Leveraging Hierarchical Parametric Networks for Skeletal Joints Based Action Segmentation and Recognition

Di Wu, Ling Shao
Department of Electronic and Electrical Engineering
The University of Sheffield, Sheffield, S1 3JD, UK
{elp10dw,ling.shao}@sheffield.ac.uk

Abstract

Over the last few years, with the immense popularity of the Kinect, there has been renewed interest in developing methods for human gesture and action recognition from 3D skeletal data. A number of approaches have been proposed to extract representative features from 3D skeletal data, most commonly hard wired geometric or bio-inspired shape context features. We propose a hierarchial dynamic framework that first extracts high level skeletal joints features and then uses the learned representation for estimating emission probability to infer action sequences. Currently gaussian mixture models are the dominant technique for modeling the emission distribution of hidden Markov models. We show that better action recognition using skeletal features can be achieved by replacing gaussian mixture models by deep neural networks that contain many layers of features to predict probability distributions over states of hidden Markov models. The framework can be easily extended to include a ergodic state to segment and recognize actions simultaneously.

1. Introduction

Recognizing human actions from video data enables applications such as video understanding, semantic retrieval, surveillance, and human-computer interaction. Depending on the application, a recognition system may be constructed in different ways. This paper focuses on action recognition given 3D joint positions. Estimating such joint positions from an image sequence is a difficult task and the traditionally vision community used MoCap systems for pose tracking until the advent of Kinect [25, 9] when consumer price device can be readily available for human body joints extraction in real time with reasonable accuracy. It may seem that the task of action recognition given 3D joint positions is trivial, but this is not the case, largely due to the high dimensionality of the pose space. [20] noted that on a major problem in content-based comparison of motion data is that logically similar motions need not be numerically similar. In other words, there are certain aspects associated with a motion class that may show significant spatio-temporal variations between different executions of the motion, while other aspects are typically consistent.

A number of new datasets [6, 7, 8, 28] have provided researchers with the opportunity to design novel representations and algorithms and test them on a much larger number of sequences. Recently the focus has shifted towards modeling the motion of individual joints or combinatorial joints that discriminate between actions. Furthermore, to achieve continuous action recognition, the sequence need to be segmented into contiguous action segments; such segmentation is as important as recognition itself and is often neglected in action recognition research.

In this paper we focus on model driven analysis and synthesis but avoid the complexities involved in imposing physics-based constraints, relying instead on a pure learning approach in which all the knowledge in the model comes from the data without sophisticated pre-processing or dimensionality reduction. Our approach can be seen as an extension to [26] in that instead of using the conditional Restricted Boltzmann Machine, a type of shallow model, to model human motion, we add layers to learn higher level features justified by a variational bound [10]; for modeling temporal information, rather than explicitly binding limited adjacent frames (3 past frames as in [26], we resort to hidden Markov model which can be well extended to long term temporal information, spanning hundreds of frames in our system.

We demonstrate that consistently better action recognition performance can be achieved using skeleton information by "pretraining" a multi-layer neural network, one layer at a time, as a generative model. The advantage of this new way of training multi-layer neural networks is that the limited amount of information in the labels is not used to design features from scratch. It is only used to change the features ever so slightly in order to adjust the class boundaries. The features themselves are discovered by building a multi-layer generative model of much richer information in the skeletal joints configurations, and this does not require labeled data.

Our approach makes three major assumptions about the nature of the relationship between the input data, which in this case is a set configurations of skeletal joints, and the labels, which are action class HMM states produced by a forced alignment. First, we assume that the discrimination we want to perform is more directly related to the underlying causes of the data than to the individual elements of the data itself (previous hardwired techniques [3, 20, 23, 24] have shown that multiple joints relational features, *e.g*. hands approaching each other, feet moving towards each other, *etc*., are more relevant for action recognition rather than a single joint spatial-temporal position). Second, we assume that a good feature-vector representation of the underlying causes can be recovered from the input data by modeling its higher-order statistical structure. Third, feature-vector produced hidden states are mostly unique, meaning sequences are non-repetitive actions as opposed to longer repetitive activities, *e.g*., walking, running, jogging, *etc*., spanning minutes or hours.

Given the structure of our model, our framework is also suited for detection of "action points" for precise temporal anchoring of human actions. Action points can be thought of as marking a specific pose conditioned on "how the user got into that pose" [23]. It makes explicit the latency/accuracy tradeoff and allow accurate evaluation of human action recognition systems in contexts where latency is important.

