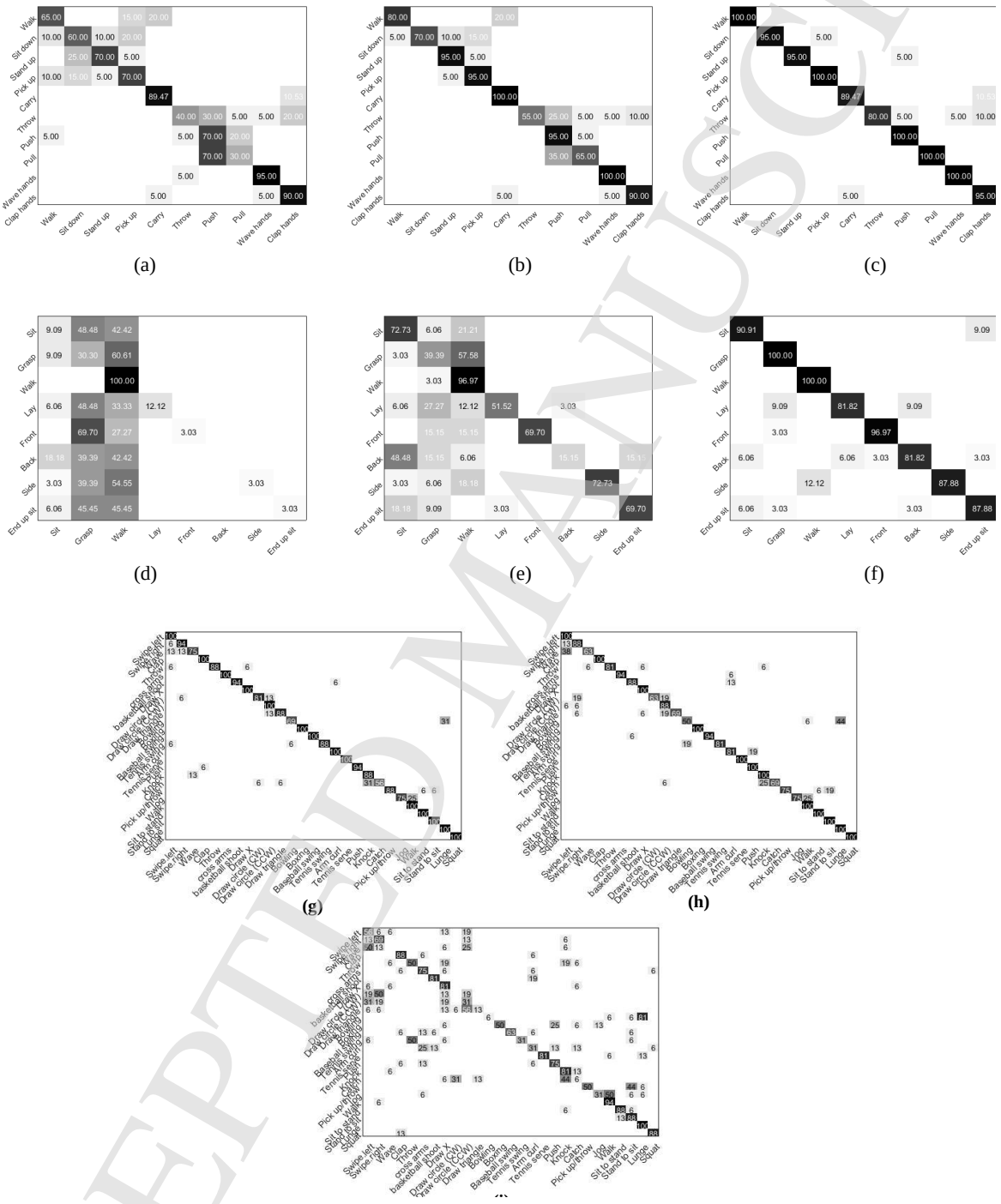

Fig. 1: Confusion matrices in different online scenarios for the (a) first 25% of query sequences in UTKinect database, (b) first 50% of query sequences in UTKinect database, (c) first 75% of query sequences in UTKinect database, (d) first 25% of query sequences in TST database, (e) first 50% of query sequences in TST database, (f) first 25% of query sequences in TST database, (g) first 25% of query sequences in UTD-MHAD database, (h) first 50% of query sequences in UTD-MHAD database, (i) first 75% of query sequences in UTD-MHAD database,

