Reconstructing Dynamics of Unobserved Joints in 3D Human Actions Using Deep Generative Priors

Suhas Lohit
Arizona State University
slohit@asu.edu

Rushil Anirudh
Lawrence Livermore National Laboratory
anirudh1@llnl.gov

Pavan Turaga
Arizona State University
pturaga@asu.edu

Abstract

Motion capture (mocap) and time-of-flight based sensing of human actions are becoming increasingly popular modalities to perform robust activity recognition. However, because of several nuisance factors such as occlusions, it may not be possible to record the movements of all joint locations, which makes it difficult to integrate it with downstream tasks like action classification easily. In this paper, we first pose the problem of reconstructing joint dynamics as an ill-posed linear inverse problem. We then propose a method based on deep generative priors to perform the reconstruction. Then, given an action with unseen joints, we complete the action by projecting it onto the manifold of human actions by optimizing the latent space representation. Experiments on both mocap and Kinect datasets clearly demonstrate that the proposed method performs very well in recovering semantics of the actions and dynamics of unseen joints. We will release all the code and models publicly.

1. Introduction

With the proliferation of low-cost sensing devices, sequential data has become ubiquitous in applications such as action and gesture recognition, health trackers, heart rate monitoring etc. In many of these applications, the essential task at hand is inferring abstract, high-level semantic quantities such as health of the patient, quality of movement, intended gestures etc. These quantities depend on the underlying dynamical process or system that is generating the time-series. Accurately estimating the dynamical process is non-trivial and would require us to completely observe the system, due to large degrees of freedom and complex interactions between sensors and humans. Traditionally, this has restricted us to using certain features of the dynamical process that can still be determined from partial observations [30], for e.g. estimating the dynamics of human movement from a few skeletal joint sequences [29]. However, more recently, with the availability of large datasets, and highly parameterized predictive models, we are able to implicitly learn the dynamics much better, leading to significantly higher performance on several benchmarks [35, 31, 26].

In this paper, we are interested in restoring dynamics of such sequential data when several of the dimensions are missing, particularly at test time. This can occur due to several factors, for e.g., identifiability issues, faulty sensors, occlusions and other environmental nuisance factors. While classical topological features may be used here, they tend to be simplistic and not predictive enough on complex datasets [3, 29]. On the other hand, conventional deep learning methods fail due to the unaccounted distribution shift between training and testing data sets, and also because when the joints are missing a lot of information regarding the dynamics is lost. Instead we pose this as a linear inverse problem that can be solved at test time, with great accuracy. We focus particularly on the problem missing joints in skeletal action recognition.

In skeletal action recognition, sensors such as motion capture (mocap), and Microsoft Kinect can directly provide approximate joint or body-part locations directly instead of standard RGB frames. These relatively novel modalities have the advantage over RGB data in that both the data and the corresponding neural network architectures are much smaller and require less memory and compute. Effective action recognition can be performed on these data directly without access to RGB frames [28, 36, 12, 35, 31, 26]. However, depth-based sensing using a device like Kinect may not be optimal when all the joints are not visible to the camera. In order to be effective, the entire body needs to be visible to the camera, and should be in relatively 'normal'

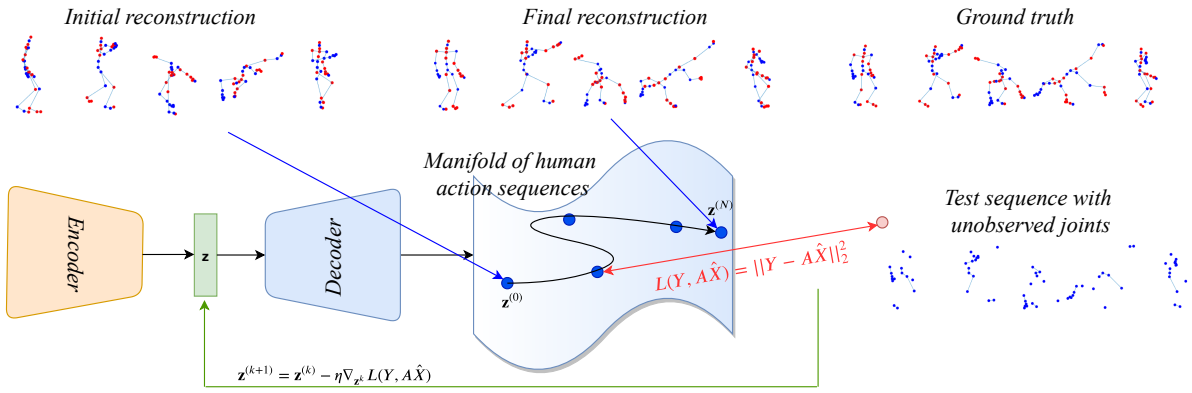


Figure 1: Block diagram illustrating the process of reconstructing dynamics of unseen joints by projecting to the range space of a generative model. In this paper, we use a temporal convolutional autoencoder as the generative model. Given a test sequence with only a subset of joints observed, we get an initial estimate of the reconstruction using the encoder representation for the test sequence $\mathbf{z}^{(0)}$. Then, we optimize over the encoder representation space in order to minimize the distance between the output of the decoder and the given test sequence, as shown. This procedure greatly improves the final reconstructed sequence.

poses. Not only is this assumption restrictive in general, but it becomes a very significant barrier in special situations such as those involving home-based physical therapy interventions, art performance, or emerging augmentations based on physical activity tracking in workplaces, where occlusions occur due common daily objects such as tables, chairs, etc. While these issues can be resolved with accurate mocap, it requires an expensive setup with a large number of cameras. Employing fewer cameras would be inexpensive, but can result in several occluded joints. Moreover, it is not clear which set of joints will be occluded, and can depend on the type of action being performed, as well as the subject position and the action trajectory.

