Sparsness embedding in bending of space and time; A case study on unsupervised 3D action recognition

Hoda Mohammadzade[a,*], Mohsen Tabejamaat[a]

[a]*Department of Electrical Engineering, Sharif University of Technology, Tehran, Iran.*

Abstract

Human action recognition from skeletal data is one of the most popular topics in computer vision which has been widely studied in the literature, occasionally with some very promising results. However, being supervised, most of the existing methods suffer from two major drawbacks; (1) too much reliance on massive labeled data and (2) high sensitivity to outliers, which in turn hinder their applications in such real-world scenarios as recognizing long-term and complex movements. In this paper, we propose a novel unsupervised 3D action recognition method called Sparseness Embedding in which the spatiotemporal representation of action sequences is nonlinearly projected into an unwarped feature representation medium, where unlike the original curved space, one can easily apply the Euclidean metrics. Our strategy can simultaneously integrate the characteristics of nonlinearity, sparsity, and space curvature of sequences into a single objective function, leading to a more robust and highly compact representation of discriminative attributes without any need to label information. Moreover, we propose a joint learning strategy for dealing with the heterogeneity of the temporal and spatial characteristics of action sequences. A set of extensive experiments on six publicly available databases, including UTKinect, TST fall, UTD-MHAD, CMU, Berkeley MHAD, and NTU RGB+D demonstrates the superiority of our method compared with the state-of-the-art algorithms.

Keywords: Unsupervised action recognition, Time series analysis, Sparseness embedding, Human computer interaction

1. Introduction

Semantic Analysis of Human Behavior (SAHB) is one of the most important aspects of visual intelligence which is institutionalized in the very first months of life. It is also one of the hot research topics of computer vision that has

*Corresponding Author

Email addresses: hoda@sharif.edu (Hoda Mohammadzade), m.tabejamaat@sharif.edu (Mohsen Tabejamaat)

widespread applications in various pragmatic fields such as abnormal event detection [1], human robot/computer interaction [2, 3], patient/prisoned behaviour monitoring [4, 5, 6] as well as the analysis of consumer's buying behavior [7]. Recent years have seen an eruption of scientific works in this area, occasionally with some very promising achievements. Yet, due to the non-rigid nature of the human body and varying styles of performing actions, it is still one of the most challenging/active research topics that should make further extensive studies.

SAHB, from a broad perspective, can be performed in three different ways; (i) recognizing actions; where the concept of a subject's movement is evaluated in a short period of time, (ii) classifying activities; where interactions between human-object or multiple individuals are considered to be analyzed, and (iii) recognizing gestures; which refers to the study of movements for some special parts of body (eg. head, hand, or foot). While there are more benefits for complicated and subtle movements, such disaggregation is not much compatible with the SAHB's horizon which is defined as recognizing all types of movements in a single interpretation framework. That is why most of the existing databases include some actions from the categories of gesture or activity.

According to the nature of sensors, motion data acquisition systems can be broadly categorized into three groups: RGB video cameras, depth sensors, and 3D motion capture systems. Traditional methods mainly focus on RGB video streams and use some off-the-shelf object detection and recognition algorithms [8, 9, 10, 11, 12]. However, despite the low costs, wide availability and user-friendliness, these methods suffer from some inherent limitations like sensitivity to illumination variations, body shape, clothing, invasive nature (due to the record of facial details), noisy background, and partial occlusions which hamper their applications in many real-world scenarios.

With the advent of depth sensing devices like Microsoft Kinect [13], Intel RealSense [14], dual camera [15], and Xtion PRO Live [16], such limitations as sensitivity to noisy, clutter background, clothing and illumination variations and privacy issue were somewhat alleviated, but some others like computational cost, heavy memory requirement, and non informativeness of background still remained. This induced the research community to turn its attention to skeleton sequences where input data is reduced to a set of 3D joint coordinates which could be estimated either directly from some wearable markers [17] or in a remote manner from a depth stream [18, 19, 20]. However, due to some imposed limitations on movements, applicability of marker based techniques are mostly limited to gaming or lab practices. In contrast, depth techniques can determine the joint's location with a competitive performance, but without sacrificing the freedom of movements [21].

Inspired by the success of deep learning in many machine vision applications such as image/speech recognition [22, 23, 24, 25, 26, 27], video captioning [28, 29, 30], and neural machine translation [31, 32, 33], some researchers were encouraged to leverage such networks for recognizing skeleton sequences. An overview on scientific papers reveals that, the overwhelming majority of the methods proposed in the last two years are established upon supervised deep learning frameworks [34, 35, 36, 37, 38, 39, 40, 41, 40, 42, 43], claiming a re-

markably better performances than any shallow learning strategies. Remember, most of these networks typically have millions of parameters, requiring to be tuned by using a proportional number of training data. However, providing such a large number of labeled sequences is a difficult and time consuming task which seems to be unrealistic under the current circumstances of accessibility to depth sensors. As a remedy, some researchers started using of data augmentation techniques. The predominant techniques used in the literature include cropping, adding up noise, re-scaling, re-sampling, rotation, flipping, and auto-encoding [44], which can even increase the size of databases up to 25 times [45]. Despite this, augmentation methods are typically susceptible to semantic deterioration, arising from the non-analytic nature of the synthesizing transformations. For example, in the case of re-sampling, a long term motion is sparsely down-sampled so as to guarantee the enough number and variety of the synthetic sequences. However, the sparser the sampling interval is, the synthetic sequences would be more distorted.

Till recent years, supervised learning has been the main stream of research on the field of action recognition. However, such a learning strategy would only be effective if action labels were fully categorical. Yet, in many applications like skill assessment, autonomous scoring of sport movements, and human robot interaction, beyond the class membership, it is particularly significant to pay enough attention on the quality of actions. This would be even more acute for the case that we seek a suitable reaction from the robotic systems. For example, a table tennis robot does not only require an understanding of how you are hitting the ball but also needs your gesture and even how much force you are approximately applying on racquet. That is why, the idea of supervised learning (pulling the training samples with the same labels closer together, no matter how close they are to the class boundary), may not always lead to an accurate semantic score, especially when the transition of the class labels occur in a soft manner (e.g. transition from walking to jogging).

In this paper, we propose a novel sparsity based 3D action recognition method, namely Sparseness Embedding (SE), that simultaneously encodes the nonlinearity and time-space relationship of action trajectories. This is much beneficial for constructing a robust spatiotemporal representation model of action sequences. Unlike the conventional sparsity based methods that can only be applied on fixed-size templates, our model can explore the structure of samples with varying sizes. Moreover, different from deep learning based models, it does not require a massive number of training sequences which is a very crucial issue in 3D modeling of a temporal data. We also introduce a constrained version of our model that enforces the projection coefficients to be shrunk, eschewing a biased estimation of the sparseness. In addition, we propose a novel joint learning strategy which simultaneously preserves the sparseness of the spatial and temporal statistics, while avoiding their well-known heterogeneity problem.

In a nutshell, the key contributions of this paper are: (i) based on our best knowledge, this is the first unsupervised 3D action recognition method established upon a non-deep learning based framework, (ii) we propose a novel encoding strategy which is capable of incorporating the data locality into an

unwarped-sparse projection space, allowing for a dual bounded mapping on the characteristics of the spatial and temporal nonlinear characteristics of action sequences. (iii) we introduce a regularized mapping function that prevents any biased sparsity estimation of a data structure in the mapping procedure of a sequence to a fixed-size sample, which may occur in some cases when there is a possible correlation in the original unwarped space of sequences, (iv) to deal with the problem of heterogeneity, we propose a novel joint sparseness embedding strategy which is capable to treat the temporal and spatial characteristics in two individual manners, preventing the projection weights to be biased towards the static or dynamic characteristics of sequences. A set of extensive experiments demonstrates the superiority of our method against the state-of-the-art techniques.

The reminder of the paper is organized as follows. In Section 2, we briefly review the related studies to our method. Section 3 describes the motivation of the research as well as our solution for a sparsity based action recognition. The experimental results is presented in Section 4. Finally, our study is concluded in Section 5.

2. Related Works

In this section, we present a brief review on the most related skeleton-based action recognition methods. Although RGB and depth map based techniques are sill among the active areas of research in the field of action analysis, they are not reviewed as being out of the scope of this paper. For more information, one can refer to the references [46, 47]. In our study, different criteria are utilized for categorizing the literature, so discussions are provided from different perspectives which allows for the possibility that a typical method can be simultaneously categorized to several groups.

2.1. Geometrical Feature Representation

Early works mostly focused on how to engineer the most discriminative attributes from the trajectory of joint positions. These methods typically measured the distance or angle between the joints or the plates passing through them. Then, differences between the extracted features over time considered as the local velocity or acceleration of sequences. For example, Muller et al. [48] used the boolean operators to describe geometric relations between the planes passing through the joints and limbs. Yao et al [49] modified these features, introducing five relational pose attributes: pairwise distance of joints, distance between a joint and a plane, distance between a joint and a normalized plane, velocity of a joint towards two other ones, velocity of a joint towards the normal vector of a plane. Li et al. [50] introduced the relative position as the sum of pairwise Euclidean distance and raw difference of joints. In Agahian et al. [51] a combination of joint positions and local normalized/unnormalized velocities was used to describe the skeletons. The works in [52, 53] utilized the position, velocity, and acceleration of raw joints as the features of skeleton sequences. [54] described each skeleton by the distances between body joints to

a reference point. Eweiwi et al. [55] proposed Joint Movement Normal (JMN) as the correlation of the joint location and velocity and utilized it alongside with raw joint positions and joint velocity to describe the skeleton. Yang et al. [56] used three pairwise measures including pairwise joint difference within current frames, between each frame and its preceding one, and between each frame and the initial skeleton. Wang et al. [57] utilized a normalized measure as the difference of pairwise joints divided by their Euclidean distance. Most of the geometrical methods were proposed to use the Cartesian coordinate system [54, 48, 49, 50, 51, 56, 57], however a few preferred to focus on the characteristics of the spherical coordinate [55, 58].

2.2. Pose Representation

Motivated by the fact that skeleton frames are not equally important in the recognition procedure, many researchers have turned attention to pose based recognition methods. From a broad perspective, these methods can be divided into two different categories: (1) Those do not much emphasize that poses require to be key [59]. These methods seek out the classes that contain the closest similar pose to each frame of a test sample. Then, the class with the maximum vote is predicted as the label of the sequence. (2) key pose based methods that seek the poses with the maximum contributions in discriminating between different actions [60, 61, 54, 62]. These poses are mainely selected in two different ways: (a) predefining the poses [23] (delineated by a human), and (b) using the clustering algorithms [61, 54, 62]. The clustering can be carried out in two different manners: (i) a categorical way in which each skeleton frame is distinctly assigned to one cluster (as in the k-means algorithm) [61, 54]. The most common constraints of these methods are enforcing cluster shapes to be pseudo-spherical, a fixed assignment weights for all the skeleton frames as well as the need for pre-specifying the number of clusters. (ii) a soft assignment of frames to each clusters (like the way of Gaussian Mixture Models (GMM)) [62]. Despite alleviation of the above mentioned issues, these methods suffer from a difficult to make parametric assumption about the data generating process.

2.3. Part based Methods

Psychological studies suggest that imitative learning in humans treats the limb movements as a whole, rather than focusing on just a series of discrete joint coordinates [63], which indicates the benefit of encoding the relative postures of body limbs to boost the accuracy of the recognition tasks. For example, no matter how the positions of elbows are, a movement is considered as clapping whenever two hands are striking together in the middle front of body. This motivated many researchers to move on hierarchical part based action recognition: partitioning human body into different compartments, encoding each part individually, and finally characterizing the relationship between the parts. However, despite prevail in multipartite tasks, such a strategy suffers from the lack of generalization ability to a wide range of subtle activities. In this context, Tao et al. [64] partitioned each skeleton into ten body parts: back, left arm,

right arm, left leg, right leg, torso, upper body, lower body, full upper body, and full body and learned one dictionary for each where the atoms of the dictionary were considered to be linear classifiers of body part movements. Zhu et al [65] divided each body skeleton into five parts: left upper limb, right upper limb, left lower limb, right lower limb, and torso. Then, Normalized Relative Orientation (NRO) was introduced for mining the attributes of each part. In Du et al. [66], each skeleton is partitioned into five compartments: two arms, two legs, and one trunk. Then, the parts were characterized by using a Bidirectional RNNs (BRRN). Then, encoding produce was extended to the dual, triple and five-way combinations of the RNN networks. Hou et al. [67] utilized a spectrum representation individually for three parts of the body, left torso, right torso and the trunk. In this scheme, the hue representation of the right part is the reverse of that presented for the left one. In addition, because of the subtle movement, they proposed to suppress the range of hue for the trunk part.

2.4. Non-Euclidean Approaches

Due to the relatively high similarity of some activities to each other, seeking their relative structures on the characteristics of the Euclidean space is very sensitive to outliers and noise. This motivates the need for a set of desired invariancies which can be ideally achieved on a set of non-linear manifolds where the geometrical variability of data distribution can be incorporated into the characteristics of discriminative statistics. In this context, Slama et al. [68] encoded the action trajectories using the Auto Regressive and Moving Average (ARMA) model and used their observationality matrices as points on the Grassman manifold. Then, the tangent bundles of this manifold were utilized to construct a discriminative representation of sequences. Amor et al. [69] used the Kendalls shape descriptor to characterize the representation of skeletons on a shape manifold. In [70], a dictionary learning framework is defined on the Kendall's shape space. Kacem et al. [71], utilized the geometry of covariance matrices for embedding the Gram representation of action trajectories into the Positive Definite (PD) manifold. Then, DTW is used for aligning the trajectories resided on PD. Similar to [67], Zhang et al. [72] used the Grame matrix to embed the actions into the PD manifold, but differently applied four well-known distance-like measures including Affine Invariant Riemannien Metric (AIRM) [73], Log-Euclidean Riemannian Metric (LERM) [74], Jensen-Bregman Log-det Divergence (JBLD) [75], KL-Divergence metric (KLDM) [76] to directly match the resided trajectories on the manifold. Rahmi et al. [77] proposed a multi-graph embedding technique which benefits from Geodesic distances between the subspaces represented by the ARMA models [78] to measure the affinity matrices of a Fisher discriminative representation. In [79], each sequence is individually represented as a curve in the Lie Group. The Group is then estimated as a collections of tangent subspaces named Lie Algebra. Dynamic Time Warping (DTW) is then utilized to handle any possible misalignments between the subspaces. Moreover, temporal modeling and classification are respectively performed by Fourier temporal pyramid [80] and linear Support Vector Machine [81]. One common problem for all these strategies is the preassumption that is

considered for the geometrical characteristics of the manifold which may not be compatible with the distribution of such wiggly data as action sequences.