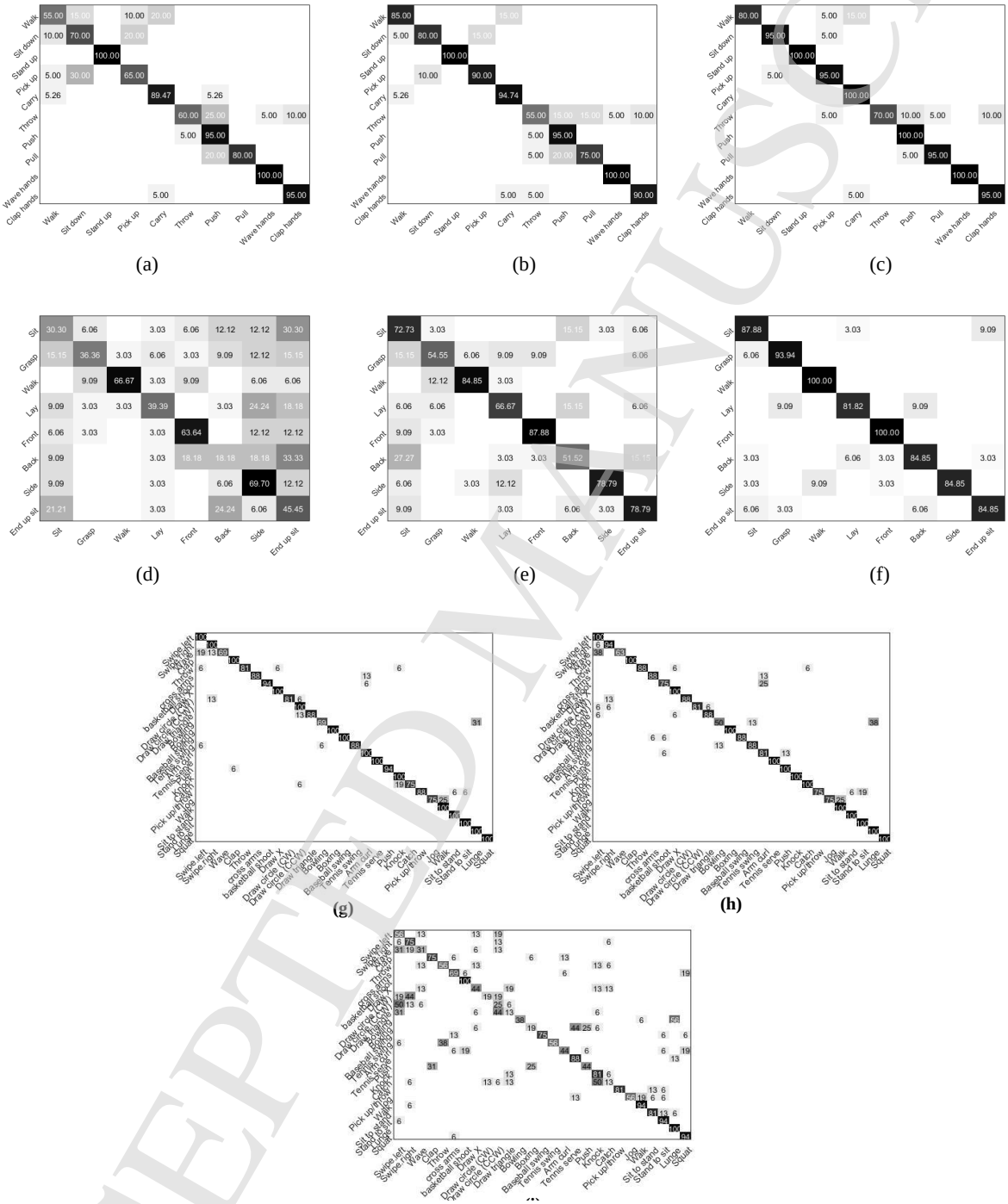

Fig. 2: Confusion matrices in different online scenarios using similar truncated training samples for (a) the first 25% of query sequences in UTKinect database, (b) the first 50% of sequences in UTKinect database, (c) the first 75% of sequences in UTKinect database, (d) the first 25% of sequences in TST database, (e) the first 50% of sequences in TST database, (f) the first 25% of sequences in TST database, (g) first 25% of sequences in UTD-MHAD database, (h) first 50% of sequences in UTD-MHAD database, (i) first 75% of sequences in UTD-MHAD database,

Table 5: Our method vs. the state-of-the-arts on UTKinect database in offline mode

Method	RR	Ntr	Ex. Pr.	Year	Alg. Pr.
ST-LSTM-FFNN [55]	97%	Offline	LOSeqO	2017	RM
GCA-LSTM [55]	97.5%	Offline	LOSeqO	2017	RM
GCA-LSTM-Att [55]	98.5%	Offline	LOSeqO	2017	RM
ST-LSTM-TGate [56]	97%	Offline	LOSeqO	2018	RM
KRP-FS [57]	99%	Offline	LOSeqO	2018	RP
LM^{3}TL [42]	98.8%	Offline	LOSeqO	2016	DTW/FP
DMIMTL [43]	99.19%	Offline	LOSeqO	2017	DTW
His3DJ[5]	90.92%	Offline	LOSeqO	2012	HMM
SDSR[44]	96.97%	Offline	Twofold	2016	PWM
FisherPose[23]	89.0%	Offline	LOSubO	2017	HMM
DNLGF[62]	96.68%	Offline	LOSubO	2018	LieCrvs
LoGM[34]	88.5%	Online	LOSeqO	2015	OBARMA
SDSR[44]	88.89%	Online	Twofold	2016	PWM
RAPTOR[45]	92.1%	Online	LOSubO	2017	PWM
KCRR+PCA(600)	95.98%	Online	LOSubO	-	LR
KCRR+LPP(600)	95.47%	Online	LOSubO	-	LR

testing and the remaining subjects are used for training, LOSeqO: Leave One Sequence Out where in each fold, one sample is excluded for test and remaining ones are used for training. Twofold: only once, half actions are used for training and other half for test, LR: Linear Representation.