In this paper, we focus on the problem of reconstructing unseen joint dynamics for 3D human action sequences at test-time. Given a human skeletal action sequence with a fraction of joints missing throughout the action, can we reconstruct the dynamics of those unseen joints and use the resulting completed action for action recognition without any modifications to the classification architectures? We answer the question affirmatively.

Contributions:

1. We show that it is possible to effectively recover semantics and dynamics about human actions while having access to as few as 50% of the joints during both training and testing.
2. We pose it as an ill-posed linear inverse problem on the space of actions, and solve it by imposing a dynamical prior using a pre-trained generative model. The solution to the inverse problem is achieved by approximately projecting the incomplete action onto the range of this generative model, as shown in Figure 1.
3. The proposed method is a test-time approach that can easily handle changes in the kind of missing joints directly at test time. We demonstrate its effectiveness using real world mocap and Kinect datasets. We validate that the reconstructions recover information about the semantics and dynamics of the actions and show that in most cases the recovered actions perform within 5% of original ground truth data.

2. Related work

In this section, we describe closely related work which we put into three categories.

2.1. 3D human action acquisition and completion

The two most common methods of capturing 3D human skeletal actions are using motion capture (mocap) systems, and using depth-sensing cameras like Kinect and RealSense and then using a algorithm to estimate the joint locations. However, both these methods suffer from drawbacks. Not all joints may be recorded mainly due to occlusions and non-standard body types/poses. Moreover, in order to reduce the cost of mocap systems, we would like to employ fewer cameras which leads to the possibility of joints being unseen throughout the actions. When the final application is action classification, it may be possible to classify directly with fewer joint trajectories, but this suffers from two disadvantages: (a) it is impractical to train a classifier for every combination of unseen joints (b) we can achieve better performance by first reconstructing the complete set of joint trajectories, exploiting information about the nature of human actions from a training dataset, and then performing the classification using a single pretrained classifier. Earlier

works have considered the problem of human action completion in different ways – human action prediction where given a few frames of a human action sequence, the future frames can be predicted employing machine learning algorithms [2], human motion synthesis [7] and editing [10]. There are also traditional methods for human action completion using k-nearest neighbors from the training set [1], and matrix factorization methods [8]. More recently, Yang et al. [32] propose using hand-crafted low-rank and sparsity priors to model spatial-temporal correlation for 3D human motion recovery problems. In this paper, we consider the problem of predicting complete joint trajectories given a subset of joints of every frame in the sequence and propose a deep learning-based solution. Kucherenko et al. [14] propose feed-forward pass through trained a LSTM autoencoder as a way of reconstructing human actions. We use a baseline similar to this method in our experiments.

2.2. Deep generative priors for inverse problems

Several important problems in imaging can be cast as ill-posed linear inverse problems. The forward operation is given by $\mathbf{y} = A\mathbf{x}$, where $\mathbf{y} \in \mathbb{R}^m, \mathbf{x} \in \mathbb{R}^n, m < n$. The goal in inverse imaging is to reconstruct $\mathbf{x}$ from $\mathbf{y}$ which is generally ill-posed by exploiting structure of the desired $\mathbf{x}$ known apriori. In recent years, the prior knowledge comes in the form of a deep generative model (autoencoders, variational autoencoder, generative adversarial networks etc.). The process of reconstruction reduces to the problem of finding the closest point on the range space of the generative model $\hat{\mathbf{x}}$ such that $A\hat{\mathbf{x}} \approx \mathbf{y}$. This idea also has thoretical guarantees as shown by Bora et al. [4] and Shah and Hegde [24]. Bora et al. [5] showed that, in some cases, the generative model can also be *trained* using the noisy images. Recent papers apply this idea for time-series imputation [18, 33]. However, these techniques are shown to work for simpler time-series classification problems using measurements where some information is available from all dimensions. In contrast, we consider human action sequences in which certain joints are completely unseen i.e., many dimensions of the time-series are missing. Thus, interpolation techniques in the time domain are not applicable. Also, current methods for skeleton completion are performed frame-wise (such as Kinect), and do not take dynamics into account, which is the goal of this paper and necessary for human action recognition.

2.3. Dynamical systems approach to deal with missing observations

In dynamical systems theory, the notion of reconstructing high-dimensional state-spaces from low-dimensional observations has been well-studied for many decades. For classical state-estimation approaches to apply, one often needs to make simplifying assumptions for state-dynamics, such as Markovian and linear dynamics. Such simplifying assumptions are not always reflective of the complexity of the task at hand, such as reconstructing human action sequences. Another approach is to avoid making such parametric assumptions, but use methods from non-linear dynamics [17] to estimate surrogate state-spaces. From standard methods in non-linear dynamics, these surrogate state-spaces are only topologically equivalent to the true state-spaces, and do not have enough predictive information for high-level inference.

3. Reconstructing unseen joint dynamics as an ill-posed linear inverse problem

Recovering the dynamics of unseen joints is an ill-posed inverse problem since we only have access to partial information of the activity. This can be considered analogous to the inverse problems in imaging such as super-resolution, image inpainting or compressive sensing. However, unlike inverse imaging problems, it is not clear what kinds of priors work for human actions. We argue that a deep generative model learned from data on human actions acts as a good approximation to the space of all possible dynamical systems for human actions. As a result, we are able to implicitly constrain the dynamics of the recovered actions by restricting the solution to lie on the *action manifold*. We formalize these ideas next.