2.5. Deep Learning

The last two years have seen a big eruption of scientific studies in the field of deep learning based action recognition, encompassing a wide spectrum of network architectures that can be broadly categorized into five kingdoms:
- *RNN/LSTM networks*: these methods manually characterize the spatial representation of the skeleton poses and relegate the encoding of their temporal characteristics to an RNN or LSTM network [34, 35, 36].
- *Attention based networks*: such methods propose to explore temporal characteristics by selectively focus on the most informative joints and/or skeleton poses [37, 38, 39, 40, 82, 83].
- *Robust LSTM*: these networks have enabled methods to implicitly suppress the noise in sequences by adding up a trust gate in the cells or a dropout layer to the networks [34, 41, 40].
- *CNN networks*: these methods manually encode the temporal dynamics of sequences in conjunction with their spatial characteristics and transform them into some two dimensional patterns. Then, one or multi-stream CNN network is applied for encoding the spatiotemporal features of the patterns [84, 42, 43].
- *3D-CNN*: unlike conventional CNNs, these methods do not require the temporal features to be manually encoded and then fed to the network rather it has the capability to directly characterize the temporal dynamics by adding up the time dimension to the kernels [88].

However, despite the advances made in this field, the superiority of such computationally demanding networks over the handcrafted attributes is not very clear, which is mainely due to the scarcity of 3D action data.

3. Proposed Method

3.1. Motivation

Sparsity Preserving Projection (SPP) was originally proposed by Qiao et al. [85] as an unsupervised strategy for face recognition. They proposed that sparse representation can efficiently characterize the topographical structure of a data set. The objective function of this strategy is formulated as follows:

$$min_{\alpha} \sum_{i=1}^{N} \|\alpha^T x_i - \alpha^T X s_i\|^2 \tag{1}$$

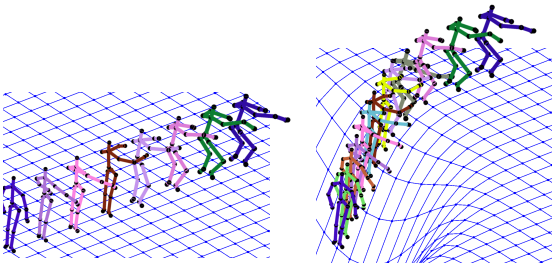
where N and l respectively stand for the number of training samples and dimensionality of data, α is the projection matrix, x_i denotes the i-th training sample from the dictionary $X \in^{l \times N}$, and s_i is the sparse reconstruction coefficients associated with the x_i. According to this formula, each sample in a low dimensional medium can be reconstructed from its sparsely connected samples

in the original input space. Unlike the methods that aim to minimize the distances between the samples belonging to the same classes (supervised learning), this strategy can preserve the structure of a soft transition between different labels, capable of dealing with interstitial samples. This would be of more interest when we need the machines to develop a sense of a qualitatively description of an action. For example, consider a smooth transition from walking to jogging, and then running. In cases like this, a categorical representation would be an unwise idea because, to make a proper reaction, machines beyond a simple prediction need to have a detailed analysis of the way that subjects interact with them. Despite the advantages, SPP suffers from its own drawbacks, most notably due to the lack of any provision for handling the varying lengths of samples which in turn hinders its application in recognizing action sequences. This limitation strongly encouraged us to propose a novel sparsity based method that can deal with the samples of varying sizes. In a nutshell, motivations of this paper are summarized as follows: (i) supervised learning strategies fail when we require to learn the style or skill scores of a sample. (ii) supervised and those unsupervised learning strategies based on the measure of the nearest neighbors, all fail when there are some outliers among the training samples. (iii) SPP can only be applied when all the training samples would be of the same length which is an unrealistic condition for such a time-varying task as recognizing actions, (iv) SPP fails to characterize the possible nonlinearity of data which is a common issue in the applications like an action recognition task.

3.2. Sparseness Embedding

From the previous section, we found that one can utilize SPP, only if the training samples are entirely of the same size. However, due to the unequal length of action sequences, arising from the variations in speed or an imperfect action detection phase, SPP can not be directly applied to human action recognition. A simple solution to this problem is the use of a sampling procedure to bind all the sequences into a fixed-length. However, there are two key issues with this solution: first, choosing such an optimal length value is a challenging task. This would be even more acute when there is a considerable difference in the length of sequences. In these cases, if the fixed value is too low, long-term sequences will have to be sparsely sampled, which results in the loss of a significant amount of crucial information. Conversely, if it is too high, temporal characteristics of short-term sequences would be roughly disappeared. The second disadvantage arises from the fact that SPP requires to convert all the sequences to a set of one-dimensional vectors. Such a flattening procedure not only increases the size of the training samples, leading to a not accurate, time-consuming eigen-decomposition, but also has a drawback of losing their temporal information. As a second remedy, we can use a dynamic matching of sequences (e.g. DTW) instead of the simple l_2-norm distance used in SPP and reformulate the objective function as follows:

$$min_{\alpha} \sum_{i=1}^{N} \sum_{j=1}^{N} dtw\big(\alpha^T \overrightarrow{x_i}, \alpha^T \overrightarrow{x_j} s_{ij}\big) \tag{2}$$



(a) An action coordinated with the sampling rate of the camera

(b) Warping of space due to the lower speed of performing the action versus the sampling rate of the camera

Figure 1: Warping of time and space in a four-dimensional fabric due to the varying speed of actions

---

Although this strategy can overcome the above mentioned problems, due to the evolutionary nature of DTW, it can not be solved in a straightforward manner. As an alternative, we propose to embed the sparsity of sequences in unwarped feature space. In this sense, input space is considered to be warped so that the rules of the Euclidean coordinate system does not hold, reminding us the Einstein's seminal theory of time dilation (Figure 1); when time is dilated, space is bent [86]. To overcome this limitation, we propose to dynamically transform all the sequences to a set of fixed-size tensors in an unwarped space which can bring us along three important advantages: it follows the Euclidean rules, unfolds the nonlinearity of original space, and also preserves the spatiotemporal characteristics of sequences. Assume a nonlinear function $\mathfrak{u}$ that maps the training sequence x_i to an unwarped space $\mathfrak{S}_{\mathfrak{w}}{}^{I_1 \times I_2 \times \ldots \times I_O}$:

$$\overrightarrow{x_i} \in \mathbb{R}^{n \times m} \xrightarrow{\mathfrak{u}} \mathfrak{u}(x_i) \in \mathfrak{S}_{\mathfrak{w}} \tag{3}$$

Applying such transformation on all the training samples, we get $X \xrightarrow{\mathfrak{u}} Span\{\overrightarrow{\mathfrak{U}}\}$. Note that, if $\mathfrak{u}$ is an O-th order tensor, $Span(\overrightarrow{\mathfrak{U}})$ indicates a concatenation of tensors at direction of $O+1$. For example, if $\mathfrak{u}$ is considered to be a vector $\mathfrak{u} \in \mathfrak{S}_{\mathfrak{w}}{}^{m \times 1}$, then $Span\{\overrightarrow{\mathfrak{U}}\} = [\mathfrak{u}(x_1), \mathfrak{u}(x_2), ..., \mathfrak{u}(x_N)]$, if is a matrix $\mathfrak{u} \in \mathfrak{S}_{\mathfrak{w}}{}^{m \times n}$, then we get $Span\{\overrightarrow{\mathfrak{U}}\} = [[\mathfrak{u}(x_1)], [\mathfrak{u}(x_2)], ..., [\mathfrak{u}(x_N)]]$, and so on. Thus, aiming at preserving the distance between each transformed sequence and its corresponding sparse representation in the uwarped space, the objective function of SE is formulated as follows:

$$min_{\alpha} \sum_{i=1}^{N} \|\alpha^T \overrightarrow{\mathfrak{u}(x_i)} - \alpha^T Span(\overrightarrow{\mathfrak{U}})|_1^N s_i\| \tag{4}$$

This way, we seek a low dimensional space that closes the similar sequences of varying sizes together, while dynamically embedding the curvature of time-space as the $(O+1)$-th dimension of the destination space. Figures 2 and 3

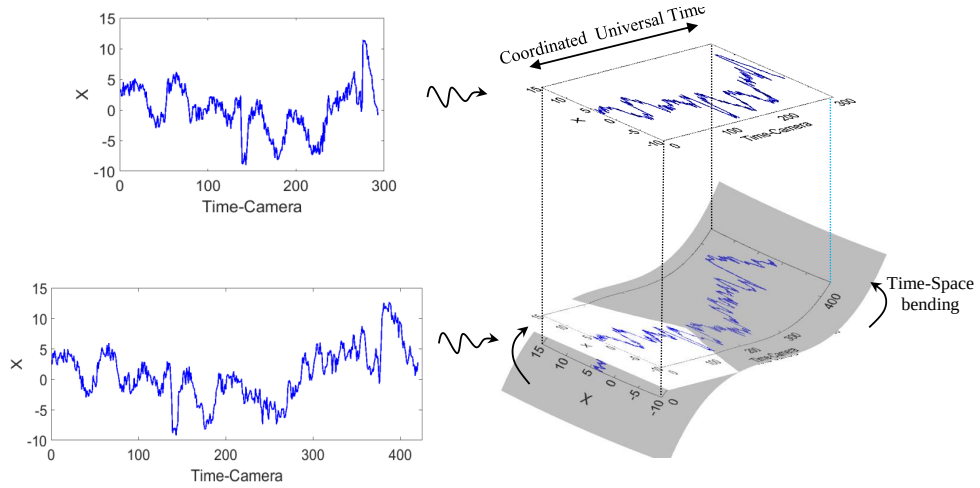


Figure 2: Time-space bending for a signal and its time dilated version and how we can align them in a coordinated universal time.

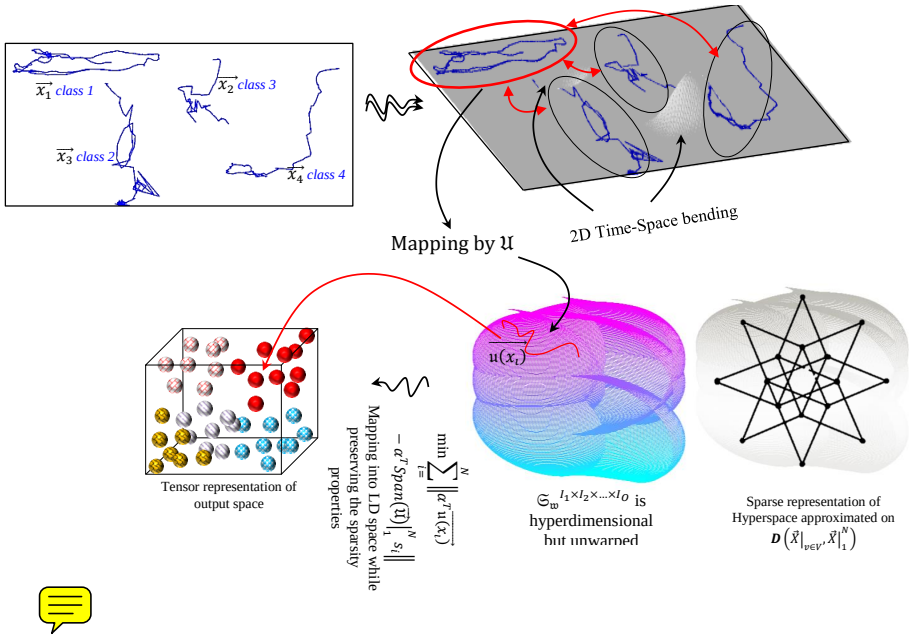


Figure 3: Sparesness embedding based on the concept of 2D time-space bending. Note that the third dimension is $(O+1)$-th dimension of the destination space, not the z-axis.

show the simplified concept of a time-space bending and its contribution to the disaggregation of space.

