Mobility21
A USDOT NATIONAL
UNIVERSITY TRANSPORTATION CENTER

Carnegie Mellon University

Penn
University of Pennsylvania

The Ohio State University

CCAC

Perception for Transportation Service Robots

Principal Investigator:
Aaron Steinfeld
Carnegie Mellon University
https://orcid.org/0000-0003-2274-0053

Report Author: **Abhijat Biswas**

FINAL RESEARCH REPORT

Contract # 69A3551747111

DISCLAIMER

The contents of this report reflect the views of the authors, who are responsible for the facts and the accuracy of the information presented herein. This document is disseminated under the sponsorship of the U.S. Department of Transportation's University Transportation Centers Program, in the interest of information exchange. The U.S. Government assumes no liability for the contents or use thereof.

Human Torso Pose Forecasting for the Real World

Abhijat Biswas

CMU-RI-TR-19-51

August 2019

The Robotics Institute
School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213

Thesis Committee:
Aaron Steinfeld, *Co-Chair*
Henny Admoni, *Co-Chair*
Kris Kitani,
Ishani Chatterjee

Submitted in partial fulfillment of the requirements for the degree of Master of Science in Robotics

Copyright © Abhijat Biswas

Keywords: pose forecasting; pose estimation; human sensing; social navigation

To my family, friends, and mentors

Abstract

Anticipatory human intent modeling is important for robots operating alongside humans in dynamic or crowded environments. Humans often telegraph intent through posture cues, such as torso or head cues. In this paper, we describe a computationally lightweight approach to human torso pose recovery and forecasting with a view towards limited sensing for easy on-board deployment. Our end-to-end system combines RGB images and point cloud information to recover 3D human pose, bridging the gap between learning-based 2D pose estimation methods and the 3D nature of the environment that robots and autonomous vehicles must reason about, with minimum overhead. In addition to pose recovery, we use a simple filter-and-polynomial fit method to forecast torso pose. We focus on rapidly generating short horizon forecasts, which is the most relevant scenario for autonomous agents that iteratively alternate between data gathering and planning steps in highly dynamic environments. While datasets suited to benchmarking multi-person 3D pose prediction in real-world scenarios are scarce, we describe an easily replicable evaluation method for benchmarking in a near real-world setting. We then assess the pose estimation performance using this evaluation procedure. Lastly, we evaluate the forecasting performance quantitatively on the Human3.6M motion capture dataset. Our simple 3D pose recovery method adds minimum overhead to 2D pose estimators, with comparable performance to 3D pose estimation baselines from a computer vision alternative. Furthermore, our uncomplicated forecasting algorithm outperforms complicated recurrent neural network methods while also being faster on the torso pose forecasting task.

Acknowledgments

I would like to thank my advisors: Prof. Aaron Steinfeld and Prof. Henny Admoni for always providing guidance about life, wisdom about paragraph breaks, and a sense of humour when appropriate. I am thankful to my labmates Reuben, Ben, Pallavi, Ada, Mike, Sarthak and Stephanie for feedback and support, both technical and emotional. I also want to thank my other committee members Prof. Kris Kitani and Ishani Chatterjee for their feedback.

Finally, I want to thank my parents, who've worked for years to give me the opportunity to live my life the way I want. For that and everything else, I am grateful in perpetuity.

I want to acknowledge funding support for this work from the National Science Foundation (IIS-1734361), and the Mobility21 National University Transportation Center which is sponsored by the US Department of Transportation.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Perceiving and anticipating human motion is increasingly important and relevant as mobile autonomous systems are steadily deployed in highly dynamic and cluttered environments with imperfect information about their surroundings. In real-world settings, a crucial aspect of human-robot interaction (HRI) is real-time anticipatory modeling of human motion. Fluid tasks such as collaborative assembly, handovers, and navigating through moving crowds require timely prediction of probable future human motion.

Consider the case where a mobile, transportation hub robot meets visitors who have requested assistance. First, it must rendezvous with the human. A strong cue that a particular human is ready for interaction is when they turn to face the oncoming robot. Second, the robot must navigate past other humans without crossing their path in a rude manner [5, 21]. Finally, the robot needs to orient itself properly as it approaches the person [3]. Timely perception of human torso pose is important for all of these steps.

More generally, to be accepted by society, mobile robots deployed in public settings need to behave in expected and predictable ways. To meet this goal, robots need to reason not only about individual humans in various trajectories, but about social groups and personal spaces for which, again, body orientation is an important feature [43].

In support of these, and similar interactions, we present a new human torso pose estimation and anticipation model. We focus specifically on the case of mobile robots with limited computational and sensing resources, operating in highly dynamic environments. The typical sensing-perceiving-acting loop in such scenarios involves alternating between data gathering and action or motion re-planning steps in rapid iterations. We show that a simple filter and polynomial fit model outperforms deep neural networks for short-to-medium horizon (under $1\ s$) predictions, which is the most important case for mobile robots expected to rapidly gather data and re-plan. We also show this method to be much faster, allowing it to be deployed for low-cost on mobile systems since it does not require significant and expensive computation.