2. Related Works

Traditionally, 3D joints data are acquire by MoCap system, and a plethora of hard wired features have been proposed: relational pose features, introduced by Müller et al. [20], have been used for indexing and retrieval of motion capture data and served as the harbinger for exploring 3D joints data. Yao et al. [31] modified a subset of the relational pose features for action recognition and showed that with these features, it is not necessary to have perfect poses to perform action recognition. [16] designed feature vectors, such that each feature corresponds to the pose of a single joint or combination of multiple joints and 7 distinct categories of hand crafted features were designed based on their analysis of the actions and features that can distinguish them. Different types characterize different levels of dynamics of an action and there are in total 141 hard wired features. During their training stage, 3 sets of ad hoc segregation of features space required laborious human involvement. [24] proposed the Sequence of Most Informative Joints (SMIJ) representation, a interpretive feature for human motion representation based on joint angle time series. [3] introduced a bio-inspired features incorporating 3D shape context into a spherical coordinate, they model a human activity using a hierarchy of 3D skeletal features in motion and learn the dynamics of these features using Linear Dynamical Systems (LDSs).

Alternative approaches to acquire discriminative features leverage statistical learning methods: [28] proposed a feature mining approach for computing discriminative actionlets from a recursively defined temporal pyramid of joint configurations. [4] proposed non-linear graphical model for structured prediction. It combines the power of deep neural networks to extract high level features with the graphical framework of Markov networks, yielding a powerful and scalable probabilistic model that was applied to signal labeling tasks.

In this paper, inspired by recent findings of [18], we propose automatic extraction of high level skeletal joints representation by using deep forward neural networks. This framework serves as a better model for estimation emission probability of hidden Markov models and achieves improved results for human action recognition amongst other well established methods. We also demonstrate that the framework can be easily adapted for simultaneous segmenting and recognizing gestures, discovering action points [23] which are precise temporal anchor points relative to the action performance. The model has been designed with human motion in mind, but should lend itself well to other high-dimensional time series.

3. Methodology

3D joint data generated via the skeleton tracking from the depth map sequences are generally more noisy than that of the MoCap data. When the difference between the actions is subtle, it is usually difficult to determine the accurate states from the observation without careful selection of the features, which undermines the performance of such generative models. Moreover, with limited amount of train ing data, training a complex generative model is easy to overfit. As generative models get better, however, the advantage of discriminative training gets smaller and is eventually outweighed by a major disadvantage: the amount of constraint that the data imposes on the parameters of a discriminative model is equal to the number of bits required to specify the correct labels of the training cases, whereas the amount of constraint for a generative model is equal to the number of bits required to specify the input vectors of the training cases. So when the input vectors contain much more structure than the labels, a generative model can learn many more parameters before it overfits.

Currently the model parameters are predominantly learnt by Gaussian mixture models using expectation maximization [1, 21]. We reason that replacing Gaussian mixture models by deep neural networks can better predict probability distributions over the states of hidden Markov models:

3.1. Problem formation

Feed forward neural networks offer several potential advantages over GMMs:

- Their estimation of the posterior probabilities of HMM states does not require detailed assumptions about the data distribution.

- They allow an easy way of combining divers features, including both discrete and continuous features.

- They use far more of the data to constrain each parameter because the output on each training case is sensitive to a large fraction of the weights.

The benefit of each weight in a neural network being constrained by a larger faction of training case than each parameter in a GMM has been masked by other differences in training. Neural networks have traditionally been training discriminatively, whereas GMMs are typically trained as generative models (even if discriminative training is performed later in the training procedure). Generative training allows the data to impose many more bits of constraints on the parameters, hence partially compensating for the fact that each component of a large GMM must be trained on a very small fraction of the data.

GMMs and HMMs co-evolved as a way of doing speech recognition when computers were too slow to explore more computationally intensive approaches. GMMs are easy to fit when they have diagonal covariance matrices, and with enough components they can model any distribution. They are, however, statistically inefficient at modeling high-dimensional data that has many kind of componential structure [18]. Suppose, for example, that $\mathcal{N}$ significantly different patterns can occur in one sub-band and $\mathcal{M}$ significantly different patterns can occur in another sub-band. Suppose also that which pattern occurs in each sub-band is approximately independent. A GMM requires $\mathcal{N} * \mathcal{M}$ components to model this structure because each component must generate both sub-bands(each piece of data has only a single latent cause). On the other hand, a model that explains the data using multiple causes only requires $\mathcal{N} + \mathcal{M}$ components, each of which is specific to a particular sub-band. This exponential inefficiency of GMMs for modeling factorial structures leads to the GMMs+HMMs systems that have a very large number of Gaussians, most of which must be estimated from a very small fraction of the data.

The benefit of learning a generative model is greatly magnified when there is a large supply of unlabeled skeletal data either acquired by motion capture systems or inferred from depth images in addition to the training data that has been labeled by a forced HMM alignment. We do not make use of unlabeled data in this paper, but it could only improve our results relative to purely discriminatively approaches.

