The Thirty-Second AAAI Conference
on Artificial Intelligence (AAAI-18)

Unsupervised Representation Learning with Long-Term Dynamics for Skeleton Based Action Recognition

Nenggan Zheng,[1] Jun Wen,[2] Risheng Liu,[3*] Liangqu Long,[2] Jianhua Dai,[4] Zhefeng Gong[5]
[1] Qiushi Academy for Advanced Studies, Zhejiang University, Hangzhou, Zhejiang, China
[2] College of Computer Science and Techology, Zhejiang University, Hangzhou, Zhejiang, China
[3] DUT-RU International School of Information Science & Engineering, Dalian University of Technology, Liaoning, China
[4] College of Information Science and Engineering, Hunan Normal University, Changsha, Hunan, China
[5] Department of Neurobiology, Zhejiang University School of Medicine, Hangzhou, Zhejiang, China
zng@cs.zju.edu.cn, {junwen, myth, zfgong}@zju.edu.cn, rsliu@dlut.edu.cn, david.joshua@qq.com

Abstract

In recent years, skeleton based action recognition is becoming an increasingly attractive alternative to existing video-based approaches, beneficial from its robust and comprehensive 3D information. In this paper, we explore an unsupervised representation learning approach for the first time to capture the long-term global motion dynamics in skeleton sequences. We design a conditional skeleton inpainting architecture for learning a fixed-dimensional representation, guided by additional adversarial training strategies. We quantitatively evaluate the effectiveness of our learning approach on three well-established action recognition datasets. Experimental results show that our learned representation is discriminative for classifying actions and can substantially reduce the sequence inpainting errors.

Introduction

As an important branch of computer vision, action recognition has been widely used in many applications, such as intelligent video surveillance, robot vision, human-computer interaction, game control and so on (Weinland, Ronfard, and Boyer 2011; Yang and Tian 2017). Traditional studies about action recognition mainly focus on videos recorded by 2D cameras. The performances are still unsatisfactory, because it is difficult to achieve viewpoint and scale invariances as 2D videos lose some information of 3D space. The other general approach is skeleton based action recognition, in which a person is represented by 3D coordinate positions of key joints. Such representations are robust to variations of positions, scales, and viewpoints.

In this paper, we focus on the problem of skeleton based action recognition. The key to this problem is to learn the discriminative body postures and their motion dynamics. Most of the traditional skeleton based action recognition methods model the temporal dynamics of skeleton joints based on Hidden Markov Models (HMMs) (Xia, Chen, and Aggarwal 2012) or Temporal Pyramids (TPs) (Vemulapalli, Arrate, and Chellappa 2014). These models usually require selecting effective features to represent human body or choosing a proper width of the sliding window to model the temporal dynamics. In past few years, end-to-end deep learning techniques (Sutskever, Vinyals, and Le 2014; LeCun, Bengio, and Hinton 2015), especially Recurrent Neural Networks (RNNs) have been used for action recognition and achieved impressively better performances (Du, Wang, and Wang 2015; Zhu et al. 2016). The insights behind these models are to extract discriminative features to represent the temporal evolutions of different actions. However, most of these works mentioned above are referred to as supervised learning methods which heavily rely on massive labeled training examples. These labeled data are usually very expensive and not available. Hence, how to effectively and efficiently learn representations from the common and easily accessible unlabeled examples is challenging and attracts increasing research attention.

Recently, a stream of unsupervised representation learning approaches have been proposed. These methods are formulated with various objectives. Some models enforce the representations to be temporally smooth and learn slowly-varying representations (Földiák 2008), while others learn representations through reconstructing past frames or predicting future frames (Srivastava, Mansimov, and Salakhutdinov 2015; Luo et al. 2017). These models receive *fixed-length* input sequences, and then reconstruct past or predict *fixed-length* future frames. Although they show promising results, most of the learned representations still focus heavily on either capturing appearance features or local motion dynamics. These approaches are not flexible enough to handle sequences of varying length and fail to take consideration of encoding the long-term global motion dependencies in skeleton sequences. Their learned representations are not discriminative enough for classifying skeleton sequences.

To address the limitations mentioned above, we propose an unsupervised representation learning framework for compactly encoding long-term global motion dynamics *conditional inpainting*. As illustrated in Figure 1, the framework consists of three sub-networks, and it works as follows. An encoder (left side) runs through an input sequence and compactly encodes it into a fixed-dimensional representation. A decoder (middle side) learns to reconstruct the randomly masked (corrupted) input sequence conditioned

[*] Coresponding author.

on the learned representation. A discriminator (right side) learns to distinguish the original from the reconstructed sequence. With only a traditional element-wise loss, the reconstructed sequence may look visually unrealistic as the filled regions may not be coherent with their context. The discriminator is responsible for guiding the decoder towards producing visually realistic sequences by giving an adversarial loss. We call our model *conditional inpainting* as the decoder inpaints sequences conditioned on the learned representation. By using effective corruption strategies and reducing inpainting error, we aim to induce the learned representation to capture the long-term global motion dynamics. Experimental results on three public datasets show that our learned representation is discriminative for classifying actions, and with the learned representation we achieve better performances than the recently proposed unsupervised models and supervised models. The contributions of our work are as follows:

