

Lab CudaVision
Learning Vision Systems on Graphics Cards (MA-INF 4308)

CudaLab Project

05.07.2022

PROF. SVEN BEHNKE, ANGEL VILLAR-CORRALES

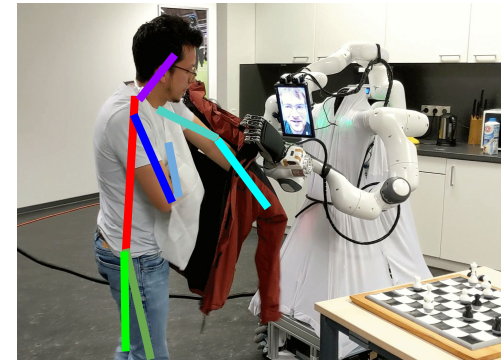
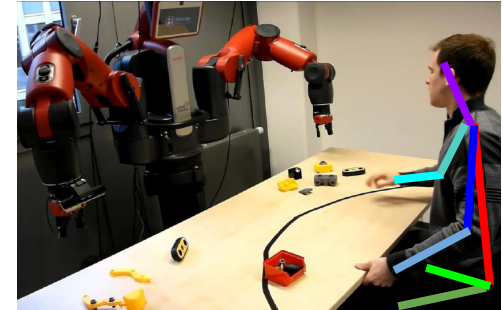
Contact: villar@ais.uni-bonn.de

Human Pose Forecasting

Motivation

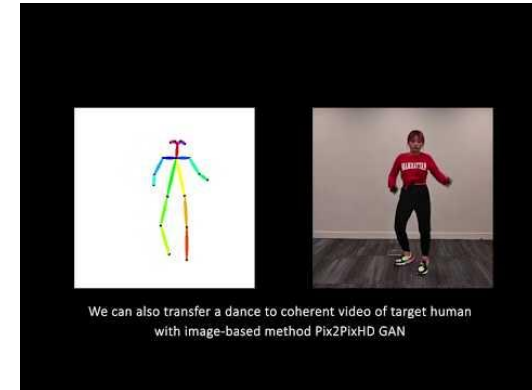
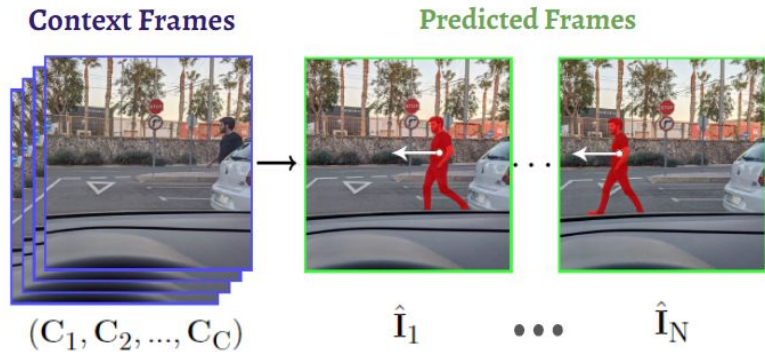
- Human-robot collaboration is a challenging task
 - Perception of the environment
 - Planning capabilities
 - **Predicting actions and behavior of nearby agents**

- Human pose is a good representation for:
 - Action recognition and prediction
 - Motion estimation
 - Planning and navigation



Human Pose Forecasting

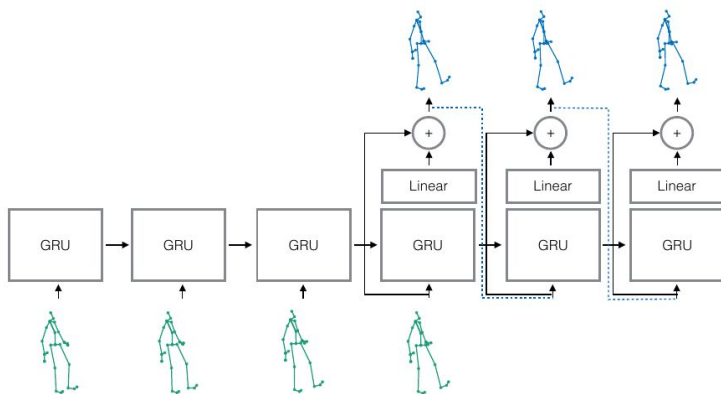
- Given a sequence of C seed poses, generate next N plausible poses
 - Predictions must be temporally consistent
 - Incorporate human motion dynamics
- Multiple applications: sports , anticipating human behavior, motion transfer, ...