Both torso pose recovery and forecasting are challenging problems, so prior approaches have involved computationally expensive solutions. As an illustration, consider that one of the preeminent 2D articulated full-body human pose detectors [4] can perform at around 18Hz using 2x Nvidia 1080 Ti GPUs [13]. Additionally, real world human perception requires the knowledge of 3D pose rather over 2D pose. We attempt to efficiently bridge this 2D to 3D gap with a focus on limited-compute, real-time operation, as well as demonstrate suitability for pose forecasting. Full body articulated human pose forecasting is also challenging due to the associated high dimensional, non-linear dynamics and inherent stochasticity of human motion.

To make the problem more tractable, researchers have approached the forecasting problem by restricting the scope to a particular part of the body relevant to the task, thereby reducing the dimensionality of the problem space. For example, some predictively model human reaching motions for a shared workspace assembly task [28], while others predict future hand locations in egocentric video to allow anticipatory motion planning and assistance [25]. We draw inspiration from this strategy and restrict the problem to modeling the spatio-temporal behaviour of the human torso. Specifically, we aim to detect and forecast the human torso plane position and orientation, the latter being an important cue correlated with motion intent and social engagement [43].

Our algorithm uses multi-modal visual input data, namely RGB with scene depth data, to estimate and forecast a 3D torso plane. This in contrast to most previous body pose forecasting work (e.g., [8, 12, 19, 29]) that either use 2D or 3D articulated pose, often with initial joint configurations obtained directly from a motion capture system. Such multi-modal sensing not only helps overcome depth ambiguity [18], but also allows us to use monocular 2D body pose estimators (which are more accurate than monocular 3D pose estimators) and project these estimations to 3D easily using an RGB image in conjunction with a registered point cloud.

All of our algorithmic design choices are made to prioritize fast running times on generic, portable hardware, such as barebones PCs or embedded systems.

Additionally, we describe a useful evaluation procedure for single-view 3D pose estimation in crowded scenes which can be used by the community for benchmarking. It is difficult to obtain ground truth pose estimates from single-view sensing in real-world scenarios due to occlusions and prior work has tended to use marker-driven motion capture data for these purposes, which inherently contains only clutter and occlusion free scenarios which are artificial. We work around this for evaluation purposes by simulating single viewpoint visual sensing in cluttered scenes using the publicly available Panoptic Studio dataset [20].

1.1 Contributions

In this thesis, we describe a computationally light-weight end-to-end 3D torso pose estimation and forecasting system combining both depth and color visual data.

Further, we show that a simple filtering and polynomial fitting algorithm outperforms more complicated recurrent neural network based pose forecasting approaches and is $45\times$ faster, trading off speed and accuracy for pose granularity. We evaluate the pose forecasting system quantitatively on the Human 3.6M (H3.6M) dataset [17]. We show superior performance for short-to-medium term forecasts and competitive results for longer term forecasts, especially for predictable activities such as Walking motions in H3.6M.

1.2 Outline

This thesis is organized as follows: The second chapter gives an overview of previous work in related areas including human pose estimation and intent prediction, especially in the context of pedestrians. In the third chapter, we introduce our simple pose recovery approach as well as our filter and spline extrapolation based torso pose forecasting method. We then present our experimental results evaluating the aforementioned methods and describe several baselines used to perform comparative quantitative evaluations. We then discuss the limitations of current evaluation datasets for properly evaluating our work in the context of pose-conditioned pedestrian forecasting and discuss future work to address this issue.

Chapter 2

Related work

Human modeling for robotics has taken various forms including estimating [4, 27, 37, 45] and forecasting [12, 19, 29] human pose from visual data, modeling human motion trajectories individually [46, 48] and in groups [44], as well as predicting human intent [23].

To this end, previous works have utilized the intrinsic kinematics of the human anatomy [15], eye gaze [1, 16], semantic information of the scene [22], and spatio-temporal structure of the task space [23]. These methods have used graphical models such as Markov Decision Processes or Conditional Random Fields to encode constraints and spatio-temporal relationships. None of these works combine a mobile robot's viewpoint with realtime forecasting.

2.1 Pedestrian intent prediction

Pedestrian tracking, modeling and trajectory prediction algorithms come in many flavors which do not always involve an autonomous agent navigation centric approach. It is common for such algorithms to take for granted information that would not be easily available to a mobile robot, such as a bird's eye view of it's surroundings or oracular annotated trajectory information [26, 36].

Among such works, sequence-to-sequence learning has emerged as a viable candidate for modeling multi-human scenarios in a predictive fashion [2, 44]. Maximal entropy inverse reinforcement learning approaches [22, 24, 48] have also been used to forecast pedestrian trajectories.

While these propose sophisticated models of human-human interactions or exploit available semantic information, they do not reflect the various challenges an on-board view from a mobile robot will experience, including incomplete scene information, imperfect data due to sensor error or occlusions, etc. Instead, we aim to study the problem using an input data distribution closer to the one expected from a mobile robot.

2.2 Pedestrian detection and tracking

Among on-board perception methods, most approaches combine 2D and 3D scene information to build a 3D, human-aware map of the autonomous agent. Some systems fuse handcrafted features from LIDAR and Histogram-of-Gradient features from vision to detect pedestrians [11]. Others align point clusters from two LIDARs and pedestrian detection bounding boxes from three RGB cameras to obtain 3D pedestrian estimates [33]. Such methods prioritize active depth sensors. Our method is agnostic to the source of the depth data and can be used with any source that provides spatial correspondences between an RGB image and points in space, including passive sensors. In our qualitative evaluation we use a stereo camera, the Stereolabs ZED.

