

# Human Pose Forecasting via Deep Markov Models

Sam Toyer\*, Anoop Cherian\*<sup>†</sup>, Tengda Han\*, and Stephen Gould\*<sup>†</sup>

\*The Australian National University

<sup>†</sup>Australian Centre for Robotic Vision

{sam.toyer, anoop.cherian, tengda.han, stephen.gould}@anu.edu.au

**Abstract**—Human pose forecasting is an important problem in computer vision with applications to human-robot interaction, visual surveillance, and autonomous driving. Usually, forecasting algorithms use 3D skeleton sequences and are trained to forecast for a few milliseconds into the future. Long-range forecasting is challenging due to the difficulty of estimating how long a person continues an activity. To this end, our contributions are threefold: (i) we propose a generative framework for poses using variational autoencoders based on Deep Markov Models (DMMs); (ii) we evaluate our pose forecasts using a pose-based action classifier, which we argue better reflects the subjective quality of pose forecasts than distance in coordinate space; (iii) last, for evaluation of the new model, we introduce a 480,000-frame video dataset called *Ikea Furniture Assembly* (Ikea FA), which depicts humans repeatedly assembling and disassembling furniture. We demonstrate promising results for our approach on both Ikea FA and the existing NTU RGB+D dataset.

## I. INTRODUCTION

Deep learning methods have enabled significant advances in a variety of problems in computer vision, including 2D human pose estimation. The impact of deep methods on this problem has been so great that state-of-the-art accuracy on existing pose estimation benchmarks is nearing human performance [1], [2], [3]. This has enabled researchers to look at more advanced scenarios than single-frame pose estimation. This includes pose estimation on video sequences [4], [5], real-time estimation of poses [6], multi-person pose estimation [7], [8], and recently human pose forecasting [9], [10], which is the central theme of this paper.

Forecasting of human poses is useful in a variety of scenarios in computer vision and robotics, including but not limited to human-robot interaction [11], action anticipation [12], [13], visual surveillance, and computer graphics. For example, consider a robot designed to help a surgeon manage their tools: it is expected that the robot forecasts the position of the limbs of the surgeon so that it can deliver the tools in time. Pose forecasting is also useful for proactive decision-making in autonomous driving systems and visual surveillance.

While the problem of human motion modelling and forecasting has been explored in the past [14], it was limited to simple scenarios such as smoothly varying poses under periodic motions [15]. As a result of the innovations in deep methods and the availability of huge datasets [16], this problem is beginning to be investigated with a renewed interest. For example, Fragkiadaki *et al.* show how to learn the dynamics of human motion using an encoder-decoder framework for recurrent neural networks [9], while Jain *et al.* present an alternative approach which is able to incorporate high-level semantics of human dynamics into a recurrent network via spatio-temporal graphs [10]. Although these newer approaches

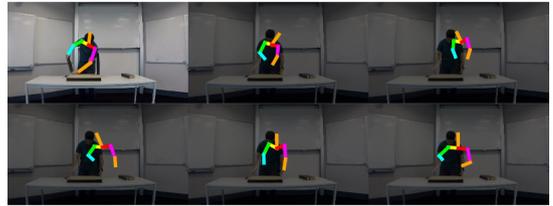


Fig. 1. Five sampled Deep Markov Model (DMM) continuations for a sequence from our proposed dataset, Ikea Furniture Assembly. The top left frame shows the ground truth pose and image 7.5s after the beginning of the forecast period. The remaining frames depict the corresponding pose from each of the five sampled continuations.

have obtained promising results, all of them are purely deterministic models which can only predict a single continuation of an observed sequence, and are thus unable to account for the stochasticity inherent in human motion. Both Fragkiadaki *et al.* and Jain *et al.* observe that this stochasticity can lead simple deterministic models to produce pose sequences which rapidly converge to the mean, but they do not attempt to resolve this issue at a fundamental level by explicitly modelling stochasticity. A similar issue is present in the evaluation of forecasted poses: most existing work compares forecasted poses to a single ground truth sequence, when in reality there are often many plausible continuations.

In this paper, we propose to address both of these problems. First, to resolve the stochasticity problem, we make use of a Deep Markov Model (DMM) [17], which allows us to sample arbitrarily many plausible continuations for an observed pose sequence. Second, on the evaluation side, we use an RNN that takes a pose sequence and returns an action label. We evaluate the quality of the forecasted pose in terms of the classification accuracy of the evaluation RNN against the ground truth forecasted action. This allows us to gauge the intuitive plausibility of the continued sequence without penalising the model for choosing reasonable continuations which do not coincide precisely with the true one.

We evaluate our proposed forecasting model on poses generated over one existing dataset and one new one. The existing dataset is NTU RGB+D [18], which includes over 5,000,000 video frames depicting a wide variety of actions, as well as Kinect poses for all subjects in each frame. While NTU RGB+D is challenging, sequences in NTU RGB+D are typically very short, and lack the kind of regularity and repetition of actions which pose prediction mechanisms ought to be able to exploit. As such, we propose a novel human pose forecasting dataset *Ikea Furniture Assembly*, that consists of much longer sequences depicting 4–12 actions, each concerned with assembling a small table. Figure 1 shows frames from one such sequence. While this dataset provides a simple baseline,

we believe the problem of pose forecasting is in its infancy, and that our dataset thus provides a good platform for systematic evaluation of the nuances of the problem. Our experiments on these datasets reveal that our proposed method demonstrates state-of-the-art accuracy.

## II. BACKGROUND AND RELATED WORK

*a) Pose estimation:* Estimating 2D human pose from monocular images is a well-studied problem. Approaches for static images include pictorial structure models [19], [20], random forests [21], and coordinate [22] or heatmap [9], [1], [2] regression with CNNs. Some approaches are able to exploit motion information by imposing high-level graphical models [23] or back-warping joint heatmaps with flow [5]; however, state-of-the-art approaches like convolutional pose machines [1] and stacked hourglass networks [2] still look at only a single frame at a time, and do not model the temporal evolution of poses.

*b) Mocap modelling and synthesis:* Modelling sequences of 3D motion capture vectors is both useful in its own right—in animation, for instance [24]—and useful as a benchmark for generic sequence learning techniques. Past approaches to this problem include Gaussian processes [24], switching linear dynamic systems [25] and Boltzmann machines [26].