Since the transformation operator $\mathfrak{u}$ is unknown, computing such an optimal

projection is not an easy task, requiring expensive labor and a set of weak approximations. As a remedy, we assume there is a series of V transformed sequences so that α lies in their span, that is:

$$\alpha = \big(Span(\overrightarrow{\mathfrak{U}})|_{\in V}\big).\beta \tag{5}$$

where β denotes the expansion coefficients. Substituting equation (5) into (4), we have:

$$\begin{aligned} min_\beta \sum_{i=1}^{N} \|\beta^T \big(Span(\overrightarrow{\mathfrak{U}})|_{\in V}\big)^T \otimes \overrightarrow{\mathfrak{u}(x_i)}... \\ - \beta^T \big(Span(\overrightarrow{\mathfrak{U}})|_{\in V}\big)^T \otimes Span(\overrightarrow{\mathfrak{U}})|_1^N s_i\| \end{aligned} \tag{6}$$

It is clear that, this model may lead to a degenerate solution. A convenient way to avoid such an issue is applying a constraint like $\|\alpha^T \big(Span(\overrightarrow{\mathfrak{U}})|_1^N\big)\| = 1$. In addition, an unconstrained β may lead to a poor estimation of the projection coefficients, arising from the possible correlations among the unwarped variables. In such cases, the influence of applying a large positive coefficient on a variable could be counteracted by a negative coefficient on its correlated counterpart. To address this issue, we introduce another constraint to shrink the coefficients towards a constant value. Considering these two stipulations, the model can be reformulated as follows:

$$\begin{aligned} & min_\beta \sum_{i=1}^{N} \|\beta^T \big(Span(\overrightarrow{\mathfrak{U}})|_{\in V}\big)^T \otimes \overrightarrow{\mathfrak{u}(x_i)}... \\ & - \beta^T \big(Span(\overrightarrow{\mathfrak{U}})|_{\in V}\big)^T \otimes Span(\overrightarrow{\mathfrak{U}})|_1^N s_i\| \\ & s.t. \quad \sum \|\alpha^T \big(Span(\overrightarrow{\mathfrak{U}})|_1^N\big)\| = 1 \qquad \sum_{t,r} \beta_{t,r}^2 = 1 \end{aligned} \tag{7}$$

Assuming $\mathfrak{S}_{\mathfrak{w}}$ is an inner product space, multiplication between each two tensors $\overrightarrow{\mathfrak{u}(x_p)}$ and $\overrightarrow{\mathfrak{u}(x_q)}$ can be approximated by the DTW distance between x_p and x_q sequences in the original warped space. This in turn allows for alleviating the influence of time dilation in the input space while simultaneously unfolding the nonlinearity of data.

$$\begin{aligned} & \big(\overrightarrow{\mathfrak{u}(x_p)}\big)^T \overrightarrow{\mathfrak{u}(x_q)} = dtw(x_p, x_q) \\ & \Rightarrow \mathbf{dtw}\big(\overrightarrow{x_p}, \overrightarrow{X}|_1^N\big) = \Big(\big(\overrightarrow{\mathfrak{u}(x_p)}\big)^T \otimes Span(\overrightarrow{\mathfrak{U}})|_1^N\Big) \\ & = \Big(dtw(x_p, x_{v_1}), dtw(x_p, x_{v_2}), ..., dtw(x_p, x_{v_N})\Big) \\ & \Rightarrow \mathbf{D}\big(\overrightarrow{X}|_{v\in V}, \overrightarrow{X}|_1^N\big) = \big(\big(Span(\overrightarrow{\mathfrak{U}})|_{\in V}\big)^T \otimes Span(\overrightarrow{\mathfrak{U}})|_1^N\big) \\ & = \Big(\Big(\mathbf{dtw}\big(\overrightarrow{x_1}, \overrightarrow{X}|_1^N\big)\Big)^T, ..., \Big(\mathbf{dtw}\big(\overrightarrow{x_V}, \overrightarrow{X}|_1^N\big)\Big)^T\Big)^T \end{aligned} \tag{8}$$

where $\overrightarrow{X}|_1^N = \{\overrightarrow{x_1}, \overrightarrow{x_2}, ..., \overrightarrow{x_N}\}$ is the set of all the training samples in the original space. Rewriting $\overrightarrow{\mathfrak{u}(x_i)}$ as $Span(\overrightarrow{\mathfrak{U}})|_1^N a_i$, where a_i is an N-dimensional unit vector in which i-th element is one and the remaining ones are zero, and then substituting equation (8) into (7), we have:

$$\begin{aligned} & min_\beta \sum_{i=1}^{N} \|\beta^T \mathbf{D}\big(\overrightarrow{X}|_{v\in V}, \overrightarrow{X}|_1^N\big)a_i... \\ & - \beta^T \mathbf{D}\big(\overrightarrow{X}|_{v\in V}, \overrightarrow{X}|_1^N\big)s_i\| \\ & s.t. \quad \sum \|\alpha^T \big(Span(\overrightarrow{\mathfrak{U}})|_1^N\big)\| = 1 \qquad \sum_{t,r} \beta_{t,r}^2 = 1 \end{aligned} \tag{9}$$

It is noteworthy that, the reciprocal dependency of DTW [87] violates the Positive Symmetric Definite (PSD) property of the similarity matrix $\mathbf{D}$ which causes the unconstraint form of this formulation to be non-convex. However, the introduced regularization parameter of β can prevent the eigenvalues of this matrix to be negative, resulting in a pseudo-PSD property which further ensures an optimal-stable solution.

To simplify notation hereafter, we will omit the arguments of $\mathbf{D}$. Then, according to the Lagrange multiplier theorem, this dual bounded model can be rewritten as an expression with only one constraint.

$$\begin{aligned} & min_\beta \sum_{i=1}^{N} \|\beta^T \mathbf{D} a_i - \beta^T \mathbf{D} s_i\| + \gamma \sum_{s,r} \beta_{s,r}^2 \\ & s.t. \quad \|\mathbf{D}^T \beta\| = 1 \\ & \Rightarrow tr\Big(\beta^T \mathbf{D}\Big(\sum a_i a_i^T - \sum a_i s_i^T - \sum s_i a_i + ... \\ & ... \sum s_i s_i^T\Big)\mathbf{D}^T \beta\Big) + \gamma tr\Big(\beta^T \beta\Big) \quad s.t. \quad \|\mathbf{D}^T \beta\| = 1 \end{aligned} \tag{10}$$

With some algebraic simplification, the model can be recast in a matrix form as:

$$max_\beta \frac{\beta^T \Big(\mathbf{D}(S + S^T - SS^T)\mathbf{D}^T + \gamma I\Big)\beta}{\beta^T \mathbf{D}\mathbf{D}^T \beta} \tag{11}$$

Then, the optimal solution of β is considered as the eigenvectors corresponding to the largest eigenvalues of the following equation:

$$\Big(\mathbf{D}(S + S^T - SS^T)\mathbf{D}^T + \gamma I\Big)\beta = \psi \mathbf{D}\mathbf{D}^T \beta \tag{12}$$

The update procedure of this method has been listed in Algorithm1.

3.3. Sparseness Embedding for 3D Action Recognition

In this paper, each action is represented by a trajectory of 3D coordinates over time, where each row of the trajectory is a time series representation for one

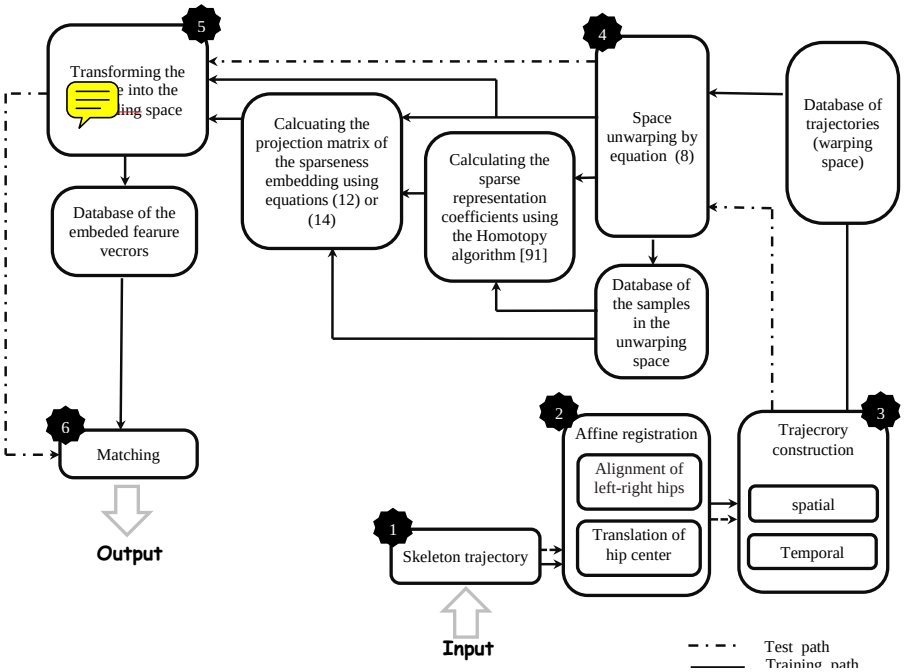


Figure 4: Block diagram of our proposed unsupervised action recognition method. Black stars indicate the necessary steps for classifying a query sequence

of the skeleton joints along the x-, y-, or z-directions. The fact that we require to recognize the actions without any user cooperation causes two fundamental problems which makes it impossible to use the raw data as the input of the classification model; first, each action can be started anywhere in the camera's field of view, not necessarily in its center, and second, one can perform the same action in different directions or even change his/her ongoing orientation while performing activities. Therefore, we first normalize all the skeletons so that to be centered, facing the camera. For this purpose, the hip joint of each skeleton is moved to the origin, and all are then rotated so that the line between the left and right hips becomes parallel to the x-axis. To alleviate the influence of the execution speed, this aligning procedure is identically applied on all the skeletons of sequences. This causes only the person-centric displacements to be involved in the recognition process which is much closer to what happens in our brain. However, such a representation method only characterizes the spatial attributes, without making any provision for the temporal features, which can be very important. Recent studies on deep learning acknowledge the correlation between the most important joints/frames and temporal variations of sequences [39]. Taking this into account, we consider the first derivative of the trajectories as the temporal expression of sequences which are referred to as temporal trajectories. Then, we propose two different protocols to simultaneously utilize

the characteristics of both the spatial $\overrightarrow{x}$ and temporal $\overrightarrow{\Im}$ representations. In the first one, DTW distances of both the trajectories are concatenated and considered as a single entity. However, because of heterogeneity, such a strategy may lead to a biased measure of the characteristics. Thus, aiming to compromise between the optimal projections of such different attributes, we propose a hybrid embedding scheme which utilizes the weighted sum strategy of sparsity preserving for these two individual measures:

$$\begin{aligned} &min_{\beta} \sum_{i=1}^{N} \|\beta^T \mathbf{D} a_i - \beta^T \mathbf{D} s_i\| + \gamma_1 \sum_{i=1}^{N} \|\beta^T \mathbf{Z} a_i - \beta^T \mathbf{Z} h_i\| + \gamma_2 \|\beta\|^2 \\ &s.t. \;\; \|\mathbf{D}^T \beta\| = 1 \end{aligned} \tag{13}$$

where $\mathbf{D} = \mathbf{D}(\overrightarrow{X}|_{v \in V}, \overrightarrow{X}|_1^N)$ and $\mathbf{Z} = \mathbf{Z}(\overrightarrow{\chi}|_{v \in Q}, \overrightarrow{\chi}|_1^N)$ are respectively the affinity approximations for the set of spatial $X = \{\overrightarrow{x_i}\}|_{i=1}^N$ and temporal $\chi = \{\overrightarrow{\Im_i}\}|_{i=1}^N$ trajectories. V and Q are those selected from the spatial and temporal trajectories in which the projection matrix β individually lies on. According to (10), equation (13) can be rewritten as follows:

$$max_{\beta} \frac{\beta^T \Big(\mathbf{D} A \mathbf{D}^T + \lambda_1 \mathbf{Z} B \mathbf{Z}^T + \lambda_2 I \Big) \beta}{\beta^T \mathbf{D} \mathbf{D}^T \beta} \tag{14}$$

where $A{=}S + S^T - S^T S$ and $B{=}P + P^T - P^T P$, S and P are respectively the sparse representation coefficients for the spatial and temporal trajectories which can be approximated by performing an l_1-norm based reconstruction on their corresponding datasets. The flowchart of our proposed method has been shown in Figure 4.

4. Experimental results

In this section, we evaluate the performance of our method compared with a set of state-of-the-art methods on six publicly available databases including UTKinect, TST fall, UTD MHAD, CMU, Berkeley MHAD and NTU RGB+D. The experiments are mostly designed in unsupervised scenarios, however a few are also conducted to demonstrate the effectiveness of our method in a supervised setting. In the following, we first briefly describe the databases, then compare the performance of our method with the state-of-the-arts, analyze the confusion matrices, and finally examine its sensitivity to different parameters.

4.1. Databases

UTKienect: This database consists of 10 human actions including walk, sit down, stand up, pick up, carry, throw, push, pull, wave, and clap hands, each performed twice by 10 different subjects. All actions were captured by a single

Algorithm 1 Sparseness Embedding

training spatial $\overrightarrow{X}|_{i=1}^{N}$ and temporal $\overrightarrow{\chi}|_{i=1}^{N}$ sequences
Embedded representation of all the sequences
#Construct affinity matrices
for p=1:N **do**
$\mathbf{d}_p$=**dtw**$\left(\overrightarrow{x_p}, \overrightarrow{X}|_1^N\right) \rightsquigarrow \left(\left(\overrightarrow{\mathfrak{u}(x_p)}\right)^T \otimes Span(\overrightarrow{\mathfrak{U}_X})|_1^N\right)$
$\mathbf{z}_p$=**dtw**$\left(\overrightarrow{\mathfrak{x}_p}, \overrightarrow{\chi}|_1^N\right) \rightsquigarrow \left(\left(\overrightarrow{\mathfrak{u}(\mathfrak{x}_p)}\right)^T \otimes Span(\overrightarrow{\mathfrak{U}_\chi})|_1^N\right)$
$\mathbf{d}_p \leftarrow \frac{\mathbf{d}_p}{\|\mathbf{d}_p\|}$
$\mathbf{z}_p \leftarrow \frac{\mathbf{z}_p}{\|\mathbf{z}_p\|}$
$\mathbf{Z} \leftarrow \mathbf{z}_p$
$\mathbf{D} \leftarrow \mathbf{d}_p$
end for
 # Concatenate both the matrices to construct a unified representation of affinity
$\mathbf{D} \leftarrow \left(\mathbf{D}, \mathbf{Z}\right)^T$
 # Determine the spanning set by applying the Greedy algorithm [88] on the affinity matrix
$\Omega = \left(\omega_1\, \omega_2, ..., \omega_N\right) \in R^{V \times N} \leftarrow \mathbf{D} \in R^{2N \times N}$
for p=1:N **do**
 # Exclude ω_p from Ω
$\mathfrak{O} \leftarrow \Omega$
 # Encode ω_p on $\mathfrak{O}$
$min_{s_i} \|\mathfrak{O} s_p - \omega_p\|_2 + \lambda \|s_p\|_1$
 # Calculate s_p by using the Homotopy algorithm [89]
end for
 # Calculate the largest eigenvalues for the objective function of SE
Solve $\left(\mathbf{D}(S + S^T - SS^T)\mathbf{D}^T + \gamma I\right)\beta = \psi \mathbf{D}\mathbf{D}^T \beta$
 # Project data into the embedding space
$\forall p \in \{1, .., N\} \;\; y_p = \beta^T \omega_p$

stationary Kinect sensor in an indoor setting with a frame rate of about 15fps. Actions are represented as a sequence of skeletons, each configured by 20 joints. Despite the small number of sequences, their high intera-class variations makes it difficult to learn these actions in an unsupervised manner.