2.3 3D Human pose estimation

Markerless human pose recovery from visual data is a challenging but useful capability that has recently seen tremendous success in the computer vision community. The focus has mostly been on joint keypoint localization using a single RGB camera in pixel space (2D pose estimation) [27] of a single individual [34, 38, 45] and, more recently, multiple individuals [4, 7, 37]. The most successful models have employed graphical or neural network models trained on large datasets.

Unlike 2D pose, large-scale data is difficult annotate for 3D pose (predictions in metric space) without dense instrumentation of the environment [20] or of the humans [17], both leading to artificial restrictions on the humans. Despite this, significant interest in 3D pose estimation exists, owing to its numerous applications. Several end-to-end models have been trained on this task that regress the individual skeletal keypoints [31, 32, 35, 40, 41, 47]. More complex methods may predict 2D and 3D methods jointly. Even though these algorithms have the advantage of being able to work with inexpensive RGB cameras, these are monocular 3D pose estimators and, as such, suffer from depth and scale ambiguities that allow multiple plausible 3D pose hypotheses given a 2D pose estimate [18]. This is best shown by the fact that these works perform evaluation by aligning each predicted 3D pose to its corresponding ground truth 3D pose by either translating the predicted pose so that the root nodes are aligned (say, the tailbone for each skeleton) or applying a Procrustes transform (aligning both poses upto a combination of translation, rotation and uniform scaling). Moreover, the real-time methods among these, such as V-Nect [30, 32], can only produce single-person pose estimates and require tracked bounding boxes for each person, making it unsuitable for use in dynamic crowds.

In this paper, we compare against a strong baseline based on the better-performing real-time method [30]. Note that these methods are not trained to account for global translation and rotation. Our baseline provides global rectification to their method in order to overcome this.

Closest to our work is [49], where the authors use a Kinect V2 sensor and a voxel-based neural network to provide 3D poses in metric space. This method achieves impressive results but is not real time and places an explicit requirement on a single type of depth sensor, which we do not. We omit discussion of multi-view pose estimation since it is not relevant to the use case of a single mobile robot in the wild.

2.4 Human pose forecasting

With human pose estimation a well-studied problem with accurate, real-time solutions in specialized use cases, predictive modeling of human pose has received considerable interest in the literature. This section covers works that forecast human pose and motions in more generic cases than the pedestrian prediction in the previous sub-section. Much work in human pose forecasting predicts articulated human poses without considering the acquisition of the pose skeletons themselves. For example, [12, 19, 29] all model and forecast human motion using recurrent neural networks (RNNs). However, these methods do not model the global position of their subjects, instead focusing on generating a continuation of observed human motion in a coordinate frame attached to the body. For forecasting, these works are the closest to ours and we compare against the best performing method in this paper [29].

The graphics community also uses deep recurrent neural networks, primarily for character motion synthesis conditioned on human user input. For example, these methods have been used to animate game characters [14].

While each of these methods generate realistic human motions, they fail to match ground truth human poses and suffer from discontinuity artifacts between the ground truth instances and first predicted instance.

In general, the intrinsic stochasticity of human motion does not allow for accurate forecasting of complete human poses over long horizons ($> 1s$) [29]. For the short horizon case, forecasts are more accurate but still suffer from unrealistic discontinuities at the beginning of the forecast. This is hypothesized to be due to the use of quantitative loss functions in training these models that penalize average error, without imposing temporal smoothness or anatomical constraints in the loss function. For this paper, we refer to short-term forecasting ($<= 400ms$) as just "forecasting" unless otherwise stated.

There has been sparse investigation of how pose estimation can benefit pedestrian perception. The autonomous and assisted driving community has also investigated the use of pedestrian pose-based features for intent prediction with encouraging results. These works restrict intent to higher-level classes such as "cross/no-cross" [9] or "start/stop/cross/bend" [10] for curbside pedestrians. These methods also use 2D pose instead of 3D which removes the ability of these systems to reason about absolute pose, which is adequate for their application, but we wish to investigate the use of finer-grained pose information.

None of the aforementioned works study real-time human pose forecasting with the sensing and computation restrictions of a typical mobile system, as we do in this paper.

Chapter 3

Approach

Keeping with our earlier example of the perception system of a mobile robot that interacts with humans, such a system would require both real-time performance and an output signal that allows human attention/intent prediction.

For both these reasons, we choose to use human torso pose as the perception output. Acquiring accurate, articulated full-body 3D pose in real-time is challenging given the constraints of on-board sensing, which is prone to occlusions because only a single view-point is available. Hence, we restrict ourselves to 3D torso pose, comprising the global Cartesian coordinates of the torso center-of-mass and the torso plane angles. Torso pose also evolves less rapidly than head pose [42], hence mitigating information loss at lower temporal sampling rates, which in turn allows lower hardware design costs.

This also allows us to incorporate a smooth temporal constraint in our model of human pose, which is a non-trivial consideration since previous learning-based methods such as [12, 19, 29], suffer from discontinuity artifacts at the beginning of the forecast, as shown in [29], which are inconsistent with human anatomical limits. For example, see the relatively large error at the start of each error graph in Figs. 5.1, 5.2, and 5.3 corresponding to the HMP method from [29].