Let the total number of joints per frame be denoted by J and the number of frames per action sequence by N. Each joint is described by its 3D co-ordinates in space. Thus, by vectorizing, each skeleton can be represented by $3J-$dimensional vector and by stacking the N frames in columns, we represent the human action as a matrix X of size $3J \times N$. Let the number of joints observed be K, so the number of unseen joints is $J - K$. The measurement operator A then is a sub-sampling operator which drops $3(J - K)$ rows of X to give us the observed action Y. As $J - K$ joints are unseen, there are $3(J - K)$ rows of Y which are unknown and we replace them with 0 before further processing. Given Y, our eventual goal is to classify the action. As an intermediate step, we first reconstruct $\hat{X}$ from Y which is the main focus of this paper. Clearly, this is an ill-posed linear inverse problem. The advantage of viewing it as such helps us adapt algorithms designed for inverse imaging problems such as image inpainting [22], super-resolution [6, 16] and compressive imaging [15]. In this paper, we adapt the most recent approaches based on generative priors [4, 5, 27]. We note that the advantage of these methods over other methods in inverse imaging such as purely data-driven approaches [6, 15], and unrolled iterative methods [23, 20, 34] is that there is no requirement of paired Y and X for training. Thus, once we have a generative model for human action sequences, the problem of reconstructing unseen joint dynamics can be solved using an optimization problem such that the output of the gener-

ative model $\hat{X}$ is closest (in some predefined sense) to the test sequence under consideration Y. Next, we describe the architecture of the generative model we construct.

4. Generative models for human actions

In order to approximate space of human actions, we employ an autoencoder architecture to construct the generative model of human action sequences. We choose an autoencoder over currently popular generative adversarial networks [9] or variational autoencoders [13] because – (1) autoencoders are much easier to train compared to the other frameworks and (2) the purpose of using the generative model in this paper is to perform reconstruction of test sequences rather than sampling new actions, which, as we will show, can be readily performed using an autoencoder.

4.1. Autoencoder architecture

As the generative model, we employ a temporal convolutional autoencoder. Both the encoder (E) and decoder (D) consists of a series of 1D convolutional layers operating in the temporal domain with ReLU non-linearity. After every convolution, we use average pooling to reduce the number of frames by half. We then use a fully-connected (FC) layer which produces the encoded/latent representation of the action, denoted by $\mathbf{z}$. The decoder reverses these operations with a series of transposed convolutional layers. The networks are trained using full/complete actions with access to information of all joint trajectories. The network is trained to minimize the Euclidean loss between the input sequence and the output of the decoder:

$$L(X, \hat{X}) = \sum_{n=1}^{N} \sum_{j=1}^{J} \left\| X_{n,j} - \hat{X}_{n,j} \right\|_2^2, \qquad (1)$$

where $X_{n,j}$ refers to the j^{th} 3D joint location in the n^{th} frame of the sequence. Other training details are provided in the supplementary material. We note that we can add an additional **adversarial loss** term to the above loss function in order to make the actions more realistic [22]. As the main focus of the paper is designing a completion algorithm, we include these results in the supplementary material due to space constraints.

4.2. Training the generative models with partially observed joint sequences

The generative models in Sections 4.1 are trained using complete actions with all joint sequences fully observed, X. However, complete actions may not be available at the training stage. *An important contribution of this paper is showing that we can construct generative models of human actions by training solely on action sequences with only a subset of the joints observed, Y.* We later show that this protocol leads to superior reconstruction compared to training with full actions. To this end, we modify the loss function as follows. The forward operator A is the sampling operator which has the effect of dropping a subset of the joints. Using the knowledge of A, we use a masked loss function between the ground-truth measured sequence $Y = AX$ and the reconstructed sequence $\hat{X}$.

$$L_(Y, \hat{X}) = \sum_{n=1}^{N} \sum_{j=1}^{J} \left\| Y_{n,j} - A\hat{X}_{n,j} \right\|_2^2 \qquad (2)$$

The network architectures and the training protocols are identical to those trained using fully observed action sequences.

5. Reconstruction via approximate projection onto the action manifold

Once we have the generative model, in our case an autoencoder, the training process is complete. The next step is to use the generative model in order to reconstruct the trajectories of unseen joints given an incomplete action sequence. To this end, we propose to project the incomplete action to the range space of the generator, which ideally is the same as the manifold of complete human action sequences.

Initialization: Feed-forward pass through the trained autoencoder As a baseline method, we can simply feed the incomplete action sequence through the autoencoder and use the output of the decoder as the reconstruction. This is likely to fail, especially in the case of the autoencoder trained with complete actions. However, in the case of the autoencoder trained on subsets of joint trajectories, even this simple method can provide a reasonable reconstructed sequence. This is used as initialization for the optimization algorithm below, $\mathbf{z}^{(0)}$.

5.1. Optimizing the latent representation

We can further improve the reconstruction quality from above by directly optimizing the encoded/latent representation, $\mathbf{z}$, such that the Euclidean distance between the reconstructed action sequence and the input incomplete sequence. This method is inspired by Bora et al. [4] where the authors propose this method for inverse problems in imaging. The optimization problem is given by

$$\mathbf{z}^* = \arg\min_{\mathbf{z}} \|Y - AD(\mathbf{z})\|_2^2, \quad \hat{X} = D(\mathbf{z}^*). \qquad (3)$$

We solve this optimization problem using a gradient descent-type method, As the optimization problem is non-convex, $\mathbf{z}^*$ is a locally optimal solution. Empirically, we find that the solutions obtained using this procedure provide high quality reconstructions.