It should be noted that our work differs from the majority of existing literature on mocap synthesis in two respects. First, we consider the problem of 2D pose estimation in image-relative coordinates, as opposed to 3D pose estimation in scene-relative coordinates. Second, our proposed benchmark does not allow algorithms to make use of a ground-truth pose sequence at test time, thereby forcing them to learn to exploit visual cues.

In the vision community, mocap modelling has sometimes been used as a motivating application of temporally aware neural network architectures. In addition to 2D pose estimation and forecasting, Fragkiadaki *et al.* showcase their RNN-based sequence learning framework by using it to extend sequences of 3D motion capture data [9]. Similarly, Jain *et al.* apply their general structural RNN framework to the same task, and report aesthetically favourable results relative to Fragkiadaki *et al.* [10]. Martinez *et al.* present a third approach specifically tailored to pose forecasting, again reporting improvements over Fragkiadaki *et al.* and Jain *et al.*, particularly in the level of error on the first few frames of a sequence [27].

*c) Pixel-wise prediction:* Recent work in motion representation has focused on anticipating future appearance, motion or other attributes at the pixel level. Structured random forests [28], feed-forward CNNs [29] and variational autoencoders [30] have all been used to predict dense pixel trajectories from single frames. Prediction of raw pixels—as opposed to pixel trajectories—has been attempted using ordinary feedforward networks, GANs [31], [32] and RNNs [33], [34].

Even over short time horizons, pixel-wise prediction remains a challenging task. State-of-the-art models are capable of predicting up to a second ahead, albeit often with significant blurring or other visual artefacts [32], [34]. We hope that, by tackling a narrower problem, pose forecasting models will be

better able to learn long-range dynamics than models for pixel-wise appearance prediction.

*d) Sequence modelling with VAEs:* Variational Autoencoders (VAEs) are a recent neural-network-based approach to modeling complex, but potentially structured, probability distributions. An ordinary regressor or autoencoder would be trained to minimise the distance between its output and the dataset  $X$ , whereas a variational autoencoder is trained to maximise the variational lower bound  $\log p(X) - \text{KL}(q(Z | X) || p(Z | X))$ , where  $q(Z | X)$  is a (learned) variational approximation to the intractable posterior  $p(Z | X)$  over latent variables  $Z$ . Not only is this lower bound to the log likelihood efficient to maximise [35], [36], but doing so confers advantages over both ordinary regressors or autoencoders:

- VAEs are able to place a fixed prior  $p(Z)$  on latent variables as part of the optimisation process. This enables a single output  $x$  to be efficiently sampled by choosing  $z \sim p(z)$ , then  $x \sim p(x | z)$ .
- In the presence of uncertainty, regressors trained to minimise  $\ell_2$  error will tend to produce the mean of all plausible outputs, which may or may not be a plausible output itself [37]. In contrast, VAEs are able model multimodal output distributions, thereby yielding crisper predictions.

The attractiveness of VAEs has led to a spate of new approaches to stochastic sequence modeling, including STORNs [38], VRNNs [39], SRNNs [40] and DMMs [17]. These approaches differ in terms of the information which their respective generators condition on between time steps, as well as the architectures of their (approximated) inference networks. We have chosen to use DMMs in this paper because of their strong performance in standard benchmark tasks and their relative simplicity.

## III. PROPOSED METHOD

In Figure 2, we show the complete pose forecasting system. First, images are passed through a pose estimation system to obtain a sequence of poses. The observed poses are then passed to a sequence learning model to infer a latent representation capturing the person’s anticipated motion after the final frame. This latent representation can then be extended out to an arbitrarily long sequence of poses using the sequence learning model’s generative capabilities. Last, the sequence of forecasted poses is evaluated by a pose-based action classifier.

### A. Pose estimation

To turn an observed sequence of images into a sequence of poses, we employ Wei *et al.*’s Convolutional Pose Machines (CPMs) [1]. Specifically, a four-stage CPM is used to centre detected person bounding boxes on their subjects, followed by a six-stage model for pose estimation. To overcome temporal instability in the estimations, the generated pose sequences are smoothed using a weighted moving average.

### B. Pose forecast model

The heart of our proposed system is a pose forecasting network based on Krishnan *et al.*’s Deep Markov Models (DMM) [17]. For the sake of understanding how the DMM

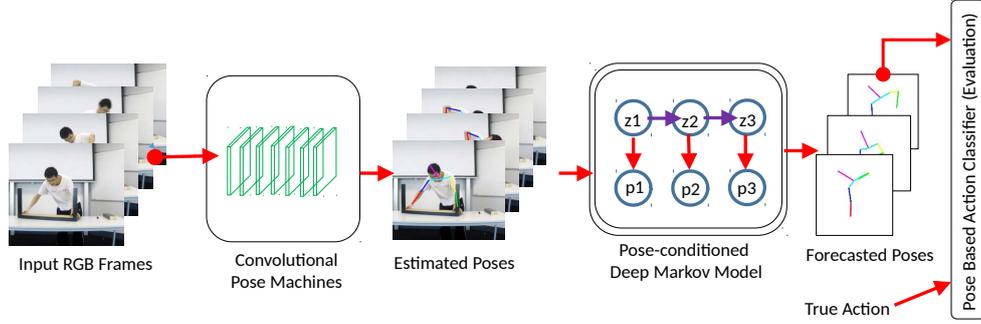


Fig. 2. System components and their relations.

theory pertains to pose estimation, we can define the task of pose forecasting formally: let  $p_{1:T}$  denote a sequence of poses  $p_1, \dots, p_T$  and  $z_{1:T}$  denote a corresponding series of latent variables  $z_1, \dots, z_T$ . Pose forecasting is the task of sampling from the joint distribution  $p(p_{1:T}, z_{1:T})$ , where  $p$  is assumed to factorise according to

$$p(p_{1:T}, z_{1:T}) = p(z_1) p(p_1 | z_1) \times \prod_{t=2}^T p(p_t | z_t) p(z_t | z_{t-1}). \quad (1)$$

The fourth panel of Figure 2 depicts this factorisation as a Bayesian network. From this Bayesian network, it follows that the posterior inference distribution  $p(z_{1:T} | p_{1:T})$  factorises to

$$p(z_{1:T} | p_{1:T}) = p(z_1 | p_{1:T}) \prod_{t=2}^T p(z_t | z_{t-1}, p_{t:T}). \quad (2)$$

