Hidden States Exploration for 3D Skeleton-based Gesture Recognition

Xin Liu[1,2], Henglin Shi[1], Xiaopeng Hong[1], Haoyu Chen[1], Dacheng Tao[2] and Guoying Zhao[1,*]

[1]Center for Machine Vision and Signal Analysis, The University of Oulu, Finland
[2]UBTech Sydney AI Institute and SIT, FEIT, The University of Sydney, Australia

[*]Corresponding author: guoying.zhao@oulu.fi

Abstract

3D skeletal data has recently attracted wide attention in human behavior analysis for its robustness to variant scenes, while accurate gesture recognition is still challenging. The main reason lies in the high intra-class variance caused by temporal dynamics. A solution is resorting to the generative models, such as the hidden Markov model (HMM). However, existing methods commonly assume fixed anchors for each hidden state, which is hard to depict the explicit temporal structure of gestures. Based on the observation that a gesture is a time series with distinctly defined phases, we propose a new formulation to build temporal compositions of gestures by the low-rank matrix decomposition. The only assumption is that the gesture's "hold" phases with static poses are linearly correlated among each other. As such, a gesture sequence could be segmented into temporal states with semantically meaningful and discriminative concepts. Furthermore, different to traditional HMMs which tend to use specific distance metric for clustering and ignore the temporal contextual information when estimating the emission probability, the Long Short-Term Memory (LSTM) is utilized to learn probability distributions over states of HMM. The proposed method is validated on two challenging datasets. Experiments demonstrate that our approach can effectively work on a wide range of gestures and actions, and achieve state-of-the-art performance.

1. Introduction

Human body gesture analysis is a fundamental study which has been widely applied in a variety of artificial intelligence applications, such as human-computer interaction, intelligent security surveillance, and video games. Recently, 3D skeletal joint-based research is attracting increasing attention in the community of human behavior understanding. One reason underlying its popularity is that the 3D skeleton data can effectively represent a gesture instance as a temporal evolution of spatial joint configurations in 3D space. Another reason is that realtime 3D data collection and skeleton extraction have become much easier [31].

Over the last few years, various 3D skeleton-based models have been developed for gesture recognition, ranging from feature representations [44, 45, 35, 7, 3, 1, 48, 27, 26, 8] to various forms of parametric approaches [36, 37, 32, 28, 22, 29, 14, 15, 41, 38, 46]; and also including many deep learning methods [43, 42, 25, 10, 4, 49, 23, 17, 19, 30, 20, 13]. Despite the encouraging progress having been made by various studies, accurately recognizing human gestures is still challenging. Especially, one open issue of human gestures recognition lies in the temporal dynamics. For instance, even the same subject may have different implementation rates and starting/ending points when performing a gesture, let alone different performers. Consequently, the variability of a category of human gestures can be very large. If those temporal dynamics being ignored, the accuracy of recognition would be deteriorated undoubtedly [1].

Recently, researchers pay more attention to modeling human behaviors by studying temporal structures, *e.g.* [7, 38, 34]. However, most of their work focus on human actions rather than body gestures. Compared with actions, the structural property of gestures is more semantically meaningful and discriminative. According to the research on gesture movements [11, 12], a gesture can be decomposed into the following "gesticular phases" (see Fig. 1): (1) *Resting*: see Fig. 1 (a). (2) *Preparation*: hands move to the initial position of the stroke, see Fig. 1 (a)→(b). (3) *Pre-stroke hold*: brief pause at the end of preparation, see Fig. 1 (b). (4) *Stroke*: hands movement that expresses the meaning of the gesture, see Fig. 1 (b)→(c)→(d). (5) *Post-stroke hold*: brief pause at the end of a stroke, maintaining the hands' configuration and position, see Fig. 1 (d). (6) *Retraction*: the hands move back to a rest position to conclude a gesture unit, see Fig. 1 (d)→(e)→(f). (7) *Resting*: see Fig. 1 (f).

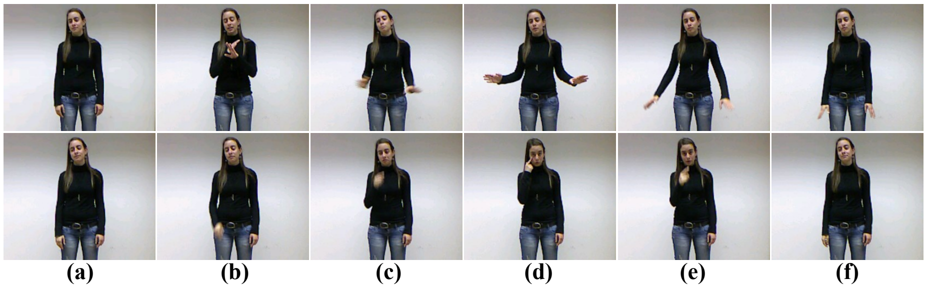

Figure 1. Frames (cropped) selected from two gestures [5] representing the meanings of "basta (enough)" and "furbo (clever)" respectively. These frames illustrate a gesture consists of a series of "gesticular phases": *Resting*→*Preparation*→*Pre-stroke hold*→*Stroke*→*Post-stroke hold* →*Retraction*→*Resting* [11, 12]. (a) *Resting*, (a)→(b) *Preparation*, (b) *Pre-stroke hold*, (b)→(c)→(d) *Stroke*, (d) *Post-stroke hold*, (d)→(e)→(f) *Retraction*, (f) *Resting*. It is noted that the gesticular phases *Preparation* and *Pre-stroke hold* are optional, and can be merged into the obligatory *Stroke* [12]. For example, the lasting time of *Pre-stroke hold* in "furbo (clever)" is too short to be determined.

It can be concluded from above definitions, three phases (2, 4, 6) with hands movement, namely *Preparation*, *Stroke*, *Retraction* are partitioned by four "hold" phases (1, 3, 5, 7) with static poses, namely *Resting* (Independent hold [12]), *Pre-stroke hold* and *Post-stroke hold*. In other words, the temporal structure of a gesture can be obtained once these "hold" phases are identified.

Based on such observation, in this paper, we develop a novel model for human gesture recognition aiming to address the difficulties of modeling temporal dynamics. We treat one human gesture as a series of separated phases, each of which is associated with a segment of arbitrary length, as illustrated in Fig. 2 (c). We propose to globally capture the temporal evolution of gestures by a generative model which is built upon a recurrent neural network to memorize contextual information for better prediction of emission probabilities. We formulate the problem in a unified framework named Hidden States Learning by Long Short-Term Memory (HSL-LSTM). The main contributions are summarized as follows:

1) We propose a new formulation to model the temporal structures based on a low-rank matrix decomposition algorithm. The only assumption is that the gesture's "hold" phases with static poses are linearly correlated with each other, which can be captured by the low rank matrix. We also explicitly consider the column-block prior of the outlier signals, the part of hand movements (phases) which cannot be fitted into the low-rank model. Thus, the temporal structure alignment is interpreted as a binary clustering problem. Different to conventional methods using fixed anchors (Fig. 2 (d)), the proposed method can segment a gesture sequence into temporal compositions (phases) with semantically meaningful and discriminative concepts (Fig. 2 (c)).