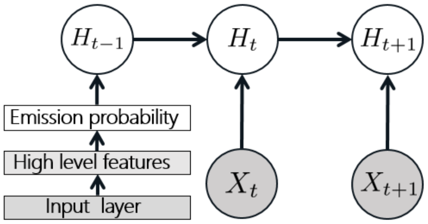


Figure 1: Per-action model: a forward-linked chain. Inputs (skeletal features) are first passed through a deep neural nets to extract high level features and output the emission probabilities of the hidden states. The deep neural net is first pre-trained using all skeletal features and then fine-tuned by the target class acquired by forced alignment for individual action class (10 hidden states for each action class in all our experiments).

Naturally, many of the high-level features learned by the generative model may be irrelevant for making the required discriminations, even though they are important for explaining the input data. However, this is a price worth paying if computation is cheap and some of the high-level features are very good for discriminating between classes of interest.

3.2. Graphical Models

We use a continuous-observation HMM with discrete hidden states. The model is constructed as follows. At each time step t we have one random observation variable X_t. Additionally we have an unobserved variable H_t taking values of a finite set $\mathcal{H} = (\bigcup_{a \in \mathcal{A}} \mathcal{H}_a)$,where $\mathcal{H}_a$ is a set of states associated to an individual action a. The intuition motivating this construction is that an action is composed of a sequence of poses where the relative duration of each pose may vary. This variance is captured by allowing flexible forward transitions within the chain. With this definitions, the full probability model is now specified as HMM:

$$p(H_{1:T}, X_{1:T}) = p(H_1)p(X_1|H_1) \prod_{t=2}^{T} p(X_t|H_t)p(H_t|H_{t-1}), \tag{1}$$

where $p(H_1)$ is the prior on the first hidden state and in all our experiments, we have a uniform prior. $p(X_t|H_t)$ is the observation model, and $p(H_t|H_{t-1})$ is the transition dynamics model. The observation domain $\mathcal{X}$ depends on the modality of the skeleton and will be described in the experimental section 4. We model the emission distribution of hidden Markov models by pre-training a multi-layer feed forward neural network [11], one layer at a time, as a generative model for a window of action frame. This pre-training makes it easier to optimize deep neural networks that have many layers of hidden units and it also allows many more parameters to be used before overfitting occurs. The graphical representation of a per-action model is shown as Fig. 1.

Because our skeleton features (*a.k.a.*observation domain $\mathcal{X}$) are continuous instead of binomial features, we use the Gaussian RBM (*GRBM*) to model the energy term of first visible layer:

$$E(v,h;\theta) = -\sum_{i=1}^{D}\frac{(v_i-b_i)^2}{2\sigma_i^2} - \sum_{i=1}^{D}\sum_{j=1}^{F} W_{ij}h_j\frac{v_i}{\sigma_i} - \sum_{j=1}^{F} a_j h_j$$

where model parameters $\theta = \{W, b, a, \sigma\}$ with W_{ij} represents the symmetric interaction term between visible unit i and hidden unit j while b_i and a_j are their bias terms with visible unit variance σ. D and F are the numbers of visible and hidden units.

Learning the higher level representation for skeleton joints features:
Neal and Hinton [22] demonstrated that the negative log probability of a single data vector, $\mathbf{v}^0$, under the multilayer generative model is bounded by a variational free energy, which is the expected energy under the approximating distribution, $Q(\mathbf{h}^0|\mathbf{v}^0)$, minus the entropy of that distribution. For a directed model, the "energy" of the configuration $\mathbf{v}^0$,$\mathbf{h}^0$ is given by $E(\mathbf{v}^0, \mathbf{h}^0) = -[\log p(\mathbf{h}^0) + \log p(\mathbf{v}^0|\mathbf{h}^0)]$. So the bound is

$$\begin{aligned}\log p(\mathbf{v}^0) \geqslant & \sum_{\mathbf{h}^0} Q(\mathbf{h}^0|\mathbf{v}^0)[\log p(\mathbf{h}^0) + \log p(\mathbf{v}^0|\mathbf{h}^0)] \\ & - \sum_{\mathbf{h}^0} Q(\mathbf{h}^0|\mathbf{v}^0)\log Q(\mathbf{h}^0|\mathbf{v}^0)\end{aligned}$$

The intuition using deep belief networks for modeling marginal distribution in skeleton joints action recognition is that by constructing multi-layer networks, semantically meaningful high level features for skeleton configuration will be extracted whilst learning the parametric prior of human pose from mass pool of skeleton joints data. In the recent work of [15], a non-parametric bayesian network is adopted for human pose prior estimation, whereas in our framework, the parametric networks are incorporated.

Using the pair wise joints features as raw input, the data-driven approach network will be able to extract relational multi-joints features which are relevant to target frame class. E.g., for "toss" action, wrist joints is rotating around shoulder joints would be extracted from the backpropagation via target frame as those task specific, *ad hoc* hard wired sets of joints configurations as in [3, 20, 23, 24].