1) *Variational approximation:* For learning of these distributions to be tractable, we must make several approximations. First, we can use the following variational approximations in place of the true generative distributions  $p(p_t | z_t)$ ,  $p(z_1)$ , and  $p(z_t | z_{t-1})$ , respectively:

$$p(p_t | z_t; \theta) = \mathcal{N}(\mu_E(z_t; \theta), \Sigma_E(z_t; \theta)) \quad (3)$$

$$p(z_1; \theta) = \mathcal{N}(\mu_{T_0}, \Sigma_{T_0}) \quad (4)$$

$$p(z_t | z_{t-1}; \theta) = \mathcal{N}(\mu_T(z_{t-1}; \theta), \Sigma_T(z_{t-1}; \theta)) \quad (5)$$

The generative model parameters  $\theta$  parametrise neural networks which are able to calculate the means ( $\mu_{T_0}$ ,  $\mu_E$ , etc.) and covariances ( $\Sigma_{T_0}$ ,  $\Sigma_E$ , etc.) of the normal transition and emission distributions ( $\mathcal{N}(\mu_{T_0}, \Sigma_{T_0})$ , etc.). Similarly, we can replace the (intractable) posterior inference distributions  $p(z_1 | p_{1:T})$  and  $p(z_t | z_{t-1}, p_{t:T})$  with two more learnt variational approximations expressed in terms of inference network parameters  $\phi$ :<sup>1</sup>

$$q(z_1 | p_{1:T}; \phi) = \mathcal{N}(\mu_{I_0}(p_{1:T}; \phi), \Sigma_{I_0}(p_{1:T}; \phi)) \quad (6)$$

$$q(z_t | z_{t-1}, p_{t:T}; \phi) = \mathcal{N}(\mu_I(z_{t-1}, p_{t:T}; \phi), \Sigma_I(z_{t-1}, p_{t:T}; \phi)) \quad (7)$$

<sup>1</sup>As there are many means and covariances produced by this model, it may help to reader to know that the subscript  $E$  has been used for the mean and covariance of the emission distribution,  $T$  for those of the transition distribution (in the generative model), and  $I$  for those of inference distribution (in the inference model).  $T_0$  and  $I_0$  refer to the initial latent distribution in the generative model and inference model, respectively.

Parameters  $\theta$  and  $\phi$  can be optimised by gradient ascent to maximise the variational lower bound  $-\mathcal{L}(p_{1:T}; \theta, \phi)$  of Kingma et al. [35]:

$$\begin{aligned} \log p(p_{1:T}; \theta) &\geq -\mathcal{L}(p_{1:T}; \theta, \phi) \\ &= \mathbb{E}_{z_{1:T} \sim q_\phi(z_{1:T} | p_{1:T}; \phi)} [\log p(p_{1:T} | z_{1:T}; \theta)] \\ &\quad - \text{KL}[q(z_{1:T} | p_{1:T}; \phi) \| p(z_{1:T}; \theta)] \end{aligned} \quad (8)$$

Where  $\text{KL}[f(\cdot) \| g(\cdot)]$  denotes the KL divergence between densities  $f$  and  $g$ , and  $\mathcal{L}(p_{1:T}; \theta, \phi)$  is shorthand for the variational lower bound itself.

We can further factorise this variational lower bound using the conditional dependencies identified above:

$$\log p(p_{1:T} | z_{1:T}; \theta) = \sum_{t=1}^T \log(p(p_t | z_t; \theta)) \quad (9)$$

$$\begin{aligned} &\text{KL}[q(z_{1:T} | p_{1:T}; \phi) \| p(z_{1:T}; \theta)] \\ &= \text{KL}[q(z_1 | p_{1:T}; \phi) \| p(z_1; \theta)] \\ &\quad + \mathbb{E}_{z_{1:T} \sim q(z_{1:T} | p_{1:T}; \phi)} \\ &\quad \sum_{t=2}^T \text{KL}[q(z_t | z_{t-1}, p_{t:T}; \phi) \| p(z_t | z_{t-1}; \theta)] \end{aligned} \quad (10)$$

Recall that all involved distributions are multivariate Gaussians, and so each KL divergence has a closed form which is amenable to optimisation with stochastic gradient descent. Further, when training with SGD, it is generally sufficient to approximate the expectation over the  $z_t$ s with a single sample from  $q(z_{1:T} | p_{1:T})$ , and it is possible to perform stochastic backpropagation *through* this sampling operation using the reparametrisation trick [35], [36].<sup>2</sup>

Taken together, these tricks make it possible to efficiently train a neural network to optimise the DMM objective  $\mathcal{L}(\mathcal{P}; \theta, \phi)$  over the full dataset  $\mathcal{P} = \{p_{1:T}^{(1)}, \dots, p_{1:T}^{(N)}\}$ :

$$\mathcal{L}(\mathcal{P}; \theta, \phi) = \sum_{i=1}^N \mathcal{L}(p_{1:T}^{(i)}; \theta, \phi) \quad (11)$$

<sup>2</sup>Readers who are completely unfamiliar with these techniques will likely appreciate the context provided by Doersch’s tutorial-style treatment of variational autoencoders [37].

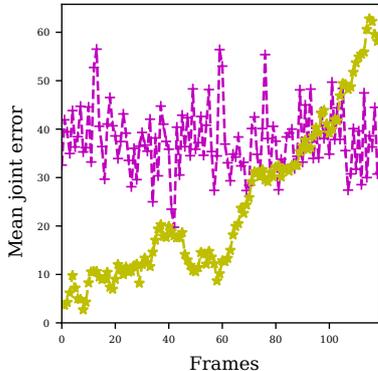
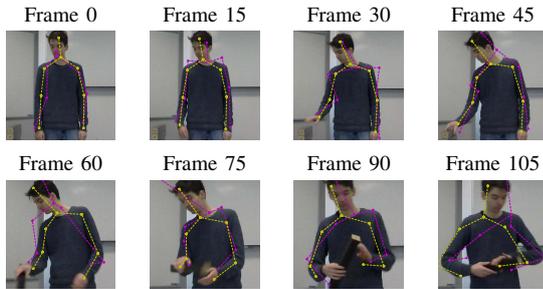


Fig. 3. Two possible continuations of a pose sequence, and the corresponding  $\ell_2$  distance between the continuations and the ground truth.