2) We propose a new hidden states learning model based on a recurrent neural network. Different temporal compositions actually correspond to the different hidden states of HMM. The usage of HMM allows to distribute heterogeneous information of one gesture class over many states (phases), and is key to improve the capability of modeling complex patterns. Different to traditional HMM using the Gaussian mixture model (GMM) [24] which ignores the temporal contextual information and uses specific distance metric for clustering, the LSTM is utilized to enhance the HMM by generating better emission probability as it provides robust classification of small temporal chunks.

3) We propose a new gesture recognition framework by absorbing the advantages of the HMM and LSTM. Rather than modelling the whole sequences (a gesture) within the LSTM as conventional RNN methods do, we feed the network by temporal compositions (hidden states) with shorter temporal length and more training samples. Therefore, the parameter learning for LSTM with large size training data is not needed. In addition, we introduce a Lie group based feature to better represent the 3D geometric relationships between various body parts. Experiments demonstrate that our approach achieves a state-of-the-art performance for 3D skeleton based human gesture recognition benchmarks.

2. Related Methods

2.1. Approaches with local temporal modeling

To account for temporal dynamics, a common treatment is the dynamic time warping (DTW), as adopted in [35, 8, 46]. DTW resorts to finding an optimal temporal alignment, then warps all sequences in the same category to a corresponding template. However, the performance of DTW heavily depends on the metric used in measuring the frame similarity. Moreover, for periodic gestures, DTW tends to produce large temporal misalignments which may harm the classification performance [36]. Wang *et al.* [36, 37] proposed the Fourier temporal pyramid (FTP) to capture local temporal patterns, which is more robust than DTW to noise and temporal misalignments. While FTP is restricted by the width of the time window and can only utilize limited contextual information [4]. In [48], Zanfir *et al.* proposed a moving pose descriptor by integrating the nor-

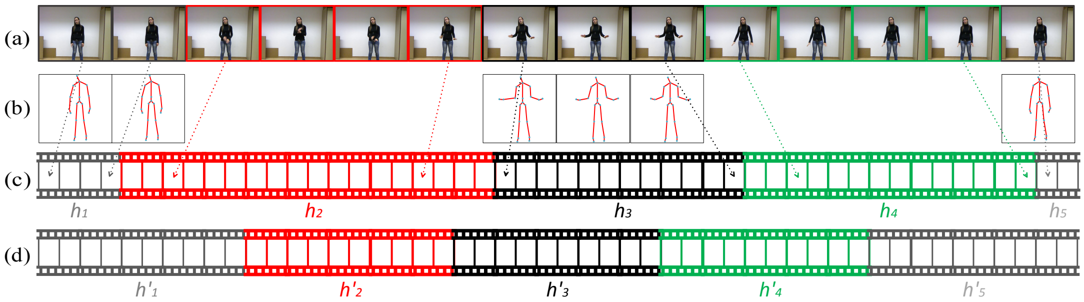

Figure 2. Illustration of phases of a gesture sequence with temporal structures. (a) Frames selected from a gesture [5] representing the meaning of "basta (enough)", (b) Skeletons (corresponding to selected frames) of static poses from "hold" phases, (c) Temporal structure (phases) segmentation by proposed, resulting hidden states h_1 (*Resting*), h_2 (*Preparation*, *Post-stroke hold*, *Stroke*), h_3 (*Post-stroke hold*), h_4 (*Retraction*), h_5 (*Resting*), (d) Fixed anchors based methods with equal-sized segmentation, resulting in hidden states $h_1^{'}$, $h_2^{'}$, $h_3^{'}$, $h_4^{'}$, $h_5^{'}$.

malized position of joints from discriminative key-frames, as well as their velocities and accelerations. Leveraging key frames can help exclude frames that are less relevant to the underlying gestures, but in comparison to the holistic based approaches, losing essential information is inevitable. In the above methods, the local temporal dynamics is generally represented within a certain time window, so they cannot globally capture the temporal evolution of gestures [4].

2.2. Approaches with generative model

Another solution to temporal dynamics is resorting to the generative models, where the time series are reorganized by a sequential prototype, and the temporal dynamics of gestures are trained as a set of transitions among these prototypes [1]. A representative is the hidden Markov model (HMM). It can globally model the temporal evolution of gestures, which is more robust than DTW, and thus it was widely utilized by [22, 44, 29, 43, 42]. It is noted that the input sequences of HMM have to be previously segmented, which itself is a challenging task. Typically, HMMs partition sequence into a fixed number of segments with equal-length for assigning the hidden states. However, they may have problems on handling complex gestures with diverse temporal durations. Another popular generative model is the conditional random field (CRF) [14, 15], while the structure of the graphs needs to be fully known, which makes this model heavily relying on the high quality annotated data. In fact, above methods tend to be around the same difficulty in determining the accurate states from observations without careful selection of the features, which undermines the performance of such generative models [36].

2.3. Approaches with recurrent neural network

With the development of deep learning technologies, plenty of recognition work addressed the problem of temporal dynamics by recurrent neural networks. Especially, the long short-term memory (LSTM) [9] carefully designs a suit of schemes to memorize (including forget, store, update, and output) contextual information observed from previous sequential inputs. Du *et al*. [4] adopted the bi-directional LSTMs for action recognition, where the entire skeleton was divided into five major groups of joints and each group was fed into a group specific LSTM sub-network, then system fused the outputs of these subnetworks hierarchically and finally fed them into another set of higher level LSTMs to capture the global body movements. Zhu *et al*. [49] added a group sparse regularization term to the cost function of LSTM, making the network to learn the co-occurrence of discriminative skeleton joints automatically. In [19], Liu *et al*. introduced the trust gate into the LSTM to learn the reliability of the inputs and accordingly adjust their confidence on updating the context information. In [17], Li *et al*. utilized a Gaussian-like curve to measure the confidences of the start and end frame of actions, and introduced a joint classification regression LSTM to solve online action detection and recognition problem. Although LSTM is powerful in modeling sequential data, it still suffer from remembering the information of the entire sequence with many time steps (states) [10, 39]. Moreover, compared with the progress in data augmentation for RGB images, research efforts on augmenting 3D skeleton data augmentation are still at a rather early stage. As such, it is still challenging to train the LSTM on limited amount of training data [36, 30]. In [13], Koller *et al*. embed a HMM into a deep CNN-BLSTM network for sign language recognition which is a problem closely related to temporal gesture segmentation. They first train a CNN using weak frame level annotations, then use a LSTM to output the Bayesian posteriors for HMM training and make use of the hidden states of each frame predicted by HMM for CNN fine-tuning. This model is based on the hypothesis that the certain boundaries can be determined by some rules in continuous sequences. Obviously, the output of temporal segment is at the "words" level but not the phase level with semantically meaningful and discriminative concepts. For example, this temporal boundary based segmentation may run into a stone wall when a gesture is composed by many different poses with temporal boundaries and a subject performs this gesture cyclically with different rates and orders. Besides, the number of hidden states is very hard to be determined by "words" based model. The method [13] use six hidden states empirically without clearly defined meanings.