Our end-to-end system comprises a torso pose recovery module followed by a forecasting module. The algorithm requires registered RGB and depth inputs with proper calibration. In its most basic form, it is agnostic to the data source. For instance, in the evaluation in Tables 5.2 and 5.1 we used a Kinect v2 (active sensing) as the input source, while our end-to-end qualitative system demonstration used a ZED camera (stereo). This allows flexibility for system designers to trade-off the requirements for their particular scenario. For example, higher fidelity 3D maps can be obtained with a 3D LIDAR at the cost of higher power consumption and overall expense.

3.1 Torso pose parametrization

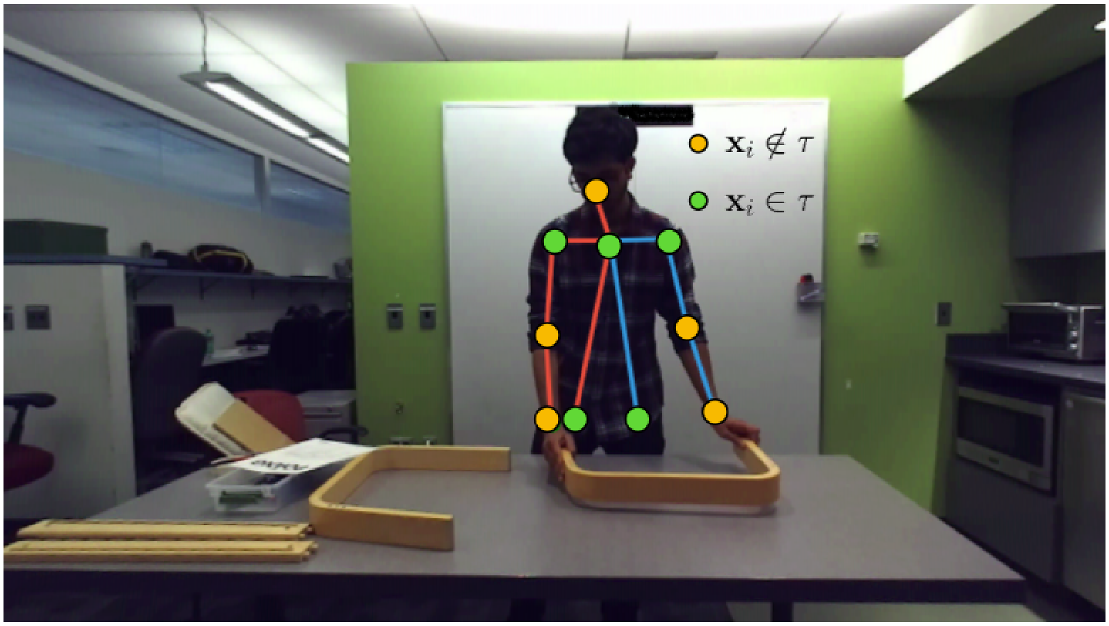

Figure 3.1: The torso pose is represented by the Least-Squares plane fit to the torso points set τ. Here, the relevant torso points are marked in green.

We parameterize torso pose by the position (x, y, z of torso center) and orientation (plane azimuth: α and elevation: θ) of an estimated torso plane. Given a pose skeleton, the torso plane is defined as the plane that minimizes sum of squared distances from each of the 3D torso joint locations. At a given pose skeleton this plane is given by:

$$\mathbf{n}^*, \mathbf{c}^* = \underset{\mathbf{n},\mathbf{c}}{\operatorname{argmin}} \sum_{i=1}^{|\tau|} |\mathbf{n} \cdot \mathbf{x_i} + \mathbf{c}| \tag{3.1}$$

where $\mathbf{n} \cdot \mathbf{x_i} + \mathbf{c} = 0$ defines the torso plane ($\mathbf{n}$ is the plane normal and $\mathbf{c}$ is a constant, both in $\mathcal{R}^3$) and $\mathbf{x_i} \in \tau \subset \mathcal{R}^3$ is the set of all torso joint locations in 3D space.