As can be seen, our method outperforms the atomistic rivals that has been evaluated under the same protocol (approximately 3.9% better than the performance of RAPTOR). However, the results exhibit a relative superiority of the performance for the holistic algorithms, which is mainly related to their ability in assimilating the information of static poses into the discriminative frames of sequences. Moreover, it has to be mentioned that, LOSubO protocol used in our experiments is far more difficult than LOSeqO used in the most of the state-of-the-art algorithms, because it evaluates the robustness of strategies against the style variation of subjects.

The confusion matrices of our method are also shown in Fig. 3. One can easily find that the main confusions occur between "throw and push", and "walk and carry" for the UTKinect database, between "sit and end up sit", and "lay and back falling" for the TST database, and between "bowling and lunge" for UTD-MHAD database, which is largely due to their motion similarities.

5.2. *Sensitivity Analysis*

A significant advantage of our linear representation based method is its capability of predicting actions even with a non-ideal estimation of skeleton joints. To validate

Table 6: Our method vs. the state-of-the-arts on TST database in offline mode

Method	RR	Ntr	Ex. Pr.	Year	Alg. Pr.
His3DJ[5]	70.83%	Offline	LOSubO[b]	2012	HMM
FisherPose[23]	88.64%	Offline	LOSubO	2017	HMM
SJOTT[47]	92.8%	Offline	LOSubO	2017	DTW
LoGM[34]	88.5%	Online	LOSeqO	2015	OBARMA
KCRR+PCA(600)	89.39%	Online	LOSubO	-	LR
KCRR+LPP(600)	91.66%	Online	LOSubO	-	LR

Table 7: Our method vs. the state-of-the-arts on UTD-MHAD database in offline mode

Method	RR	Ntr	Ex. Pr.	Year	Alg. Pr.
SOS [60]	86.97%	Offline	Two-fold	2018	MoP
JTM [38]	85.81%	Offline	Two-fold	2016	MoP
CNN-stream [61]	69.90%	Offline	Two-fold	2018	MoP
GMGE [63]	90.47%	Offline	Two-fold	2018	ARMA
KCRR + PCA(600)	90.05%	Online	Two-fold	-	LR
KCRR + LPP(600)	90.05%	Online	Two-fold	-	LR

this intuition, we evaluate our method on the UTKinect database. The experiments are performed in online scenario where only the beginning 75% portion of each query action has been observed. The setting used for data partitioning and the threshold of non repetitive frames is the same as the previous section. The impact of the non-ideal joint localization is simulated by adding up some Gaussian White Noise (GWN) with zero mean and standard deviation ranging from $0.01L$ to $0.15L$, where L is the average length of the bones in the first frame of each action sequence, to all the joints of query samples. Note that, the dictionary is constructed using the original training samples. As can be seen from Fig. 4, the average recognition rate of our method decreases with the increase of noise, however the amount of the drop is quite different for each class of actions. According to the drop value, the classes can be categorized into three groups: very sensitive (pull, clap hands), almost sensitive (wave hands, walk, sit down, throw, push), and almost non-sensitive (pick up, carry, stand up) to noise. This categorization indicates that, the more the movement of skeletal center of gravity over the frames, the corresponding action will be more robust to noise. How skeletal mass changes and how it associates with the robustness of the method are shown in detail in Fig. 5. As can be seen, in such activities like picking up and carrying objects, the added noise is negligible compared to the movement of skeletal mass (at least in one direction), causing a high signal-to-noise ratio in that direction which in turn allows for correct recognition of those noisy

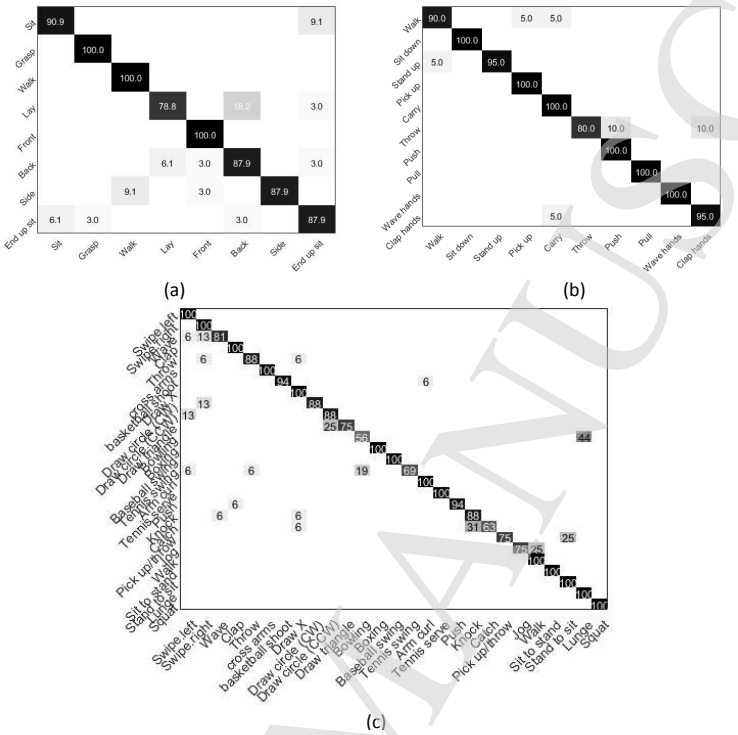

Fig. 3: Confusion matrices for (a) UTKinect database, and (b) TST database, (c) UTD-MHAD database in offline mode.

sequences. However, such activities as pulling and hand clapping do not follow this kind of behaviour.