2) *Network architecture*: To calculate the various  $\mu$ s and  $\Sigma$ s required by the variational approximation, we use a set of networks based on Krishnan et al.’s ST-LR model [17]:

$\{\mu, \Sigma\}_E(z_t; \theta)$ : The mean and covariance of the emission Gaussian are calculated by a simple multilayer perceptron.

$\{\mu, \Sigma\}_{T_0}$ : The mean and covariance of the first latent vector can be learnt directly, without need for an MLP.

$\{\mu, \Sigma\}_T(z_t; \theta)$ : The transition function is reminiscent of a GRU: it uses  $z_{t-1}$  to compute hidden activations  $h_t$  and gating activations  $g_t$ , then uses elementwise multiplication with the  $g_t$  to trade off the contributions of  $z_{t-1}$  and  $h_t$  to the new hidden state. See [17] for details.

$\{\mu, \Sigma\}_{I_0}(z_{t-1}, p_{t:T}; \phi)$  and  $\{\mu, \Sigma\}_I(p_{1:T}; \phi)$ : Both pairs of means and covariances are calculated using the same Bidirectional Recurrent Neural Network (BRNN), which processes a pose sequence and then yields  $z_1, z_2, \dots, z_T$  in order. Note that the BRNN outputs  $z_t$  after processing actions and poses from both the future *and* the past: that is,  $z_t$  is calculated from all of  $p_{1:T}$ , rather than just the  $p_{t:T}$  and  $z_{t-1}$  on which it is probabilistically dependent. We retain the  $p_{t:T}$  notation in the discussion above for theoretical consistency.

### C. Classifier-evaluator

Evaluation of forecasted poses is difficult. The most natural approach is to adopt a geometric measure of error and use that to compare different approaches. For instance, Fragkiadaki et al. [9] uses a standard location-based measure of pose estimation accuracy to report the performance of a 2D pose forecasting system, while Jain et al. [10] evaluate a 3D pose forecasting system using differences in joint angles. Unfortunately, neither of these approaches is adequate to capture human intuition about what constitutes a natural—or even plausible—continuation of a pose sequence.

Method	Accuracy	
	Ikea FA	NTU RGB+D
Ground truth	91.0%	83.4%
Zero-velocity	81.8%	73.3%
DMM	<b>75.8%</b>	<b>70.3%</b>
ERD	75.2%	66.9%
LSTM-3LR	61.9%	53.5%
LSTM	54.5%	67.9%

Table I. Action classifier results for both forecasted and ground-truth pose sequences. “Ground truth” shows the accuracy of the learnt classifier on the true pose sequences, and thus serves as an ideal for predictors to meet.

As an example of this issue, take the two pose continuation strategies in Figure 3: the magenta continuation shifts each joint of each pose independently by a number of pixels chosen from  $\mathcal{N}(0, 20^2)$  (again independently in both  $x$  and  $y$  dimensions). On the other hand, the yellow continuation is simply the ground truth with  $\mathcal{N}(0, 3^2)$  additional pixels of drift accumulated at each time step. The latter clearly produces a more coherent pose sequence, but is quickly exceeded in “accuracy” by the former. This plainly illustrates the limitations of the  $\ell_2$  distance metric for capturing the subjective quality of pose sequence continuations in the presence of drift. Further, although it is not illustrated in Figure 3,  $\ell_2$  distance can also penalise plausible, but incorrect predictions; for instance, if a subject begins to move to the left during a forecast was instead predicted to move to the right, it may still be possible to produce an intuitively plausible continuation, but the continuation would be penalised heavily by  $\ell_2$  distance.

Instead of measuring  $\ell_2$  distance, we score forecasted poses by checking whether they are consistent with the action(s) they represent. This is achieved by passing the poses through a recurrent classifier-evaluator which is trained to recover action labels from pose sequences. The recovered action labels associated with the forecasted pose sequences can be compared with the true labels, providing a robust measure of the realism of the pose sequence. The classifier-evaluator is trained using ground truth poses and action labels from the same dataset used to train the corresponding pose forecast model(s). However, it is *not* trained adversarially, as training it in tandem with any one forecast model would compromise its ability to provide a model-independent measure of pose realism.

## IV. EXPERIMENTS

### A. Datasets

a) *Ikea Furniture Assembly*: As alluded to earlier, there are several challenging benchmark datasets for both action recognition and pose estimation. However, we believe these datasets are too demanding for a problem such as pose forecasting. This is because most of the existing video benchmarks that include both poses and actions (such as Penn Actions [41], the MPII Continuous Pose Dataset [42], or the JHMDB dataset [43]) have either sequences that undergo significant occlusions of body-parts, or include activities that are hard to discriminate. As a result, it may be difficult to understand or separate the challenge imposed by pose forecasting from that imposed by other tasks.

Towards this end, we propose a new dataset called *Ikea Furniture Assembly* (Ikea FA) that consists of basic actions of individuals assembling a small piece of Ikea furniture. Our

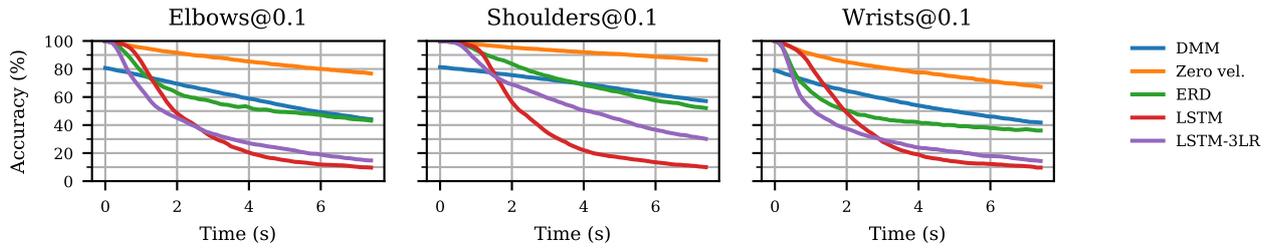


Fig. 4. Percentage Correct Keypoints (PCK) at different times and a fixed threshold on Ikea FA, for a range of methods.