TST fall: It consists of 264 sequences of eight actions, each performed three times by 11 subjects. The sequences are mainly performed in two different categories: daily living activities (sit, grasp, walk, and lay) and falling actions (falling front, back, side and falling backward while ends up sitting). Each action includes a time-series of 3D positions for 24 joints estimated from a depth stream of a Kinect V2 sensor. The main challenge of this database is the very own falling style of each subject which is critical for real-world elderly/patient monitoring systems.

UTD-MHAD: This database include 861 sequences from 8 participants, each performing 27 actions from three different groups; sport, daily living, and hand gesture activities. The actions include 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, basketball shoot, right hand draw x, right hand draw circle (clockwise), right hand draw circle (counter clockwise), draw triangle, bowling (right hand), front boxing, baseball swing from right, tennis right hand forehand swing, arm curl (two arms), tennis serve, two hand push, right hand knock on door, right hand catch an object, right hand pick up and throw, jogging in place, walking in place, sit to stand, stand to sit, forward lunge (left foot forward), and squat, all captured in a real indoor environment by a Kinect sensor at a frame rate of 30fps.

CMU: This database contains 2235 sequences of 45 actions performed by 144 participants. Unlike the previous databases, CMU includes long term activities with quite varying lengths which in turn allows for evaluating algorithms under more realistic conditions. However, these also make the main challenges of this database. Each action is represented by the 3D coordinate of 31 joints. Following the protocol designed in [90], we use only a subset of 664 sequences for 8 more common actions of daily routine including jump, walk back, run, sit, getup, pickup, basketball, and cartwheel.

Berkeley MHAD: This database contains 659 sequences for 11 actions from 12 subjects where each action is repeated five times. The actions include jumping, jumping jacks, bending, punching, waving two hands, waving one hand, clapping, throwing, sit down/stand up, sit down, and stand up. For each skeleton, 35 body joints are provided by the database employing an optical motion capture system. Due to high resolution of the joint coordinates, this database provide much more clean information in comparison with the other databases.

NTU RGB+D: NTU RGB+D was originally established for use in data-hungry algorithms like deep learning based approaches. It contains 56880 action sequences from 40 different individuals with the age range from 10 to 35 years. Number of joints and their configuration are similar to the TST database, but the settings of x-y-z- axis are slightly different. The activities are mainly categorized into 60 groups including 40 types of daily routines, 11 mutual interactions and 9 health-related activities. All actions have been collected by a Kinect v2

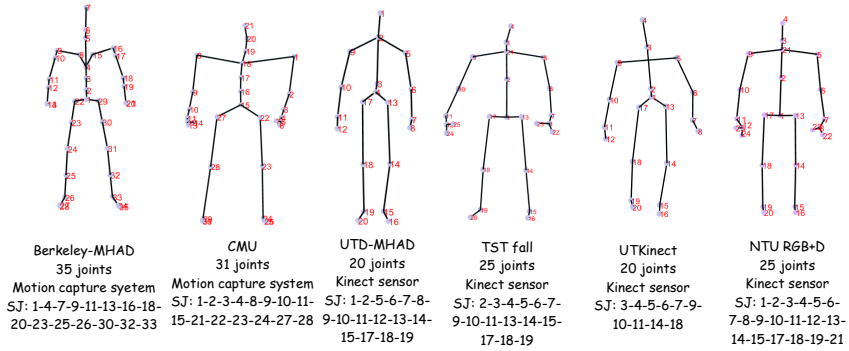


Figure 5: Skeleton configuration in different databases as well as the most informative joints used in our proposed method.

sensor under a variety of setups for camera height and distance. Each setting consists of two individual performances captured at three different angles.

4.2. Implementation Details

In this section, we describe the evaluation protocol and details of the alignment procedure and warping of sequences. For all the databases, we utilize the hip center as the origin and align the skeletons w.r.t the angle of the line passing through the left-right hips. As body parts are not entirely involved in all activities, similar to [91] we utilize a joint selection strategy to focus on more discriminative ones. This prevents a significant attention being directed on the non-salient units of body, avoiding any possible redundancy in the representation of sequences. The selected joints for each database are reported in Figure 5. This not only improves the discriminative ability of the measures, but also decreases the computational cost of the warping procedure in equation (8) which can be estimated by $O(o\mathfrak{G}\mathfrak{F})$, where o, $\mathfrak{G}$, and $\mathfrak{F}$ respectively stand for the number of joints, and the length of the first and second sequences. For UTKinect, we use two different evaluation protocols; Leave-One-Sequence-Out (LOSeqO), and Leave-One-Subject-Out (LOSubO). For each fold of LOSeqO, one sequence is excluded for test and the remaining ones are regarded as the training samples. This procedure is repeated k times while every fold is excluded once. This ensures that the training samples of true class and test sequences are roughly of the same length. Therefore, it is more beneficial to first down-sample all the sequences so as to have a predefined number of frames (which is empirically set to 15). Although more common in the literature, LOSeqO does not have any provision for the study of subject-to-subject motion style variations and may bias the results towards the tendency of performing actions in the same manner. In contrast, LOSubO excludes all the sequences belonging to one subject when selecting one of its samples for test. The recognition accuracies for different methods are reported in Table 1. To our best knowledge, there is not any unsupervised method evaluated on this database.

Table 1: Recognition Accuracy for UTKinect database

Method	Year	Strategy	Protocol	Acc rate	# Seq
3DKPM [92]	2016	Supervised	LOSeqO	93.47%	199
LARP+mfPCA [93]	2015	Supervised	Two-fold	94.87%	199
ST-LSTM+TG [94]	2018	Supervised	Two-fold	95.00%	200
ST-LSTM+TG [94]	2018	Supervised	LOSeqO	97.00%	200
mLSTM+JLD [36]	2017	Supervised	Two-fold	95.96%	199
GCA-LSTM [95]	2017	Supervised	LOSeqO	98.5%	200
GCA-LSTM+SWT [94]	2017	Supervised	LOSeqO	99.00%	200
Lie Algebra [79]	2014	Supervised	Two-fold	97.08%	199
Lie Algebra+DL [96]	2018	Supervised	LOSeqO	98.5%	199
GSR [97]	2018	Supervised	Two-fold	94.30%	199
GM+LTB [68]	2015	Supervised	LOSeqO	88.50%	200
HODV [98]	2014	Supervised	LOSeqO	91.96%	199
HOJ3D [99]	2012	Supervised	LOSeqO	90.09%	199
RAPToR [100]	2017	Supervised	LOSeqO	92.01%	199
FisherPose [60]	2018	Supervised	LOSubO	89.00%	200
LM3TL [101]	2017	Supervised	LOSubO	98.9%	199
DMIMTL [102]	2017	Supervised	LOSubO	99.19%	199
Our method	-	**Unsupervised**	LOSubO	92.00%	200
Our method	-	**Unsupervised**	LOSeqO	94.50%	200

Table 2: Recognition Accuracy for TST fall Database.

Method	Year	Strategy	Protocol	Acc rate
FisherPose [60]	2018	Supervised	LOSubO	88.6%
MDTW [103]	2018	Supervised	LOSubO	92.3%
HOJ3D [99]	2012	Supervised	LOSubO	70.83%
PKE [104]	2014	Supervised	LOSubO	84.09%
Our method	-	**Unsupervised**	LOSubO	94.27%

For TST fall, we conduct the experiments using the LOSubO protocol. Therefore, in each fold, action sequences of 10 subjects are used for training and the remaining ones regarded for test. The comparison results for this database are listed in Table 2. As can be seen, there is a noticeable lack of studies on this database which might be due to its originality which, in its true sprit, has been established for studying the problem of falling actions.

For UTD-MHAD, we follow the cross validation protocol designed by Chen et al. [94] which is also the overwhelming trend in the competing algorithms. On this basis, subjects with odd indices are used for training and the remaining ones are selected for test. Table 3 shows the results of recognition accuracies on this database.

For CMU, there is a standard protocol suggested by Zhu et al. [92] in

Table 3: Recognition Accuracy for UTD-MHAD Database

Method	Year	Strategy	Protocol	Acc rate
Multiview+CNN+LSTM [28]	2019	Supervised	Two-fold	95.58%
CNN+OSF [67]	2018	Supervised	Two-fold	86.97%
DLF [105]	2018	Supervised	Two-fold	92.8%
JTM+CNN [106]	2016	Supervised	Two-fold	85.81%
ModJTM+CNN [40]	2018	Supervised	Two-fold	87.90%
JDM+CNN [107]	2017	Supervised	Two-fold	88.10%
AG [108]	2017	Supervised	Two-fold	81.00%
Skepxel+L+V [109]	2018	Supervised	Two-fold	97.2%
GME [77]	2018	Supervised	Two-fold	90.74%
Our method	-	**Unsupervised**	Two-fold	93.29%

Table 4: Recognition Accuracy for CMU Database

Method	Year	Strategy	Protocol	Acc rate
Hierarchical RNN [66]	2015	Supervised	Three-fold	83.13%
Deep LSTM [90]	2016	Supervised	Three-fold	86.00%
Deep LSTM + Co-occurrence [90]	2016	Supervised	Three-fold	88.40%
Coordinates + FTP [110]	2017	Supervised	Three-fold	83.44%
Frames + CNN [110]	2017	Supervised	Three-fold	91.53%
Clips + CNN + Concatenation [110]	2017	Supervised	Three-fold	90.97%
Clips + CNN + Pooling [110]	2017	Supervised	Three-fold	90.66%
Clips + CNN+ MTLN [110]	2017	Supervised	Three-fold	93.22%
Encoder-Decoder+GAN [45]	2018	**Unsupervised**	Three-fold	84.57%
Autoencoder [111]	2015	**Unsupervised**	Three-fold	77.03%
Our method	-	**Unsupervised**	Three-fold	85.69%

which a three-fold class validation strategy is applied on a 664-sequence subset of sequences. The experimental results on this database is reported in Table 4.

Similar to CMU, Berkeley MHAD takes advantage of a standard evaluation protocol provided by the database creators in which first 7 subjects are selected for training and last 5 ones are used for test. The recognition accuracies on this database are listed in Table 5.

For NTU RGB+D, we utilize its two well-known protocols, cross-subject and cross-view. In cross subject, half of the subjects are utilized for training and other half for test. For cross-view evaluation, the sequences taken by the first camera are considered as test samples while the remaining ones are used for training. To dealing with the large size of database, we propose to divide the training sequences into seventeen categorizes. Then, mapping function of each category is individually calculated according to equation (12). Then, test samples are classified based on the least matching scores of the individual mappings.

Table 5: Recognition Accuracy for Berkely MHAD Database

Method	Year	Strategy	Protocol	Acc rate
SMIJ [112]	2013	Supervised	Two-fold	95.40%
DBRNN [66]	2015	Supervised	Two-fold	99.64%
HBRNN [66]	2015	Supervised	Two-fold	100%
Deep LSTM [90]	2016	Supervised	Two-fold	100%
Co-deep LSTM [90]	2016	Supervised	Two-fold	100%
Encoder-Decoder+GAN [45]	2018	**Unsupervised**	Two-fold	100%
Autoencoder [111]	2015	**Unsupervised**	Two-fold	99.56%
Our method	-	**Unsupervised**	Two-fold	100%

Table 6: Recognition Accuracy for NTU RGB+D Database

Method	Year	Strategy	Protocol	Acc rate
LS-LSTM+TG [94]	2018	Supervised	CS	69.2%
LS-LSTM+TG [94]	2018	Supervised	CV	77.7%
Lie Algebra [79]	2014	Supervised	CS	50.1%
Lie Algebra [79]	2014	Supervised	CV	52.8%
PA-LSTM [113]	2016	Supervised	CS	62.9%
PA-LSTM [113]	2016	Supervised	CV	70.3%
Deep RNN [113]	2016	Supervised	CS	56.3%
Deep RNN [113]	2016	Supervised	CV	64.1%
Hierarchical RNN [66]	2015	Supervised	CS	59.1%
Hierarchical RNN [66]	2015	Supervised	CV	64.0%
GCA-LSTM [52]	2018	Supervised	CS	76.1%
GCA-LSTM [52]	2018	Supervised	CV	84.0%
SPMF Inception-ResNet-222 [114]	2018	Supervised	CS	78.9%
SPMF Inception-ResNet-222 [114]	2018	Supervised	CV	86.1%
Skeletal Quads [115]	2014	Supervised	CS	38.6%
Skeletal Quads [115]	2014	Supervised	CV	41.4%
FTP Dynamic Skeletons [116]	2015	Supervised	CS	60.2%
FTP Dynamic Skeletons [116]	2015	Supervised	CV	65.2%
DPRL+GCNN [82]	2018	Supervised	CS	83.5%
DPRL+GCNN [82]	2018	Supervised	CV	89.8%
Our method	-	**Unsupervised**	CS	71.8%
Our method	-	**Unsupervised**	CV	79.3%

4.3. Discussion

Based on the experimental results summarized in Tables 1 to 5, we have the following observations:

No matter how the sequences are modeled, supervised learning strategies have usually better overall performance than their corresponding unsupervised cousins. However, this comes with two major drawbacks: First, they are limited to a heavy reliance on massive labeled training samples which may not

Table 7: Recognition Rate of Our Method Compared with Some Supervised Learning Strategies When Outliers are Present in the Database.