We also conduct another experiment to validate the usefulness of our method against any abnormality in detecting actions (starting point and occurrence time of an action in a video stream). For this purpose, we consider a very challenging task, where on the one hand the starting point is mistakenly set $k \in 1, .., 50$ frames before the occurrence time of actions, and on the other hand, only 75% of the actions is considered to be performed. Therefore, we concatenate L noisy frames to the beginning of query action sequences, but reconstruct them using the original training samples. The noisy frames are exactly the same as the first frame of query samples but with some GWN added. Figure 6 shows accuracy rates for different numbers of noisy frames. As in the previous experiment, action sequences fall into three categories: very sensitive (sit down, pull, and throw), almost sensitive (pick up, walk, push), almost non-sensitive (stand up, carry, and clap hands, wave hands). However, their categorization story is quite different from the joints localization problem. Here, most of the recognition errors occur because (i) some frames of

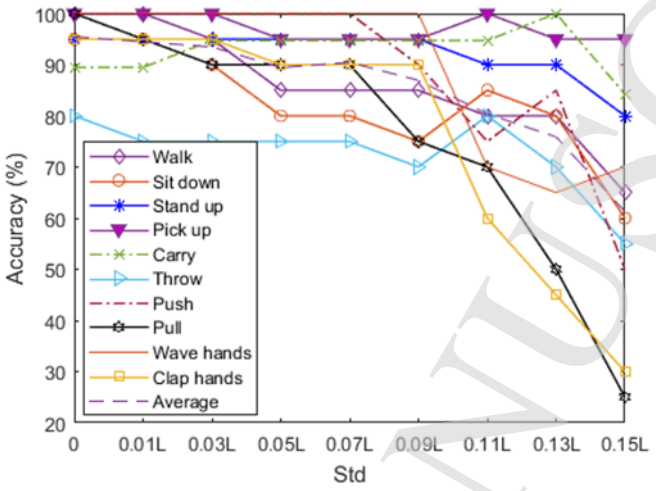

Fig. 4: The impact of imperfect estimation of joint locations on the performance of CRR algorithm

an action are unintentionally created within the shifted version of an action from another class (Fig. 7(a)), or (ii) assimilating the dynamic information of the main frames to the faulty detected sequence, leading to no influence of the constraint of equation (14) on its reconstruction coefficients (Fig. 7(b)).

6. Conclusion

This paper presented a novel 3D activity recognition scheme using the contributive representations of skeletal information. A frame-to-frame linear reconstruction strategy termed CRR was exploited to compare time series with different lengthes which provides an alternative for the traditional aligning methods such as DTW or PWM. Unlike the traditional methods that combine the posture characteristics and temporal statistics of skeletal data in the feature extraction phase, CRR incorporates the temporal information as a constraint on the reconstruction coefficients of the posture information and therefore prevent a raw combination of two set of data with different natures. Moreover, CRR represents a query time series collaboratively over all the training sets and therefore provides a one-to-all matching strategy for different classes leading to an ability to recognize the innovative actions with the first part similar to the first part of a training action, and second part similar to the second part of another training action in the same class. Experimental results on three publicly available benchmark databases demonstrated the superiority of our method compared to a set of state-of-the-art methods especially in online mode.