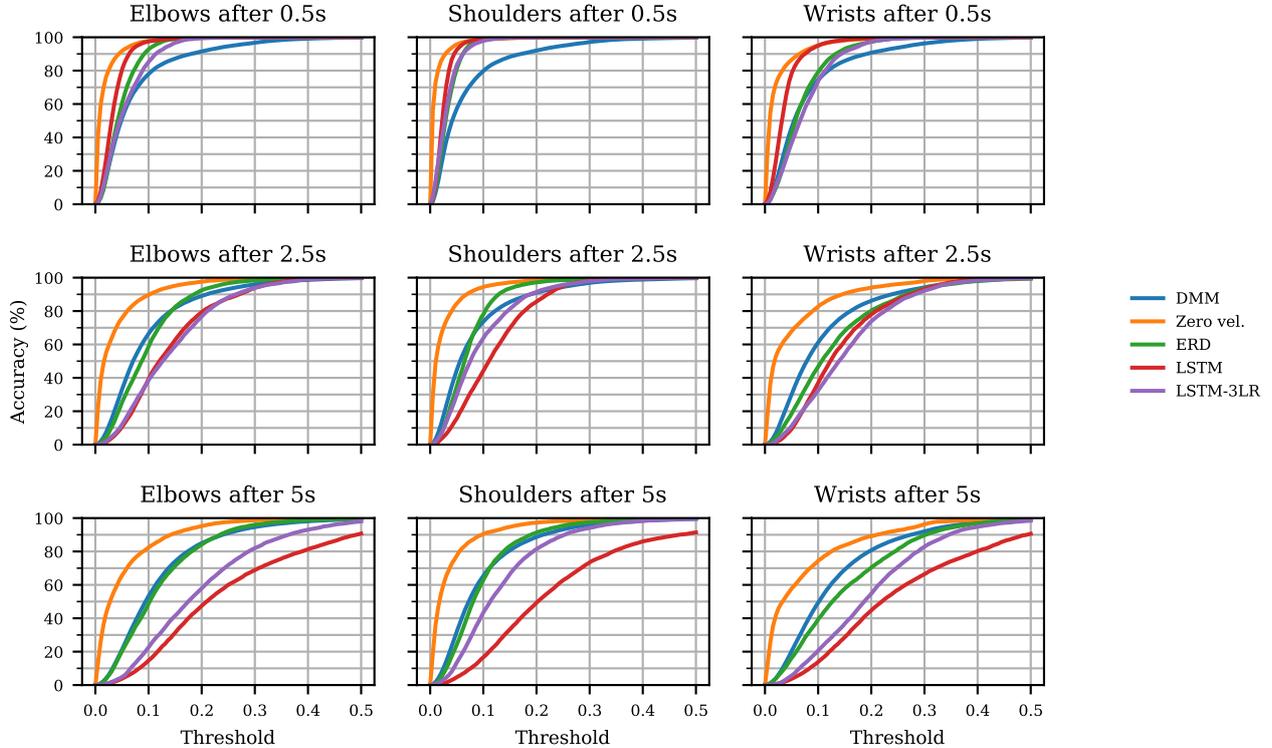


Fig. 5. PCK at different thresholds and a handful of fixed times on Ikea FA, again with a range of methods.

goal with this dataset is to predict the pose of a subject after a few observed frames containing an assembling action. There are 101 videos in the dataset, each about 2–4 minutes long, and shot at 30 frames per second (although we downsample to 10fps in our experiments to maximise the duration for which it is computationally feasible to predict). There are 14 actors in the videos, and we use sequences from 11 actors for training and validation while testing on the rest. Half of the sequences show assembly on the floor, while the other half show assembly on a workbench. As the frames in Figure 9 indicate, floor sequences tend to provide more challenging pose variations than table ones. The dataset is available for download on the web.<sup>3</sup>

The original Ikea FA action labels includes four “attach leg” actions (one for each leg), four “detach leg” actions, a “pick leg” action, a “flip table” action, “spin in” and “spin out” actions, and a null action for frames which were not labelled or could not be labelled. Since several of these actions are indistinguishable from pose alone, we merged all attach and detach actions into a single super-action, discarded the null

actions, and merged “spin in” with “spin out”, yielding only four actions.

Ikea FA does not include ground truth poses for all frames, so we used poses estimated by CPM for our experiments. We have checked the CPM-labelled poses against a small subset of hand-labelled poses using (strict) Percentage Correct Parts (PCP), which measures the number of limb instances in the dataset where both endpoints of the limb were estimated accurately to within half the true length of the limb [44]. By this criterion, upper arms were localised correctly 83.0% of the time, and lower arms 76.6% of the time.

*b) NTU RGB+D (2D):* NTU RGB+D [18] is an action recognition dataset of over 56,000 short videos, which collectively include over 4 million frames. Each sequence was recorded with a Kinect sensor, and consequently includes RGB imagery, depth maps and 2D/3D skeletons. Each sequence is also given a single action label from one of 60 classes, allowing us to perform an action-classifier-based evaluation. Instead of using the full 3D skeletons supplied with NTU RGB+D, we limit ourselves to 2D skeletons for easier comparison with Ikea Furniture Assembly.

<sup>3</sup><http://users.cecs.anu.edu.au/~u5568237/ikea/>

Because NTU RGB+D splits each action into a separate video, most of its sequences are only a few seconds each. Further, since the actions are largely unrelated, it is not possible to produce meaningful sequences of actions by stitching the videos together—one would only end up with a long sequence of seemingly random actions that neither humans nor computers could be expected to anticipate. However, we still wish to test the performance of our system on long videos, so we have limited our NTU RGB+D evaluation to subsequences of 5s or more. Unlike evaluation, training does not require a consistent sequence length, so we still train on all sequences not used for evaluation. For both training and evaluation, we downsample to 15fps to minimise the computational overhead of prediction on long sequences.

As with Ikea FA, we merge NTU RGB+D’s 60 actions into only seven classes for action-based evaluation. These classes constitute “super-actions” which are reasonably close in appearance. They include a super-action for a subject moving their hand to their head, another for moving their whole body up or down on the spot, one for walking, one for stretching their hands out in front of them, another for kicking, a super-action for engaging in other fine manipulation with the hands, and finally a catch-all class for actions which do not fit into the aforementioned categories.