Method	Lie Alg.	RP	AQ	3DT	Our method
Orig. database	94.5%	97.5%	94.5%	94.0%	94.00%
Database with Outliers	89.5%	93.5%	92.00%	87.00	94.00%

always be available especially when dealing with complex/chaotic actions such as sport movements. Second, and more importantly, this superiority is only truly guaranteed when there is no outlier among the training samples. To facilitate further analysis, we conduct another experiment to validate the robustness of our method against some incorrectly labeled sequences, and compare the results with a few supervised learning strategies including Lie Algebra [34], Relative Position [34], Angle Quaternion [34], and 3D Trajectories [34]. For this purpose, we replace the first sequence of each class with the first sample of the next one, and reevaluate the methods on this new database. All the experiments are performed on UTKinect using LOSeqO strategy with the same setting as the previous section, both for splitting the database and the regularization parameters. The results of the experiment are reported in Table 6. As can be seen, despite the superiority of the supervised algorithms for ideally labeled database, 97.5% achieved by Relative Position versus 94.5% for our unsupervised strategy, these methods can not achieve the best results in the presence of outliers, according to a 93.5% accuracy rate for Relative Position versus 94.5% achieved by our method, which is a great challenge for the practical realization of such strategies. Note that, as we focus on the semantics of responses, the outcome of outliers are considered to be the same as their original indices.

Deep features of sequences does not always outperform the shallow representations. For example, among the supervised algorithms examined on 199-sequence UTKinect database using LOSeqO protocol, DMIMTL can achieve 99.19% accuracy which is much better than that of many deep learning based methods such as mLSTM+JLD [36] and Lie Algebra+DL [96]. A consistent trend can also be observed among the unsupervised approaches, where the performance of our method is respectively 0.52% and 8.06% better than those obtained by the recently proposed deep learning based AutoLSTM [45], and EDGAN [111] methods, while it also ~~benefit~~ from a far less computational cost. Note that, each layer of the encoder, decoder, and discriminator of EDGAN has respectively 800, 800, and 200 hidden units, requiring massive amounts of computational power. Besides, due to scarcity of 3D sequences, such a large number of parameters not only does not contribute to a good generalization but also makes the network prone to overfitting. That is why deep learning strategies can not repeat their previous huge success for 3D action recognition.

Our method outperforms the manifold learning based algorithms including GM+LTB [68] and GME [77]. Specifically, about 6 and 2.5% improvement is respectively archived on UTKinect and UTD-MHAD databases, arising from one

of the following advantages: (1) SE does not require any assumption regarding the geometrical distribution of data. (2) unlike manifold learning methods that mostly focus on the second-order statistics, our model only utilize the first-order correlations which are less susceptible to noise. (3) it does not involve any statistical modeling of dynamic reversions, leading to less uncertainty and degenerated solutions.

Similar to SE, some other rivals like the methods in [71, 117], take advantages of the time warping to dealing with the latency issue of sequences. However, due to the monotonicity and the boundary constraints [68], dtw fails to be directly applicable for classifying complex sequences. In contrast, our method differs from these techniques in two different ways: (1) SE utilizes the warping strategy to capture the underlying distribution of data rather than aligning the sequences. (2) it encodes the nonlinearity in a sparse manner, leading to a higher degree of robustness against the partial noise or corruption of trajectories.

Different from the multi-stream, multi-step deep learning based approaches, our method has a unified structure that fulfils the need to any further correlational or causal analysis of hierarchies.

Our method achieves a significant improvement over the current unsupervised 3D action recognition methods, AutoLSTM and EDGAN, mainely from two different aspects: a higher recognition rate and lower computational cost. In addition, it does not rely on any augmentation procedure.

Unlike the attention mechanism in deep learning strategies, our method utilize the temporal characteristics of joint trajectories as a measure of selective concentration on informative joints, leading to a far much lower computational cost.

Unlike the graph-based and tree-traversal algorithms [56, 39], our method neglects the spatial relations between the joints, which causes the loss of some valuable information of data, leading to a decreased accuracy. However, this way it allows for a simple interpretation of the single joint's role in the recognition process which is much crucial for debugging the algorithms. Additionally, such graph based networks suffer from two major drawbacks: with a shallow structure they lack the ability of generalization and in a deep form has the problem of over-smoothing, both hinder their application in 3D model characterisation.

4.4. Confusability

In this section, we further explore which action classes are more prone to be confused with each other. For this purpose, confusion matrices for UTKinect, TST fall, UTD-MHAD, CMU, and NTU RGB+D databases are respectively shown in Figures 6 to 12. However, due to demonstrating a perfect recognition rate, the matrix are not depicted for Berkeley MHAD. For all the measures, diagonal elements show the per-class accuracy rates, while those off-diagonal represent the misclassification percentages.

For UTKinect, the matrix is depicted for a classification rate of 92% achieved when setting the parameters to $\{\varepsilon{=}10^{-6}, \lambda{=}10^{-3}, \gamma{=}10^{-5}, \#GFS{=}30, \#PC{=}30\}$. As can be seen, most of the confusions occur between the throwing with pushing actions, and also picking up and walking. In the first case, the reason mostly

lies in the aligning way of the skeletons. Note that, for aligned-free strategies similar to that found in humans, ground is considered to be the reference surface of comparing different actions, which in turn allows for distinguishing between the almost similar actions which have different motion patterns for their centers of gravity. However, single-joint based mining of such information from hip-aligned skeletons would not be an easy task, leading to a reduced efficiency of our method. But for the second case, the confusions are mostly related to the similarities between the actions, where, even the visual recognition mainly results from such subtle cues as the existence of an object or difference of accelerations.

For the TST database, the confusion matrix is given for a recognition accuracy of 94.27% which is achieved by the following parameters: $\{\varepsilon{=}10^{-4}, \lambda{=}10^{-7}, \gamma{=}10^{-2}, \#GFS{=}60, \#PC{=}60\}$. The matrix implies that misclassifications mainly occur in the falling actions, especially between falling side and falling back which share a set of significant initial motion similarities. Another reason is related to the greater impact of varying styles on performing falling actions.

For UTD-MHAD, the confusion matrix is shown for a recognition rate of 93.29% ($\{\varepsilon{=}10^{-1}, \lambda{=}10^{-5}, \gamma{=}10^{-2}, \#GFS{=}130, \#PC{=}130\}$). As can be seen, our method can achieve a perfect recognition rate of 100% for 15 classes, while only one class experiences a rate lower than 80% accuracy. Obviously, the most difficulty of the database lies in misclassifications between push and basketball shoot, cross arm and clap, as well as draw circle and triangle which are mainly due to their visual consistency in hierarchy of employing different body parts. While such cases as confusions in the drawing actions can be alleviated by using the relative information of joints, in other cases, the error is mostly related to the failure of Euclidean based DTW measure in distinguishing between the actions that only differ in small parts.

The confusion matrices of the CMU subsets at the recognition rates of 85.52%, 86.88%, and 84.68% are respectively illustrated in Figures 7 to 9. As mentioned before, the best parameter setting of each subset is individually calculated using the gride search algorithm and the cross validation strategy. It is easy to see that, the worst performance of our method is achieved on this database which is mostly due to the large sequence length variations in this database. The ratio of longest to shortest length of sequences in this database is about 99.1% which is much greater than that in UTKinect with 22.8%, TST fall with 6.17%, and UTD-MHAD with 3%. The lowest accuracies on this database are achieved for the actions sit and get up which, due to the displacement of the gravity point, are very sensitive to the alignment procedure. The noteworthy thing about this database is the presence of three imposter actions (pick up, run, and jump) whose aligned form exhibit high levels of resemblance to other sequences, such that 11.47%, 28.98%, and 20.67% of false positives in the database are respectively assigned to these actions.

For NTU RGB+D, the confusion matrices are individually illustrated for both the cross-subject and cross-view protocols (Figure 12 (a) and (b)), respectively for the recognition rates of 71.8% and 79.3% ($\{\varepsilon{=}10^{-3}; \lambda{=}10^{-3}; \gamma{=}10^{-3}; \#GFS{=}95; \#PC{=}95\}$). As other databases, most of the confusions also

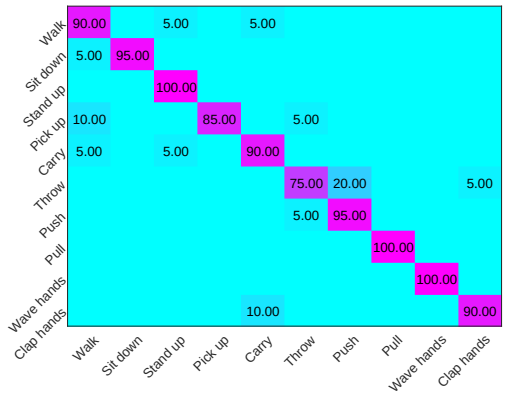


Figure 6: Confusion Matrix for UTKineact Database

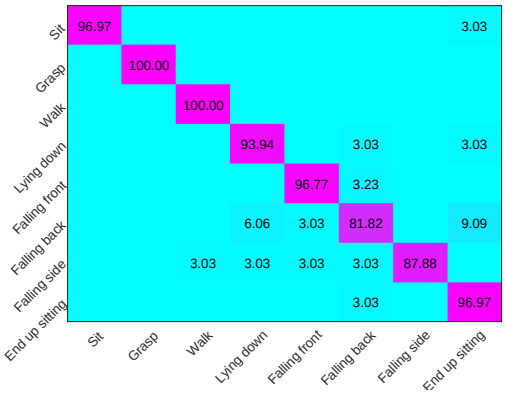


Figure 7: Confusion Matrix for TST fall Database

occur in those similar motions like reading/writing vs playing with phone, pat on patch/touch batch vs shaking hand, and brush teeth vs phone call. Moreover, one can observe that the confusions of both the proctors are largely similar, however their number for cross-view is significantly less than those occurred in the cross-subject protocol. Interestingly, despite some confusions, the reverse motions like "putting on something" and "taking off it" have been well discriminated (specially for cross-view protocol) by our method which is mostly due to the use of a temporal encoding procedure in its mapping function.

4.5. Evaluation of joint Learning Strategy

In this section, we evaluate how well our joint learning strategy works for a sparseness embedding task. For this purpose, a set of experiments are conducted on UTD-MHAD database and results are compared with the baseline method. Therefore, the splitting protocol and experimental setting are selected as the previous section (λ=10^{-16}, ε=10^{-1}, and γ=10^{-2}). Moreover, the parame-

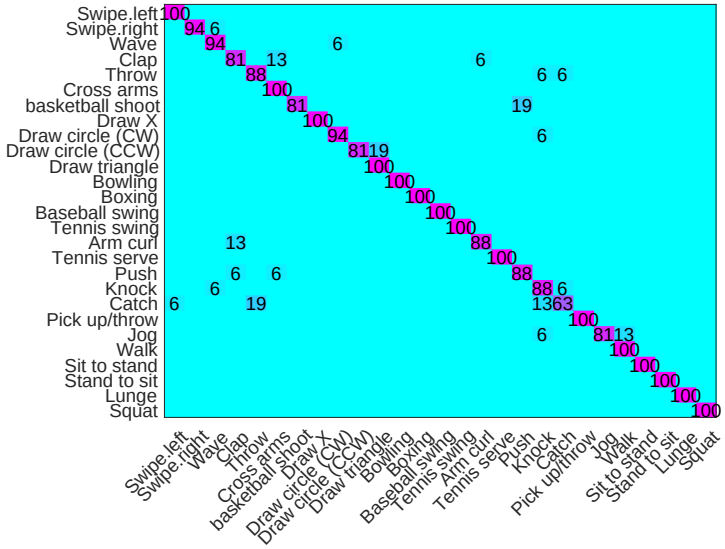


Figure 8: Confusion Matrix for UTD-MHAD Database

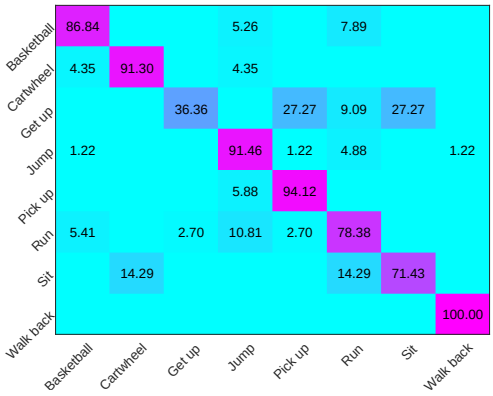


Figure 9: Confusion Matrix for CMU-subset1

ters γ_1 and γ_2 are respectively selected as 0.4 and 0.6, suggested by a grid-search strategy. The experimental results of the unsupervised mode of this strategy with different number of the spanning set determined by the Greedy algorithm [88] are summarized in Table 7. When comparing the results with Table 3, one can infer that our method is distinctly better than the baseline model. More specifically, the strategy leads to a 2.3% and 3.97% increase respectively in the maximum and average recognition accuracies of the method which demonstrate the benefit of jointly representing the temporal and spatial characteristics of sequences in two different spaces, but with an unified objective function.