3. HMM for Gesture Modeling

In this paper, the gesture modeling problem in HMM is formulated by following definitions, given a set $\Theta = \{\theta_1, \theta_2, \cdots, \theta_{K-1}, \theta_K\}$ which contains K gesture sequences with arbitrary lengths. Any gesture sequence θ_k can be denoted as $\theta_k = \{f_{k,1}, f_{k,2}, \cdots, f_{k,T_k-1}, f_{k,T_k}\}$, where $f_{k,t}$ is the t^{th} frame (or the representation of a frame) of θ_k and T_k denotes its length. For any θ_k from Θ, its label δ_c satisfies $\delta_c \in \Delta$, where Δ is the set of C gesture labels which is denoted as $\Delta = \{\delta_1, \delta_2, \cdots, \delta_{C-1}, \delta_C\}$.

Specifically, given an observation of gesture sequences as $X = \{x_1, x_2, \cdots x_{T-1}, x_T\}$, where $X \in \Theta$, we utilize the HMM to infer a hidden state sequence $H = \{h_1, h_2, \cdots h_{T-1}, h_T\}$. Any state h_t from H fulfills $h_t \in \Psi$ ($1 \leq t \leq T$), where Ψ denotes an universal set which contains all possible Markov hidden states.

Typically, the states alignment is conducted based on a hypothesis that gestures are completed by uniformly performing Z defined hidden states in order, and hidden states from different gesture classes do not overlap. Given a gesture with class δ_c, the corresponding hidden state sequence is defined as $\{\psi_{c,1}, \psi_{c,2}, \cdots, \psi_{c,Z-1}, \psi_{c,Z}\}$, then we can generalize this concept for all gesture classes, and define the universal set of hidden states for all gesture classes:

$$\Psi = \left\{ \begin{array}{ccccc} \psi_{1,1} & \psi_{1,2} & \cdots & \psi_{1,Z-1} & \psi_{1,Z} \\ \psi_{2,1} & \psi_{2,2} & \cdots & \psi_{2,Z-1} & \psi_{2,Z} \\ \cdots & \cdots & \psi_{c,z} & \cdots & \cdots \\ \psi_{C-1,1} & \psi_{C-1,2} & \cdots & \psi_{C-1,Z-1} & \psi_{C-1,Z} \\ \psi_{C,1} & \psi_{C,2} & \cdots & \psi_{C,Z-1} & \psi_{C,Z} \end{array} \right\}$$

where $\psi_{c,z}$ denotes the z^{th} hidden state of gesture class δ_c, $1 \leq \delta_c \leq C$ and $1 \leq z \leq Z$.

Thus, according to the HMM full probability model:

$$P(H, X) = P(h_1)P(x_1|h_1) \prod_{t=2}^{T} P(h_t|h_{t-1})P(x_t|h_t) \tag{1}$$

the goal of the gesture modeling problem is to find an optimal hidden state sequence $\hat{H}$ which maximize the joint probability $P(\hat{H}, X)$, given a set of observations X, the optimization problem of solving $\hat{H}$ can be given as:

$$\hat{H} = \arg\max_{H} P(H|X) \underset{X}{\propto} \arg\max_{H} P(H, X) \tag{2}$$

It can be concluded that HMM-based gesture recognition methods have two key issues required to be carefully solved:

* Given the observation of a gesture sequence, how to select a corresponding hidden states sequence that is optimal in some meaningful sense to best explain the observation?

* Three sets of parameters need to be estimated to complete the specification of a HMM, namely the prior distribution of the state at the first frame $P(h_1)$, the hidden state transition probability $P(h_t|h_{t-1})$, and the emission probability $P(x_t|h_t)$ for generating an observation at time t when given the hidden state h_t. How to efficiently compute these parameters (distributions)?

For the first problem, Wu *et al.* [43, 42] adopted a deep belief network (DBN) to estimate the emission probability, while a forced alignment scheme is used to divide video sequences temporally equal to obtain hidden states for supervised training. In [44], the posture words were learned by the linear discriminant analysis, and each gesture is modeled as a time series of these words (hidden states). However, the success of clustering heavily relys on specific distance metric. Moreover, this one frame one pose tactic cannot fully characterize the motion temporality since it ignores the contextual information, which is hard to handle high intra-class variance of human gestures. In fact, for the task of gesture recognition, the gestures themselves exhibit internal temporal structure. As defined in Section 1, gestures typically have definite gesticular phases with varied durations and starting/ending times. For illustration, two examples are given in Fig. 1 and 2. Based on this observation, in this paper, gestures are modeled as compositions of different gesticular phases, when gestures have distinct phases, models that exploit hidden states are advantageous. As such, different gesticular phases actually correspond to the different hidden states of HMM, the usage of HMM allows to distribute heterogeneous information of one gesture class over many states (phases). Therefore, the mission of uncovering the hidden states thus can be transformed to identifying the starting and ending frames of each phases.

For the second issue, the Gaussian mixture model (GMM) [22, 29, 24] has been widely utilized as the dominant technique for estimating the emission distribution of HMM. In [43, 42], a DBN is used as a generative model to replace the traditional GMM for estimating the emission probability. However, there still exists a conflict that any frame within a sequence usually has contextual information and correlated with previous frames, however, which is ignored by previous works. Both of the DBN and GMM treat input frames from different time steps as independent variables so that output emission probability in the current time step only relies on the current input. To solve this issue and acquire the emission probability more appropriately, the LSTM [9] is utilized because of its stronger contextual information modeling ability, which uses memory cells to store information learned from previous sequential inputs and stored information can affect the output of the network.