- Different from most existing sequential feature learning methods, which only learn appearances or short-term local motions, we introduce a novel conditional skeleton inpainting network to capture the long-term global motion dynamics in sequences with varying length. Moreover, to our best knowledge, we are the first to explore unsupervised representation learning approaches for skeleton-based action recognition.
- By designing the additional adversarial training strategy, we enhance the encoder-decoder model for learning more discriminative representations and reduce errors in skeleton sequence inpainting.
- Exhaustive experiments on real-world benchmarks verify the efficiency of our approach against both recently proposed unsupervised and supervised networks. As a non-trivial byproduct, we give comprehensive evaluations and ablation studies on the representations learned by different approaches.

Related Works

We first present related works on unsupervised representation learning for sequences. Then, we give a brief overview on existing skeleton based action recognition approaches.

Sequential Representation Learning

Wiskott and Sejnowski proposed the slow feature analysis framework for exploiting temporal structure in sequences and attempted to learn slowly-varying representations (Földiák 2008). Memisevic and Hinton approached this problem with a generative model by learning transformations between pairs of consecutive moments (Memisevic and Hinton 2010). Recently, a stream of reconstruction and prediction based models have been proposed. Ranzato et al. proposed a generative model that predicts the next frame or interpolates between frames using a recurrent neural network (Ranzato et al. 2014). This work was extended by Srivastava et al. with an LSTM Encoder-Decoder architecture that reconstructs fixed-length past frames or predicts fixed-length future frames (Srivastava, Mansimov, and Salakhudinov 2015). A further work is proposed by Luo et al. to learn representations by predicting the 3D motions of videos (Luo et al. 2017). These models can learn useful semantic features for some specific tasks. However, one common disadvantage of them is that these models only read in fixed-length sequence and can not flexibly handle sequences of varying length. These models inherently suffer from the inability to model long-term temporal dependencies.

Skeleton-based Action Recognition

Models for skeleton-based action recognition use body postures and motion dynamics to represent human actions. Most of the previous skeleton based action recognition methods explicitly model the temporal dynamics of skeleton joints by using TPs (Wang et al. 2012) or HMMs (Wu and Shao 2014). Recently, inspired by the success of deep recurrent neural networks for sequence modeling (Graves and others 2012), many end-to-end models have been proposed. The hierarchical recurrent neural network of (Du, Wang, and Wang 2015) divides the skeleton joints of a human body into five sets, combines their features hierarchically, and makes a final prediction with the fused joint information. The work of (Zhu et al. 2016) leverages on the intuition that co-occurrence of joints is a strong discriminative feature for human action recognition, and learns the mappings between co-occurring joints and the human action by enforcing a group sparsity constraint on the connection matrix. To learn features both in the temporal and spatial domains, deep LSTMs with Trust Gates (Liu et al. 2016) and a spatiotemporal attention model (Liu et al. 2016) are introduced. Though these models achieve cracking performances, they are usually limited by the heavy dependency on expensive labels.

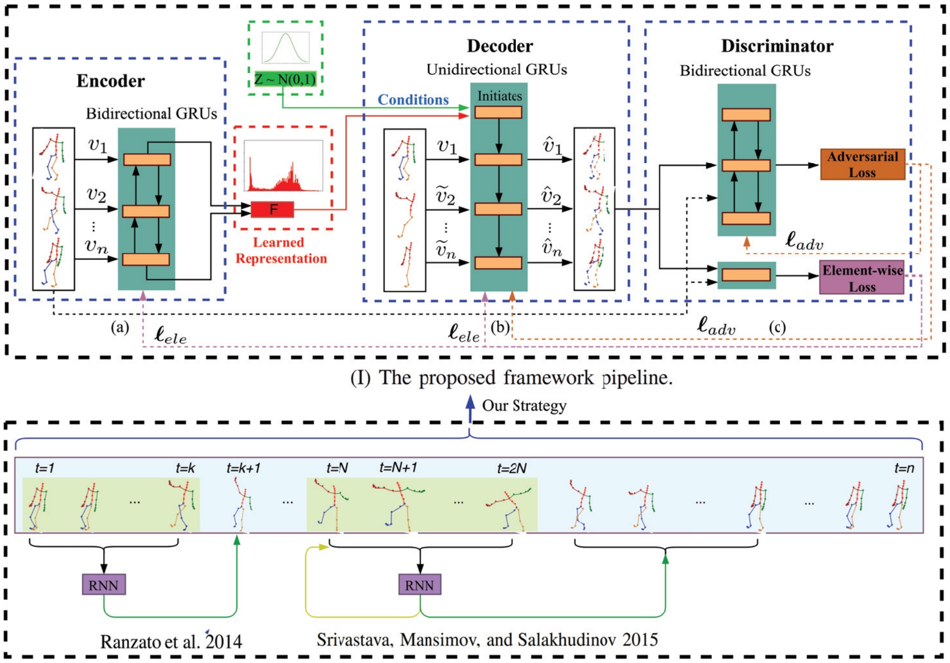
The Proposed Framework

In this section, we describe the proposed framework for unsupervised learning of long-term motion dynamics. The proposed framework is illustrated in Figure 1. It consists of three sub-networks: encoder network (*Enc*), decoder network (*Dec*) and Discriminator network (*Dis*). Our framework is based on the recurrent neural network (RNN) and the generative adversarial network (GAN) (Goodfellow et al. 2014). We firstly give brief reviews of both.