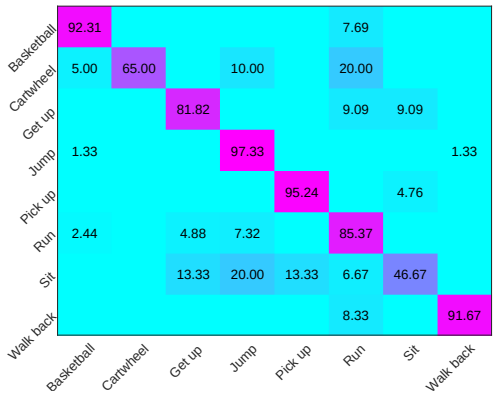


Figure 10: Confusion Matrix for CMU-subset2

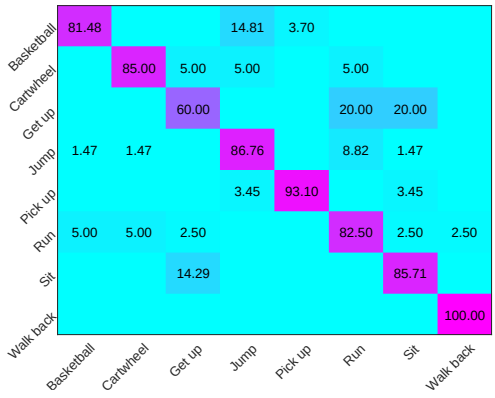
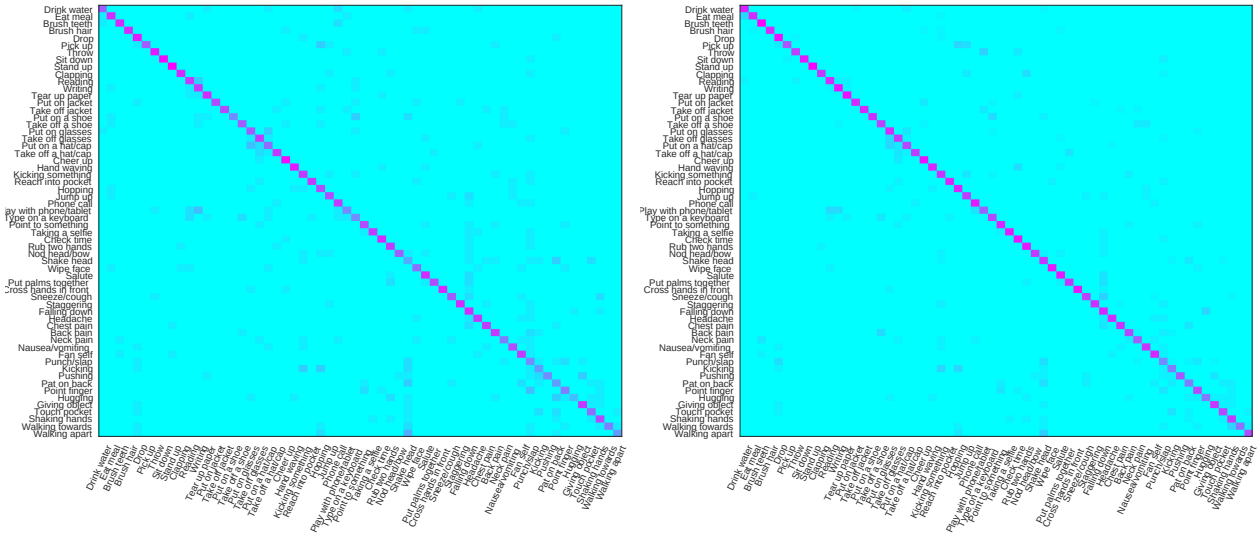


Figure 11: Confusion Matrix for CMU-subset3

We also conduct another experiment to validate the performance of this strategy in a supervised mode. To make a fair comparison, the parameter setting is assumed to be as the unsupervised strategy. Putting together the results of Table 7 and 3 demonstrates the competitive performance of our method compared with the best performing algorithms. However, in comparison, these rivals all suffer from an expensive training effort caused by the deep hierarchical or multi-stream framework of their structures. Without their code, it would not be possible for us to quantitatively compare the computational times of the methods. However, comparing the number of parameters (5 in our method versus more than millions in Skepxel [109]) can reveal the computational advantage of each strategies with further confirms the superiority of our algorithm.

We also visualize the representation results of the baseline and joint learning strategies of our method in Figure 10. For each strategy, sequences are initially transformed into the unwarped feature space. Then, PCA is applied to project the resultant trajectories into a compact three dimensional space. We also il-



(a) Cross-subject evaluation

(b) Cross-view evaluation

Figure 12: Confusion matrix for cross-view and cross-subject evaluations of NTU RGB+D database

Table 8: Recognition accuracy of our method using the joint learning strategy.

#Features	60	90	120	150	180	210
Unsupervised	96.67%	92.13%	93.52%	94.21%	95.37%	92.13%
Supervised	93.29%	96.53%	96.30%	96.53%	97.23%	96.30%

lustrate the sparsity of data from the viewpoint of the first, second, and third principal components. The brighter the lines, the more closely connection would be established among the trajectories. Each color in the distribution space is a label corresponding to one of the action classes in the database. As can be seen, the discriminative representations of the joint learning strategy is much better than that of the baseline algorithm.

Despite the promising performance on such a large number of actions, this strategy can be remarkably influenced by the challenging task of choosing the regularization parameters γ_1 and γ_2, imposing the need for an expensive grad search before applying the main body of the algorithm. Besides, this strategy is worth considering only if the temporal characteristics has competitive discriminative information compared to the spatial ones, which mostly occur in databases with clean movements.

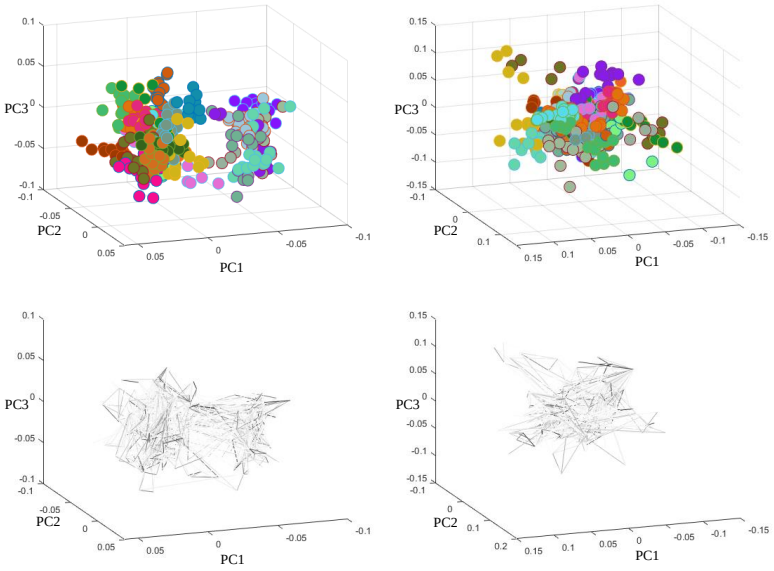


Figure 13: A Visual Interpretation of the Joint Learning Strategy (left column) versus the Baseline Algorithm (right column). The visualization of data (first row, each color represents a class of actions) vs. the sparsity coefficients between the samples (second row)

4.6. Sensitivity Analysis

In this section, we evaluate the parameter sensitivity of our method. According to whether the effects are direct or indirect, there are two categories of parameters to be tuned: (i) Those imposed by SRC to mine the intrinsic structure of data. As we use the Homotopy algorithm, such parameters are limited to ε and λ. (ii) The parameter γ, which is directly introduced in our proposed objective function in equation (10). In addition, there are two other factors that can significantly influence the performance of our method. First, the number of spanning sequences (V in Equation 4, which is referred to as the 'Number of Features') determined by applying the Greedy algorithm on the feature vectors derived from Equation (7) and the number of the principal components in the eigen-decomposition problem of Equation (11). In practice, to find the best values, we ran an exhaustive grid search over the five dimensional space of the parameters which is much more reliable than other hyper-parameter optimisation algorithms. However, illustrating such an extensive results could present a confusing picture that will not be properly informative. Therefore, to simplify the comparisons, the influence of each parameter is evaluated under the assumption that others are set to some fixed values, chosen to best illustrate the influence of the unknown parameter. The experimental results for different databases are listed in Table 8. It is noteworthy that the evaluation of a dictionary based learning algorithm like SE on such a large scale database as NTU RGB+D requires the training samples to be disaggregated into some dis-

tinctive categories. Due to the individual trend of sensitivity for each category (sometimes with a contradictory behaviour), it would not be informative to analyze them as a whole. However, empirical evaluation shows that a setting of $\{\varepsilon{=}10^{-3}, \lambda{=}10^{-3}, \gamma{=}10^{-3}, \#F{=}95, \#PC{=}95\}$ could achieve a compromise on all the categories of this database.

For ε, the values of the remaining parameters are fixed as $\{\lambda{=}10^{-3}, \gamma{=}10^{-5}, \#F{=}30\}$ for UTKinect, $\{\lambda{=}10^{-7}, \gamma{=}10^{-2}, \#F{=}50\}$ for TST fall, $\{\lambda{=}10^{-5}, \gamma{=}10^{-2}, \#F{=}130\}$ for UTD-MHAD, and $\{\lambda{=}10^{-5}, \gamma{=}10^{-6}, \#F{=}60\}$ for the CMU database. For λ, other parameters are set to be $\{\varepsilon{=}10^{-6}, \gamma{=}10^{-5}, \#F{=}30\}$ for UTKinect, $\{\varepsilon{=}10^{-3}, \gamma{=}10^{-2}, \#F{=}50\}$ for TST fall, $\{\varepsilon{=}10^{-1}, \gamma{=}10^{-2}, \#F{=}130\}$ for UTD-MHAD, $\{\varepsilon{=}10^{-1}, \gamma{=}10^{-5}, \#F{=}60\}$ for CMU. For γ, we set the other parameters as $\{\varepsilon{=}10^{-6}, \gamma{=}10^{-3}, \#F{=}30\}$ for UTKinect, $\{\varepsilon{=}10^{-7}, \gamma{=}10^{-4}, \#F{=}50\}$ for TST fall, $\{\varepsilon{=}10^{-1}, \gamma{=}10^{-2}, \#F{=}130\}$ for UTD-MHAD, AND $\{\varepsilon{=}10^{-1}, \gamma{=}10^{-5}, \#F{=}60\}$ for the CMU database. Note that, for all the experiments, number of the principal components in Equation (11) is assumed to be the same as $\{\#F\}$.

We also conduct another experiment to analyzes the influence of the dimensionality of the V space $\#F$ as well as the number of principal components on the performance of our model. For this purpose, the parameter setting of each database is considered to be the same as the Section IV.D. The performance variations with respect to these parameters are depicted in Figures 14 to 19.

Table 9: Sensitivity Analysis of Our Method Against the Parameters γ, λ, and ε.

	UTKinect				TST fall				UTD-MHAD			
ε	10^{-2}	10^{-4}	10^{-6}	10^{-8}	10^{-1}	10^{-2}	10^{-3}	10^{-4}	10^{-1}	10^{-2}	10^{-3}	10^{-4}
Recognition Rate	90.00	91.5%	92.00%	92.00%	88.55%	90.84%	92.37%	93.89%	93.29%	89.35%	86.81%	85.19%
λ	10^{-2}	10^{-3}	10^{-4}	10^{-5}	10^{-3}	10^{-5}	10^{-7}	10^{-9}	10^{-3}	10^{-5}	10^{-7}	10^{-9}
Recognition Rate	90.50%	92.00%	89.50%	89.50%	91.98%	91.22%	92.37%	91.60%	91.44%	93.29%	93.29%	93.29%
γ	10^{-3}	10^{-4}	10^{-5}	10^{-6}	10^{-1}	10^{-2}	10^{-3}	10^{-4}	10^{-1}	10^{-2}	10^{-3}	10^{-4}
Recognition Rate	86.50%	88.50%	92.00%	91.00%	90.46%	93.89%	90.08%	82.44%	86.81%	93.29%	86.81%	74.04%
	CMU-Subset1				CMU-Subset2				CMU-Subset3			
ε	10^{-1}	10^{-2}	10^{-3}	10^{-4}	10^{-1}	10^{-2}	10^{-3}	10^{-4}	10^{-1}	10^{-2}	10^{-3}	10^{-4}
Recognition Rate	85.07%	80.09%	82.81%	81.90%	84.62%	86.43%	83.26%	83.26%	82.43%	81.98%	83.78%	84.23%
λ	10^{-3}	10^{-4}	10^{-5}	10^{-6}	10^{-3}	10^{-4}	10^{-5}	10^{-6}	10^{-3}	10^{-4}	10^{-5}	10^{-6}
Recognition Rate	84.16%	85.07%	85.52%	85.52%	82.81%	85.07%	83.26%	83.26%	84.23%	82.88%	83.33%	84.68%
γ	10^{-2}	10^{-3}	10^{-4}	10^{-5}	10^{-2}	10^{-3}	10^{-4}	10^{-5}	10^{-2}	10^{-3}	10^{-4}	10^{-5}
Recognition Rate	81.00%	80.09%	82.81%	85.52%	85.07%	84.62%	83.71%	83.25%	81.98%	83.33%	81.53%	84.68%

Based on the results of both the experiments, we can reach the following observations:

- There is no absolute best parameter setting that can achieve the best performance on all the databases which is due to that the sparsity is a data-driven concept whose definition varies according to the scattering of data.

- The influence of variation in the parameter λ does not results in as large

performance variations as the parameters ε and γ.

- The more the number of training sequences (more dense space), the lower value the parameter ε may lead to more promising results.
- The best performance are usually obtained when the number of the number of principal components is equal to the dimensionality of the spanning set V.
- The more the number of action classes, the better results will be achieved using the higher dimensionality of the spanning set.