4. Low-rank Decomposition for Exploring Gesture Temporal Structures

One important issue is the choice of features to capture the variability of 3D skeletons, within and across gesture classes. Inspired by [35], the Lie group-based representation is introduced. Instead of using the absolute coordinate, we utilize the relative geometry between different body parts to characterize the body movement. More specifically, the human skeleton can be modeled by an articulated system of rigid segments connected by joints. Mathematically, any rigid body displacement can be implemented by a rotation about an axis combined with a translation parallel to that axis. This 3D rigid body displacement forms a $SE(3)$, the special Euclidean group in three dimensions. Thus, given a pair of bones (body parts), their relative geometry can be represented in a local coordinate system attached to the other, which can be formulated by the $SE(3)$. As a result, an entire human skeleton can be represented by the relative geometry between all pairs of bones, as a point on the product space of $SE(3) \times \cdots \times SE(3)$, which is a Lie group. This relative geometry has a natural stability and consistency. For example, if a pair of bones undergoes the same rotation, their relative geometry matrix would not be altered. Also, this feature has the property of view-invariance such that can guarantee the uniqueness of motion. Finally, a human skeleton can be characterized by a Lie algebra (mapped from Lie group) feature vector with dimension of G.

Given an observed sequences (T frames), for a subject, we can construct a matrix D by stacking (Lie group based) skeletal representations of every frame horizontally (column wise), then $D \in \mathbb{R}^{G \times T}$. Since the gesture's "hold" phases are with static poses, we can assume that these static poses (in the form of Lie group-based features) are linearly correlated with each other, as the Lie group-based representation reflects the relative geometry of body parts which are independent of the subject's position. In other words, these "hold" phases should be captured by a low-rank matrix, and the hand movements (phases) mean gesture changes which cannot be fitted into the low-rank model of static poses, and thus should be treated as outliers. Based on this observation, we consider the hidden states exploration from the viewpoint of matrix decomposition problem, which can be expressed as follows:

$$D = L + S \tag{3}$$

where L and S denote the "hold" states (phases) and hand movements signals respectively. We assume that the static poses of "hold" states forming a low-rank matrix L. And component S should be a column-block sparse matrix with non-zero columns corresponding to the outliers. In order to eliminate ambiguity, the columns of the low-rank matrix L corresponding to the outlier columns are assumed to be zeros. To formalize column-block priors, we introduce the $\ell_{2,1}$-norm and then propose a Low-rank and Column-Block sparsity matrix Decomposition (LCBD) method, as

$$\min_{L,S} \|L\|_* + \kappa\lambda\|S\|_{2,1} + \kappa\left(1-\lambda\right)\|L\|_{2,1} \quad s.t.\ D = L + S \tag{4}$$

where $\|L\|_*$ means the nuclear norm of matrix L, the sum of its singular values, and $\|S\|_{2,1}$ means ℓ_1-norm of the vector formed by taking the ℓ_2-norms of the columns of matrix S

$$\|S\|_{2,1} = \sum_{i=1}^{T} \|S_i\|_2 \tag{5}$$

where S_i denotes the i^{th} column of S.

Inspired by methods [33, 18, 47], the extra introduced term $\kappa\left(1-\lambda\right)\|L\|_{2,1}$ ensures that recovered matrix L has exact zero columns correspond to the non-zero ones of S. The Eq. (4) is an optimization problem and we could solve it based on the augmented Lagrange multiplier (ALM) method [18] [40] [21], which can be defined as

$$\begin{aligned}\mathcal{L}\left(L,S,Y;\mu\right) = \|L\|_* + \kappa\lambda\|S\|_{2,1} + \kappa\left(1-\lambda\right)\|L\|_{2,1} + \\ \langle Y, D-L-S\rangle + \frac{\mu}{2}\|D-L-S\|_F^2\end{aligned} \tag{6}$$

where Y is a vector of Lagrange multipliers, μ is a positive scalar. ALM solves (6) by alternating between optimizing the primal variables L, S and updating the dual variable Y, which solves the following three sub-problems

$$\begin{cases} L_{k+1} = \arg\min_L \mathcal{L}_1\left(L, S_k, Y_k; \mu\right) \\ S_{k+1} = \arg\min_S \mathcal{L}_1\left(L_{k+1}, S, Y_k; \mu\right) \\ Y_{k+1} = Y_k + \mu\left(D - L_{k+1} - S_{k+1}\right) \end{cases} \tag{7}$$

The first problem in (7) which solves for L at fixed S, Y can be explicitly expressed as the following form

$$\min_L \|L\|_* + \kappa\left(1-\lambda\right)\|L\|_{2,1} + \frac{\mu}{2}\left\|\left(D - S_k + \mu^{-1}Y_k\right) - L\right\|_F^2 \tag{8}$$

In each iteration, the Eq. (8) can be rewritten as

$$L_{k+1} = \arg\min_L\left\{\|L\|_* + \kappa\left(1-\lambda\right)\|L\|_{2,1} + \frac{\mu_k}{2}\left\|G^L - L\right\|_F^2\right\} \tag{9}$$

where $G^L = D - S_k + \mu^{-1}Y_k$. We use the Douglas/Peaceman Rachford (DR) monotone operator splitting method [2] [6] to iteratively solve (9).

Define $f_1\left(L\right) = \kappa\left(1-\lambda\right)\|L\|_{2,1} + \frac{\mu_k}{2}\left\|G^L - L\right\|_F^2$ and $f_2\left(L\right) = \|L\|_*$. For $\beta > 0$ and a sequence $\alpha_j \in (0, 2)$, the DR iteration for (9) is expressed as

$$\begin{aligned} &L^{(j+1/2)} = prox_{\beta f_2}\left(L^{(j)}\right), \\ &L^{(j+1)} = L^{(j)} + \alpha_j\left(prox_{\beta f_1}\left(2L^{(j+1/2)} - L^{(j)}\right) - L^{(j+1/2)}\right) \end{aligned} \tag{10}$$

where the two proximity operators involved in DR iteration are defined as

$$\begin{array}{l} prox_{\beta f_1}\left(L\right) = \tau_{\frac{\beta\kappa(1-\lambda)}{1+\beta\mu_k}}\left(\frac{L+\beta\mu_k G^L}{1+\beta\mu_k}\right) \\ prox_{\beta f_2}\left(L\right) = US_\beta\left(\sum\right)V^T \\ \tau_\eta\left(G_p\right) = G_p \max\left(0, 1-\frac{\eta}{\|G_p\|_2}\right), p=1,2,...,n. \\ S_\beta\left(x\right) = \max\left(0, x-\beta\right), x \geq 0, \beta > 0 \end{array} \tag{11}$$

With the same idea of developing (8), the second problem in (7) can be shown as the following formula:

$$\min_S \frac{\mu}{2}\left\|\left(D - L_{k+1} + \mu^{-1}Y_k\right) - S\right\|_F^2 + \kappa\lambda\|S\|_{2,1} \tag{12}$$

Similar, note $G^S = D - L_k + \mu^{-1}Y_k$. Then,

$$S = \tau_{\frac{\kappa\lambda}{\mu_k}}\left(G^S\right) \tag{13}$$

In the processing of iteration, the error in outer loop is computed as $\|D - L_k - S_k\|_F / \|D\|_F$. The outer loop stops when it reaches the value lower than 10^{-7} or the maximal iteration number 500 is reached. The error in the inner loop stops when the difference between successive matrices L_k^j equals to 10^{-6} or a maximal iteration equals to 20. The tuning parameters κ and λ are set to 0.041 and 0.73, respectively. For the DR iteration, α and β are set to 1 and 0.57, respectively. Please refer to [18, 33, 2, 6] for more details.