6. Measuring reconstruction performance

Our main goal in this paper is to recover trajectories of unseen joints from incomplete human actions. Hence, in order to evaluate the effectiveness of the proposed methods, we need to choose proper metrics to compute how well the reconstructed actions are. We choose the following three methods for measuring the quality of reconstructed actions.

6.1. Classification performance

An important reason for reconstructing action sequences is to employ predefined classification pipelines without any modification. Therefore, we train a single action classifier on sequences with information of all joints, feed the reconstructed sequences as test inputs, and use the classification performance as a metric for quality of reconstructed actions.

Classifier architecture: In all our experiments, we employ a popular architecture for 3D human action recognition based on temporal convolutional networks (TCNs) [12]. The classifer consists of a series of temporal convolutional blocks. Each block consists of layers of 1D convolutional layers operating in the temporal domain with ReLU non-linearity. We employ batch normalization for each layer. Residual connections are employed from one block to the next. After every block, average pooling is employed to reduce the number of frames by half. Finally a fully-connected (FC) layer with softmax is used to map to a probability distribution over the classes. As there may be a domain shift between the original actions and the reconstructed actions from the autoencoder, we train a single classifier on the reconstructed actions from the autoencoder trained on complete action sequences. Other training details are provided in the supplementary material.

Visualization of high-level semantic features: We also visualize the effectiveness of the reconstructions for downstream applications, like classification, with t-SNE embeddings in 2D [19]. We use the feature maps of the penultimate layer of the trained classifier.

6.2. Self-similarity matrix

In order to better measure the differences in the dynamics of the reconstructed actions for the baseline and proposed methods compared to the ground-truth, we propose to use self-similarity matrices (SSMs) [11]. SSMs capture dynamics better than using just classification accuracy, and at the same time, can be easily visualized. Once the SSMs are constructed, we use Euclidean distance between the respective SSMs to measure the difference between ground-truth and reconstructed actions. That is, for two sequences X_1, X_2, the SSM distance is $\|SSM(X_1) - SSM(X_2)\|_2$. $SSM(X) \in \mathbb{R}^{N\times N}$, $SSM(X)_{i,j} = e^{-\|X_i - X_j\|}$, where X_i is the i^{th} frame of X.

7. Datasets and experimental results

7.1. HDM05 mocap dataset [21]

Dataset details: HDM05 is a large publicly available and challenging database of 3D human actions with 2337 action samples. There are 130 different types of actions performed by 5 subjects and recorded in a laboratory setting using an optical motion capture system. Each skeleton is made up of 31 joints. For our experiments, we resample all the actions such that the length of the every action sequence $N = 100$. Thus $X, Y, \hat{X} \in \mathbb{R}^{93\times 100}$. All sequences are normalized so that the hip joint is fixed at the origin in 3D space. We perform 5-fold cross-validation. For each run, we use 4 subjects for training and the remaining subject for testing with about 1850 samples for training and the rest for testing.

Network architectures: The encoder consists of 4 temporal convolution layers with filter size of 4, and the number of feature maps in each layer is set to 75 (equal to the number of channels at the input layer). We use a latent space dimension of 200. The decoder consists of 4 temporal transposed convolutional layers. For our experiments, we train multiple autoencoders with actions consisting of random subsets of joint sequences sampled from the actions. Different fractions of joints included for training each autoencoder: $100\%, 90\%, 75\%, 50\%$. We will use the term Observed-to-Total Percentage (OTP) = $\frac{K}{J} \times 100$, to denote this quantity. For classification, we use a TCN classifier similar [12]. It consists of 3 TCN blocks with one convolutional layer each.

Reconstruction performance and visualization: A simple feed-forward pass through the trained autoencoders (trained with different fractions of observed joints) serves as a coarse reconstruction and is used as a baseline. As explained in Section 5.1, using our proposed method, we can achieve significantly better reconstructions by using an optimization procedure over the latent space of the autoencoder model in order to minimize the Euclidean distance between the reconstruction and input test sequence with only a subset of observed joint trajectories. Note that the parameters of the generative model, the decoder D in our case, are held fixed for this optimization. We use Adam optimizer for 500 iterations with an initial learning rate of 1.0. As the initialization, we use the latent representation of the incomplete action obtained by a feed-forward pass through the encoder. We carry out reconstruction experiments on the test set for autoencoders trained with different fractions of observed joints.

In order to better test the generalization ability, we use test-time OTPs that are different from training-time OTPs.