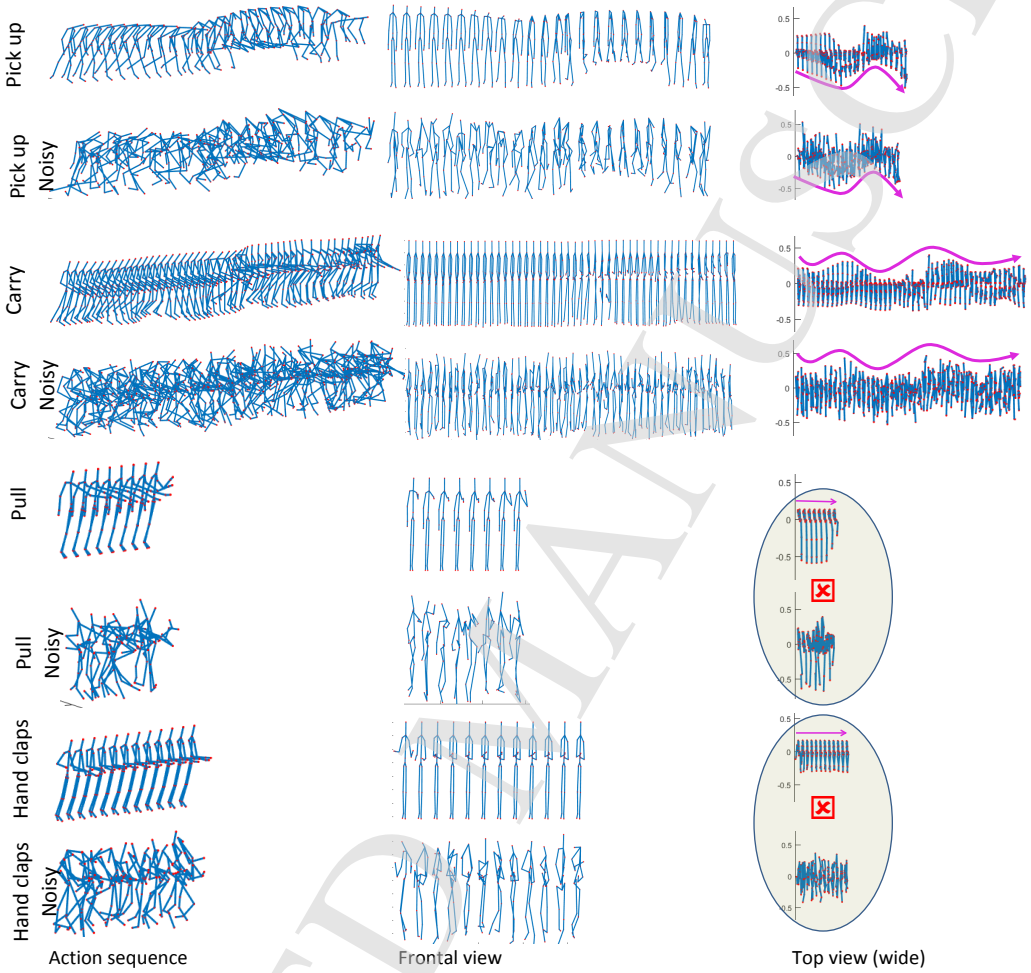

Fig. 5: Which class of actions are more robust against the noise of the joint localization. For pick up and carry the movement of skeletal mass in y-direction creates distinct motion patterns that are not easily affected by the localization noise. In contrast, lack of such a distinctive pattern makes pull and hand claps susceptible to the noise.

Acknowledgments

This work was supported by a grant from Iran National Science Foundation (INSF).

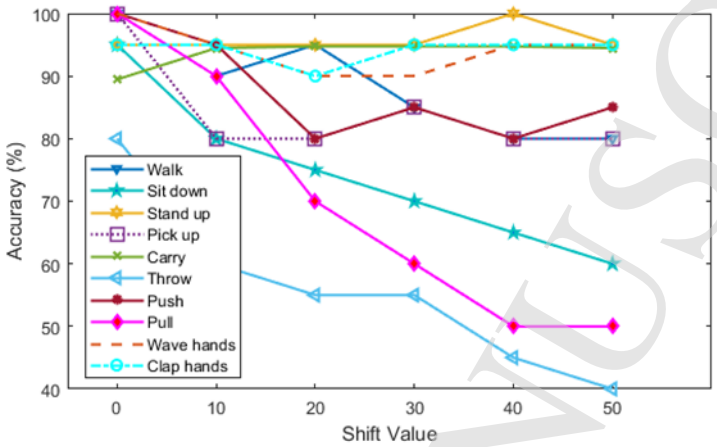

Fig. 6: The impact of imperfect action detection on the performance of CRR algorithm

References

1. J. Shotton, A. Fitzgibbon, M. Cook, T. Sharp, M. Finocchio, R. Moore, A. Kipman, and A. Blake. Real-time human pose recognition in parts from single depth images. In IEEE Conference on Computer Vision and Pattern Recognition, 2011, pages 12971304, 2011.
2. 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 im- ages, Commun. ACM 56 (1) (2013) 116124 .
3. Campbell, L.W. and Bobick, A.F., 1995, June. Recognition of human body motion using phase space constraints. In Computer vision, 1995. proceedings., fifth international conference on (pp. 624-630). IEEE.
4. Hussein, M.E., Torki, M., Gowayyed, M.A. and El-Saban, M., 2013, August. Human Action Recognition Using a Temporal Hierarchy of Covariance Descriptors on 3D Joint Locations. In IJCAI (Vol. 13, pp. 2466-2472).
5. Xia, L., Chen, C.C. and Aggarwal, J.K., 2012, June. View invariant human action recognition using histograms of 3d joints. In Computer Vision and Pattern Recognition Workshops (CVPRW), 2012 IEEE Computer Society Conference on (pp. 20-27). IEEE.
6. Zanfir, M., Leordeanu, M. and Sminchisescu, C., 2013. The moving pose: An efficient 3d kinematics descriptor for low-latency action recognition and detection. In Proceedings of the IEEE international conference on computer vision (pp. 2752-2759).
7. Qiao, R., Liu, L., Shen, C. and van den Hengel, A., 2017. Learning discriminative trajectorylet detector sets for accurate skeleton-based action recognition. Pattern Recognition, 66, pp.202-212.
8. Yang, W., Lyons, T., Ni, H., Schmid, C., Jin, L. and Chang, J., 2017. Leveraging the Path Signature for Skeleton-based Human Action Recognition. arXiv preprint arXiv:1707.03993.
9. Seddik, B., Gazzah, S. and Amara, N.E.B., 2017. Human-action recognition using a multi-layered fusion scheme of Kinect modalities. IET Computer Vision, 11(7), pp.530-