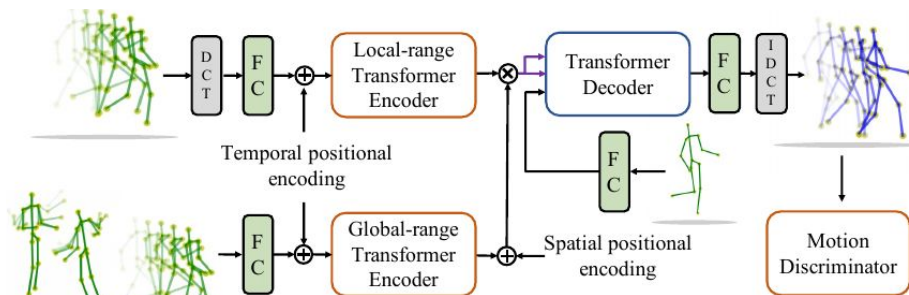
Related Work

- Human motion prediction is an ongoing research topic (2015 - ...)

Recurrent neural networks



Transformers



Graph neural networks, 3D-Convolutions, ...

Model

Model Inspiration



This CVPR paper is the Open Access version, provided by the Computer Vision Foundation. Except for this watermark, it is identical to the version available on IEEE Xplore.

On human motion prediction using recurrent neural networks

Julietta Martinez¹, Michael J. Black², and Javier Romero¹

¹University of British Columbia, Vancouver, Canada

²MPI for Intelligent Systems, Tübingen, Germany

³Body Labs Inc., New York, NY

joel@cs.ubc.ca, black@cs.ubc.ca, romero@bodylabs.com

Abstract

Human motion modeling is a classical problem at the intersection of graphics and computer vision, with applications spanning human-computer interaction, motion synthesis, and motion prediction for virtual and augmented reality. Following the success of deep learning methods in several computer vision tasks, recent work has focused on using deep recurrent neural networks (RNNs) to model human motion, with the goal of learning time-dependent representations that perform tasks such as short-term motion prediction and long-term human motion synthesis. We examine recent work, with a focus on the evaluation methodology commonly used in the literature, and show that, surprisingly, state-of-the-art performance can be achieved by a simple baseline that does not attempt to model motion at all. We investigate this result, and analyze recent RNN methods by looking at the architectures, loss functions, and training procedures used in state-of-the-art approaches. We propose three changes to the standard RNN models typically used for human motion, which result in a simple and scalable RNN architecture that achieves state-of-the-art performance on human motion prediction.

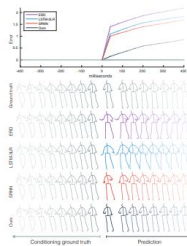


Figure 1. Top Mean average prediction error for different motion prediction methods. Ground truth (gray) and proposed method (red) are shown in gray, and short-term motion prediction are shown in color. Previous work, based on deep RNNs, produces strong performance at the state of the prediction (middle columns). Our method predicts motion, low error prediction.

1. Introduction

An important component of our capacity to interact with the world resides in the ability to predict its evolution over time. Handling an object in another person, playing sports, or simply walking in a crowded street would be extremely challenging without our understanding of how people move, and our ability to predict what they are likely to do in the following instants. Similarly, machines that are able to perceive and interact with moving people, either in physical or virtual environments, must have a notion of how people move. Since human motion is the result of both physical limitations (i.e. a torque exerted by muscles, gravity, momentum preservation) and the intentions of subjects (how to perform

an intentional motion), motion modeling is a complex task that should be ideally learned from observations.

Our focus in this paper is to learn models of human motion from motion capture (mocap) data. More specifically, we are interested in subjects who move freely (without preservation) and the intentions of subjects (how to perform

Research carried out while Julietta was an intern at MPI.