The outputs of the neural net are the hidden states learned by force alignment during the supervised training process. Once we have model, we can use the normal online or offline smoothing, inferring the hidden marginal distributions $p(H_t|X_t)$ of every node (frame) of the test video. Because the graph for the hidden Markov model is a directed tree, this problem can be solved exactly using the max-sum algorithm. The number of possible paths through the lattice grows exponentially with the length of the chain. The Viterbi algorithm searches this space of paths efficiently to find the most probable path with a computational cost that grows only linearly with the length of the chain [1]. We can infer the action presence in a new sequence by Viterbi decoding as:

$$V_{t,\mathcal{H}} = P(H_t|X_t) + \log(\max_{\mathcal{H}\in\mathcal{H}_a}(V_{t-1,\mathcal{H}})) \qquad (2)$$

where initial state $V_{1,\mathcal{H}} = \log(P(H_1|X_1))$. From the inference results, we define the probability of an action $a \in \mathcal{A}$ as $p(y_t = a|x_{1:t}) = V_{T,\mathcal{H}}$.

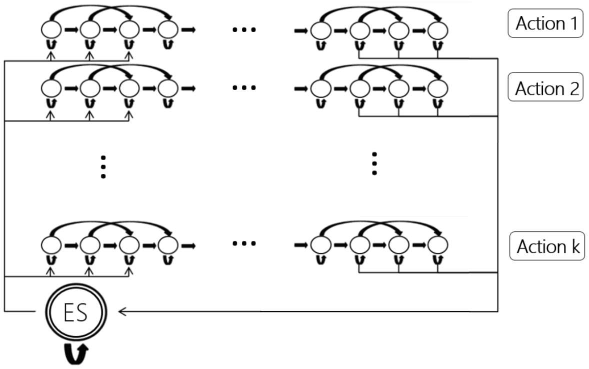


Figure 2: State diagram of the *ES-HMM* model for low-latency action segmentation and recognition. An ergodic states (ES) shows the resting position between action sequence. Each node represents a single frame and each row represents a single action model. The arrows indicate possible transitions between states.

3.3. Simultaneous Segmentation and Recognition

The aforementioned framework can be easily adapted for simultaneous action segmentation and recognition by adding an ergodic states-$\mathcal{ES}$ which resembles the silence state for speech recognition. Hence, the unobserved variable H_t takes an extra finite set $\mathcal{H} = (\bigcup_{a\in\mathcal{A}}\mathcal{H}_a)\bigcup\mathcal{ES}$, where $\mathcal{ES}$ is the ergodic state as the resting position between actions and we refer the model as *ES-HMM*.

Since want to capture the variation in speed of performing gestures, we set the transitions in the following way: when being in a particular node n in time t, moving to time $t+1$ we can either stay in the same node (slower performance), move to node $n+1$ (the same speed of performance), or move to node $n+2$ (faster performance). From the $\mathcal{ES}$ we can move to the first three nodes of any video, and from the last three nodes of every video we can move to the $\mathcal{ES}$ as shown in Fig. 2. The *ES-HMM* framework differs from the Firing Hidden Markov Model of [23] in that we strictly follow the temporal independent assump-

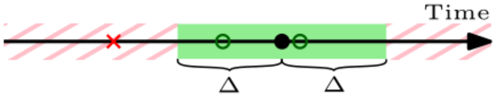


Figure 3: Latency-aware measure of predictive performance for a single action: a fixed time window of size 2Δ is centered around the ground truth action point annotation (marked ●) and used to partition the three predicted firing events into correct (marked ○) and incorrect predictions (marked ×). If there is more than one firing event within a ground truth window only one prediction is counted, the remaining ones are ignored. All incorrect detections are counted. The number of correct and incorrect detections determines $prec(\Delta)$ and $rec(\Delta)$.

tion, forbidding inter-states transverse, preconditioned that a non-repetitive sequence would maintain its unique states throughout its performing cycle.

The emission probability of the trained model is represented as a matrix of size $N_{\mathcal{TC}} * N_{\mathcal{F}}$ where $N_{\mathcal{F}}$ is the number of frames in a test sequence and output target class $N_{\mathcal{TC}} = N_{\mathcal{A}} * N_{\mathcal{H}_a} + 1$ where $N_{\mathcal{A}}$ is the number of action class and $N_{\mathcal{H}_a}$ is number of states associated to an individual action a and one $\mathcal{ES}$ state. Result of the Viterbi algorithm is a path–sequence of nodes which corresponds to states. From this path we can infer the class of the gesture as shown in Fig. 4b.

Performance Measure: F-score@Δ The performance of the system is measured in terms of precision and recall, defined in [7] as:

$$F - score(\Delta) = 2\frac{prec(\Delta) * rec(\Delta)}{prec(\Delta) + rec(\Delta)}$$

To achieve a high precision, the training data should only contain movements that users of the deployed system will associate with the gesture. To achieve a high recall, the training data should contain all movements that the designer wants to associate with a gesture. For a specified amount of tolerated latency (Δms) we measure the precision and recall as shown in Fig. 3.