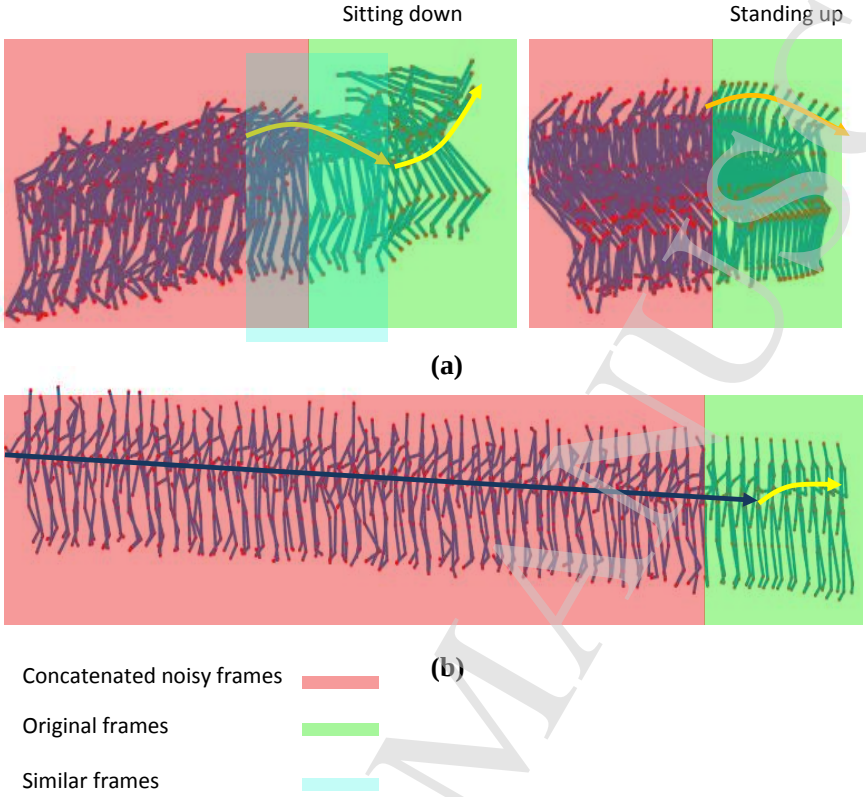

Fig. 7: A visual interpretation of the impact of faulty action detection, (a) some of the standing up frames are unintentionally created inside a faulty-detected standing up sequence. (b) dynamic information of the original frames of a pull action is assimilated to the noisy frames of a faulty detected sequence.

540.
10. Luvizon, D.C., Tabia, H. and Picard, D., 2017. Learning features combination for human action recognition from skeleton sequences. Pattern Recognition Letters, 99, pp.13-20.
11. Huang, M., Cai, G.R., Zhang, H.B., Yu, S., Gong, D.Y., Cao, D.L., Li, S. and Su, S.Z., 2018. Discriminative parts learning for 3D human action recognition. Neurocomputing, 291, pp.84-96.
12. Jiang, M., Kong, J., Bebis, G. and Huo, H., 2015. Informative joints based human action recognition using skeleton contexts. Signal Processing: Image Communication, 33, pp.29-40.
13. Ding, W., Liu, K., Belyaev, E. and Cheng, F., 2018. Tensor-based linear dynamical systems for action recognition from 3D skeletons. Pattern Recognition, 77, pp.75-86.
14. Zhu, G., Zhang, L., Shen, P. and Song, J., 2016. Human action recognition using multi-layer codebooks of key poses and atomic motions. Signal Processing: Image Com-