5. Hidden States Learning via LSTM

In previous HMM based methods [43][42], they assumed fixed anchors for each hidden state, and divide a gesture sequence θ_k into equal-length segments, then assign frames located in a segment with the same hidden state as the frame label. While the HMM imposes a geometric distribution on the time within a state, under that scheme of fixed anchors, HMM suffers because a state may transit to itself (two contiguously segmented states but with same state in real).

Different to previous approaches, our method is completely model-based to learn all HMM parameters for transition and duration distributions adaptively. More specifically, we initialize the hidden states of the temporal segments for each of training samples, according to the most discriminative portions (phases) of sequences as presented in Section 4. Based on these hidden states we can calculate three sets of HMM parameters with more meaningful sense than previous.

For representing the probability of the first hidden state prior, we use $\pi = (\pi_i)_{E\times 1}$, where $\pi_i = P(h_1 = \psi_i)$, and ψ_i is the i^{th} state of hidden states set Ψ. Then we can estimate π_i by calculating

$$\pi_i = \frac{\sum_{k=1}^{K}(h_{k,1} == \psi_i)}{K} \tag{14}$$

where k denotes the index of an observation, and K is the total number of observations (gesture sequences).

Next, the hidden states transition parameter (matrix) is denoted using $A = [a_{i,j}]_{E\times E}$, where $a_{i,j} = P(h_t = \psi_j | h_{t-1} = \psi_i)$. We can calculate $a_{i,j}$ by

$$a_{i,j} = \frac{\sum_{k=1}^{K}\sum_{t=2}^{T_k}((h_{k,t-1} == \psi_i)\mathbf{AND}(h_{k,t} == \psi_j))}{\sum_{k=1}^{K}\sum_{t=2}^{T_k}(h_{k,t-1} == \psi_i)} \tag{15}$$

Another important parameter is the emission probability. Compared with DBN and GMM which are widely used in pervious methods, LSTM can learn the contextual information from sequential data, which is powerful for sequential data modeling. On one hand, it receives the output from the previous time step and use it as a part of the input of current time step. On the other hand, it uses memory cells to store contextual information learned from the input and uses gate units to maintain the stored contextual information. In order to let LSTM generate outputs in the form which is closer to the emission probability $P(x_t|h_t)$, we use a softmax loss function to train the network. It can instruct the LSTM network to generate a posterior distribution $P(h_t|x_t, \zeta)$, where ζ is the network parameter which is shared in all time steps. Thus, we can use such network outputs to infer the emission probability according to

$$P(x_t|h_t) = \frac{P(h_t|x_t)P(x_t)}{P(h_t)} \underset{\zeta, x_t}{\propto} \frac{P(h_t|x_t)}{P(h_t)} \tag{16}$$

Lastly, by combining (1), (2), and (16), we can get our final objective function as follows

$$\hat{H} = \arg\max_H P(h_1|x_1)\prod_{t=2}^{T} P(h_t|h_{t-1})\frac{P(h_t|x_t)}{P(h_t)} \tag{17}$$

where $\hat{H}$ denotes the optimal hidden state sequence.

As we know, most of the gesture (action) recognition algorithms feed a whole gesture instance into the LSTM network (frames with the same labeling). In contrast, we feed the LSTM with a hidden state (segments) of the gesture instance. More specifically, in the training pipeline (see Fig. 3 (a)) of the proposed model, the input of LSTM is the Lie group based representation f_i of frame x_i (an 3D skeleton), and its label is a hidden state $\psi_{c,z}$ which is obtained by the proposed LCBD method, where the subscript c is the gesture category and z is the hidden states index. The purpose of training is to force the LSTM to generate the posterior probabilities for modelling the HMM emission probabilities. In the testing pipeline (see Fig. 3 (b)), a first order hidden Markov is adopted, and the Viterbi [24] algorithm (optimization problem of (17)) can be utilized to find the most likely path.

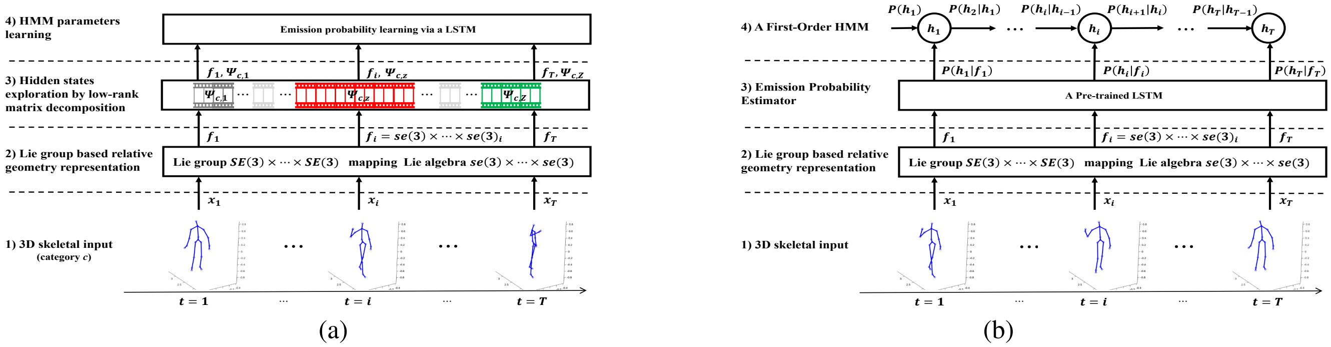
4) HMM parameters learning
Emission probability learning via a LSTM
$f_1, \Psi_{c,1}$
$f_i, \Psi_{c,z}$
$f_T, \Psi_{c,Z}$
3) Hidden states exploration by low-rank matrix decomposition
$\Psi_{c,1}$
$\Psi_{c,z}$
$\Psi_{c,Z}$
f_1
$f_i = se(3) \times \cdots \times se(3)_i$
f_T
2) Lie group based relative geometry representation
Lie group $SE(3) \times \cdots \times SE(3)$ mapping Lie algebra $se(3) \times \cdots \times se(3)$
x_1
x_i
x_T
1) 3D skeletal input (category c)
$t = 1$
$t = i$
$t = T$
(a)
4) A First-Order HMM
$P(h_1)$
$P(h_2|h_1)$
$P(h_i|h_{i-1})$
$P(h_{i+1}|h_i)$
$P(h_T|h_{T-1})$
h_1
h_i
h_T
$P(h_1|f_1)$
$P(h_i|f_i)$
$P(h_T|f_T)$
3) Emission Probability Estimator
A Pre-trained LSTM
f_1
$f_i = se(3) \times \cdots \times se(3)_i$
f_T
2) Lie group based relative geometry representation
Lie group $SE(3) \times \cdots \times SE(3)$ mapping Lie algebra $se(3) \times \cdots \times se(3)$
x_1
x_i
x_T
1) 3D skeletal input
$t = 1$
$t = i$
$t = T$
(b)