4. Experiments

Features

The 3D coordinates of N joints of current frame c are given as: $X_c = \{x_1^c, x_2^c, \ldots, x_N^c\}$. We deploy 3D positional pairwise differences of joints [30] for observation domain $\mathcal{X}$. They capture posture features, motion features and offset features by direction concatenation: $\mathcal{X} = [f_{cc}, f_{cp}, f_{ci}]$ as demonstrated in Fig. 4a.

$$f_{cc} = \{x_i^c - x_j^c | i, j = 1, 2, \ldots, N; i \neq j\}$$
$$f_{cp} = \{x_i^c - x_j^p | x_i^c \in X_c; x_j^p \in X_p\}$$
$$f_{ci} = \{x_i^c - x_j^I | x_i^c \in X_c; x_j^I \in X_I\}$$

Resulting in a raw dimension of $N_{\mathcal{X}} = N_{joints} * (N_{joints} - 1)/2 + N_{joints}^2 + N_{joints}^2) * 3$ where N_{joints} is the number of joints used. Note that before extracting any features, all the 3D joint coordinates are transformed from the world coordinate system to a person centric coordinate system by placing the HipCenter (or ShoulderCenter if applied) at the origin. By including temporal differences f_{cp}, f_{ci} partially overcomes the very strong conditional independence assumption of HMMs, *i.e.* successive frames are independent given the hidden state of the HMM.

Admittedly, we do not completely negate human prior knowledge about information extraction for relevant static posture, velocity and offset overall dynamics of motion data. Nevertheless, aforementioned three attributes are all very crude pairwise features without any tweak into the data set or handpick the most relevant pairwise, triple wise, *etc*., designed features [3, 16, 20, 23, 24, 31]. Similar data driven approach has been adopted in [7] where random forest classifiers were adapted to the problem of recognizing gestures using a bundle of 35 frames. These sets of features extraction processes resemble the *Mel Frequency Cepstral Coefficients (MFCCs)* for speech recognition community [18].

Experimental setup

For high level skeleton feature extraction, we fix network architecture as $[N_{\mathcal{X}}, N_2, 1000, 1000, 1000, 1000, N_{\mathcal{TC}}]$ where $N_{\mathcal{X}}$ is the observation domain dimension and N_2 is the number of hidden nodes in *GRBM*, depending on the used joints set and is chosen as 2000 for upper body joints and 4000 for full body skeletal joints; $N_{\mathcal{TC}}$ is the output target class. And in all our experiments number of states associated to an individual action $N_{\mathcal{H}_a}$ is chosen as 10 for modeling the states of an action class. The feed forward networks are pre-trained with a fixed recipe using stochastic gradient decent with a mini-batch size of 100 training cases. Unsupervised initializations tend to avoid local minima and increase the networks performance stability and we have run 100 epochs for unsupervised pre-training. For Gaussian-binary RBMs, learning rate is fixed at 0.001 while for binary-binary RBMs as 0.01. For fine-tuning, the learning rate starts at 0.1 with 0.998 scaling after each epoch. To prevent complex co-adaptations in which a feature detector is only helpful in the context of several other specific feature detectors, we dropout [12] half of the feature detectors. Though we believe further carefully fine-tuned parameters would lead to more competitive results, in order not to “creeping overfitting”, as

Data Set Method	ChaLearn Gesture	MSR Action3D
EigenJoints+NBNN [30]	0.593	0.720
GMM+HMM [21]	0.408	0.704
NN+DTW [29]	0.599	-
DBN+HMM (this work)	***0.628***	***0.735***

Table 1: Baseline comparisons: first row (EigenJoints+NBNN) adopts same sets of features: showing that our model's efficacy in temporal incorporation; second row method (GMM+HMM) has the same graphical representation except that the Deep Belief Network is used to extract high level skeletal features, proving DBN is more effective for estimating the emission probability. And our model achieves better recognition rate than the winner of the challenge [29] that uses variant of nearest neighbour and dynamic time warping in the ChaLearn Gesture dataset.

algorithms over time become too adapted to the dataset, essentially memorizing all its idiosyncrasies, and losing ability to generalize [27], we would like to treat the model as the aforementioned more generic approach.

Baseline

We perform the sanity check for our algorithm as an effective way of comparing against two baselines: in order to verify that the model is a more powerful alternative to GMM for relating HMM states, we compare our approach against the *GMM+HMM* paradigm [21] for modeling the observation states $p(X_t|H_t)$; to verify that the temporal incorporation in our model is a more effective approach for action recognition against the Bag-of-Visual-Word approach, we compare against the *EigenJoint-Naive Bayes Nearest Neighbour* [30] where the same set of raw features have been used.