*c) Pose parametrisation:* We found that the choice of representation for poses heavily influenced the subjective plausibility of pose sequences. Giving the DMM and baselines an absolute  $(x, y)$  coordinate for each joint resulted in wildly implausible continuations and poor generalisation. All experiments in this paper were instead carried out with a relative parameterisation: the location of the head is encoded as a series of frame-to-frame displacements over a sequence (i.e. its velocity), while the locations of each other joint was given relative to its parent. All sequences were mean-centred and scaled to have poses of the same height before applying this reparametrisation; after reparametrisation, each feature was again re-scaled to have zero mean and unit variance over the whole dataset.

### B. DMM and baseline configurations

Each DMM experiment used a 50 dimensional latent space for the DMM’s generative network, and 50 dimensional state vectors for the inference network’s bidirectional recurrent units.<sup>4</sup> The architecture is otherwise identical to the ST-LR architecture of [17].

We compare with four baseline predictors. The first baseline is a zero-velocity model which merely assumes that all future poses will be identical to the last observed pose. The second is a neural network consisting of a single, unidirectional, 128 dimensional LSTM followed by a Fully Connected (FC) output layer. The third and fourth are Encoder-Recurrent-Decoder (ERD) and three-layer LSTM networks (LSTM-3LR) with the same architectures as those presented in [9]. Note that the latter two models have significantly higher capacity than the former two: the ERD has two 500 dimensional FC layers, followed by two 1000 dimensional LSTMs, then another two

<sup>4</sup>At evaluation time, these bidirectional units are responsible for processing *only* the observed sequence of poses, and not the subsequent ground truth to be predicted.

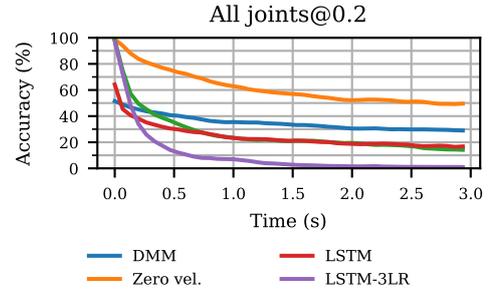


Fig. 6. PCK at different times and a fixed threshold on NTU RGB+D; statistics for all joints have been merged together.

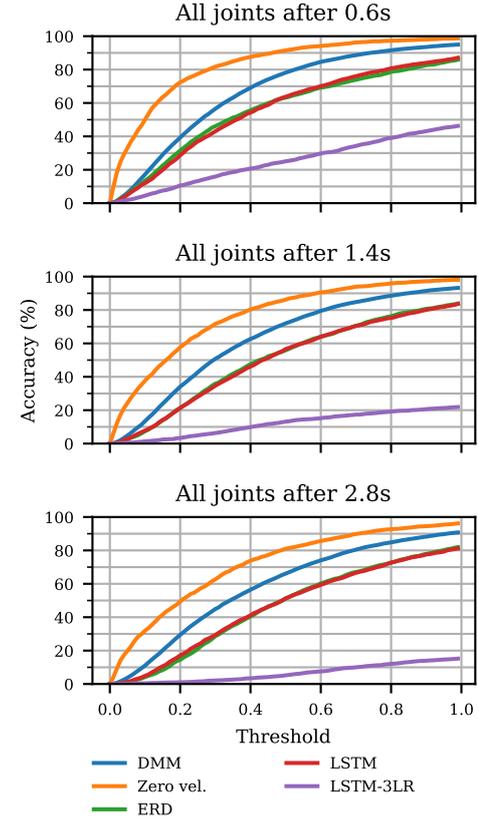


Fig. 7. PCK at different thresholds and a handful of fixed times on NTU RGB+D. Statistics for all joints have been merged together.

hidden 500 and 100 dimensional FC layers, before the FC output layer. Likewise, the LSTM-3LR has a 500 dimensional hidden FC layer, followed by three 1000 dimensional LSTMs, before the FC output layer.

### C. Evaluation protocols

Our first set of experiments focuses on displacements between forecasted poses and the ground truth. Figure 4 and Figure 5 depict the accuracy of predicted poses on Ikea FA, while Figure 6 and Figure 7 depict corresponding statistics for NTU RGB+D. Accuracies are reported as Percentage Correct Keypoints (PCK); this shows, for a given joint or collection of joints, the proportion of instances in which the predicted joint was within a given distance threshold of the ground truth. On Ikea FA, these distances are normalised to the length of the diagonal of a tight bounding box around each pose, while on