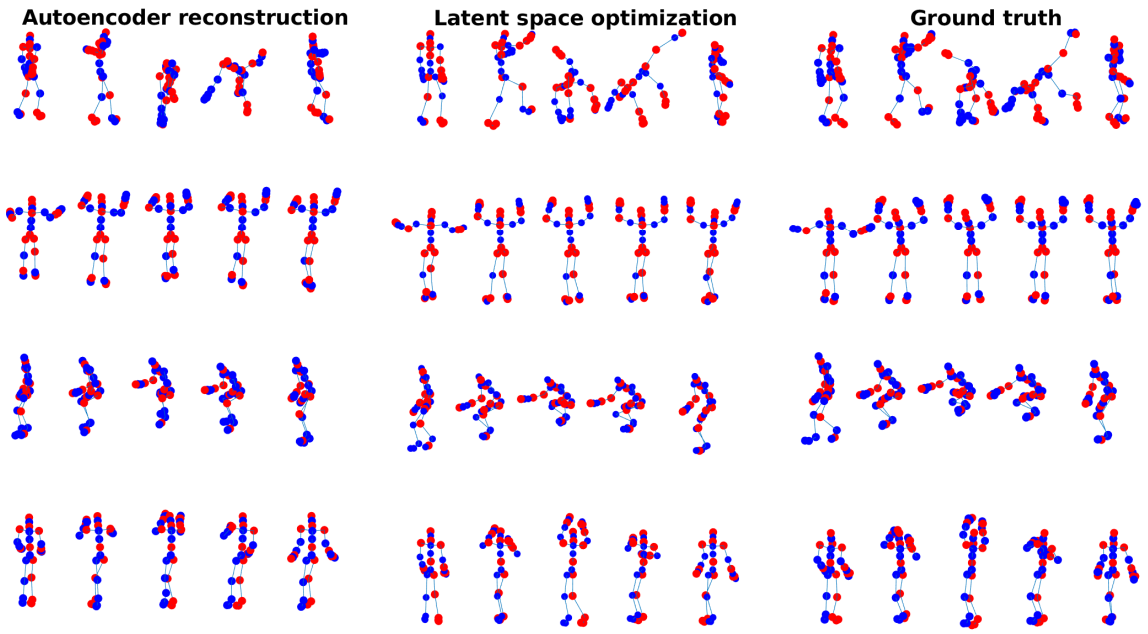
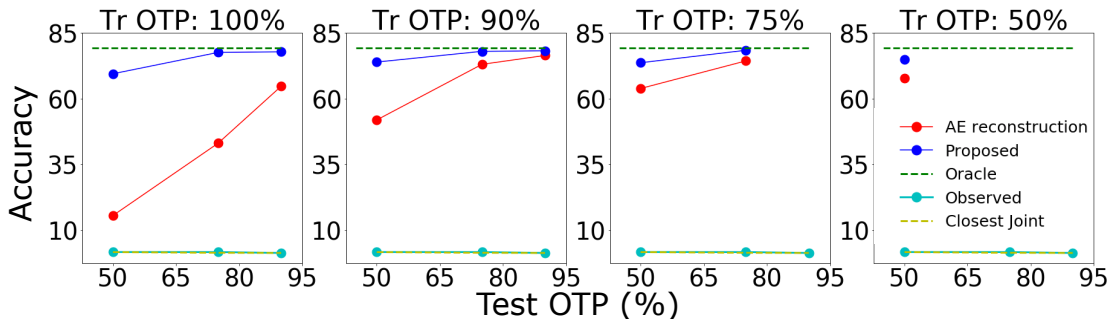


Figure 2: Reconstructed actions for the HDM05 database for train OTP / test OTP = $90/50$. From the top row, the actions shown are "Cartwheel" ,"Hand waving", "Grab low" and "Throw a basketball". The first column shows the reconstructions obtained by a simple feedforward pass through the trained autoencoder: $D(E(Y))$. The middle column shows the output of the proposed approach which solves the optimization problem in Equation 3: $D(\mathbf{z}^*)$. Blue dots represent the observed joints and red dots represent the unobserved. We clearly observe that the optimization approach produces superior reconstructions.

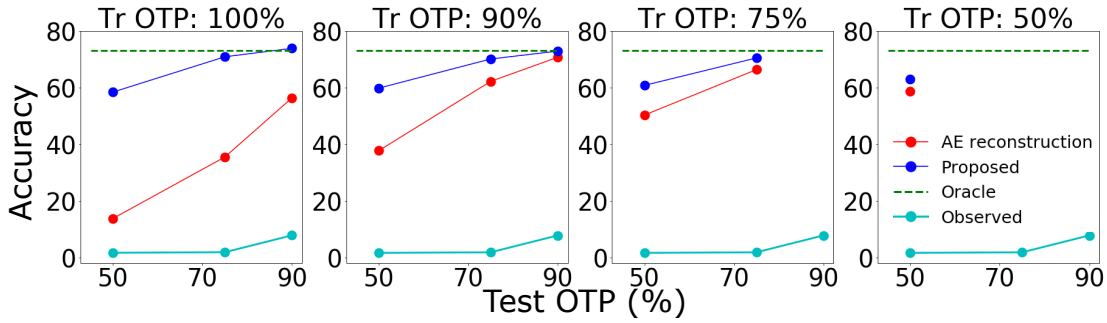
For the reconstructed sequences thus obtained, we use a pre-trained classifier to classify the test set reconstruction into one of the pre-defined 130 classes for different variants and test-time OTPs of the autoencoders. The results for recognition performance are shown in Figure 3a, and a few sample reconstructed skeletal sequences are shown in Figure 2. We also compute self similarity matrices (SSMs) for (1) baseline reconstruction, (2) output of the proposed method and (3) ground-truth sequences and compute distances between them. The SSM metrics thus obtained averaged over all the folds are shown in Figure 5a and visualized in Figure 6a. Finally, using the penultimate layer output of the classifier, we compute t-SNE embeddings for (1) autoencoder reconstructions (2) final reconstruction after optimizing the latent space, and (3) ground-truth. The results are shown for the test set of one of the folds in Figure 4.