4.1. ChaLearn Italian Gesture Recognition

This challenge is on "multiple instance, user independent learning" [6] of gestures. We focus only on the skeletal modularity. There are 20 Italian cultural/anthropological signs, *i.e.,vattene, vieniqui, perfetto, furbo, cheduepalle, chevuoi, daccordo, seipazzo, combinato, freganiente , ok, cosatifarei, basta, prendere, noncenepiu, fame, tantotempo, buonissimo, messidaccordo, sonostufo.* We use the subset where the label data are provided during our evaluation process. The set contains 393 labeled sequences with a total of 7754 gestures. We used 350 sequences for training and the rest 43 sequences for testing, each sequence contains 20 unique gestures. For training set, there are in total 339,700 frames (20 fps). Note that large number of frames (up to hundred thousands of frames) is advantageous in our model settings over other nonparametric models for estimating skeletal human poses (*e.g.* GPLVM [14], Kernel Methods [13] could not be readily scaled up). Due to the parametric structure, once the training set is learned, testing time will be trivial compares with memory based method [2, 30]. Because this is a gesture recognition data set, only upper body joints are relevant to our discriminative tasks. Therefor, we consider only the upper 9 body for our task (full body joints have been compared, but as expected, results in inferior results compared to upper body 9 joints). The 9 upper body joints used are *"ShoulderCenter, ShoulderLeft, ElbowLeft, WristLeft, HandLeft, ShoulderRight, ElbowRight, WristRight, HandRight"*. The consistent improvement of recognition accuracy against two baseline methods under the same experimental settings in Table 1 shows the efficacy of the proposed framework in better estimating observation model and parsing temporal domain knowledge.

Method	Classification rate
Sequence of Most Informative Joints [24]	0.29
Recurrent Neural Network [17]	0.425
Dynamic Temporal Warping [19]	0.54
Multiple Instance Learning [5]	0.657
Structured Streaming Skeletons [32]	0.817
Actionlet Ensemble [28]	*0.88*
DBN+HMM (this work)	*0.82*

Table 2: Recognition accuracy on MSR-Action3D dataset compared to state-of-the-art approaches.

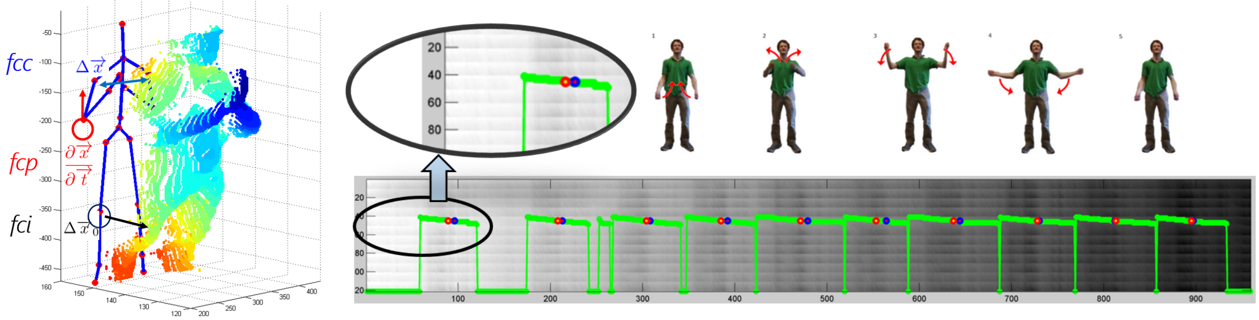
4.2. MSR Action3D

MSR Action3D dataset [28] is an action dataset of depth sequences captured by a depth camera. This dataset contains twenty actions, Each action was performed by ten subjects for three times. We compared the methods using only skeleton joints module in Table 1. Though the model still consistently outperforms two other baselines, the margins become smaller because only less than 10,000 frames in MSR Action3D dataset so that limited frames would not bring the advantages of generative pre-training into play.

We compare our model with the state-of-the-art methods on the cross-subject test setting as in [28] where training and testing sets are split by half of the actors. Though various idiosyncratic experimental set ups make it hard to have a fair comparison and generally render our generic 20 classes model at a disadvantage, (*e.g.* [28] with parameters that are empirically selected with data set dependent further tuning), the performance in Table 2 still exhibits the reasonable effectiveness of our model for this small frame number dataset.