Figure 3. Illustration of the pipelines of the proposed method. (a) training pipeline, (b) testing pipeline. Please note the purpose we use f_i rather than x_i in $P(h_i|f_i)$ is to emphasize the Lie group based representation.

6. Experiments

In this section, two benchmarks, ChaLearn 2014 gesture [5] and MSR Action3D [16] are utilized to evaluate the proposed approach. The proposed method is compared with eighteen state-of-the-arts. We simply divided them into three groups. The first group's methods are most related to us, including four HMM related methods, namely HMM with GMM (HMM-GMM) [24], HMM with AdaBoost (HMM-AdaBoost) [22], HMM with DBN (HMM-DBN) [43] and its extension (HMM-DBN-ext) [42]. The methods in second group are based on classic feature representations, including histogram of 3D joints (HOJ3D) [44], EigenJoints [45], actionlet ensemble (Actionlet) [36, 37], histogram of oriented 4D normals (HON4D) [27], discriminative key-frames (Key-frames) [48], Lie group [35], Riemannian manifold (Manifold) [3], rotation and relative velocity with DTW (RVV+DTW) [8], latent max-margin multitask learning (LM^{3}TL) [46], spatio-temporal naive-bayes nearest-neighbor (ST-NBNN) [38]. The last group including four deep networks, namely the convolutional neural network based ModDrop (CNN) [25], LSTM [9], hierarchical recurrent neural network (HBRNN) [4], spatio-temporal LSTM with trust gates (ST-LSTM-TG) [19]. The baseline results are reported from original papers. Note that some of the compared methods were developed for multi-modal datasets such as the HMM-DBN-ext [42] utilized RGB and skeleton, while the proposed method is only based on 3D skeleton data. All tests are performed on an Intel Xeon CPU E5-2650 with a NVIDIA Tesla K80 GPU.

In the proposed method, the emission probability estimator is defined as a recurrent neural network with 4 layers which are connected in the following order: one LSTM layer with 512 units, a fully connected layer with 256 neurons, a dropout layer with the dropout ratio of 50%, and a softmax loss layer to force the network to generate the likelihood $P(h_t|x_t, \zeta)$. When training the network, we set the batch size to 400. The learning rate is fixed to 0.01 for the ChaLearn 2014 gesture dataset and 0.002 for MSR Action3D dataset. The network is trained till the validation accuracy and the loss is stable after a number epochs of iterations depending on the size of training data. We set 70 as the max training epoch for the ChaLearn 2014 gesture dataset due to its large training data size. For the MSR Action3D dataset, the setting is 200. An important parameter is the number of hidden states. In our experiments, we found that almost none of gesture instances have the *Pre-stroke hold* phase (or the lasting time of that static pose is too short to be determined). In most cases, there is a large confusion among *Preparation*, *Pre-stroke hold*, and *Stroke*. In fact, according to the study on movement in gestures, some gesticular phases such as the *Preparation* and *Pre-stroke hold* are optional, and can be merged into the obligatory *Stroke* [12]. Based on this observation, we choose 5 rather than 7 gesticular phases for dividing hidden states (see Fig. 2).

6.1. ChaLearn 2014 gesture Dataset

The ChaLearn 2014 is a gesture dataset of Looking At People (LAP) challenge [5] with multi-modality, including data of RGB frames, depth maps, user body masks, and 3D skeletal joint positions. This dataset collects 940 videos and each one contains 10 to 20 Italian cultural gesture instances. In total, there are 13585 gesture instances from 20 classes. We use the protocol provided by the dataset which assigns fixed 7754 gesture sequences for training, 3362 sequences for validating, and 2742 sequences for testing. It is noted that the Jaccard index score recommended by the publisher of Chalearn 2014 dataset is a frame-level metric. However, the proposed is a sequential based model. In the comparison shown in Table 1, all the results reported are in accuracy, making the comparison fair. To verify the effectiveness of the hidden states exploration, we compared the proposed method with three HMM-based state-of-the-arts, it can be seen that the recognition accuracies of HMM with GMM [24] and with DBN (HMM-DBN) [43] are only 49.1% and 83.6%, this is due to both of the DBN and GMM treat input frames at each time step as independent variables, the contextual information is ignored when learning the emission probability. The HMM-DBN-ext [42] can reach up to 86.4%, while it used both skeleton, RGB, and depth information. It also can be observed that the accuracy of the LSTM [9] is 11 percents less than the proposed method. As discussed in the introduction, LSTM is designed to explore the long-term temporal dependency, but it is still challenging for LSTM to memorize the information of the entire sequence with many states [10, 39]. Moreover, with limited amount of training data, training a LSTM is prone to overfitting [37, 30]. In the proposed method, the shorter gesture segments (states) are fed into the network to bypass the difficulty of LSTM when modeling multi-states gestures with temporal dynamics. Furthermore, this states-based feeding enlarges the number of training samples but without any data augmentation operations. Take the Chalern 2014 dataset for example, we obtained 38770 gesture (hidden states) segments for training, which is five times more training samples than LSTM with raw (7754) gesture sequences. The method [35] utilized the same Lie group to represent the 3D skeletons as ours, and it employed the DTW to deal with the temporal dynamics issue. However, DTW cannot globally capture the temporal evolution of whole sequences, so its performance is inferior to the proposed. It is notable that the ModDrop [25] was the winner of the 2014 LAP Challenge (track 3). The proposed method can achieve the similar performance to ModDrop but without using the RGB-D data.