Results: We observe in all cases where either the training or test OTP is less than 100%, the proposed method of solving the optimization problem in Equation (3) i.e., using $D(\mathbf{z}^*)$ leads to significantly better results compared to using a single feed-forward pass through the autoencoder i.e., $D(E(Y))$, where D and E are the decoder and encoder respectively. In almost all cases, using the optimization approach yields accuracies within 5% points of the oracle performance. We also observe for the cases with training OTP = 100%, using a different test OTP causes more degradation in performance than when the train OTP = $90\%, 75\%, 50\%$. This shows that training the autoencoder with incomplete action sequences actually leads to performance improvements in classification, attributed to increased robustness of the classifier, similar to the dropout strategy. We note that we also compared with two additional simpler baselines: (1) perform reconstruction of skeletons per frame by replacing every unseen joint with the closest observed joint in the skeleton, and then use the pre-trained classifier and (2) train the classifier directly on the action sequences with only a subset of joints observed. Irrespective of the train/test OTP, both these baselines fail and yield only accuracies which are close to chance.

Reconstruction using structured masks: In the above, we trained and tested autoencoders with random subsets of joints. In this experiment, we drop joints in a structured fashion during test time. We carry out four sets of experiments with the joints corresponding to the following body parts dropped: right arm (6 joints), left arm (6), right leg (4) and left leg (4). This demonstrates how occlusion of different limbs can affect the performance of our algorithm. Note that the autoencoders were trained on random subsets of joints, as before. The results are shown in Table 1. We see once again that the proposed algorithm yields good recognition performance compared to the baselines considered.



(a) HDM dataset [21] recognition performance



(b) NTU dataset [25] recognition performance

Figure 3: Measuring the performance of recovering dynamic information with classification accuracy for NTU and HDM datasets. We observe that the proposed optimization-based reconstruction is far superior to a feed-forward pass through the autoencoder (AE). As the train Observed-to-Total Percentage (OTP) of joints is reduced, performance degrades more gracefully in the case of the optimization-based approach. In almost all cases, we can get to within 5% points of the oracle performance (train OTP/test OTP = $100/100$) "Observed" refers to passing the observed sequence Y with fewer joints through the classifier. Best viewed in color.

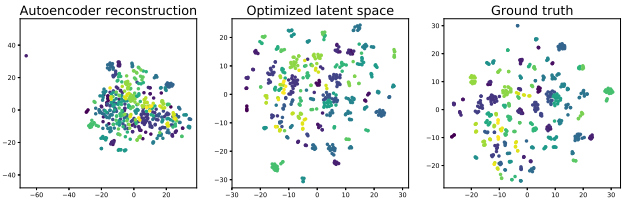


Figure 4: t-SNE embeddings of the penultimate layer of the action classifier for the HDM05 dataset. We see that a lot of semantic information is lost when the joints are dropped, but can be recovered most effectively with an optimized latent space.

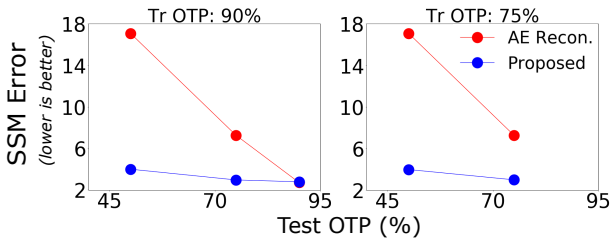
7.2. NTU RGB-D dataset [25]

Dataset details: This is a large database of about 56000 3D skeletal action sequences obtained from Kinect of actions belonging to 60 classes and performed by 45 subjects. For each skeleton, 25 joint locations are provided. We resample all the sequences to have $N = 50$ frames. Thus $X, Y, \hat{X} \in \mathbb{R}^{75\times 50}$. We perform the experiments in the cross-subject setting and use the train-test split as suggested by the authors of the dataset. The training set consists of about 40000 examples and the remaining are in the test set. All sequences are normalized so that the hip is fixed at the origin in 3D space.

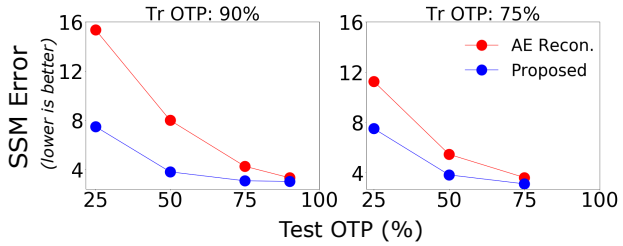
Joints dropped	Train OTP	Method	Accuracy (%)
Right arm	90	$D(E(Y))$	50.05
		$D(\mathbf{z}^*)$	**59.77**
	75	$D(E(Y))$	51.08
		$D(\mathbf{z}^*)$	**63.79**
Left arm	90	$D(E(Y))$	54.48
		$D(\mathbf{z}^*)$	**64.15**
	75	$D(E(Y))$	55.59
		$D(\mathbf{z}^*)$	**66.54**
Right leg	90	$D(E(Y))$	57.50
		$D(\mathbf{z}^*)$	**69.43**
Left leg	90	$D(E(Y))$	57.97
		$D(\mathbf{z}^*)$	**66.29**

Table 1: Average classification results (over 5 folds) of reconstructed actions on the HDM05 database. The inputs are actions with contiguous body parts that are hidden or unobserved. Here, $D(E(Y))$ is the baseline and $D(\mathbf{z}^*)$ is the proposed optimization strategy.