Building Block Network Structures

RNN and GRUs. We use RNN for discriminative feature learning and temporal dependency modeling, as it is a successful model for sequential information learning (Sutskever, Vinyals, and Le 2014). Comparing with the Long Short-Term Memory (LSTMs) (Hochreiter and Schmidhuber 1997), the Gated Recurrent Units (GRUs) (Cho et al. 2014) is easier to train and has fewer parameters. We use GRUs in our framework and achieve better performances, but it also works well with LSTMs or other units.

Adversarial Loss. The GAN model is a framework for training generative models. We incorporate it for conditional skeleton sequence inpainting. It consists of two competing networks: the *generator* is trained to map a latent variable



(I) The proposed framework pipeline.

(II) Strategy comparisons.

Figure 1: (I) Pipeline of our proposed method. (a) Encoder, (b) Decoder and (c) Discriminator. (II) Different learning strategies. Green: Strategies in (Ranzato et al. 2014; Srivastava, Mansimov, and Salakhudinov 2015), and Blue: our strategy. Previous works handle fixed-length sequences and model short-term temporal dependencies. Our model receives the whole input sequences with varying length and learns long-term global dynamics.

to the data space while the *discriminator* is trained to distinguish the generated data from samples from the training data. By learning a best possible *discriminator*, we aim to encourage the reconstructed sequences to best resemble the original sequences.

Motion Dynamics Encoding

The *Enc* is designed to learn a compact representation that encodes the long-term global motion dynamics and its structure is illustrated in Figure.2. The input is a sequence of vectors, each of which corresponds to a frame of 3D skeleton coordinates. Unlike other prediction based models that only read in parts of the input sequence, our *Enc* reads in the whole input sequence. After the last frame is read in, the last states of the GRUs cells in the last layer, acting as the holistic summary, are fully connected to a hidden layer to get the fixed-dimensional representation. To utilize the context information of each step and facilitate learning global temporal dependencies, the *Enc* uses bidirectional connections (Schuster and Paliwal 1997).

Decoding by Conditional Skeleton Inpainting

The *Dec* receives both the learned representation and the randomly corrupted input sequence. The goal of the *Dec* is to fill the masked regions in the input sequence conditioned on the learned representation. The decoder GRUs reads in the learned representation as the first-frame data to initiate its states. From the second step, it reads in the masked (corrupted) input sequence, of which a random number parts of a human body are masked from the second frame (a human body is divided into five parts, namely four limbs and a trunk). We observe that it is better to keep the first frame unmasked so as to give an initial inpainting reference for the *Dec*. The input value of the masked part of human body is set to be zero. As shown in Figure 1, we mask the same body parts for each frame of the input sequence, namely the masked regions are unknown from the second frame till the last frame. Otherwise, the *Dec* can do inpainting easily by merely referring to the history values and copying these values when predicting masked regions of current frames. This practice substantially increases the inpainting difficulty for the *Dec* but at the same time enhances the importance of the learned representation in inpainting. Meanwhile, we observe that it is better for the *Dec* not to refer to future frames when inpainting the current frame. Hence we use unidirectional GRUs for the *Dec*.

Coherence Driven Discriminator

The *Dec* is trained to fill the masked regions with small reconstruction errors. However, it does not ensure that the filled regions are coherent with its contexts, namely other unmasked human body parts of the current frame or the neighboring frames. Consequently, the inpainted sequence may look blurry and visually unrealistic. To encourage realistic inpainting results, we adopt the *Dis* that serves as a binary classifier to distinguish between real and fake sequences. The goal of the *Dis* is to point the *Dec* towards coherent sequence inpainting by giving the adversarial loss. The *(*Dis) shares a same bidirectional GRUs architecture as the *Enc*.

Training Loss with Adversarial Regularization

Our model combines the encoder-decoder architecture (Sutskever, Vinyals, and Le 2014) and the GAN model (Goodfellow et al. 2014). The decoder network is shared by both, acting as the *generator* of the GAN model. We train our model with a joint objective function:

$$\ell_{joint} = \ell_{ele} + \lambda_{adv}\ell_{adv}, \tag{1}$$

where ℓ_{ele} is the element-wise loss and ℓ_{adv} is the adversarial loss. λ_{adv} controls the weight of adversarial loss. We use the L_2 distance between the inpainted sequence and the original sequence as the element-wise loss. The element-wise loss can be interpreted as the *content* error of inpainting. It is responsible for capturing the overall structure of the missing regions but tends to point the model towards averaging the multiple modes in inpainting predictions. Consequently, the produced sequences tend to look blurry and visually unrealistic. The adversarial loss can be regarded as the *style* error of inpainting, and has the effect of picking a particular one from the multiple prediction modes and encourages the *Dec* to produce sequences that look visually realistic, namely fitting the distribution of the input sequences.