munication, 42, pp.19-30.
15. Ohn-Bar, E. and Trivedi, M.M., 2013, June. Joint angles similarities and HOG2 for action recognition. In Computer vision and pattern recognition workshops (CVPRW), 2013 IEEE conference on (pp. 465-470). IEEE.
16. Evangelidis, G., Singh, G. and Horaud, R., 2014, August. Skeletal quads: Human action recognition using joint quadruples. In Pattern Recognition (ICPR), 2014 22nd International Conference on (pp. 4513-4518). IEEE.
17. Lin, S.Y., Shie, C.K., Chen, S.C., Lee, M.S. and Hung, Y.P., 2012, November. Human action recognition using action trait code. In Pattern Recognition (ICPR), 2012 21st International Conference on (pp. 3456-3459). IEEE.
18. Yang, X. and Tian, Y., 2014. Effective 3d action recognition using eigenjoints. Journal of Visual Communication and Image Representation, 25(1), pp.2-11.
19. Azis, N.A., Jeong, Y.S., Choi, H.J. and Iraqi, Y., 2016. Weighted averaging fusion for multi-view skeletal data and its application in action recognition. IET Computer Vision, 10(2), pp.134-142.
20. Ofli, F., Chaudhry, R., Kurillo, G., Vidal, R. and Bajcsy, R., 2014. Sequence of the most informative joints (smij): A new representation for human skeletal action recognition. Journal of Visual Communication and Image Representation, 25(1), pp.24-38
21. Cippitelli, E., Gasparrini, S., Gambi, E. and Spinsante, S., 2016. A human activity recognition system using skeleton data from RGBD sensors. Computational intelligence and neuroscience, 2016, p.21.
22. Mokari, M., Mohammadzade, H. and Ghojogh, B., 2017. Recognizing Involuntary Actions from 3D Skeleton Data Using Body States. arXiv preprint arXiv:1708.06227.
23. Ghojogh, B., Mohammadzade, H. and Mokari, M., 2017. Fisherposes for Human Action Recognition Using Kinect Sensor Data. IEEE Sensors Journal.
24. Wang, Y., Shi, Y. and Wei, G., 2017. A novel local feature descriptor based on energy information for human activity recognition. Neurocomputing, 228, pp.19-28.
25. Veeraraghavan, A., Roy-Chowdhury, A.K. and Chellappa, R., 2005. Matching shape sequences in video with applications in human movement analysis. IEEE Transactions on Pattern Analysis and Machine Intelligence, 27(12), pp.1896-1909.
26. Abdelkader, M.F., Abd-Almageed, W., Srivastava, A. and Chellappa, R., 2011. Silhouette-based gesture and action recognition via modeling trajectories on Riemannian shape manifolds. Computer Vision and Image Understanding, 115(3), pp.439-455.
27. Joshi, S.H., Klassen, E., Srivastava, A. and Jermyn, I., 2007, June. A novel representation for Riemannian analysis of elastic curves in Rn. In Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on (pp. 1-7). IEEE.
28. Joshi, S.H., Klassen, E., Srivastava, A. and Jermyn, I., 2007, August. Removing shape-preserving transformations in square-root elastic (SRE) framework for shape analysis of curves. In International Workshop on Energy Minimization Methods in Computer Vision and Pattern Recognition (pp. 387-398). Springer, Berlin, Heidelberg.
29. Turaga, P., Veeraraghavan, A., Srivastava, A. and Chellappa, R., 2011. Statistical computations on Grassmann and Stiefel manifolds for image and video-based recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence, 33(11), pp.2273-2286.
30. Harandi, M., Salzmann, M. and Porikli, F., 2014. Bregman divergences for infinite dimensional covariance matrices. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 1003-1010).
31. Kulis, B., Sustik, M.A. and Dhillon, I.S., 2009. Low-rank kernel learning with Bregman matrix divergences. Journal of Machine Learning Research, 10(Feb), pp.341-376.
32. Zhang, X., Wang, Y., Gou, M., Sznaier, M. and Camps, O., 2016. Efficient temporal

sequence comparison and classification using gram matrix embeddings on a riemannian manifold. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 4498-4507).

33. Lui, Y.M. and Beveridge, J.R., 2011, March. Tangent bundle for human action recognition. In Automatic Face and Gesture Recognition and Workshops (FG 2011), 2011 IEEE International Conference on (pp. 97-102). IEEE.
34. Slama, R., Wannous, H., Daoudi, M. and Srivastava, A., 2015. Accurate 3D action recognition using learning on the Grassmann manifold. Pattern Recognition, 48(2), pp.556-567.
35. Du, Y., Wang, W., Wang, L.: Hierarchical recurrent neural network for skeleton based action recognition. In: CVPR. (2015)
36. Zhu, W., Lan, C., Xing, J., Zeng, W., Li, Y., Shen, L., Xie, X.: Co-occurrence feature learning for skeleton based action recognition using regularized deep lstm networks. In: AAAI. (2016)
37. Veeriah, V., Zhuang, N., Qi, G.J.: Differential recurrent neural networks for action recognition. In: ICCV. (2015)
38. Wang, P., Li, Z., Hou, Y. and Li, W., 2016, October. Action recognition based on joint trajectory maps using convolutional neural networks. In Proceedings of the 2016 ACM on Multimedia Conference (pp. 102-106). ACM.
39. Hou, Y., Li, Z., Wang, P. and Li, W., 2016. Skeleton optical spectra based action recognition using convolutional neural networks. IEEE Transactions on Circuits and Systems for Video Technology.
40. Du, Y., Fu, Y. and Wang, L., 2015, November. Skeleton based action recognition with convolutional neural network. In Pattern Recognition (ACPR), 2015 3rd IAPR Asian Conference on (pp. 579-583). IEEE.
41. Li, C., Zhong, Q., Xie, D. and Pu, S., 2017, July. Skeleton-based action recognition with convolutional neural networks. In Multimedia and Expo Workshops (ICMEW), 2017 IEEE International Conference on (pp. 597-600). IEEE.
42. Yang, Y., Deng, C., Tao, D., Zhang, S., Liu, W. and Gao, X., 2017. Latent max-margin multitask learning with skelets for 3-D action recognition. IEEE transactions on cybernetics, 47(2), pp.439-448.
43. Yang, Y., Deng, C., Gao, S., Liu, W., Tao, D. and Gao, X., 2017. Discriminative multi-instance multitask learning for 3D action recognition. IEEE Transactions on Multimedia, 19(3), pp.519-529.
44. Annadani, Y., Rakshith, D.L. and Biswas, S., 2016. Sliding Dictionary Based Sparse Representation For Action Recognition. arXiv preprint arXiv:1611.00218.
45. Hayes, B. and Shah, J.A., 2017, May. Interpretable models for fast activity recognition and anomaly explanation during collaborative robotics tasks. In Robotics and Automation (ICRA), 2017 IEEE International Conference on (pp. 6586-6593). IEEE.
46. Shan, J. and Akella, S., 2014, September. 3D human action segmentation and recognition using pose kinetic energy. In Advanced Robotics and its Social Impacts (ARSO), 2014 IEEE Workshop on (pp. 69-75). IEEE.
47. Ghodsi, S., Mohammadzade, H. and Korki, E., 2017. Simultaneous Joint and Object Trajectory Templates for Human Activity Recognition from 3-D Data. arXiv preprint arXiv:1711.01589.
48. Ellis, C., Masood, S.Z., Tappen, M.F., LaViola, J.J. and Sukthankar, R., 2013. Exploring the trade-off between accuracy and observational latency in action recognition. International Journal of Computer Vision, 101(3), pp.420-436.
49. Perez-DArpino, C. and Shah, J.A., 2015, May. Fast target prediction of human reaching motion for cooperative human-robot manipulation tasks using time series classifi-