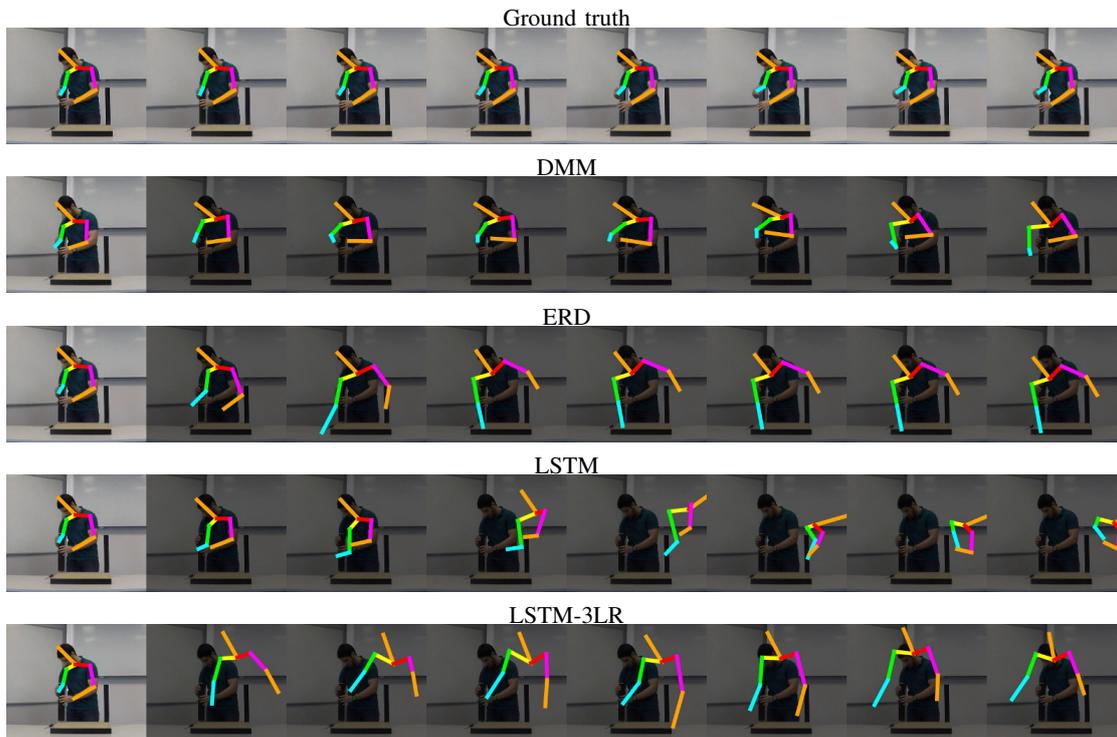


Fig. 8. Qualitative comparison of four models and the ground truth over a 7.5s forecast. Frames on the far left show the final observed poses, and subsequent frames show forecasted poses.

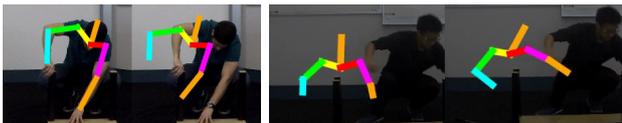


Fig. 9. Typical DMM errors: on the left, the velocity-based parameterisation has led to drift between the original, observed pose (far left) and the corresponding pose recovered from the DMM (second from left). On the right, the DMM has made an implausible transition between two independently plausible poses in adjacent frames.

NTU RGB+D, the distances are instead normalised using the average of displacements between a given hip (left or right) and the opposite shoulder (right or left). Further, for the DMM, we sampled five random continuations of each sequence and reported *expected* PCK instead of ordinary PCK; this was not necessary for the other baselines, which are deterministic.

Our second set of experiments focuses on the degree to which forecasted pose sequences resemble the ground truth sequence of actions for the forecast. To recover an action sequence from a forecasted pose sequence, we apply a recurrent classifier network consisting of a pair of 50 dimensional bidirectional GRUs, followed by a pair of 50 dimensional FC hidden layers and an FC output layer. Weights for this network are learnt from the training sets of each dataset. As with the first set of experiments, we averaged the DMM’s performance over five sampled continuations per pose sequence. Table I shows the results of these experiments.

## V. DISCUSSION

The joint-position-based evaluations on Ikea FA and NTU RGB+D show that the DMM performs best on longer forecasting horizons of several seconds or more. Further, Table I

shows that we also improve on the three “smart” baselines (ERD, LSTM, LSTM-3LR) in terms of the consistency of our produced poses with the ground truth action sequence. Qualitatively, we found that the ERD, LSTM and LSTM-3LR baselines tend to start out with little error, but quickly either converge to a mean pose or diverge to an implausible set of poses. In contrast, the individual plausibility of poses produced by the DMM tends not to decrease over time, and the DMM’s output does not rapidly converge to an obvious fixed point. Forecasts for each model on a single example sequence are included in Figure 8.

One surprising—and troubling—outcome of our experiments is the high performance of the zero velocity model. Not only does it dominate all baselines in both types of experiment, but it also beats our DMM model. This kind of performance has been observed by other authors, as well: in introducing the ERD, Fragkiadaki *et al.* reported that their model did not consistently outperform a zero-velocity baseline on forecasting of upper-body joints [9], and more recent work shows that this is also true of other “state-of-the-art” models [27].

There are some factors specific to the DMM which may be causing it to fall short of the zero-velocity baseline. In large part, we expect that its inaccuracy is due to drift in the predicted pose: as noted in Section III-C, small errors in the predicted centre-of-mass of a person can rapidly build up to destroy prediction accuracy, even when the motion of limbs relative to one another is small. This problem is exacerbated by our choice of a velocity-based parametrisation for head location, which leads to very rapid build-up of error during fast motions. Even when the DMM is able to observe ground truth poses, before beginning to forecast, it still accumulates error as the velocity-based input features do not provide the necessary

information for it to correct the drift which it accumulates. The left side of Figure 9 illustrates the issue.

As Figure 1 demonstrates, the DMM is able to produce a good diversity of continuations for a given pose sequence. To some extent, though, this diversity comes at the cost of temporal consistency: during evaluation, we found that the DMM would sometimes flip between poses which were independently plausible, but not temporally consistent, as demonstrated on the right side of Figure 9. In contrast, the deterministic baselines offer much smoother continuations, but at the cost of poorer accuracy and a rapid drift toward implausible poses. In principle, it ought to be possible to obtain the best of both methods by adding a temporal smoothness penalty to the DMM, although we have so far been unable to improve results this way.

## VI. CONCLUSION

We have proposed a novel application of Deep Markov Models (DMMs) to human pose prediction, and shown that they are better able to make long-range pose forecasts than state-of-the-art models. Further, we have introduced a new action recognition and pose forecasting dataset called Ikea Furniture Assembly, and proposed a mechanism for action-based evaluation of pose forecasts. Given the inherent difficulty of making long-range motion forecasts from skeletons alone, we believe that the most fertile ground for future research lies in the use of visual context to enable more meaningful predictions over horizons of several seconds and beyond.