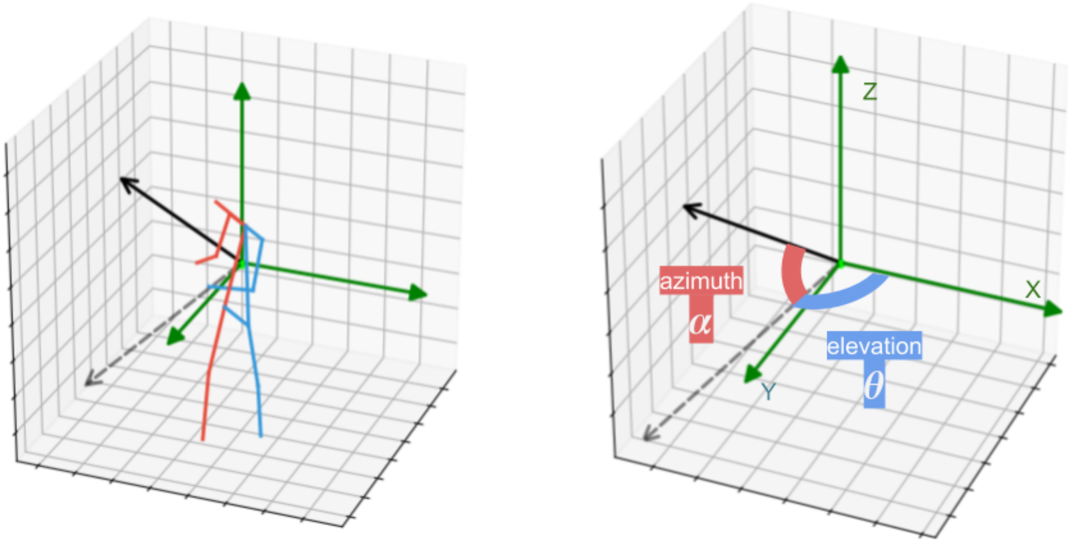

Figure 3.2: Torso pose from 3D joints. (Left) Normal to the torso plane is shown in solid black and its projection to the horizontal plane in dashed gray. Torso center is plotted in fluorescent green. (Right) The plane azimuth (α) and plane elevation (θ) are shown in blue and red, respectively.

Hence, the torso center is:

$$C_{torso} = \frac{\sum_{i=1}^{|\tau|} \mathbf{x_i}}{|\tau|} \tag{3.2}$$

Once τ is constructed, the plane is calculated using Equation 3.1, giving the plane azimuth (α) and elevation (θ) directly (Fig. 3.2).

$$\alpha = \arctan \frac{\mathbf{n}_y}{\mathbf{n}_x} \tag{3.3}$$

$$\theta = \arccos \frac{\mathbf{n}_z}{||\mathbf{n}||_2} \tag{3.4}$$

3.2 Torso pose recovery

We use an off-the-shelf 2D human pose detection system [4] in conjunction with registered depth information in a two-step process. The input to the 2D pose detector is an RGB image, which is used to obtain joint locations for humans in the scene. Once 2D joint locations are known, they are projected onto a registered point cloud obtained by triangulation in a separate step, giving us 3D joint locations.

For our method, we compose the set τ comprising solely the torso points available from the 2D pose detector. For annotated ground-truth skeletons from the datasets used in our evaluation, τ contains the shoulder joints, two hip joints, the mid-spine, and the tip of the tailbone (see Fig 3.1). In this formulation, the registered point-cloud is constructed at every time-step and the points $\mathbf{x_i}$ corresponding to the detected joints in the RGB image are picked from the corresponding point cloud. Hence, $\mathbf{x_i}$s are in metric 3D space. For each $\mathbf{x_i}$, temporal consistency is enforced by discarding values that deviate over $10\ cm$ between consecutive time-steps, lending some robustness to temporary occlusion. The discarded values are replaced by the corresponding $\mathbf{x_i}$ from the previous time-step. See Fig. 3.3 for an overview of this method.

3.3 Torso pose forecasting

For forecasting, we predict elevation, azimuth, and absolute position of the torso plane for a variable time lookahead, from a $2\ s$ history. Once the torso plane is acquired from the pose estimation module, we apply a low pass filter to the two orientation components. This is followed by fitting an Nth order polynomial smoothing spline, which is used to extrapolate

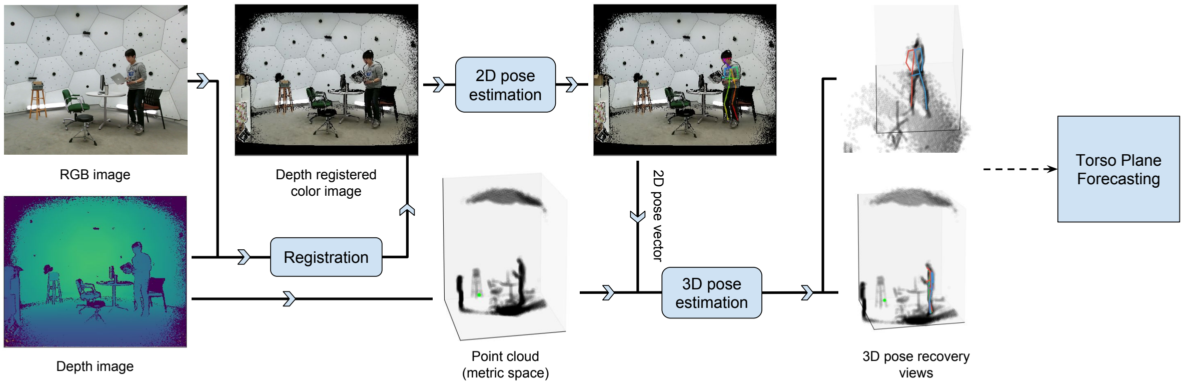

Figure 3.3: 3D torso pose estimation algorithm overview. For details see Section 3.2.

a forecast for each individual component in a univariate fashion. For more details about this univariate spline fit, please see Section 3.3.2. Error analyses across various orders of the fitted polynomial (N) as well as several baselines (see Section 4.1.2) are presented in Table 5.3 .

3.3.1 Low-pass filter

The low-pass filter is a standard second-order Butterworth filter, with the cutoff frequency empirically set to 5Hz. A low pass filter was chosen so that we model only the macro-level orientation of a human subject, which is the most relevant signal for many activities. We also wish to avoid jitter in the torso pose since regression-like methods such as the polynomial fit are sensitive to outliers. We choose a Butterworth filter since it guarantees maximal flatness in the passband of the frequency response.