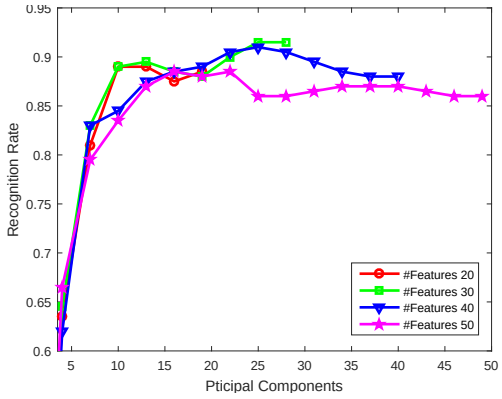


Figure 14: Performance of Our Method with the Varying Size of panning Set and Principal Components on UTKInect Database

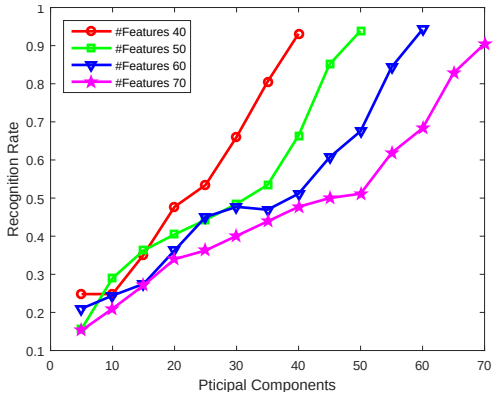


Figure 15: Performance of Our Method with the Varying Size of panning Set and Principal Components on TST fall Database

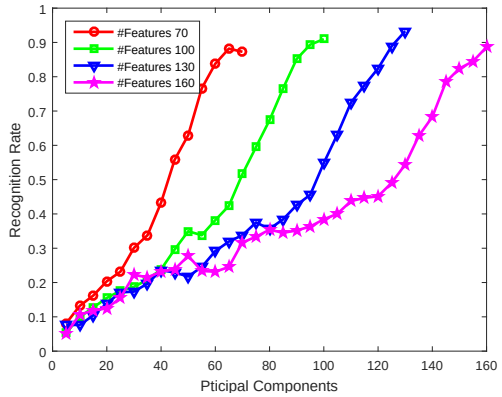


Figure 16: Performance of Our Method with the Varying Size of panning Set and Principal Components on UTS-MHAD Database

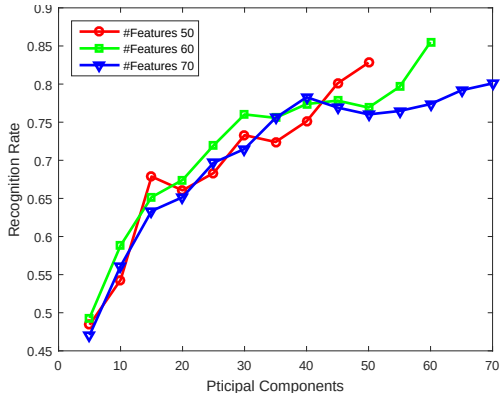


Figure 17: Performance of Our Method with the Varying Size of panning Set and Principal Components on CMU-subset1

5. Conclusion

This paper proposed a novel unsupervised 3D action recognition algorithm based on the coordinate information of skeletal data. The idea was to unwarp the original space such that the sparse neighborhood structure of sequences is preserved into a low dimensional unwarped space. This way, the idea of sparsity could be subtly introduced into the framework of high dimensional time-variant tensor analysis. Our scheme also provided a novel unified structure to integrate the nonlinearity and space curvature characteristics into a low dimensional sparse representation, allowing for a more reliable modeling of spatiotemporal attributes. Moreover, we proposed a novel joint learning strategy for dealing with the heterogeneity of temporal and spatial characteristics of action sequences which has been overlooked in the literature. The experimental results on six publicly available databases; UTKinect, TST fall, UTD-MHAD,

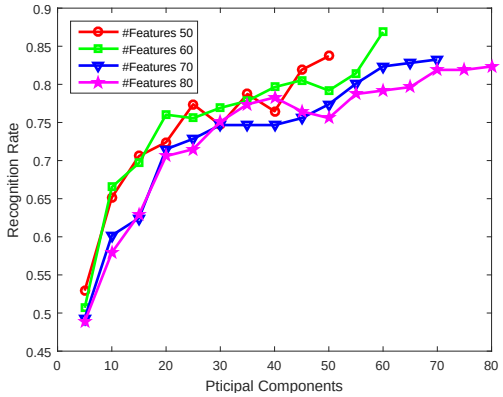


Figure 18: Performance of Our Method with the Varying Size of panning Set and Principal Components on CMU-subset2

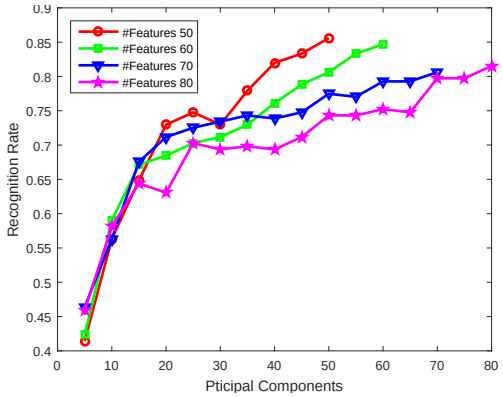


Figure 19: Performance of Our Method with the Varying Size of panning Set and Principal Components on CMU-subset3

CMU, Berkeley MHAD, and NTU RGB+D demonstrated the effectiveness of our method compared with a set of recently proposed shallow and even deep learning based strategies.

References

[1] A. B. Mabrouk, E. Zagrouba, Abnormal behavior recognition for intelligent video surveillance systems: A review, Expert Systems with Applications 91 (2018) 480–491.

[2] B. Reily, F. Han, L. E. Parker, H. Zhang, Skeleton-based bio-inspired human activity prediction for real-time human–robot interaction, Autonomous Robots 42 (6) (2018) 1281–1298.

[3] O. Mazhar, S. Ramdani, B. Navarro, R. Passama, A. Cherubini, Towards real-time physical human-robot interaction using skeleton information and hand gestures, in: 2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), IEEE, 2018, pp. 1–6.

[4] Y. Gao, X. Xiang, N. Xiong, B. Huang, H. J. Lee, R. Alrifai, X. Jiang, Z. Fang, Human action monitoring for healthcare based on deep learning, IEEE Access 6 (2018) 52277–52285.

[5] A. Prati, C. Shan, K. I.-K. Wang, Sensors, vision and networks: From video surveillance to activity recognition and health monitoring, Journal of Ambient Intelligence and Smart Environments 11 (1) (2019) 5–22.

[6] W. Bouachir, R. Gouiaa, B. Li, R. Noumeir, Intelligent video surveillance for real-time detection of suicide attempts, Pattern Recognition Letters 110 (2018) 1–7.

[7] J. Liu, Y. Gu, S. Kamijo, Customer behavior classification using surveillance camera for marketing, Multimedia Tools and Applications 76 (5) (2017) 6595–6622.

[8] F. Souza, E. Valle, G. Cámara-Chávez, A. Araújo, An evaluation on color invariant based local spatiotemporal features for action recognition, IEEE SIBGRAPI.

[9] A. A. Salah, B. Lepri, Second international workshop on human behavior understanding: inducing behavioral change, in: International Joint Conference on Ambient Intelligence, Springer, 2011, pp. 376–377.

[10] B. Yao, X. Jiang, A. Khosla, A. L. Lin, L. Guibas, L. Fei-Fei, Human action recognition by learning bases of action attributes and parts, in: 2011 International Conference on Computer Vision, IEEE, 2011, pp. 1331–1338.

[11] A. Kovashka, K. Grauman, Learning a hierarchy of discriminative space-time neighborhood features for human action recognition, in: 2010 IEEE computer society conference on computer vision and pattern recognition, IEEE, 2010, pp. 2046–2053.

[12] S. Ali, M. Shah, Human action recognition in videos using kinematic features and multiple instance learning, IEEE transactions on pattern analysis and machine intelligence 32 (2) (2010) 288–303.

[13] Z. Zhang, Microsoft kinect sensor and its effect, IEEE multimedia 19 (2) (2012) 4–10.

[14] L. Keselman, J. Iselin Woodfill, A. Grunnet-Jepsen, A. Bhowmik, Intel realsense stereoscopic depth cameras, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2017, pp. 1–10.

[15] H. Yasin, U. Iqbal, B. Kruger, A. Weber, J. Gall, A dual-source approach for 3d pose estimation from a single image, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 4948–4956.

[16] C. Migniot, F. Ababsa, 3d human tracking in a top view using depth information recorded by the xtion pro-live camera, in: International Symposium on Visual Computing, Springer, 2013, pp. 603–612.

[17] E. Velloso, A. Bulling, H. Gellersen, Autobap: Automatic coding of body action and posture units from wearable sensors, in: 2013 Humaine Association Conference on Affective Computing and Intelligent Interaction, IEEE, 2013, pp. 135–140.

[18] L. A. Schwarz, A. Mkhitaryan, D. Mateus, N. Navab, Human skeleton tracking from depth data using geodesic distances and optical flow, Image and Vision Computing 30 (3) (2012) 217–226.

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

[20] E. A. Suma, B. Lange, A. S. Rizzo, D. M. Krum, M. Bolas, Faast: The flexible action and articulated skeleton toolkit, in: 2011 IEEE Virtual Reality Conference, IEEE, 2011, pp. 247–248.

[21] J. Shotton, A. W. Fitzgibbon, M. Cook, T. Sharp, M. Finocchio, R. Moore, A. Kipman, A. Blake, Real-time human pose recognition in parts from single depth images., in: Cvpr, Vol. 2, 2011, p. 3.

[22] W. Chan, N. Jaitly, Q. V. Le, O. Vinyals, N. M. Shazeer, Speech recognition with attention-based recurrent neural networks, uS Patent 9,799,327 (Oct. 24 2017).

[23] C.-C. Chiu, T. N. Sainath, Y. Wu, R. Prabhavalkar, P. Nguyen, Z. Chen, A. Kannan, R. J. Weiss, K. Rao, E. Gonina, et al., State-of-the-art speech recognition with sequence-to-sequence models, in: 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2018, pp. 4774–4778.

[24] A. B. Graves, System and method for speech recognition using deep recurrent neural networks, uS Patent App. 15/043,341 (Feb. 5 2019).

[25] J. Hu, L. Shen, G. Sun, Squeeze-and-excitation networks, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 7132–7141.

[26] G. Huang, Z. Liu, L. Van Der Maaten, K. Q. Weinberger, Densely connected convolutional networks, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 4700–4708.

[27] B. Zoph, V. Vasudevan, J. Shlens, Q. V. Le, Learning transferable architectures for scalable image recognition, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 8697–8710.

[28] H. Yu, J. Wang, Z. Huang, Y. Yang, W. Xu, Video paragraph captioning using hierarchical recurrent neural networks, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 4584–4593.

[29] L. Gao, Z. Guo, H. Zhang, X. Xu, H. T. Shen, Video captioning with attention-based lstm and semantic consistency, IEEE Transactions on Multimedia 19 (9) (2017) 2045–2055.

[30] S. Venugopalan, L. A. Hendricks, R. Mooney, K. Saenko, Improving lstm-based video description with linguistic knowledge mined from text, arXiv preprint arXiv:1604.01729.

[31] D. Bahdanau, K. Cho, Y. Bengio, Neural machine translation by jointly learning to align and translate, arXiv preprint arXiv:1409.0473.

[32] Y. Wu, M. Schuster, Z. Chen, Q. V. Le, M. Norouzi, W. Macherey, M. Krikun, Y. Cao, Q. Gao, K. Macherey, et al., Google's neural machine translation system: Bridging the gap between human and machine translation, arXiv preprint arXiv:1609.08144.

[33] M.-T. Luong, H. Pham, C. D. Manning, Effective approaches to attention-based neural machine translation, arXiv preprint arXiv:1508.04025.

[34] H. Wang, L. Wang, Beyond joints: Learning representations from primitive geometries for skeleton-based action recognition and detection, IEEE Transactions on Image Processing 27 (9) (2018) 4382–4394.

[35] L. Wang, X. Zhao, Y. Liu, Skeleton feature fusion based on multi-stream lstm for action recognition, IEEE Access 6 (2018) 50788–50800.

[36] S. Zhang, X. Liu, J. Xiao, On geometric features for skeleton-based action recognition using multilayer lstm networks, in: 2017 IEEE Winter Conference on Applications of Computer Vision (WACV), IEEE, 2017, pp. 148–157.

[37] Z. Fan, X. Zhao, T. Lin, H. Su, Attention-based multiview re-observation fusion network for skeletal action recognition, IEEE Transactions on Multimedia 21 (2) (2019) 363–374.

[38] Y. Han, S.-L. Chung, A. Ambikapathi, J.-S. Chan, W.-Y. Lin, S.-F. Su, Robust human action recognition using global spatial-temporal attention for human skeleton data, in: 2018 International Joint Conference on Neural Networks (IJCNN), IEEE, 2018, pp. 1–8.

[39] C. Li, Z. Cui, W. Zheng, C. Xu, R. Ji, J. Yang, Action-attending graphic neural network, IEEE Transactions on Image Processing 27 (7) (2018) 3657–3670.

[40] T. Liu, J. Wang, S. Hutchinson, M. Q.-H. Meng, Skeleton-based human action recognition by pose specificity and weighted voting, International Journal of Social Robotics (2018) 1–16.

[41] J. Liu, A. Shahroudy, D. Xu, G. Wang, Spatio-temporal lstm with trust gates for 3d human action recognition, in: European Conference on Computer Vision, Springer, 2016, pp. 816–833.