Intention-based Long-Term Human Motion Anticipation

Julian Tanke¹, Chintan Zaveri¹, Juergen Gall¹

University of Bonn

{tanke|gall}@iai.uni-bonn.de

Abstract

Recently, a few works have been proposed to model the uncertainty of the future human motion. These works do not forecast a single sequence, but multiple sequences for the same observation. While these works focus on increasing the diversity, this work focuses on keeping a high quality of the forecast sequences even for very long time horizons of up to 30 seconds. In order to achieve this goal, we propose to forecast the intentions of the person ahead of time. This has the advantage that the generated human motion remain goal oriented and that the motion transitions between two actions are smooth and highly realistic. We furthermore propose a new quality score for evaluation that correlates better with human perception than other metrics. The results and a user study show that our approach forecasts multiple sequences that are more plausible compared to the state-of-the-art.

1. Introduction

Anticipating human motion is highly relevant for many interactive activities such as sports, manufacturing, or navigation [25] and significant progress has been made in forecasting human motion [8, 9, 10, 11, 15, 17, 23, 26, 35]. Most progress has been made in anticipating motion over a short time horizon of around one second. However, deep methods fail when anticipating longer time horizons as they either produce unrealistic poses or the motion freezes. As another issue that occurs when the time horizon gets larger is the fact that there are more than one future sequence that are plausible for a single observed sequence of human motion as is shown in Figure 2. Going from a short time horizon of less than one second to a larger time horizon of a few seconds therefore implies the following challenges: (a) How can we model the uncertainty of the future? (b) How can we ensure that the motion remains plausible? (c) How can we measure the quality of methods that generate more than one sequence?

Handling the uncertainty of the future has been so far addressed by very few recent works [4, 28, 37] for ha-

uman motion anticipation. These approaches are able to forecast diverse sequences from the same observation, but the quality of the sequences decreases for longer time horizons beyond 1 second. In this work, we also propose a network that generates multiple sequences as shown in Figure 2, but our goal is to generate more plausible sequences for time horizons of multiple seconds. In order to achieve this goal, we not only model the human motion but also the intention of the person as illustrated in Figure 1. In fact, human motion anticipation depends on two factors, namely the past motion and the intention. The latter, which is ignored by existing works, is very important for longer sequences since a motion without a goal is perceived as random and unnatural. We therefore model the intention as discrete actions and propose to forecast the intention as well as the human motion. The key aspect is that our model forecasts the intention ahead of time and that the forecast human motion is conditioned on the past motion and on the forecast intention as shown in Figure 1.

It however, remains an open issue how methods that generate multiple sequences are best compared. Recent works suggest to evaluate both the quality of the generated motion as well as the sample diversity. While diversity is measured by the number of nearby agents and model predictions about their actions and behavior. Depending on the desired prediction time-horizon, the level of abstraction of the predicted sequences may differ. Furthermore, the body pose forecasting is more relevant for shorter time horizons, while the intention forecasting is more relevant for longer time horizons. In this paper, we compare the quality of the generated motion as well as the sample diversity. For longer time horizons, it is no longer possible to predict these exact details, hence it can be advantageous to predict only abstract representations but maintain a high level of semantics. Finally, for planning longer into the future, only representations of a higher level of abstraction, such as actions or locations, might be reliably predicted.

In the last few years, several deep-learning-based approaches [12, 21, 23, 19, 14] have been proposed to predict future video frames. These methods, which often combine

¹equal contribution

Video Prediction at Multiple Scales with Hierarchical Recurrent Networks

Ani Karapantzi¹, Angel Villar-Corralés¹, Andreas Böhres¹ and Sven Behnke¹

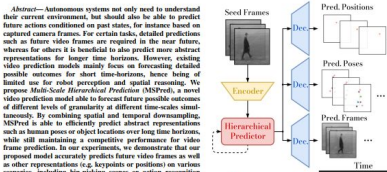


Fig. 1: Given a sequence of seed frames, MSPred predicts representations of different levels of granularity at distinct time-scales. Low-level representations, such as video frames, are predicted for short time horizons with a fine temporal resolution. Conversely, higher-level representations, such as human poses or locations, are forecasted longer into the future using coarser time resolutions, hence allowing for long-term predictions with a small number of iterations.

1. INTRODUCTION

For effective human-robot collaboration, autonomous systems, such as domestic robots, need not only to perceive and understand their surroundings, but should also be able to estimate the intentions of nearby agents and make predictions about their actions and behavior. Depending on the desired prediction time-horizon, the level of abstraction of the predicted sequences may differ. Furthermore, the body pose forecasting is more relevant for shorter time horizons, while the intention forecasting is more relevant for longer time horizons. In this paper, we compare the quality of the generated motion as well as the sample diversity. For longer time horizons, it is no longer possible to predict these exact details, hence it can be advantageous to predict only abstract representations but maintain a high level of semantics. Finally, for planning longer into the future, only representations of a higher level of abstraction, such as actions or locations, might be reliably predicted.