This is an essential step, as can be seen from the ablation in Tables 5.4 and 5.5 where the unfiltered signal is used directly to fit the spline and extrapolate for forecasting, leading to exploding error.

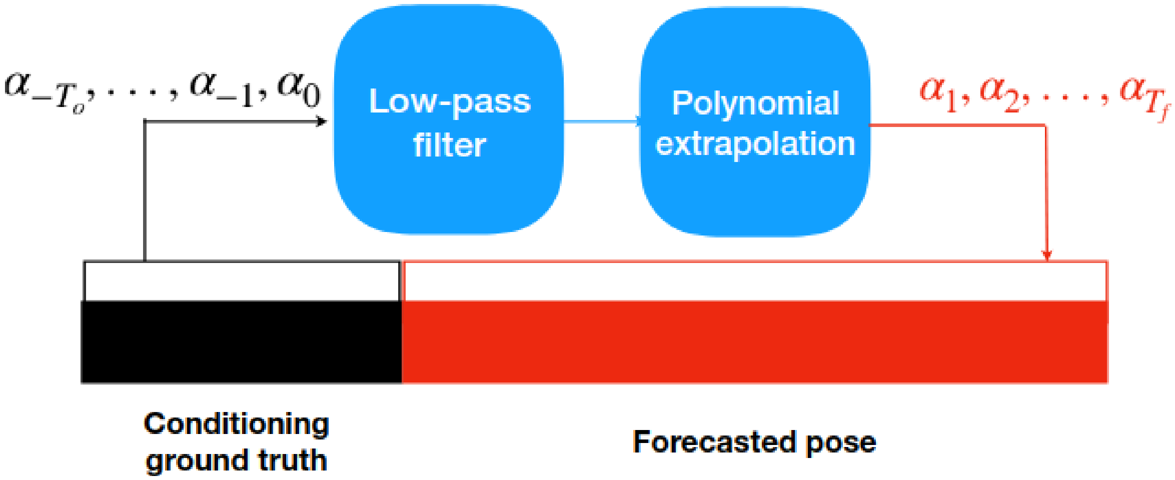
$\alpha_{-T_o}, \ldots, \alpha_{-1}, \alpha_0$
Low-pass filter
Polynomial extrapolation
$\alpha_1, \alpha_2, \ldots, \alpha_{T_f}$
Conditioning ground truth
Forecasted pose

Figure 3.4: Overview of the pose forecasting paradigm, for the azimuth (α) component of torso pose. The same paradigm applies to all other components. See Algorithm 1 for an algorithmic overview and Section 3.3.2 for mathematical details

3.3.2 Smoothing spline fit

The individual components of the torso pose forecasting system are then fit via a smoothing spline.

A smoothing spline is chosen over other spline fit methods because smoothing splines often result in similar fits to other more complex fitting methods, such as kernel regression, but are much more efficient [6]. This efficiency comes from bypassing the knot selection problem and using the input points, directly, as knots. However, since this may lead to overfitting, the smoothing spline uses a regularization term to counteract this.

We cover the forecasting procedure below in detail for one of the torso pose components, the torso plane azimuth, α. This procedure is similar for the other 4 components: elevation (θ), and the three position components (x, y, z).

To reiterate the aforementioned forecasting method mathematically, we first observe the torso pose signal samples, $\alpha(t)$ for a history of discrete time steps $t = -(T_o - 1), -(T_o - 2), \ldots, -1, 0$. These observed poses are referred to as the conditioning ground truth, which consists of T_o samples. In our Human3.6M dataset experiments in Section 4, the conditioning window is 2 seconds long with samples drawn at 25 Hz. Hence, in this cases $T_o = 50$

We then fit a spline, $\hat{\alpha}(t)$ to this observed data and then extrapolate to compute the forecast for all samples in the future, $t > 0$. The spline function, $\hat{\alpha}(t)$ which is of order n may be represented as a sum of normalised B-splines:

$$\hat{\alpha}(t) = \sum_{i}^{M} c_i B_{i,n}(t) \qquad \text{where } n \text{ is the B-spline order} \tag{3.5}$$

The number of basis functions used, N_b depends on the order of the polynomial and the number of knots. For a sequence of knots (arranged in ascending order), $k_1, k_2, ..., k_M$ the B-spline of order 1 is given by:

$$B_{i,1}(t) = \begin{cases} 1 & \text{if } k_i \leq t < k_{i+1} \\ 0 & \text{else} \end{cases} \tag{3.6}$$

A B-spline of order $n + 1$ is defined recursively as:

$$B_{i,n+1}(t) = \frac{t - k_i}{k_{i+n} - k_i} B_{i,n}(t) + \frac{k_{i+n+1} - t}{k_{i+n+1} - k_{i+1}} B_{i+1,n}(t) \tag{3.7}$$

The relationship between the number of spline basis functions, N_b, the polynomial order, n, and the number of knots, M is given by:

$$N_b = M - (n + 1) \tag{3.8}$$

So, our azimuth spline, of order n, may similarly be represented as a linear combination of B-splines:

$$\hat{\alpha}(t) = \sum_{i}^{N_b} c_i B_{i,n}(t) \tag{3.9}$$

Then, all that remains to find the unique spline that best fits a given set of data points is to find the coefficients of the linear combination by optimizing an objective function.