Table 1. Comparison Of Recognition Accuracy (%) With Skeletal-Based Methods on Datasets ChaLearn 2014 (ChaL) [5] and MSR Action3D (MSR) [16] (best: bold, second best: underline).

Methods	Accuracy	
	ChaL	MSR
HMM-GMM [24]	49.1	81.5
HMM-AdaBoost[22]	-	63.0
HMM-DBN [43]	83.6	82.0
HMM-DBN-ext [42]*	86.4	-
EigenJoints [45]	59.3	82.3
Actionlet [36][37]*	-	88.2
HOJ3D [44]	-	78.9
HON4D [27]*	-	88.9
Key-frames [48]	-	91.7
Lie group [35]	79.2	92.5
Manifold [3]	-	92.1
RVV+DTW [8]	-	93.4
LM^{3}TL [46]	-	95.6
ST-NBNN [38]	-	94.8
ModDrop (CNN) [25]*	93.1	-
LSTM [9]	82.0	88.9
HBRNN [4]	-	94.5
ST-LSTM-TG [19]	-	94.8
Proposed	**93.8**	**96.3**

* The methods use skeleton and RGB-D data.

6.2. MSR Action3D Dataset

The MSR Action3D [16] is a commonly used actions recognition dataset, especially for evaluating the effectiveness of temporal dynamics modeling techniques, since this dataset is challenging where actions are highly similar to each other and have typical large temporal misalignments. MSR Action3D dataset comprises of 567 pre-segmented action instances. There are 10 subjects performing 20 classes of actions. This dataset is so popular that many researchers have reported their results on it. For a fair comparison, the same evaluation protocol, namely the cross-subject test as described in [16] is followed, where half of the subjects are used for training (subjects number 1, 3, 5, 7, 9) and the remainder for testing (2, 4, 6, 8, 10). The recognition accuracies are recorded in Table 1. It can be seen the proposed method achieves better performance than DTW-based recognition approaches, such as Lie group [35] and RVV+DTW [8]. In [48], the authors emphasized the importance of discriminative key-frames for action recognition. However, the key frames selection itself is a difficult task, which usually suffers from an issue of information losing. The HMM-DBN [43] employed a deep neural network to learn the parameters of HMM, while it utilized the fixed anchors for obtaining the hidden states. On the contrary, we formulate a model over the temporal domain that is able to capture the static poses between sub-gestures, thus, a gesture sequences could be segmented into temporal compositions (states) with semantically meaningful and discriminative concepts. Compared with HMM-DBN, the experimental results on MSR Action3D dataset verifies the effectiveness of the proposed method again. As can be seen, compared with all of the 16 methods (including some most recent methods, such as RVV+DTW [8], LM^{3}TL [46], ST-LSTM-TG [19], and ST-NBNN [38]), our model achieves the highest recognition accuracy.

7. Conclusion

In the study of human movement, a gesture could be explained as a sequence of separated sub-gestures or phases. Based on this observation, this paper focuses on studying HMM-based approaches to explore more appropriate hidden states alignment. Possible directions for future work include studying the embedding problem of the Lie group. Typically, the embedding is obtained by flattening the manifold via tangent spaces, such as the Lie algebra. However, in that way, only distances between points to the tangent pole are equal to true geodesic distances, which may lead to an inaccurate modeling issue. So, a novel embedding method will be explored to keep the distances estimation being performed in the framework of Riemannian geometry.

Acknowledgment. This work is supported by Academy of Finland, and in part by Endeavour Research Fellowship.

References

[1] B. B. Amor, J. Su, and A. Srivastava. Action recognition using rate-invariant analysis of skeletal shape trajectories. *IEEE Trans. Pattern Anal. Mach. Intell.*, 38(1):1–13, 2016.

[2] P. L. Combettes and J.-C. Pesquet. A Douglas–Rachford splitting approach to nonsmooth convex variational signal recovery. *IEEE J. Sel. Topics Signal Process.*, 1(4):564–574, 2007.

[3] M. Devanne, H. Wannous, S. Berretti, P. Pala, M. Daoudi, and A. Del Bimbo. 3D human action recognition by shape analysis of motion trajectories on Riemannian manifold. *IEEE Trans. Cybern.*, 45(7):1340–1352, 2015.

[4] Y. Du, W. Wang, and L. Wang. Hierarchical recurrent neural network for skeleton based action recognition. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pages 1110–1118. IEEE, 2015.

[5] S. Escalera, X. Baró, J. Gonzalez, M. A. Bautista, M. Madadi, M. Reyes, V. Ponce-López, H. J. Escalante, J. Shotton, and I. Guyon. Chalearn looking at people challenge 2014: Dataset and results. In *Proc. Eur. Conf. Comput. Vis. Workshops*, pages 459–473. Springer, 2014.

[6] M.-J. Fadili and J.-L. Starck. Monotone operator splitting for optimization problems in sparse recovery. In *Proc. IEEE Int. Conf. Image Process.*, pages 1461–1464. IEEE, 2009.

[7] D. Gong, G. Medioni, and X. Zhao. Structured time series analysis for human action segmentation and recognition. *IEEE Trans. Pattern Anal. Mach. Intell.*, 36(7):1414–1427, 2014.

[8] Y. Guo, Y. Li, and Z. Shao. RRV: A spatiotemporal descriptor for rigid body motion recognition. *IEEE Trans. Cybern.*, 2017.

[9] S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural Comput.*, 9(8):1735–1780, 1997.

[10] Q. Ke, M. Bennamoun, S. An, F. Sohel, and F. Boussaid. A new representation of skeleton sequences for 3D action recognition. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017.

[11] A. Kendon. Gesticulation and speech: Two aspects of the process of utterance. *The relationship of verbal and nonverbal communication*, 25(1980):207–227, 1980.

[12] S. Kita, I. Van Gijn, and H. Van der Hulst. Movement phases in signs and co-speech gestures, and their transcription by human coders. In *International Gesture Workshop*, pages 23–35. Springer, 1997.

[13] O. Koller, S. Zargaran, and H. Ney. Re-sign: Re-aligned end-to-end sequence modelling with deep recurrent CNN-HMMs. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017.

[14] H. Koppula and A. Saxena. Learning spatio-temporal structure from RGB-D videos for human activity detection and anticipation. In *Proc. Int. Conf. Mach. Learn.*, pages 792–800, 2013.

[15] H. Koppula and A. Saxena. Anticipating human activities using object affordances for reactive robotic response. *IEEE Trans. Pattern Anal. Mach. Intell.*, 38(1):14–29, 2016.

[16] W. Li, Z. Zhang, and Z. Liu. Action recognition based on a bag of 3D points. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, pages 9–14. IEEE, 2010.