In the last few years, several deep-learning-based approaches [12, 21, 23, 19, 14] have been proposed to predict future video frames. These methods, which often combine

This work was funded by grant 25016/2 (Research Unit FOR 2501) Anticipating Human Behavior of the German Research Foundation (DFG) ¹equal contribution ²corresponding author ³researcher at the University of Bonn, Germany ⁴researcher at the University of Bonn, Germany



This CVPR working paper is the Open Access version, provided by the Computer Vision Foundation. Except for this watermark, it is identical to the accepted version. The final published version of the proceedings is available on IEEE Xplore.

Learning Decoupled Representations for Human Pose Forecasting

Behnam Parsaeifard^{1,2,*}, Saeed Saadehjalal², Yuezhang Liu², Taylor Mordan², Alexandre Alahi¹

¹University of Basel, Switzerland ²École Polytechnique Fédérale de Lausanne (EPFL), Switzerland

behnam.parsaeifard@epfl.ch, saeed.saadehjalal@epfl.ch

Abstract

Human pose forecasting involves complex spatiotemporal interactions between body parts (e.g., arms, legs, spine). State-of-the-art approaches use Long Short-Term Memory (LSTMs) or Variational Autoencoders (VAEs) to solve the problem. Yet, they do not effectively predict human motions when both global trajectory and local pose movements exist. We propose to learn decoupled representations for the global and local pose forecasting tasks. We also show that it is better to separate the prediction when the uncertainty in human motion increases. Our forecasting model outperforms all existing methods on the human forecasting benchmark to date by over 30%. The code is available online¹.



Figure 1: Decoupling the human pose into trajectory and local pose. The dashed arrows indicate the trajectory of the human pose.

1. Introduction

Human pose forecasting is defined as predicting future human keypoints' locations—the body parts (e.g., legs, arms, spine)—given a sequence of observed ones. It has attracted more attention in recent years due to its critical applications in self-driving cars [15], robotics [11, 12] and healthcare [8, 45, 44, 46]. For example, in self-driving cars, it is very important to predict the location of pedestrians to avoid collisions [12, 46]. Furthermore, the body pose of pedestrians often provide useful information about whether or not they intend to cross the street [16]. Unfortunately, the high uncertainty in this problem makes it challenging that even in humans, are often not able to exactly predict the next motion. In this work, we want to learn a representation of human pose dynamics to effectively predict plausible motions and potentially solve problems when the uncertainty is too high.

The human pose forecasting task can be decomposed into a global (sequence) trajectory forecasting task and a local (fine-grained) pose forecasting task. At the coarse level, the large-scale movements of humans with respect to the camera are of interest.

¹Equal contribution, unless stated otherwise ²https://github.com/ani-karapantzi/epfl-humans-pose-forecasting