Network architectures: The generative mode is a temporal convolutional autoencoder. The encoder consists of 3 temporal convolution layers with filter size of 8, and the number of feature maps in each layer is set to 75 (equal to



(a) SSM distances for HDM05 dataset



(b) SSM distances for NTU dataset

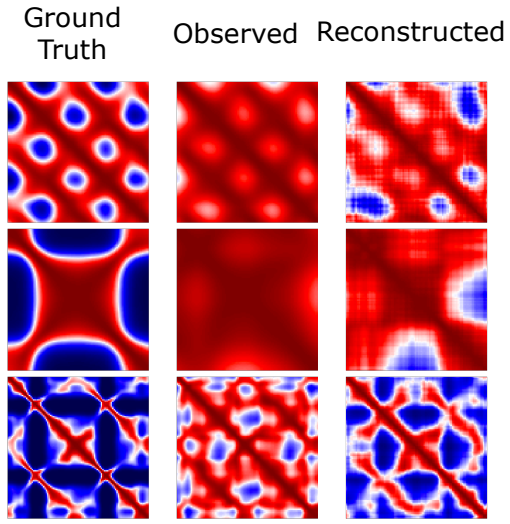
Figure 5: Evaluating the quality of reconstructions on the two datasets considered in this paper, in the self-similarity matrix (SSM) domain for different train-test OTPs. We see a significant improvement in the proposed method to recover dynamics over the baseline: a feed-forward pass through the autoencoder.

the number of channels at the input layer). We use a latent space dimension of 200. The decoder consists of 3 temporal transposed convolutional layers. As the action classifier, we use a TCN classifier identical to that proposed by Kim and Reiter [12]. It consists of 3 TCN blocks with 3 convolutional layers each.

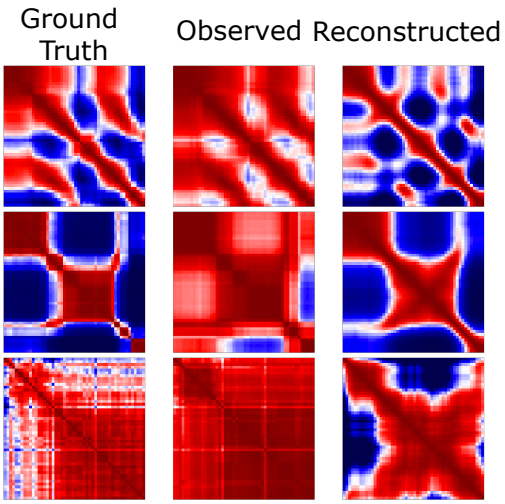
Reconstruction performance: We conduct an identical set of experiments as in the case of the HDM05 dataset. The classfication results are shown in Figure 3b and the SSM-based metrics are shown in Figure 5b and visualized in Figure 6b. Skeletal visualizations are provided in the supplement. The trends observed are the same as those in HDM05. Compared to the baselines of (1) directy passing the observed sequence through the classifier and (2) using the autoencoder reconstruction, the proposed method of latent space optimization achieves far superior results especially when train and test OTPs are considerably different, and gets close to oracle classification performance and SSM distances even with just 75% of observed joints.

8. Conclusion

In this paper, we consider the problem of reconstructing completely unseen dimensions of a multi-variate time series. The problem is traditionally studied in the framework of system identification and non-linear dynamics. However, for tractability, such methods make strong assumptions on the data such as linearity of the underlying dynamical system, sparsity of observations in transform domains etc. In this paper, we consider the specific example of reconstructing unseen joint dynamics from 3D human actions, for which we cannot easily construct hand-crafted priors. Instead, we propose first construct a generative model of complete actions, even when the training data has up to 50% of the joints missing. The reconstruction problem then can be solved by projecting the observed action onto the range of the generative model, which is done via optimization in the latent space. Through extensive experiments and different metrics, we show that the proposed approach can



(a) SSMs for 3 different actions from the HDM dataset [21]



(b) SSMs for 3 different actions from the NTU dataset [25]

Figure 6: We visualize the dynamics of actions using the self-similarity matrices (SSMs) on the two datasets. We see that even though a lot of dynamics are lost in the observed action with missing joints, the proposed method recovers them effectively.

effectively recover the dynamics of unseen joints. An interesting extension of this idea for human actions is to design stronger priors using spatio-temporal graph convolutional autoencoders which can better take into account the skeletal graph structure into account for representation learning.

References

[1] Andreas Aristidou, Daniel Cohen-Or, Jessica K Hodgins, and Ariel Shamir. Self-similarity analysis for motion capture cleaning. In *Computer Graphics Forum*, volume 37, pages 297–309. Wiley Online Library, 2018. 3

[2] Emad Barsoum, John Kender, and Zicheng Liu. HP-GAN: Probabilistic 3d human motion prediction via gan. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 1418–1427, 2018. 3

[3] Alessandro Bissacco. Modeling and learning contact dynamics in human motion. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, volume 1, pages 421–428. IEEE, 2005. 1

[4] Ashish Bora, Ajil Jalal, Eric Price, and Alexandros G Dimakis. Compressed sensing using generative models. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 537–546. JMLR. org, 2017. 3, 4

[5] Ashish Bora, Eric Price, and Alexandros G Dimakis. Ambientgan: Generative models from lossy measurements. *ICLR*, 2:5, 2018. 3

[6] Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang. Image super-resolution using deep convolutional networks. *IEEE transactions on pattern analysis and machine intelligence*, 38(2):295–307, 2015. 3

[7] Katerina Fragkiadaki, Sergey Levine, Panna Felsen, and Jitendra Malik. Recurrent network models for human dynamics. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4346–4354, 2015. 3