[42] N. E. El Madany, Y. He, L. Guan, Integrating entropy skeleton motion maps and convolutional neural networks for human action recognition, in: 2018 IEEE International Conference on Multimedia and Expo (ICME), IEEE, 2018, pp. 1–6.

[43] J. Ren, N. Reyes, A. Barczak, C. Scogings, M. Liu, An investigation of skeleton-based optical flow-guided features for 3d action recognition using a multi-stream cnn model, in: 2018 IEEE 3rd International Conference on Image, Vision and Computing (ICIVC), IEEE, 2018, pp. 199–203.

[44] J. Tu, H. Liu, F. Meng, M. Liu, R. Ding, Spatial-temporal data augmentation based on lstm autoencoder network for skeleton-based human action recognition, in: 2018 25th IEEE International Conference on Image Processing (ICIP), IEEE, 2018, pp. 3478–3482.

[45] N. Zheng, J. Wen, R. Liu, L. Long, J. Dai, Z. Gong, Unsupervised representation learning with long-term dynamics for skeleton based action recognition, in: Thirty-Second AAAI Conference on Artificial Intelligence, 2018.

[46] H.-B. Zhang, Y.-X. Zhang, B. Zhong, Q. Lei, L. Yang, J.-X. Du, D.-S. Chen, A comprehensive survey of vision-based human action recognition methods, Sensors 19 (5) (2019) 1005.

[47] P. Wang, W. Li, P. Ogunbona, J. Wan, S. Escalera, Rgb-d-based human motion recognition with deep learning: A survey, Computer Vision and Image Understanding 171 (2018) 118–139.

[48] M. Müller, T. Röder, M. Clausen, Efficient content-based retrieval of motion capture data, in: ACM Transactions on Graphics (ToG), Vol. 24, ACM, 2005, pp. 677–685.

[49] A. Yao, J. Gall, G. Fanelli, L. Van Gool, Does human action recognition benefit from pose estimation?, in: BMVC 2011-Proceedings of the British Machine Vision Conference 2011, 2011.

[50] X. Li, Y. Zhang, J. Zhang, Improved key poses model for skeleton-based action recognition, in: Pacific Rim Conference on Multimedia, Springer, 2017, pp. 358–367.

[51] S. Agahian, F. Negin, C. Köse, Improving bag-of-poses with semi-temporal pose descriptors for skeleton-based action recognition, The Visual Computer 35 (4) (2019) 591–607.

[52] J. Liu, G. Wang, L.-Y. Duan, K. Abdiyeva, A. C. Kot, Skeleton-based human action recognition with global context-aware attention lstm networks, IEEE Transactions on Image Processing 27 (4) (2018) 1586–1599.

[53] M. Zanfir, M. Leordeanu, C. Sminchisescu, 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, 2013, pp. 2752–2759.

[54] E. Cippitelli, S. Gasparrini, E. Gambi, S. Spinsante, A human activity recognition system using skeleton data from rgbd sensors, Computational intelligence and neuroscience 2016 (2016) 21.

[55] A. Eweiwi, M. S. Cheema, C. Bauckhage, J. Gall, Efficient pose-based action recognition, in: Asian conference on computer vision, Springer, 2014, pp. 428–443.

[56] X. Yang, Y. Tian, Effective 3d action recognition using eigenjoints, Journal of Visual Communication and Image Representation 25 (1) (2014) 2–11.

[57] P. Wang, W. Li, C. Li, Y. Hou, Action recognition based on joint trajectory maps with convolutional neural networks, Knowledge-Based Systems 158 (2018) 43–53.

[58] K. N. Tran, I. A. Kakadiaris, S. K. Shah, Modeling motion of body parts for action recognition., in: BMVC, Vol. 11, Citeseer, 2011, pp. 1–12.

[59] S. Baysal, M. C. Kurt, P. Duygulu, Recognizing human actions using key poses, in: 2010 20th International Conference on Pattern Recognition, IEEE, 2010, pp. 1727–1730.

[60] B. Ghojogh, H. Mohammadzade, M. Mokari, Fisherposes for human action recognition using kinect sensor data, IEEE Sensors Journal 18 (4) (2017) 1612–1627.

[61] X. Li, Y. Zhang, D. Liao, Mining key skeleton poses with latent svm for action recognition, Applied Computational Intelligence and Soft Computing 2017.

[62] X. Peng, L. Wang, X. Wang, Y. Qiao, Bag of visual words and fusion methods for action recognition: Comprehensive study and good practice, Computer Vision and Image Understanding 150 (2016) 109–125.

[63] C. B. Trevarthen, Brain circuits and functions of the mind: Essays in honor of Roger W. Sperry., Cambridge University Press, 1990.

[64] L. Tao, R. Vidal, Moving poselets: A discriminative and interpretable skeletal motion representation for action recognition, in: Proceedings of the IEEE International Conference on Computer Vision Workshops, 2015, pp. 61–69.

[65] G. Zhu, L. Zhang, P. Shen, J. Song, Human action recognition using multi-layer codebooks of key poses and atomic motions, Signal Processing: Image Communication 42 (2016) 19–30.

[66] Y. Du, W. Wang, L. Wang, Hierarchical recurrent neural network for skeleton based action recognition, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2015, pp. 1110–1118.

[67] Y. Hou, Z. Li, P. Wang, W. Li, Skeleton optical spectra-based action recognition using convolutional neural networks, IEEE Transactions on Circuits and Systems for Video Technology 28 (3) (2018) 807–811.

[68] 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.

[69] 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.

[70] A. Ben Tanfous, H. Drira, B. Ben Amor, Coding kendall's shape trajectories for 3d action recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 2840–2849.

[71] A. Kacem, M. Daoudi, B. B. Amor, S. Berretti, J. C. Alvarez-Paiva, A novel geometric framework on gram matrix trajectories for human behavior understanding, IEEE transactions on pattern analysis and machine intelligence.

[72] X. Zhang, Y. Wang, M. Gou, M. Sznaier, O. Camps, 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, 2016, pp. 4498–4507.

[73] X. Pennec, P. Fillard, N. Ayache, A riemannian framework for tensor computing, International Journal of computer vision 66 (1) (2006) 41–66.

[74] V. Arsigny, P. Fillard, X. Pennec, N. Ayache, Log-euclidean metrics for fast and simple calculus on diffusion tensors, Magnetic Resonance in Medicine: An Official Journal of the International Society for Magnetic Resonance in Medicine 56 (2) (2006) 411–421.

[75] A. Cherian, S. Sra, A. Banerjee, N. Papanikolopoulos, Jensen-bregman logdet divergence with application to efficient similarity search for covariance matrices, IEEE transactions on pattern analysis and machine intelligence 35 (9) (2013) 2161–2174.

[76] M. Moakher, P. G. Batchelor, Symmetric positive-definite matrices: From geometry to applications and visualization, in: Visualization and Processing of Tensor Fields, Springer, 2006, pp. 285–298.

[77] S. Rahimi, A. Aghagolzadeh, M. Ezoji, Human action recognition based on the grassmann multi-graph embedding, Signal, Image and Video Processing 13 (2) (2019) 271–279.

[78] B. Choi, ARMA model identification, Springer Science & Business Media, 2012.

[79] 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.

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

[81] J. A. Suykens, J. Vandewalle, Least squares support vector machine classifiers, Neural processing letters 9 (3) (1999) 293–300.

[82] Y. Tang, Y. Tian, J. Lu, P. Li, J. Zhou, Deep progressive reinforcement learning for skeleton-based action recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 5323–5332.

[83] L. Chen, J. Lu, Z. Song, J. Zhou, Part-activated deep reinforcement learning for action prediction, in: Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 421–436.

[84] C. Li, Y. Hou, P. Wang, W. Li, Multiview-based 3-d action recognition using deep networks, IEEE Transactions on Human-Machine Systems 49 (1) (2019) 95–104.

[85] L. Qiao, S. Chen, X. Tan, Sparsity preserving projections with applications to face recognition, Pattern Recognition 43 (1) (2010) 331–341.

[86] A. Einstein, Zur elektrodynamik bewegter körper, Annalen der physik 322 (10) (1905) 891–921.

[87] H. Lei, B. Sun, A study on the dynamic time warping in kernel machines, in: 2007 Third International IEEE Conference on Signal-Image Technologies and Internet-Based System, IEEE, 2007, pp. 839–845.

[88] A. K. Farahat, A. Ghodsi, M. S. Kamel, Efficient greedy feature selection for unsupervised learning, Knowledge and information systems 35 (2) (2013) 285–310.

[89] M. S. Asif, J. Romberg, Sparse recovery of streaming signals using l1-homotopy, IEEE Transactions on Signal Processing 62 (16) (2014) 4209–4223.

[90] W. Zhu, C. Lan, J. Xing, W. Zeng, Y. Li, L. Shen, X. Xie, Co-occurrence feature learning for skeleton based action recognition using regularized deep lstm networks, in: Thirtieth AAAI Conference on Artificial Intelligence, 2016.

[91] M. L. Anjum, S. Rosa, B. Bona, Tracking a subset of skeleton joints: An effective approach towards complex human activity recognition, Journal of Robotics 2017.

[92] C. Wang, Y. Wang, A. L. Yuille, Mining 3d key-pose-motifs for action recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 2639–2647.

[93] 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.

[94] J. Liu, A. Shahroudy, D. Xu, A. C. Kot, G. Wang, Skeleton-based action recognition using spatio-temporal lstm network with trust gates, IEEE transactions on pattern analysis and machine intelligence 40 (12) (2018) 3007–3021.

[95] J. Liu, G. Wang, P. Hu, L.-Y. Duan, A. C. Kot, Global context-aware attention lstm networks for 3d action recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 1647–1656.

[96] Z. Huang, C. Wan, T. Probst, L. Van Gool, Deep learning on lie groups for skeleton-based action recognition, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 6099–6108.

[97] M. Li, L. Yan, Q. Wang, Group sparse regression-based learning model for real-time depth-based human action prediction, Mathematical Problems in Engineering 2018.

[98] A. Chrungoo, S. Manimaran, B. Ravindran, Activity recognition for natural human robot interaction, in: International Conference on Social Robotics, Springer, 2014, pp. 84–94.

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

[100] B. Hayes, J. A. Shah, Interpretable models for fast activity recognition and anomaly explanation during collaborative robotics tasks, in: 2017 IEEE International Conference on Robotics and Automation (ICRA), IEEE, 2017, pp. 6586–6593.

[101] Y. Yang, C. Deng, D. Tao, S. Zhang, W. Liu, X. Gao, Latent max-margin multitask learning with skelets for 3-d action recognition, IEEE transactions on cybernetics 47 (2) (2017) 439–448.

[102] Y. Yang, C. Deng, S. Gao, W. Liu, D. Tao, X. Gao, Discriminative multi-instance multitask learning for 3d action recognition, IEEE Transactions on Multimedia 19 (3) (2017) 519–529.

[103] S. Ghodsi, H. Mohammadzade, E. Korki, Simultaneous joint and object trajectory templates for human activity recognition from 3-d data, Journal of Visual Communication and Image Representation 55 (2018) 729–741.

[104] J. Shan, S. Akella, 3d human action segmentation and recognition using pose kinetic energy, in: 2014 IEEE international workshop on advanced robotics and its social impacts, IEEE, 2014, pp. 69–75.

[105] N. Dawar, N. Kehtarnavaz, Action detection and recognition in continuous action streams by deep learning-based sensing fusion, IEEE Sensors Journal 18 (23) (2018) 9660–9668.

[106] P. Wang, Z. Li, Y. Hou, W. Li, Action recognition based on joint trajectory maps using convolutional neural networks, in: Proceedings of the 24th ACM international conference on Multimedia, ACM, 2016, pp. 102–106.

[107] C. Li, Y. Hou, P. Wang, W. Li, Joint distance maps based action recognition with convolutional neural networks, IEEE Signal Processing Letters 24 (5) (2017) 624–628.

[108] K. Chen, K. D. Forbus, Action recognition from skeleton data via analogical generalization, in: Proc. 30th International Workshop on Qualitative Reasoning, 2017.

[109] J. Liu, N. Akhtar, A. Mian, Skepxels: Spatio-temporal image representation of human skeleton joints for action recognition, arXiv preprint arXiv:1711.05941.

[110] Q. Ke, M. Bennamoun, S. An, F. Sohel, F. Boussaid, A new representation of skeleton sequences for 3d action recognition, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 3288–3297.

[111] N. Srivastava, E. Mansimov, R. Salakhudinov, Unsupervised learning of video representations using lstms, in: International conference on machine learning, 2015, pp. 843–852.

[112] F. Ofli, R. Chaudhry, G. Kurillo, R. Vidal, R. Bajcsy, Sequence of the most informative joints (smij): A new representation for human skeletal action recognition, Journal of Visual Communication and Image Representation 25 (1) (2014) 24–38.

[113] A. Shahroudy, J. Liu, T.-T. Ng, G. Wang, Ntu rgb+ d: A large scale dataset for 3d human activity analysis, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 1010–1019.

[114] H.-H. Pham, L. Khoudour, A. Crouzil, P. Zegers, S. A. Velastin, Skeletal movement to color map: A novel representation for 3d action recognition with inception residual networks, in: 2018 25th IEEE International Conference on Image Processing (ICIP), IEEE, 2018, pp. 3483–3487.

[115] G. Evangelidis, G. Singh, R. Horaud, Skeletal quads: Human action recognition using joint quadruples, in: 2014 22nd International Conference on Pattern Recognition, IEEE, 2014, pp. 4513–4518.

[116] J.-F. Hu, W.-S. Zheng, J. Lai, J. Zhang, Jointly learning heterogeneous features for rgb-d activity recognition, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2015, pp. 5344–5352.

[117] 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.