The *Dec* generates inpainted sequences, conditioned on the learned representation F. We observe better results when the inpainting of *Dec* is also conditioned on a random variable z, and the adversarial loss is:

$$\begin{aligned} \ell_{adv} = \log\left((Dis(x)\right) + \log\left(1 - Dis(Dec(F, \widehat{x}))\right) \\ + \lambda_z \log\left(1 - Dis(Dec(z, \widehat{x}))\right), \end{aligned} \tag{2}$$

where z is sampled from the prior distribution $N(0, I)$. For a masked input sequence x, $Dec(F, \widehat{x})$ has a much lower inpainting loss than $Dec(Z, \widehat{x})$ which is not included in the element-wise loss. Parameters λ_z controls the weight of z in the total adversarial loss. It is introduced to reduce the coupling of the *Enc* and *Dec*, and improve the generalization of the *Dec* as a generator of the GAN model and the distinguishing ability of the *Dis*.

Important Issues in Training

Due to the mutual influence of the three networks, to learn a non-trivial representation, the optimization is rather difficult. We therefore provide three practical considerations in this section. We refer to Figure 1 and Algorithm 1 for overviews of the training procedure.

- **Limit Error Signals to Relevant Networks.** With the joint loss function in Equation 1, we train both an Encoder-Decoder model and a GAN. This is possible because we do not update all network parameters wrt. the joint loss. The *Dis* only minimizes the adversarial loss, while the *Dec* receives error signals from both the element-wise loss and the adversarial loss. We observe that the *Enc* should not try to minimize the adversarial loss, otherwise the *Enc* tends to encode information useful for producing visually realistic sequences instead of encoding the motion dynamics for sequence inpainting. As a result, the learned representation performs poorly in classifying actions.

Algorithm 1 Training the conditional inpainting model.

```
Require: Training dataset $D$;
Ensure: Optimal Enc, Dec, Dis;
$\theta_{Enc}$,$\theta_{Dec}$,$\theta_{Dis} \leftarrow$ initialize network parameters;
repeat
    $\mathbf{x} \leftarrow$ random mini-batch from $D$;
    $\mathbf{F} \leftarrow Enc(\mathbf{x})$;
    $\widetilde{\mathbf{x}} \leftarrow$ randomly masked $\mathbf{x}$;
    $\hat{\mathbf{x}} \leftarrow Dec(\widetilde{\mathbf{x}}, \mathbf{F})$;
    $\ell_{ele} \leftarrow ||\mathbf{x} - \hat{\mathbf{x}}||_2$;
    $\mathbf{z} \leftarrow$ samples from prior $N(0, I)$;
    $\hat{\mathbf{x}}_z \leftarrow Dec(\widetilde{\mathbf{x}}, \mathbf{z})$;
    $\ell_{adv} \leftarrow \log((Dis(x)) + \log(1 - Dis(\hat{\mathbf{x}})) + \lambda_z \log(1 - Dis(\hat{\mathbf{x}}_z))$;
    // Update parameters according to gradients;
    $\theta_{Enc} \xleftarrow{+} -\nabla_{\theta_{Enc}} \ell_{ele}$;
    $\theta_{Dec} \xleftarrow{+} -\nabla_{\theta_{Dec}} (\ell_{ele} + \lambda_{adv}\ell_{adv})$;
    $\theta_{Dis} \xleftarrow{+} -\nabla_{\theta_{Dis}} \ell_{adv}$;
until deadline
```

- **Weighting Adversarial Loss.** The *Dec* minimizes the joint loss. We use a parameters λ_{adv} to weight the *content* error vs. the *style* error. We find that smaller λ_{adv} helps the *Enc* to learn a more effective representation, and we set it 0.1 in experiments. Meanwhile, the capability of the *Dis* network should also be small, otherwise it tends to focus on some trivial distinctions between the inpainted results and the original sequences.
- **Weighting of Conditional Inpainting on A Random Variable.** Parameters λ_z controls the weight of z in the total adversarial loss, which can be viewed as the weight of the *Dec* as a general conditional Generator of the GAN model. λ_z should not be set large, and in experiments, we set it 0.1. Otherwise, the *Dec* fails to cooperate with *Enc* to learn an effective representation for sequence inpainting. Larger λ_z tends to induce the *Enc* to learn a representation that only helps the *Dec* to deceive the *Dis*.