[17] Y. Li, C. Lan, J. Xing, W. Zeng, C. Yuan, and J. Liu. Online human action detection using joint classification-regression recurrent neural networks. In *Proc. Eur. Conf. Comput. Vis.*, pages 203–220. Springer, 2016.

[18] Z. Lin, M. Chen, and Y. Ma. The augmented lagrange multiplier method for exact recovery of corrupted low-rank matrices. *arXiv preprint arXiv:1009.5055*, 2010.

[19] J. Liu, A. Shahroudy, D. Xu, and G. Wang. Spatio-temporal LSTM with trust gates for 3D human action recognition. In *Proc. Eur. Conf. Comput. Vis.*, pages 816–833. Springer, 2016.

[20] J. Liu, G. Wang, P. Hu, L.-Y. Duan, and A. C. Kot. Global context-aware attention LSTM networks for 3D action recognition. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pages 1647–1656, 2017.

[21] X. Liu, G. Zhao, J. Yao, and C. Qi. Background subtraction based on low-rank and structured sparse decomposition. *IEEE Trans. Image Process.*, 24(8):2502–2514, 2015.

[22] F. Lv and R. Nevatia. Recognition and segmentation of 3D human action using HMM and multi-class adaboost. *Proc. Eur. Conf. Comput. Vis.*, pages 359–372, 2006.

[23] B. Mahasseni and S. Todorovic. Regularizing long short term memory with 3D human-skeleton sequences for action recognition. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pages 3054–3062. IEEE, June 2016.

[24] K. P. Murphy. *Machine Learning: A Probabilistic Perspective*. MIT Press, 2012.

[25] N. Neverova, C. Wolf, G. Taylor, and F. Nebout. Moddrop: adaptive multi-modal gesture recognition. *IEEE Trans. Pattern Anal. Mach. Intell.*, 38(8):1692–1706, 2016.

[26] F. Ofli, R. Chaudhry, G. Kurillo, R. Vidal, and R. Bajcsy. Sequence of the most informative joints (SMIJ): A new representation for human skeletal action recognition. *J Vis. Commun. Image Represent.*, 25(1):24–38, 2014.

[27] O. Oreifej and Z. Liu. HON4D: Histogram of oriented 4D normals for activity recognition from depth sequences. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pages 716–723, 2013.

[28] B. Packer, K. Saenko, and D. Koller. A combined pose, object, and feature model for action understanding. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pages 1378–1385, 2012.

[29] L. Piyathilaka and S. Kodagoda. Gaussian mixture based HMM for human daily activity recognition using 3D skeleton features. In *Proc. IEEE Conf. Ind. Electron. Appl.*, pages 567–572. IEEE, 2013.

[30] A. Shahroudy, J. Liu, T.-T. Ng, and G. Wang. NTU RGB+D: A large scale dataset for 3D human activity analysis. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pages 1010–1019, 2016.

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

[32] J. Sung, C. Ponce, B. Selman, and A. Saxena. Unstructured human activity detection from RBGD images. In *Proc. IEEE Conf. Robot. Autom.*, pages 842–849. IEEE, 2012.

[33] G. Tang and A. Nehorai. Robust principal component analysis based on low-rank and block-sparse matrix decomposition. In *Proc. Conf. Infor, Sci. Syst.*, pages 1–5. IEEE, 2011.

[34] K. Tang, L. Fei-Fei, and D. Koller. Learning latent temporal structure for complex event detection. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pages 1250–1257. IEEE, 2012.

[35] R. Vemulapalli, F. Arrate, and R. Chellappa. Human action recognition by representing 3D skeletons as points in a Lie group. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pages 588–595. IEEE, 2014.

[36] J. Wang, Z. Liu, Y. Wu, and J. Yuan. Mining actionlet ensemble for action recognition with depth cameras. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pages 1290–1297. IEEE, 2012.

[37] J. Wang, Z. Liu, Y. Wu, and J. Yuan. Learning actionlet ensemble for 3D human action recognition. *IEEE Trans. Pattern Anal. Mach. Intell.*, 36(5):914–927, 2014.

[38] J. Weng, C. Weng, and J. Yuan. Spatio-temporal naive-bayes nearest-neighbor (ST-NBNN) for skeleton-based action recognition. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017.

[39] J. Weston, S. Chopra, and A. Bordes. Memory networks. *arXiv preprint arXiv:1410.3916*, 2014.

[40] J. Wright, A. Ganesh, S. Rao, Y. Peng, and Y. Ma. Robust principal component analysis: Exact recovery of corrupted low-rank matrices via convex optimization. In *Proc. Adv. Neural Inf. Process. Syst.*, pages 2080–2088, 2009.

[41] C. Wu, J. Zhang, S. Savarese, and A. Saxena. Watch-n-patch: Unsupervised understanding of actions and relations. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pages 4362–4370, 2015.

[42] D. Wu, L. Pigou, P. J. Kindermans, N. Le, L. Shao, J. Dambre, and J. M. Odobez. Deep dynamic neural networks for multimodal gesture segmentation and recognition. *IEEE Trans. Pattern Anal. Mach. Intell.*, 38(8):1583–1597, 2016.

[43] D. Wu and L. Shao. Leveraging hierarchical parametric networks for skeletal joints based action segmentation and recognition. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pages 724–731. IEEE, 2014.

[44] L. Xia, C. C. Chen, and J. K. Aggarwal. View invariant human action recognition using histograms of 3D joints. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, pages 20–27. IEEE, 2012.

[45] X. Yang and Y. Tian. Eigenjoints-based action recognition using naive-bayes-nearest-neighbor. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, pages 14–19. IEEE, 2012.

[46] Y. Yang, C. Deng, D. Tao, S. Zhang, W. Liu, and X. Gao. Latent max-margin multitask learning with skelets for 3-D action recognition. *IEEE Trans. Cybern.*, 47(2):439–448, 2017.

[47] J. Yao, X. Liu, and C. Qi. Foreground detection using low rank and structured sparsity. In *Proc. IEEE Int. Conf. Multimed. Expo.*, pages 1–6, 2014.

[48] M. Zanfir, M. Leordeanu, and C. Sminchisescu. The moving pose: An efficient 3D kinematics descriptor for low-latency action recognition and detection. In *Proc. IEEE Int. Conf. Comput. Vis.*, pages 2752–2759, 2013.

[49] W. Zhu, C. Lan, J. Xing, W. Zeng, Y. Li, L. Shen, X. Xie, et al. Co-occurrence feature learning for skeleton based action recognition using regularized deep LSTM networks. In *Proc. AAAI Conf. Artif. Intell.*, volume 2, page 8, 2016.