4.3. MSRC12 dataset

The MSRC12 dataset [7] is originally proposed to investigate the question of what is the most appropriate semiotic



(a) (b)

Figure 4: (a) point cloud projection of depth image from ChaLearn Italian Gesture dataset and the 3D positional features. (b) top right: a "Wind up the music" metaphoric action instance from MSRC12 dataset; bottom: global score matrix via accumulating emission probabilities - fluorescent green is the Viterbi path from back tracking, a zoom in (top left) shows a path from states 41-50 (y-axis) indicating the gesture number 5 because we assign 10 states for each hidden Markov Model. Blue circles are the oracle ground truth action points and red circles are the predicted action points (middle frame of the Viterbi best path).

modality of instructions for conveying to human subjects the movements the system developer needs them to perform. Two categories of gesticulation, *i.e.*, Iconic - those imbue a correspondence between the gesture and the reference and Metaphoric - those that represent an abstract content, were investigated. Specifically they are: *lift outstretched arms, Duck, Push right, Goggles, Wind it up, Shoot, Bow, Throw, Had enough, Change weapon, Beat both, Kick.* The dataset includes 594 sequences and 719,359 frames-approximately six hours and 40 minutes-collected from 30 people performing 12 gestures. In total, there are 6,244 gesture instances. The motion files contain tracks of 20 joints estimated using the Kinect Pose Estimation pipeline. The body poses are captured at a sample rate of 30Hz with an accuracy of about 10 centimeters in joint positions. We conduct our experiments on sequences with a compound semiotic modality, (*i.e.* tagstream with letter *"A"*, such as Video + Text or Image + Text) and follow a "leave-persons-out" protocol, using 14 sequences of each gesture class for training (note that each sequence contains multiple gesture instances), leaving 4-6 sequences for intra-modality testing or 29-30 sequence for inter-modality testing.

For training the network, we set the ground truth action point annotation as the middle state of the target class, encoding a window of 100 frames centered around the action point, with the rest of frames encoded as the $\mathcal{ES}$. We assess the dual-modality generalization performance for all 12 gestures and compare against the random forest recognition system which has been successfully integrated into a game title that is currently being sold in retail stores and most recently proposed Structured Streaming Skeletons in the metric of F-score in Table 3. Fig. 4b illustrates a "Wind up the music" metaphoric action instance and the visualization of action point detection. Fig 5 plots the F-score at tolerated latency for the ES-HMM.

Modality F-score	intra-modality	inter-modality
Randomized Forest [7]	0.621	0.576
Structured Streaming Skeletons [32]	0.718	-
DBN-ES-HMM (this work)	***0.7243***	***0.7098***

Table 3: F-score at $\Delta = 333ms$ for intra-modality and inter-modality test of MSRC12 dataset.

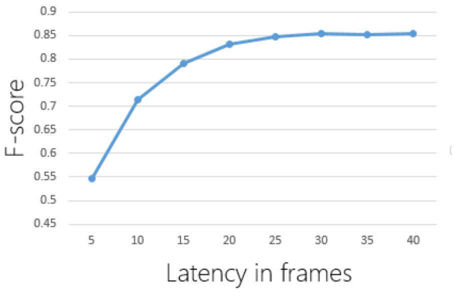


Figure 5: Latency profile of the MSRC 12 action point recognition task. We show F-score as a function of the tolerated latency for the *DBN+ES-HMM*. Each frame lasts 33ms. The plateau is largely due to low recall (missing detection).

Computational complexity

Though learning the network using stochastic gradient descent is tediously lengthy, once the model finishes training, with low inference cost, our framework can perform in realtime action segmentation/recognition. Specifically, a single feed forward neural network incurs trivial computation time, linearly in $\mathcal{O}(T)$ and the complexity of Viterbi algorithm is $\mathcal{O}(T * |S|^2)$ with number of frames T and state number S.

5. Conclusion

We have made feature extraction from skeletal joints data an implicit approach utilizing deep belief networks. By encoding dynamic structure into a HMM-based model, our discriminative trained, hierarchical parametric model excelled the GMM paradigm at better estimating emission probabilities for the directed graphical model. Further, we have introduced an ergodic states, rendered the framework being able to anchor the precise temporal locations of actions that are momentary, voluntarily performed and discrete in nature. Experiments have confirmed the efficacy of the framework at better estimating observation model than the GMM and integration of temporal domain knowledge exceeds the Bag-of-Visual-Word approach.

Acknowledgments: The authors would like to thank Ben Glocker, Antonio Criminisi and Sebastian Nowozin for their helpful suggestions and discussions about temporal modeling.

References

[1] C. Bishop. *Pattern recognition and machine learning*. Springer, 2006. 2, 4

[2] O. Boiman, E. Shechtman, and M. Irani. In defense of nearest-neighbor based image classification. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2008. 6

[3] R. Chaudhry, F. Ofli, G. Kurillo, R. Bajcsy, and R. Vidal. Bio-inspired dynamic 3d discriminative skeletal features for human action recognition. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2013. 2, 4, 5

[4] T. Do, T. Arti, et al. Neural conditional random fields. In *International Conference on Artificial Intelligence and Statistics*, 2010. 2

[5] C. Ellis, S. Z. Masood, M. F. Tappen, J. J. LaViola Jr, and R. Sukthankar. Exploring the trade-off between accuracy and observational latency in action recognition. *International Journal of Computer Vision*, 2013. 6