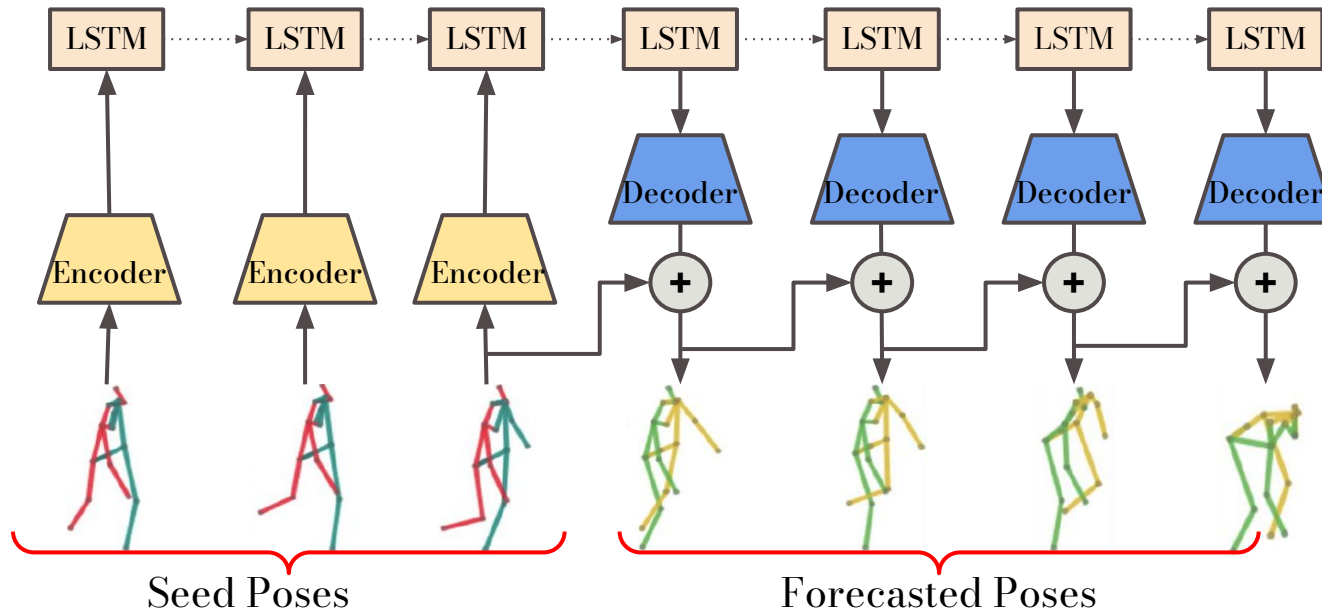
modelled. However, at the fine-grained level, all the detailed local movements of different keypoints are modelled. Previous works showed promising results for trajectory forecasting [1, 10] and local pose forecasting (i.e., excluding the global trajectory movement [36, 3]). They used Long Short-Term Memories (LSTMs) because of their ability to capture temporal dependencies or Variational Autoencoders (VAEs) because of their ability in generating a new pose considering the non-deterministic task. While they achieved outstanding results for each of these separate tasks, they have limited performance to predict the human pose dynamics when both trajectory and local pose move.

Considering the complexity of this task, we propose to decompose it into trajectory forecasting and local pose forecasting tasks (see Figure 1). When a person moves, their global coordinates and the local coordinates of their keypoints (with respect to their trajectory) change in different ways and the distinction helps to exploit different approaches for both.

We propose an LSTM encoder-decoder network for trajectory forecasting and a VAE-decoder to solve this local pose forecasting task. Moreover, if the network is not confident about the future, it stops predicting and takes the last prediction. We show that using this approach results in a

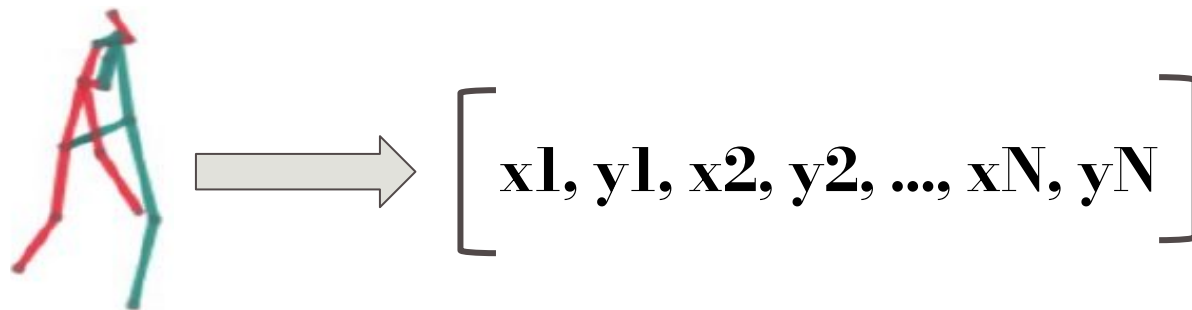
Proposed Model 1

- Skeleton-based pose prediction



Pose Representation

- Pose is parameterized as list of coordinates
 - One dimensional representation
 - Shape is $2N$, where N is the number of joints
- Pose can be preprocessed:
 - a. Normalize each coordinate by the maximum x & y values respectively
 - b. Normalize each coordinate by the image size