[8] Øyvind Gløersen and Peter Federolf. Predicting missing marker trajectories in human motion data using marker intercorrelations. *PloS one*, 11(3):e0152616, 2016. 3

[9] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014. 4

[10] Daniel Holden, Jun Saito, and Taku Komura. A deep learning framework for character motion synthesis and editing. *ACM Transactions on Graphics (TOG)*, 35(4):138, 2016. 3

[11] Imran N Junejo, Emilie Dexter, Ivan Laptev, and Patrick PÚrez. Cross-view action recognition from temporal self-similarities. In *European Conference on Computer Vision*, pages 293–306. Springer, 2008. 5

[12] Tae Soo Kim and Austin Reiter. Interpretable 3d human action analysis with temporal convolutional networks. In *2017 IEEE conference on computer vision and pattern recognition workshops (CVPRW)*, pages 1623–1631. IEEE, 2017. 1, 5, 8

[13] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013. 4

[14] Taras Kucherenko, Jonas Beskow, and Hedvig Kjellström. A neural network approach to missing marker reconstruction in human motion capture. *arXiv preprint arXiv:1803.02665*, 2018. 3

[15] Kuldeep Kulkarni, Suhas Lohit, Pavan Turaga, Ronan Kerviche, and Amit Ashok. Reconnet: Non-iterative reconstruction of images from compressively sensed measurements. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 449–458, 2016. 3

[16] Christian Ledig, Lucas Theis, Ferenc Huszár, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, et al. Photo-realistic single image super-resolution using a generative adversarial network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4681–4690, 2017. 3

[17] Lennart Ljung. System identification. *Wiley Encyclopedia of Electrical and Electronics Engineering*, 2001. 3

[18] Yonghong Luo, Xiangrui Cai, Ying Zhang, Jun Xu, et al. Multivariate time series imputation with generative adversarial networks. In *Advances in Neural Information Processing Systems*, pages 1596–1607, 2018. 3

[19] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(Nov):2579–2605, 2008. 5

[20] Chris Metzler, Ali Mousavi, and Richard Baraniuk. Learned d-amp: Principled neural network based compressive image recovery. In *Advances in Neural Information Processing Systems*, pages 1772–1783, 2017. 3

[21] M. Müller, T. Röder, M. Clausen, B. Eberhardt, B. Krüger, and A. Weber. Documentation mocap database HDM05. Technical Report CG-2007-2, Universität Bonn, June 2007. 5, 7, 8

[22] Deepak Pathak, Philipp Krahenbuhl, Jeff Donahue, Trevor Darrell, and Alexei A Efros. Context encoders: Feature learning by inpainting. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2536–2544, 2016. 3, 4

[23] JH Rick Chang, Chun-Liang Li, Barnabas Poczos, BVK Vijaya Kumar, and Aswin C Sankaranarayanan. One network to solve them all–solving linear inverse problems using deep projection models. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 5888–5897, 2017. 3

[24] Viraj Shah and Chinmay Hegde. Solving linear inverse problems using gan priors: An algorithm with provable guarantees. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4609–4613. IEEE, 2018. 3

[25] Amir Shahroudy, Jun Liu, Tian-Tsong Ng, and Gang Wang. NTU RGB+D: A large scale dataset for 3D human activity analysis. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1010–1019, 2016. 7, 8

[26] Yansong Tang, Yi Tian, Jiwen Lu, Peiyang Li, and Jie Zhou. Deep progressive reinforcement learning for skeleton-based action recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5323–5332, 2018. 1

[27] Dmitry Ulyanov, Andrea Vedaldi, and Victor Lempitsky. Deep image prior. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 9446–9454, 2018. 3

[28] Raviteja Vemulapalli, Felipe Arrate, and Rama Chellappa. Human action recognition by representing 3d skeletons as points in a lie group. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 588–595, 2014. 1

[29] Vinay Venkataraman and Pavan Turaga. Shape distributions of nonlinear dynamical systems for video-based inference. *IEEE transactions on pattern analysis and machine intelligence*, 38(12):2531–2543, 2016. 1

[30] Alan Wolf, Jack B Swift, Harry L Swinney, and John A Vastano. Determining lyapunov exponents from a time series. *Physica D: Nonlinear Phenomena*, 16(3):285–317, 1985. 1

[31] Sijie Yan, Yuanjun Xiong, and Dahua Lin. Spatial temporal graph convolutional networks for skeleton-based action recognition. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018. 1

[32] Jingyu Yang, Xin Guo, Kun Li, Meiyuan Wang, Yu-Kun Lai, and Feng Wu. Spatio-temporal reconstruction for 3d motion recovery. *IEEE Transactions on Circuits and Systems for Video Technology*, 2019. 3

[33] Jinsung Yoon, James Jordon, and Mihaela Van Der Schaar. Gain: Missing data imputation using generative adversarial nets. *International Conference on Machine Learning*, 2018. 3

[34] Jian Zhang and Bernard Ghanem. ISTA-net: Interpretable optimization-inspired deep network for image compressive sensing. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1828–1837, 2018. 3

[35] Pengfei Zhang, Cuiling Lan, Junliang Xing, Wenjun Zeng, Jianru Xue, and Nanning Zheng. View adaptive neural networks for high performance skeleton-based human action recognition. *IEEE transactions on pattern analysis and machine intelligence*, 2019. 1

[36] Xikang Zhang, Yin Wang, Mengran Gou, Mario Sznaier, and Octavia Camps. Efficient temporal sequence comparison and classification using gram matrix embeddings on a riemannian manifold. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4498–4507, 2016. 1