Experiments

The final goal of learning long-term motion dynamics is to classify actions in skeleton sequences. We learn representations with our unsupervised learning approach. A final classification layer is added on the top of the learned representation to classify actions, as shown in Figure 2. To study the effectiveness of our unsupervised learning approach, we consider the following three scenarios:

- **Unsupervised.** Fix the learned encoder and only fine-tune the last classifier layer with the available labels. This setting is to verify the effectiveness of the learned representation on action recognition.

- **Supervised+Pretraining.** Initialize the encoder with our learned weights and fine-tune the whole network with the available labels. We aim to explore if the learned representation is useful for supervised models.
- **Supervised.** Initialize the weight of our encoder randomly and learn them with labels available for the supervision task. This is our baseline model for comparison.

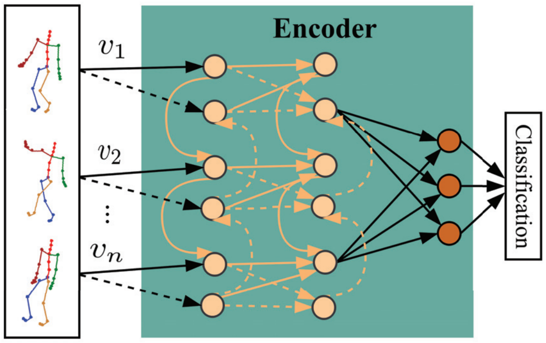


Figure 2: Detailed network architecture for action recognition. Each input skeleton sequence is encoded into a fixed-dimensional representation with the learned encoder (with the weights fixed) which is a two-layer bidirectional GRUs. Then, a classification layer is trained to infer the action.

Datasets

We perform our experiments on the following three datasets: the CMU dataset (CMU 2003), the HDM05 dataset (Müller et al. 2007), and the Berkeley MHAD dataset (Ofli et al. 2013).

CMU Dataset. This dataset contains 2,235 sequences. These sequences are performed by 144 non-professional actors. For each frame, the 3D coordinates of 31 joints are provided. The entire dataset has been categorized into 45 classes (Zhu et al. 2016). The dataset is very challenging due to the large sequence length variations and intra-class diversities. As in (Zhu et al. 2016), the evaluation is conducted on both the entire dataset and a selected subset of 664 sequences. For the entire dataset, the testing protocol is 4-fold cross validation, and for the subset, it is evaluated with 3-fold cross validation.

HDM05. This dataset contains 2,337 sequences for 130 actions. These sequences are performed by 5 actors. For each frame, 31 skeleton joints coordinates are recorded. As stated in (Cho and Chen 2014), some samples of these 130 actions should be classified into the same category. After combination, the actions are reduced to 65 categories. We follow the experimental protocol proposed in (Du, Wang, and Wang 2015) and perform 10-fold cross validation on this dataset.

Berkeley MHAD. There are 659 valid samples in this dataset, which consists of 11 actions performed by 12 subjects with 5 repetitions of each action. For each frame in a sequence, 35 joints coordinates are recorded. We follow the experimental protocol proposed in (Ofli et al. 2013) on this dataset.

Experiments Setup and Implementation Details

We firstly down-sample all datasets to *15 frames/s*. For the CMU dataset and Berkeley MHAD, to reduce the effects of large sequence length variations and computation expenses, sequences are sub-sampled again to ensure that the sequence length is below 36 frames. We augment the original training datasets by 25 times for unsupervised training. We sample an average of 25 sub-sequences from each original sequence. The length of these sub-subsequences randomly ranges from 7 frames to 35 frames. For all datasets, we scale the skeleton joints so that their 3D coordinates are in the interval of [-1, 1].

We implement our model in Tensorflow (Abadi et al. 2016) and optimize it with ADAM (Kingma and Ba 2014). Dropout regularization is used, and we only drop the activations that are communicated across layers, as proposed in (Zaremba, Sutskever, and Vinyals 2014). We set dropout ratio to be 0.2. Both the *Enc* and the *Dis* is a two-layer bidirectional GRUs. The *Dec* is a two-layer unidirectional GRUs. Each layer of the *Enc* and *Dec* has 800 hidden units. The *Dis* network is smaller, with 200 hidden units each layer. The dimensionality of the learned representation, namely the number of the hidden units of the fully connected layer, is the same as the input frames.

Comparisons to the State-of-the-Art

The aim of this set of experiments is to see if the representations learned by unsupervised learning are useful for action recognition. We compare our method with recently proposed unsupervised learning approaches and supervised models. The performances are summarized in Table 1 and Table 2. The two tables are divided into three sets. The first set shows the performances of the representations learned by two different unsupervised learning approaches. The second set presents the results of state-of-the-art supervised models. The third set compares supervised models that are pretrained with different unsupervised learning methods.