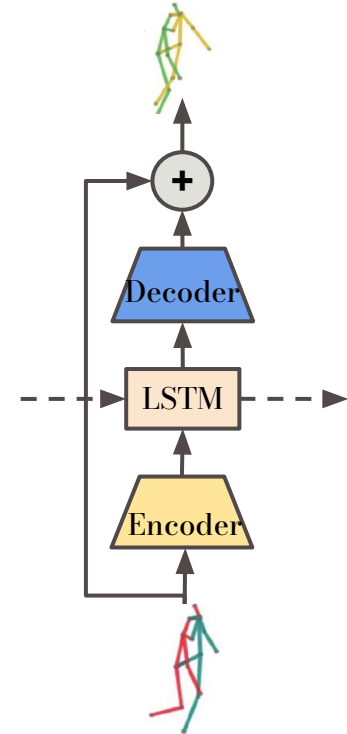


Model

- Encoder:
 - Maps input poses into higher-dimensional representation
 - One single fully connected layer might suffice
- Recurrent model:
 - Learns motion dynamics
 - One RNN (possibly with multiple cells) or Seq-to-Seq architecture
 - RNN can be either LSTM or GRU
- Decoder:
 - Maps output of predictor back to pose space
 - One single fully connected layer might suffice

Residual Connection

- Poses cannot change much between two consecutive time steps
- Recurrent connection between time steps
 - Network must only model changes
 - Faster learning



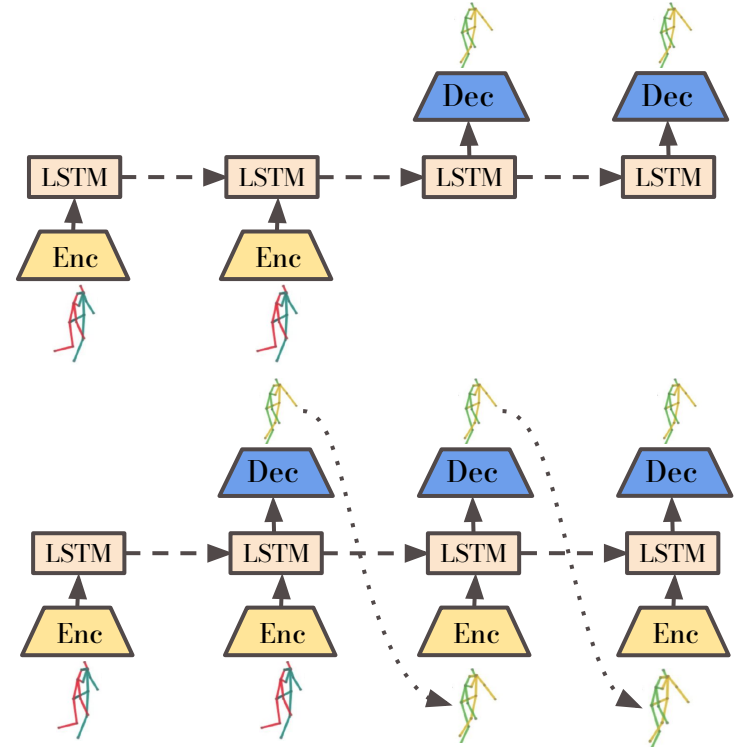
Model Flow

State space model:

- Autoregressive in feature space
- RNN carries pose and motion representations

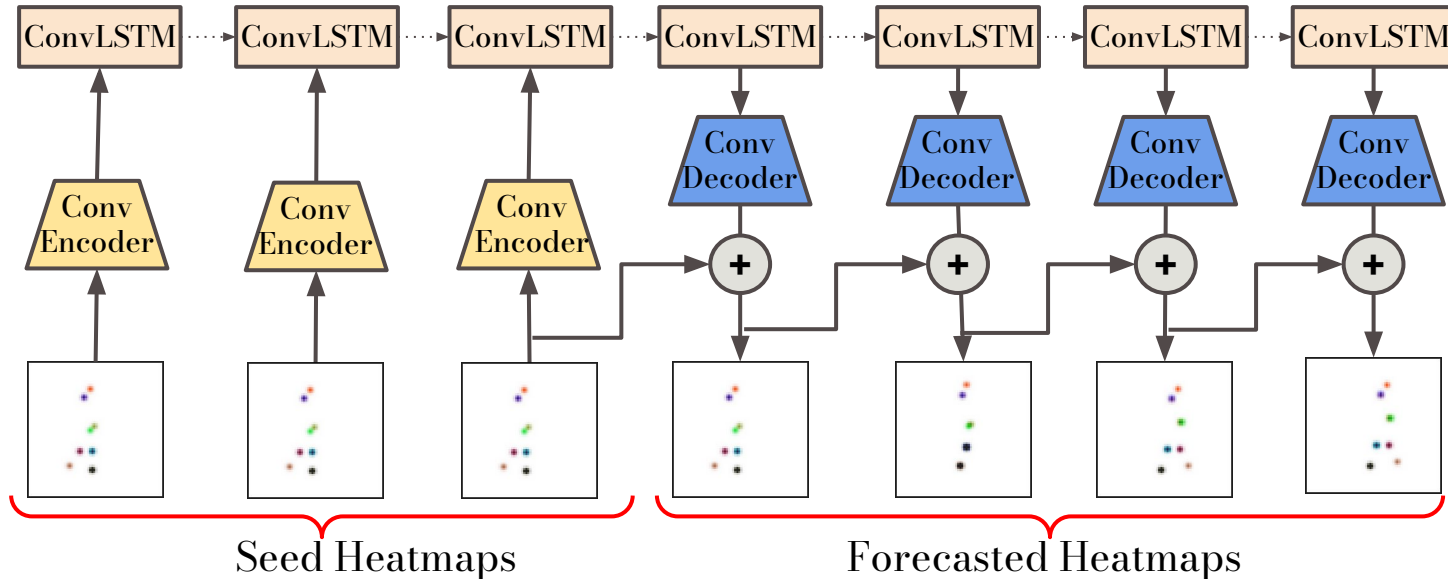
Autoregressive model:

- Predictions are re-encoded and used as inputs
- RNN must only model motion



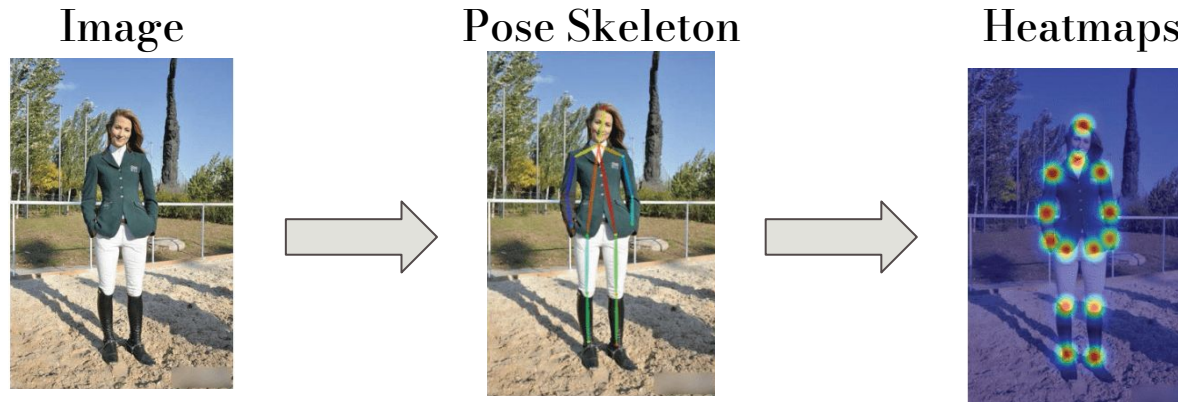
Proposed Model 2

- Heatmap-based pose prediction



Heatmap Representation

- Pose is parameterized as multi-channel heatmaps
 - One channel for each joint
 - Shape is (N, H, W) , where N is the number of joints
 - Heatmaps are generated by fitting a Gaussian on the joint center



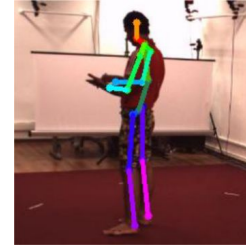
Model

- Encoder:
 - Maps input heatmaps into better representation for prediction
 - Convolutional encoder, e.g., VGG-ish or ResNet-like
- Recurrent model:
 - Learns motion dynamics
 - One ConvRNN (possibly with multiple cells) or Seq-to-Seq architecture
 - ConvRNN can be either ConvLSTM or ConvGRU
- Decoder:
 - Maps output of predictor back to pose space
 - Convolutional encoder, e.g., VGG-ish or ResNet-like
 - Mirrored version of encoder
- Model flow and residual connections are the same as for model 1