## REFERENCES

- [1] S.-E. Wei, V. Ramakrishna, T. Kanade, and Y. Sheikh, "Convolutional pose machines," *arXiv:1602.00134*, 2016. 1, 2
- [2] A. Newell, K. Yang, and J. Deng, "Stacked hourglass networks for human pose estimation," *arXiv:1603.06937*, 2016. 1, 2
- [3] J. Tompson, R. Goroshin, A. Jain, Y. LeCun, and C. Bregler, "Efficient object localization using convolutional networks," in *CVPR*, 2015. 1
- [4] J. Charles, T. Pfister, D. Magee, D. Hogg, and A. Zisserman, "Personalizing human video pose estimation," in *CVPR*, 2016. 1
- [5] T. Pfister, J. Charles, and A. Zisserman, "Flowing convnets for human pose estimation in videos," in *ICCV*, 2015. 1, 2
- [6] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh, "Realtime multi-person 2D pose estimation using part affinity fields," *arXiv:1611.08050*, 2016. 1
- [7] E. Insafutdinov, L. Pishchulin, B. Andres, M. Andriluka, and B. Schiele, "Deepcut: A deeper, stronger, and faster multi-person pose estimation model," in *ECCV*, 2016. 1
- [8] U. Iqbal and J. Gall, "Multi-person pose estimation with local joint-to-person associations," in *ECCV*, 2016. 1
- [9] K. Fragkiadaki, S. Levine, P. Felsen, and J. Malik, "Recurrent network models for human dynamics," in *ICCV*, 2015. 1, 2, 4, 6, 7
- [10] A. Jain, A. R. Zamir, S. Savarese, and A. Saxena, "Structural-RNN: Deep learning on spatio-temporal graphs," in *CVPR*, 2016. 1, 2, 4
- [11] H. S. Koppula and A. Saxena, "Anticipating human activities using object affordances for reactive robotic response," *TPAMI*, 2016. 1
- [12] —, "Learning spatio-temporal structure from RGB-D videos for human activity detection and anticipation," in *ICML*, 2013. 1
- [13] D.-A. Huang and K. M. Kitani, "Action-reaction: Forecasting the dynamics of human interaction," in *ECCV*, 2014. 1
- [14] F. Han, B. Reily, W. Hoff, and H. Zhang, "Space-time representation of people based on 3D skeletal data: A review," *Computer Vision and Image Understanding*, 2017. 1
- [15] G. W. Taylor, G. E. Hinton, and S. T. Roweis, "Modeling human motion using binary latent variables," in *NIPS*, 2006. 1
- [16] C. Ionescu, D. Papava, V. Olaru, and C. Sminchisescu, "Human 3.6M: Large scale datasets and predictive methods for 3d human sensing in natural environments," *PAMI*, 2014. 1
- [17] R. G. Krishnan, U. Shalit, and D. Sontag, "Structured inference networks for nonlinear state space models," *arXiv:1609.09869*, 2016. 1, 2, 3, 4, 6
- [18] A. Shahroudy, J. Liu, T.-T. Ng, and G. Wang, "NTU RGB+D: A large scale dataset for 3D human activity analysis," in *CVPR*, 2016. 1, 5
- [19] Y. Yang and D. Ramanan, "Articulated pose estimation with flexible mixtures-of-parts," in *CVPR*, 2011. 2
- [20] X. Chen and A. Yuille, "Articulated pose estimation by a graphical model with image dependent pairwise relations," in *NIPS*, 2014. 2
- [21] M. Sun, P. Kohli, and J. Shotton, "Conditional regression forests for human pose estimation," in *CVPR*, 2012. 2
- [22] A. Toshev and C. Szegedy, "DeepPose: Human pose estimation via deep neural networks," in *CVPR*, 2014. 2
- [23] A. Cherian, J. Mairal, K. Alahari, and C. Schmid, "Mixing body-part sequences for human pose estimation," in *CVPR*, 2014. 2
- [24] J. M. Wang, D. J. Fleet, and A. Hertzmann, "Gaussian process dynamical models for human motion," *TPAMI*, 2008. 2
- [25] V. Pavlovic, J. M. Rehg, and J. MacCormick, "Learning switching linear models of human motion," in *NIPS*, 2000. 2
- [26] I. Sutskever, G. E. Hinton, and G. W. Taylor, "The recurrent temporal restricted boltzmann machine," in *NIPS*, 2008. 2
- [27] J. Martinez, M. J. Black, and J. Romero, "On human motion prediction using recurrent neural networks," in *CVPR*, 2017. 2, 7
- [28] S. L. Pntea, J. C. van Gemert, and A. W. Smeulders, "Déjà vu," in *ECCV*, 2014. 2
- [29] J. Walker, A. Gupta, and M. Hebert, "Dense optical flow prediction from a static image," in *ICCV*, 2015. 2
- [30] J. Walker, C. Doersch, A. Gupta, and M. Hebert, "An uncertain future: Forecasting from static images using variational autoencoders," in *ECCV*, 2016. 2
- [31] C. Vondrick, H. Pirsiavash, and A. Torralba, "Generating videos with scene dynamics," *arXiv:1609.02612*, 2016. 2
- [32] M. Mathieu, C. Couprie, and Y. LeCun, "Deep multi-scale video prediction beyond mean square error," *arXiv:1511.05440*, 2015. 2
- [33] R. Mahjourian, M. Wicke, and A. Angelova, "Geometry-based next frame prediction from monocular video," *arXiv:1609.06377*, 2016. 2
- [34] C. Finn, I. Goodfellow, and S. Levine, "Unsupervised learning for physical interaction through video prediction," *arXiv:1605.07157*, 2016. 2
- [35] D. P. Kingma and M. Welling, "Auto-encoding variational Bayes," *arXiv:1312.6114*, 2013. 2, 3
- [36] D. J. Rezende, S. Mohamed, and D. Wierstra, "Stochastic back-propagation and approximate inference in deep generative models," *arXiv:1401.4082*, 2014. 2, 3
- [37] C. Doersch, "Tutorial on variational autoencoders," *arXiv:1606.05908*, 2016. 2, 3
- [38] J. Bayer and C. Osendorfer, "Learning stochastic recurrent networks," *arXiv:1411.7610*, 2014. 2
- [39] J. Chung, K. Kastner, L. Dinh, K. Goel, A. C. Courville, and Y. Bengio, "A recurrent latent variable model for sequential data," in *NIPS*, 2015. 2
- [40] M. Fraccaro, S. K. Sønderby, U. Paquet, and O. Winther, "Sequential neural models with stochastic layers," in *NIPS*, 2016. 2
- [41] L. Wang, Y. Qiao, and X. Tang, "Mofap: A multi-level representation for action recognition," *IJCV*, 2016. 4
- [42] M. Rohrbach, S. Amin, M. Andriluka, and B. Schiele, "A database for fine grained activity detection of cooking activities," in *CVPR*, 2012. 4
- [43] H. Jhuang, J. Gall, S. Zuffi, C. Schmid, and M. J. Black, "Towards understanding action recognition," in *ICCV*, 2013. 4
- [44] M. Eichner, M. Marin-Jimenez, A. Zisserman, and V. Ferrari, "2D articulated human pose estimation and retrieval in (almost) unconstrained still images," *IJCV*, 2012. 5