Supervised Models. Our unsupervised model achieves considerable performances. On the CMU subset, it achieves an accuracy of 84.57% which even beats the supervised *HBRNN* model by 1.44%. On Berkeley MHAD, it shows an advantage over the supervised RNN-based baseline model *DBRNN*, and achieves an accuracy of 100.00%. On the entire CMU dataset and CMU subset, the pretrained supervised model performs much better than the unsupervised model with an improvement of 15.19% and 6.62%, respectively. The performance gaps show the necessity of fine-tuning the whole network with labeled data to achieve further improvements. This is due to the fact that our model targets at only learning the global motion dynamics, ignoring the body postures. Fine-tuning the whole network with labeled data helps to learn more discriminative body poses and hence boost the performances.

Our baseline model, namely the supervised model without pretraining, achieves better performances than the *Deep LSTM* and is comparable with the state-of-the-art model. It shows that it is a very strong baseline. On the entire CMU dataset, CMU subset and HDM05 dataset, the learned representation by unsupervised learning succeeds in giving a further improvement of 1.53%, 3.36% and 0.63%, respectively. Our pretrained model achieves the best performances on all datasets comparing to the recently proposed supervised models.

Table 1: Comparisons on the CMU dataset in accuracy (%). The first group reports unsupervised (U) methods; the second group presents state-of-the-art supervised (S) methods; the third shows pretrained supervised methods(S+P). We report results averaged over 10 different samples of training sets.

Methods	Subset	CMU
U: Ours	84.57	66.23
U: Autoencoder[1]	77.03	61.43
S: HBRNN[2]	83.13	75.02
S: Deep LSTM (Zhu et al. 2016)	86.00	79.53
S: Co-deep LSTM (Zhu et al. 2016)	88.40	81.04
S: Ours	87.83	79.89
S+P: Unsupervised LSTM [1]	88.89	78.63
S+P: Ours	**91.19**	**81.42**

[1] (Srivastava, Mansimov, and Salakhudinov 2015)
[2] (Du, Wang, and Wang 2015)

Unsupervised Approaches. We compare to the state-of-the-art unsupervised learning approach *Unsupervised LSTM* (Srivastava, Mansimov, and Salakhudinov 2015) and the related model *Autoencoder*. We re-implement the two models and achieve better performances with GRUs cells. Meanwhile, since the *Unsupervised LSTM* only works on fixed-length input sequences, hence we report the performances of models initiated with it, as the authors did in the paper. Both models share the same baseline network and parameter settings with our model. The only difference lies in different learning approaches. As shown in Table 1 and Table 2, the representations learned with the *Unsupervised LSTM* does not necessarily help to boost the performances of supervised models. The performance even deteriorates on CMU dataset when pretrained on it. Our unsupervised approach consistently gives further improvements over the supervised models. The *Unsupervised LSTM* model and the *Autonencoder* do learn some semantic features, such as body postures or short-term motions, through reconstruction or prediction. However, these features can be easily obtained by the supervised models even with a small set of labeled data. The better performances verify the effectiveness of learning long-term global motion dynamics which needs plenty of examples to be captured.

Comparing to the *Autoencoder*, our unsupervised learning approach also performs much better. On the CMU subset, CMU dataset and HDM05 dataset , the performance gaps are 7.54%, 4.80%, and 0.72%, respectively. The advantages of our approach are more distinct on the challenging CMU dataset. Actually, due to the larger sequence length variations and intra-class diversities in the CMU dataset, modelling the long-term global temporal dependencies is critical for distinguishing some confusing actions and boost the performances. For example, the actions “basketball”,“run”, and “jump” are often misclassified by the *Autoencoder*, as they share some basic short-term motions, such as raising and lowering of legs. The performance of the generative model proposed in (Ranzato et al. 2014) is not reported due to the fact that it is heavily dependent on the quantization of inputs into a large dictionary and can’t work with our loss function.

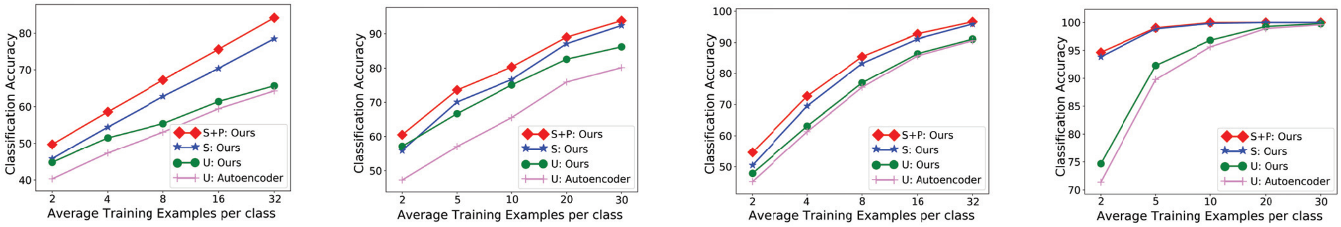
Table 2: Comparisons on the HDM05 and Berkeley MHAD (B-MHAD) datasets in accuracy (%). The first group reports unsupervised (U) methods; the second group presents state-of-the-art supervised (S) methods; the third shows pretrained supervised methods(S+P). We report results averaged over 10 different samples of training sets.