[6] S. Escalera, J. Gonzlez, X. Bar, M. Reyes, O. Lops, I. Guyon, V. Athitsos, and H. J. Escalante. Multi-modal gesture recognition challenge 2013: Dataset and results. In *ACM ChaLearn Multi-Modal Gesture Recognition Grand Challenge and Workshop*, 2013. 1, 6

[7] S. Fothergill, H. M. Mentis, P. Kohli, and S. Nowozin. Instructing people for training gestural interactive systems. In *ACM CHI*, 2012. 1, 5, 6, 7

[8] I. Guyon, V. Athitsos, P. Jangyodsuk, B. Hamner, and H. J. Escalante. Chalearn gesture challenge: Design and first results. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2012. 1

[9] J. Han, L. Shao, D. Xu, and J. Shotton. Enhanced computer vision with microsoft kinect sensor: A review. *IEEE Transactions on Cybernetics*, 2013. 1

[10] G. E. Hinton, S. Osindero, and Y.-W. Teh. A fast learning algorithm for deep belief nets. *Neural computation*, 2006. 1

[11] G. E. Hinton and R. R. Salakhutdinov. Reducing the dimensionality of data with neural networks. *Science*, 2006. 3

[12] G. E. Hinton, N. Srivastava, A. Krizhevsky, I. Sutskever, and R. R. Salakhutdinov. Improving neural networks by preventing co-adaptation of feature detectors. *arXiv preprint arXiv:1207.0580*, 2012. 5

[13] T. Hofmann, B. Schölkopf, and A. J. Smola. Kernel methods in machine learning. *The annals of statistics*, 2008. 6

[14] D. N. Lawrence. Gaussian process models for visualisation of high dimensional data. In *Advances in Neural Information Processing Systems*, 2004. 6

[15] A. Lehrmann, P. Gehler, and S. Nowozin. A non-parametric bayesian network prior of human pose. In *International Conference on Computer Vision*, 2013. 4

[16] F. Lv and R. Nevatia. Recognition and segmentation of 3-d human action using hmm and multi-class adaboost. In *European Conference on Computer Vision*. Springer, 2006. 2, 5

[17] J. Martens and I. Sutskever. Learning recurrent neural networks with hessian-free optimization. In *International Conference on Machine Learning*, 2011. 6

[18] A. Mohamed, G. E. Dahl, and G. Hinton. Acoustic modeling using deep belief networks. *IEEE Transactions on Audio, Speech, and Language Processing*, 2012. 2, 3, 5

[19] M. Müller, A. Baak, and H.-P. Seidel. Efficient and robust annotation of motion capture data. In *ACM SIGGRAPH/Eurographics Symposium on Computer Animation*. 6

[20] M. Müller and T. Röder. Motion templates for automatic classification and retrieval of motion capture data. In *SIGGRAPH/Eurographics symposium on Computer animation*. Eurographics Association, 2006. 1, 2, 4, 5

[21] K. P. Murphy. *Machine learning: a probabilistic perspective*. The MIT Press, 2012. 2, 6

[22] R. M. Neal and G. E. Hinton. A view of the em algorithm that justifies incremental, sparse, and other variants. In *Learning in graphical models*. Springer, 1998. 4

[23] S. Nowozin and J. Shotton. Action points: A representation for low-latency online human action recognition. Technical report, 2012. 2, 4, 5

[24] 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. *Journal of Visual Communication and Image Representation*, 2013. 2, 4, 5, 6

[25] J. Shotton, A. Fitzgibbon, M. Cook, T. Sharp, M. Finocchio, R. Moore, A. Kipman, and A. Blake. Real-time human pose recognition in parts from a single depth image. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2011. 1

[26] G. W. Taylor, G. E. Hinton, and S. T. Roweis. Modeling human motion using binary latent variables. In *Advances in neural information processing systems*, 2006. 1

[27] A. Torralba and A. A. Efros. Unbiased look at dataset bias. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2011. 6

[28] J. Wang, Z. Liu, Y. Wu, and J. Yuan. Mining actionlet ensemble for action recognition with depth cameras. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2012. 1, 2, 6

[29] J. Wu, J. Cheng, C. Zhao, and H. Lu. Fusing multi-modal features for gesture recognition. In *ACM International Conference on Multimodal Interaction*, 2013. 6

[30] X. Yang and Y. Tian. Eigenjoints-based action recognition using naive-bayes-nearest-neighbor. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2012. 5, 6

[31] A. Yao, J. Gall, and L. Van Gool. Coupled action recognition and pose estimation from multiple views. *International journal of computer vision*, 2012. 2, 5

[32] X. Zhao, X. Li, C. Pang, X. Zhu, and Q. Z. Sheng. Online human gesture recognition from motion data streams. In *ACM International conference on Multimedia*, 2013. 6, 7