Datasets

Human 3.6M Dataset

- Dataset used as a benchmark for many tasks
 - Pose estimation and forecasting
 - Video prediction
- Stats and characteristics
 - 3.6 million 3D human poses and images
 - 11 professional actors (6 male, 5 female)
 - 17 scenarios (discussion, smoking, ...)
- Data is available in:
 - `beast2:/home/cache/H36`
 - Images + Annotations
 - Jupyter notebook with an example



Training & Evaluation

Training and Prediction

- Datasets:
 - Train & evaluate on Humans3.6M, both with poses and heatmaps
 - Use the official train-test splits
 - Use a downsampling factor of 8
 - $[f_1, f_2, f_3, \dots, f_7, f_8, f_9, \dots, f_{65}] \rightarrow [f_1, f_9, f_{17}, f_{25}, f_{33}, \dots, f_{65}]$
- Train and evaluate with sequence of size:
 - For pose vectors: (Batch_size, 20, 2 * N_kpts)
 - For heatmaps: (Batch_size, 20, N_kpts, 64, 64)
- Use 10 frames as seed frames, and predict the next 10 frames.

Training and Prediction

- Model:
 - Train & evaluate on pose-based and heatmap-based models
 - Choose your design choices: model flow, residual connections, ...
 - Teacher forcing vs. no teacher forcing
- Criterion:
 - Use MSE or MAE as loss functions
 - (Optional) Add an addition perceptual loss (SSIM, LPIPS, ...) or adversarial loss

Evaluation

- Measure performance only on the 10 predicted frames
- Evaluate using the following metrics:
 - MSE
 - MAE
 - PDJ
 - PCK
 - MPJPE
- Qualitative evaluation by observing predicted frames

Project Goals and Deliverables

Passing Requirements

1. Implement both models, pipelines and utils
2. Train your models to achieve best possible results on Humans3.6M
 - You must implement and train the described models
 - Make changes and train further models to achieve better results
3. Create overview notebook
4. Write project report

Deliverables

- Complete codebase
 - Clean and structured
 - Not just a notebook!
- Trained model checkpoint and (tensorboard, WandB, ...) logs
- Overview notebook (.ipynb & .html) showing main functionalities:
 - Load data
 - Load pretrained model
 - Display some results
- Project report

Grading

- Results and Experiments **55%-60%:**
 - Performing several experiments and obtaining good results
 - Additional experiments: ablation study, changes in the model, ...
- Codebase & Overview Notebook **20%:**
 - Implement all functionalities
 - Modularity and structure
- Report **20%-25%**

Project Report

- Document your work in the project report
- Try to be brief, but readable and informative
- Include figures and tables
- Use *BibTex* for the references
- I expect 6-12 pages, but highly depends on number and size of imgs/tables
- Use the following template
 - <https://www.overleaf.com/read/tmnvhrsdmjrp>

Additional Experiment Ideas

- Try your own ideas!
- Investigate data preprocessing
- Tweak the model
 - Change modules (num. layers, num. kernels, ...)
 - Investigate different predictors (ConvLSTM, Seq-to-Seq)
- Investigate different training strategies or transfer learning:
 - Use additional loss functions
 - Use adversarial supervision
- Make changes to the model
 - Stochastic model: <https://arxiv.org/abs/1802.07687>
 - Transformer based: <https://arxiv.org/abs/2111.12073>

Important Dates

- **05.07:** Starting date
- **29.08-09.09:** Revision session
- **15.09:** Draft submission due
- **30.09:** Final submission:

Questions?



References

1. Parsaeifard, Behnam, et al. "Learning decoupled representations for human pose forecasting." IEEE/CVF International Conference on Computer Vision (ICCV). 2021.
2. Tanke, Julian, Chintan Zaveri, and Juergen Gall. "Intention-based Long-Term Human Motion Anticipation." 2021 International Conference on 3D Vision (3DV). IEEE, 2021.
3. Karapetyan, Villar-Corrales et al. "Video Prediction at Multiple Scales with Hierarchical Recurrent Networks." arXiv preprint arXiv:2203.09303 (2022).
4. Martinez, Julieta, Michael J. Black, and Javier Romero. "On human motion prediction using recurrent neural networks." IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2017.
5. Catalin Ionescu, et al., "Human3.6M: Large Scale Datasets and Predictive Methods for 3D Human Sensing in Natural Environments", IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI). 2014