Methods	HDM05	B-MHAD
U: Ours	93.47	**100.00**
U: Autoencoder[1]	92.75	99.56
S: SMIJ (Ofli et al. 2014)	-	95.40
S: MLP (Cho and Chen 2014)	95.59	-
S: DBRNN[2]	96.70	99.64
S: HBRNN[2]	96.92	**100.00**
S: Deep LSTM[3]	96.80	**100.00**
S: Co-deep LSTM[3]	97.25	**100.00**
S: Ours	96.89	**100.00**
S+P: Unsupervised LSTM[1]	97.05	**100.00**
S+P: Our method	**97.52**	**100.00**

[1] (Srivastava, Mansimov, and Salakhudinov 2015)
[2] (Du, Wang, and Wang 2015)
[3] (Zhu et al. 2016)

Varying Size of Labeled Dataset

The aim of this set of experiments is to see the effects of the representations learned by unsupervised learning on the performances of supervised models with varying size of labeled training sets. The results are shown in Figure 3. We can see that for the case of very few training examples, unsupervised learning can give substantial improvements. For example, for the CMU dataset, the performance improves from 45.8% to 49.73% when trained with only 2 examples per category on average. On the CMU subset, the improvement is from 55.89% to 60.51%. As the size of the labelled dataset grows, the improvements gradually reduce. We believe that the decreasing improvement is due to the fact that the long-term motion dynamics learned by the unsupervised learning approach need plenty of data to be captured. When the labeled dataset is small, it is difficult for supervised mod-



(a) CMU dataset

(b) CMU subset

(c) HDM05

(d) Berkeley MHAD

Figure 3: Effectiveness of the learned representations with changes in the size of labelled training set. The results are averaged over 10 different samples of training sets.

els to capture the long-term motion dynamics, hence the improvements bringed by the unsupervised learning is larger. As the size of the labelled dataset grows, the advantages of unsupervised learning weaken.

Ablation Study

In this section, we aim to explore the importance of incorporating the GAN model for conditional sequence inpainting. We experiment with the CMU dataset with two different loss functions: the element-wise loss and the joint loss. The results are evaluated in two aspects: the inpainting error, measured with the mean square root (MSE), and the effectiveness of the learned representation, evaluated in classification accuracy and clustering accuracy. The clustering accuracy is evaluated with *ACC* (Xie, Girshick, and Farhadi 2016), which ranges in [0, 1] and larger value indicates better clustering performance. The results are summarized in Table 3. The table is divided into two sets. The first set shows results of the inpainting model conditioned on the learned representation, *Ours (w/ F)*, while the second set conditioned on zeros, *Ours (w/o F)*. It shows that adding the adversarial loss consistently reduces inpainting errors irrespective of the using of the learned representation. Adding the adversarial loss reduces the inpainting error from 0.20 to 0.16 when conditioned on the learned representation. Meanwhile, the adversarial loss improves the classification accuracy from 62.56% to 66.23% and the clustering accuracy from 45.27% to 51.35%, respectively.

Table 3: Ablation study on the CMU dataset. The last two columns report supervised accuracy, Acc.(S), and unsupervised accuracy, Acc.(U).

Method	Objective	MSE	Acc.(S)	Acc.(U)
Ours (w/ F)	ℓ_{ele}	0.20	62.56	45.27
Ours (w/ F)	ℓ_{joint}	0.16	66.23	51.35
Ours (w/o F)	ℓ_{ele}	0.26	-	-
Ours (w/o F)	ℓ_{joint}	0.23	-	-

In this section, we verify the effectiveness of the learned representation on helping the *Dec* with inpaining. The results are summarized in Table 3. It shows that when conditioned on the learned representation, *Ours (w/ F)*, the inpainting error is consistently much lower irrespective of the loss functions. Even without the learned representation, our basic model, *Ours (w/o F)*, can work well in inpainting, and the mean square root (*MSE*) is only 0.23 when trained with the joint loss. When conditioned on F, the inpainting error reduces from 0.23 to 0.16. The much lower *MSE* proves the effectiveness of the learned representation on helping the Dec with inpainting.

Conclusions

We presents a general framework for unsupervised learning of long-term motion dynamics in skeleton sequences of varying length. By combining the Encoder-Decoder model and the GAN model, we use the inpainting error as the supervision to learn a discriminative representation. We prove the effectiveness of our learned representation from multiple aspects on three well-established action recognition datasets, and achieve better performances than recently proposed unsupervised and supervised models.

For future work, we aim to explore the performances of our method on more challenging skeleton datasets and generalize it to multiple persons. Further, we aim to explore its performances on noisy datasets as the element-wise construction loss can be sensitive to the noise of skeleton coordinates. We also want to explore its effectiveness on RGB based datasets such as the *UCF-101*, *ActivityNet* or other supervised tasks beyond action recognition, and exploit other free labels from sequences such as the smooth constraints.

Acknowledgments

This work was supported by National Key Basic Research Program of China (973 program 2013CB329504), National Natural Science Foundation of China (61572433, 61672125, 31671074, 61572433, 61473259) and Zhejiang Provincial Natural Science Foundation of China (LZ14F020002, LR13C090001).