cation. In Robotics and Automation (ICRA), 2015 IEEE International Conference on (pp. 6175-6182). IEEE.
50. Silva, D.F., Batista, G.E.D.A.P.A. and Keogh, E., 2016. On the effect of endpoints on dynamic time warping. In SIGKDD Workshop on Mining and Learning from Time Series, II. Association for Computing Machinery-ACM.
51. Hayashi, A., Mizuhara, Y. and Suematsu, N., 2005, July. Embedding time series data for classification. In International Workshop on Machine Learning and Data Mining in Pattern Recognition (pp. 356-365). Springer, Berlin, Heidelberg
52. Zhang, Z., Xu, Y., Yang, J., Li, X. and Zhang, D., 2015. A survey of sparse representation: algorithms and applications. IEEE access, 3, pp.490-530.
53. Wright, J., Yang, A.Y., Ganesh, A., Sastry, S.S. and Ma, Y., 2009. Robust face recognition via sparse representation. IEEE transactions on pattern analysis and machine intelligence, 31(2), pp.210-227
54. Zhang, L., Yang, M. and Feng, X., 2011, November. Sparse representation or collaborative representation: Which helps face recognition?. In Computer vision (ICCV), 2011 IEEE international conference on (pp. 471-478). IEEE
55. Liu, J., Wang, G., Hu, P., Duan, L.Y. and Kot, A.C., 2017, July. Global context-aware attention lstm networks for 3d action recognition. In The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (Vol. 7, p. 43).
56. Liu, J., Shahroudy, A., Xu, D., Kot, A.C. and Wang, G., 2018. Skeleton-based action recognition using spatio-temporal LSTM network with trust gates. IEEE transactions on pattern analysis and machine intelligence, 40(12), pp.3007-3021.
57. Cherian, A., Sra, S., Gould, S. and Hartley, R., 2018, March. Non-Linear Temporal Subspace Representations for Activity Recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 2197-2206).
58. Gasparrini, S., Cippitelli, E., Gambi, E., Spinsante, S. and Flrez-Revuelta, F., 2015. Performance analysis of self-organising neural networks tracking algorithms for intake monitoring using kinect.
59. Chen, C., Jafari, R. and Kehtarnavaz, N., 2015, September. Utd-mhad: A multimodal dataset for human action recognition utilizing a depth camera and a wearable inertial sensor. In Image Processing (ICIP), 2015 IEEE International Conference on (pp. 168-172). IEEE.
60. Hou, Y., Li, Z., Wang, P. and Li, W., 2018. Skeleton optical spectra-based action recognition using convolutional neural networks. IEEE Transactions on Circuits and Systems for Video Technology, 28(3), pp.807-811.
61. Khaire, P., Kumar, P. and Imran, J., 2018. Combining CNN streams of RGB-D and skeletal data for human activity recognition. Pattern Recognition Letters.
62. Rhif, M., Wannous, H. and Farah, I.R., 2018, August. Action Recognition from 3D Skeleton Sequences using Deep Networks on Lie Group Features. In 2018 24th International Conference on Pattern Recognition (ICPR) (pp. 3427-3432). IEEE.
63. Rahimi, S., Aghagolzadeh, A. and Ezoji, M., 2018. Human action recognition based on the Grassmann multi-graph embedding. Signal, Image and Video Processing, pp.1-9.
64. Miranda, L., Vieira, T., Martnez, D., Lewiner, T., Vieira, A.W. and Campos, M.F., 2014. Online gesture recognition from pose kernel learning and decision forests. Pattern Recognition Letters, 39, pp.65-73.
65. Kulkarni, K., Evangelidis, G., Cech, J. and Horaud, R., 2015. Continuous action recognition based on sequence alignment. International Journal of Computer Vision, 112(1), pp.90-114.
66. Tang, C., Li, W., Wang, P. and Wang, L., 2018. Online human action recognition based on incremental learning of weighted covariance descriptors. Information Sciences, 467,

pp.219-237.
67. Zhang, S., Yang, Y., Xiao, J., Liu, X., Yang, Y., Xie, D. and Zhuang, Y., 2018. Fusing geometric features for skeleton-based action recognition using multilayer LSTM networks. IEEE Transactions on Multimedia, 20(9), pp.2330-2343.