We use the following standard smoothing spline objective[6], which is minimized to obtain the coefficients.

$$\begin{aligned} J(\mathbf{c}) &= \sum_{t=-(T_o-1)}^{0} w(t) * (\alpha(t) - (\hat{\alpha}(t)))^2 + \lambda \int (\hat{\alpha}''(t))^2 dt && (3.10) \\ &= \sum_{t=-(T_o-1)}^{0} w(t) * \left(\alpha(t) - \left(\sum_{i} c_i B_{i,n}(t) \right) \right)^2 + \lambda \int (\hat{\alpha}''(t))^2 dt && (3.11) \end{aligned}$$

Here, λ is a smoothing hyperparameter (we use $\lambda = 1$) and the second term is a roughness penalty which is used to control the smoothness of the spline. This term is relevant only for higher order splines ($n > 3$). For lower orders, this objective reduces to just a least-squares problem with a solution that can be obtained efficiently via SVD and the smoothing spline reduces to an interpolating spline. For higher orders, the solution can also be found in a closed form [6] and is presented here for convenience:

$$C = (B^T B + \lambda \Omega)^{-1} B^T \alpha \tag{3.12}$$

Here, C is the column vector of all coefficients, c_is. The matrix B is defined as $B_{i,j} = B_{j,n}(t_i)$ and α is th column matrix of all observed poses $\alpha_t \forall t \in [-(T_p - 1), \ldots, -1, 0]$. The matrix Ω encompasses the penalty terms coming from the second term in the objective funciton, from Eq 3.10.

Additionally, this formulation allows us to incorporate different levels of confidence in different torso pose detections directly into the forecast, by weighing the importance of the different input points in proportion to their detected confidences (w_t in Eq 3.10). This will result in the spline giving more importance to the torso poses detected with higher confidence and less to those with lower detection confidences. For our evaluation experiments in Chapter 4 we used uniform weighting across all input points. This is done because the conditioning torso poses are all acquired from ground truth motion capture rather than detected from the real world via some algorithm.

Note that the two orientation components (azimuth and elevation), require a phase unwrapping step at the end, since we have treated them as linear variables while performing the spline fit and extrapolation. Phase unwrapping is performed by correcting the sequence of forecasted angles by adding multiples of $\pm 2\pi$ when absolute jumps between consecutive elements are greater than or equal to π radians.

We presented a filter-and-fit method to forecast torso pose components, individually, in a univariate fashion that is summarized in the following algorithm:

Algorithm 1: Azimuth forecasting via spline fit

```
Input: $\alpha(t)$: $\alpha_{-T_o+1}, \ldots, \alpha_{-1}, \alpha_0$
        /* Require history of torso plane azimuth angles */
Output: $\hat{\alpha}(t)$: $\forall t > 0$
        /* Forecast of torso plane azimuth angles */
begin
    Low-pass filter the obtained pose angles $\alpha_{filt}(t) = b(\alpha(t))$
    Set knots at all input points, $\alpha_{-T_o+1}, \ldots, \alpha_{-1}, \alpha_0$;
    Use knots and minimize $\sum_{t=-T_o+1}^{0} w_t * \left(\alpha(t) - \left(\sum_i c_i B_{i,n}(t)\right)\right)^2$ for all $c_i$ to fit
      spline
    Compute the forecast $\hat{\alpha}(t) = \sum_i c_i B_{i,n}(t) \forall t > 0$
    Correct the obtained $\hat{\alpha}(t)$ for phase by phase unwrapping
end
```

Chapter 4

Experimental setup

4.1 Evaluation procedures

4.1.1 Torso pose recovery

To evaluate this component of our algorithm, we wanted to simulate single view-point visual sensing (e.g. a mobile robot with on-board sensing) by using inputs from a single Kinect v2 RGB-D sensor in the Panoptic Studio [20]. This allows us to test pose recovery in the presence of occlusions, which is important for applications like dynamic pedestrian tracking in busy environments.

For each frame during a sequence, the pose recovery component of the algorithm in Fig. 3.3 is used. Ground truth articulated pose (which is reconstructed with a combination of over 500 camera views and a 2D pose estimation method [4, 20]) is used to compute the ground truth torso plane, as in Equation 3.1. The body center-of-mass and plane angle errors are shown in Table 5.3.

As a competitive baseline, we use the method of [30]. This is a 4-layer shallow neural network trained on the Human 3.6M dataset [17] to "lift" a given 2D pose detection into 3D space. Being a 4-layer neural network with relatively low dimensional inputs and outputs, this only takes about $5\ ms$ per forward pass, making it suitable for real-time deployment, as opposed to other more computationally intensive 3D human pose predictors discussed in Section 2.3. Comparing against this method is also fair since it also tries to bridge the gap between 2D and 3D pose estimation rather than attempting monolithic 3D pose estimation.

In the method from [29], the 3D poses are not guaranteed to be recovered in global-scale. To transform their output pose into the global coordinate frame, we find and apply the best least-squares rigid transform between the 3D predicted pose and 3D ground truth pose. Note that this represents an unattainable gold-standard performance for this method, since ground-truth pose is never available in a real-world setting.