References

Abadi, M.; Agarwal, A.; Barham, P.; Brevdo, E.; Chen, Z.; Citro, C.; Corrado, G. S.; Davis, A.; Dean, J.; Devin, M.; et al. 2016. Tensorflow: Large-scale machine learning on heterogeneous distributed systems. *arXiv preprint arXiv:1603.04467*.

Cho, K., and Chen, X. 2014. Classifying and visualizing motion capture sequences using deep neural networks. In *Computer Vision Theory and Applications (VISAPP), 2014 International Conference on*, volume 2, 122–130. IEEE.

Cho, K.; Van Merriënboer, B.; Gulcehre, C.; Bahdanau, D.; Bougares, F.; Schwenk, H.; and Bengio, Y. 2014. Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*.

CMU. 2003. Cmu graphics lab motion capture database. *http://mocap.cs.cmu.edu*.

Du, Y.; Wang, W.; and Wang, L. 2015. Hierarchical recurrent neural network for skeleton based action recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1110–1118.

Földiák, P. 2008. Learning invariance from transformation sequences. *Learning* 3(2).

Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; and Bengio, Y. 2014. Generative adversarial nets. In *Advances in neural information processing systems*, 2672–2680.

Graves, A., et al. 2012. *Supervised sequence labelling with recurrent neural networks*, volume 385. Springer.

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

Kingma, D., and Ba, J. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

LeCun, Y.; Bengio, Y.; and Hinton, G. 2015. Deep learning. *Nature* 521(7553):436–444.

Liu, J.; Shahroudy, A.; Xu, D.; and Wang, G. 2016. Spatio-temporal lstm with trust gates for 3d human action recognition. In *European Conference on Computer Vision*, 816–833. Springer.

Luo, Z.; Peng, B.; Huang, D.-A.; Alahi, A.; and Fei-Fei, L. 2017. Unsupervised learning of long-term motion dynamics for videos. *arXiv preprint arXiv:1701.01821*.

Memisevic, R., and Hinton, G. E. 2010. Learning to represent spatial transformations with factored higher-order boltzmann machines. *Neural computation* 22(6):1473–1492.

Müller, M.; Röder, T.; Clausen, M.; Eberhardt, B.; Krüger, B.; and Weber, A. 2007. Documentation mocap database hdm05.

Ofli, F.; Chaudhry, R.; Kurillo, G.; Vidal, R.; and Bajcsy, R. 2013. Berkeley mhad: A comprehensive multimodal human action database. In *Applications of Computer Vision (WACV), 2013 IEEE Workshop on*, 53–60. IEEE.

Ofli, F.; Chaudhry, R.; Kurillo, G.; Vidal, R.; and Bajcsy, R. 2014. Sequence of the most informative joints (smij): A new representation for human skeletal action recognition. *Journal of Visual Communication and Image Representation*.

Ranzato, M.; Szlam, A.; Bruna, J.; Mathieu, M.; Collobert, R.; and Chopra, S. 2014. Video (language) modeling: a baseline for generative models of natural videos. *arXiv preprint arXiv:1412.6604*.

Schuster, M., and Paliwal, K. K. 1997. Bidirectional recurrent neural networks. *IEEE Transactions on Signal Processing* 45(11):2673–2681.

Srivastava, N.; Mansimov, E.; and Salakhudinov, R. 2015. Unsupervised learning of video representations using lstms. In *International Conference on Machine Learning*, 843–852.

Sutskever, I.; Vinyals, O.; and Le, Q. V. 2014. Sequence to sequence learning with neural networks. In *Advances in neural information processing systems*, 3104–3112.

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

Wang, J.; Liu, Z.; Wu, Y.; and Yuan, J. 2012. Mining actionlet ensemble for action recognition with depth cameras. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, 1290–1297. IEEE.

Weinland, D.; Ronfard, R.; and Boyer, E. 2011. A survey of vision-based methods for action representation, segmentation and recognition. *Computer vision and image understanding* 115(2):224–241.

Wu, D., and Shao, L. 2014. Leveraging hierarchical parametric networks for skeletal joints based action segmentation and recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 724–731.

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

Xie, J.; Girshick, R.; and Farhadi, A. 2016. Unsupervised deep embedding for clustering analysis. In *International Conference on Machine Learning*, 478–487.

Yang, X., and Tian, Y. 2017. Super normal vector for human activity recognition with depth cameras. *IEEE transactions on pattern analysis and machine intelligence* 39(5):1028–1039.

Zaremba, W.; Sutskever, I.; and Vinyals, O. 2014. Recurrent neural network regularization. *arXiv preprint arXiv:1409.2329*.

Zhu, W.; Lan, C.; Xing, J.; Zeng, W.; Li, Y.; Shen, L.; Xie, X.; et al. 2016. Co-occurrence feature learning for skeleton based action recognition using regularized deep lstm networks. In *AAAI*, volume 2, 8.