4.1.2 Torso pose forecasting

For quantitative evaluation on Human 3.6M, we used the same train-test split as [12, 19, 29] and compared against [29] since it is the quantitatively best performing model of the three. In [29], the MoCap data was down-sampled to 25 Hz. During testing, skeletal poses over a $2\ s$ sample (50 frames) were fed to a recurrent neural network (single-layer), which then generated samples over a forecast window of $400\ ms$ (10 frames) sample. The initial 50 frames are referred to as the conditioning ground truth. Their method also has the advantage over previous work [12, 19] in that it trains one-model across all actions in the dataset. We retain this advantage by using the same set of filter parameters for the entire dataset, eliminating the need to tune for every individual action.

The choice of the $400ms$ forecasting method follows from previous work [12, 19, 29]. Further, to properly characterize the properties of our method and HMP [29] as well as to enable comparison of the two, we present the analysis for multiple forecast windows.

The aforementioned methods do not estimate the 3D pose of a human from visual data. Rather, they acquire the ground truth 3D poses directly obtained from the MoCap data accompanying Human3.6M. For evaluating our pose forecasting method in this experiment, our pose estimation module was bypassed to keep the quantitative comparison of our forecasting system with [29] fair.

Since our method focuses on torso planes rather than full body articulated pose, we must obtain ground truth planes from the MoCap data. This was done by fitting a least squares plane to hip, shoulder, and neck joints of an articulated pose obtained from the MoCap data, as described in Equation 3.1.

Additionally, instead of using the Euclidean distance in Euler angle space for all body joints (as in previous work [29]), we computed the angle error of the plane orientation forecast. We chose this measure since it is most indicative of the macro-level expression of torso pose. See Table 5.3 for average azimuth and elevation angle error for each of the 15 Human3.6M activities as well as within the subcategories described in section 4.2.

Baseline methods

We use the following torso pose forecasting baselines:

1. **Human motion prediction (HMP)** [29]: A seq-to-seq GRU based method, that conditions a model on an observed pose sequence to generate a forecasted pose sequence.

2. **Zero velocity**: The last observed torso pose during the conditioning window is predicted for the entire prediction window.

3. **Constant velocity (full $2s$ window)**: The average velocity over the entire conditioning window is integrated during the forecasting window to get torso pose.

4. **Constant velocity (0.4s window)**: The average velocity over the last 0.4s of the conditioning window is integrated during the forecasting window to get torso pose.

4.2 Datasets

We evaluated the two modules, pose recovery and forecasting, on two datasets respectively:

- **Panoptic Studio** [20]: We used this for quantitative evaluation of the pose recovery system. It contains RGB-D inputs and multi-person scenarios, representing the closest available data to our target application.

- **Human 3.6M** [17]: We used this for quantitative evaluation of the pose forecasting system. 3D pose in world-coordinates can directly be obtained from their marker-based MoCap system. The lack of RGB-D views prevents us from testing the system end-to-end.

While we would ideally evaluate our work end-to-end, most datasets with grounding for 3D pose are marker-based and do not have associated RGBD data. The Panoptic Studio does have marker-less grounding for 3D pose but the contained activities (e.g. ”Office”:sitting at a desk or ”Range of Motion”: arm movements that keep the torso mostly stationary) are unsuited to evaluation of a torso pose forecasting method, hence we perform modular evaluation.

The motivation of this work is to provide a fast method to recover and forecast multi-person 3D pose in the real world, with a focus towards social navigation. Hence, ideal evaluation would be data-driven and with said data collected in the wild. However, instrumenting to recover accurate 3D pose in such scenarios is difficult due to financial, computational, and privacy concerns. Unfortunately, 3D pose datasets with multiple, simultaneous humans and RGBD inputs are rare.

The best effort in this domain is the Panoptic Studio [20], which uses advances in 2D pose recognition with 500 RGB cameras and 10 Kinect sensors to recover the 3D pose of observed humans accurately. While still being an artificial environment, housed in a geodesic sphere of diameter $5.49\ m$, it solves the occluded pose recovery problem by dense instrumentation of the environment rather than equipping humans, allowing for more naturalistic movement. This dataset has the added benefit of multi-person capture sequences that present several types of occlusion challenges likely to also be found in dynamic crowds. We use relevant sequences with Kinect inputs and ground-truth 3D human pose present. This amounts to about 100 minutes of data at 30 Hz.

Although it seems that this dataset is highly appropriate for evaluating our end-to-end system, most of these sequences involve largely stationary participants involved in social activities such as meetings and lunches. This would make any forecasting evaluation on such data unrepresentative of performance on dynamic scenarios where forecasting is most useful. Hence, we restrict ourselves to pose recovery evaluation on the Panoptic Studio data, rather than end-to-end (recovery and forecasting) evaluation. An additional consideration is that the input sensor is always stationary, as opposed to being on a mobile robot.

To evaluate the pose forecasting module we chose the Human 3.6M [17] dataset. This is currently the largest publicly available dataset of motion capture data, containing 7 actors performing 15 varied activities such as walking, taking photos or giving directions, with only a single person per task. We group the tasks into three categories: Pedestrian, Constrained, and High Variance. The Pedestrian group contains Walking, Walking Dog, Walking Together activities. The Constrained group comprises of Eating, Smoking, Phoning, Sitting, Sitting Down which all involve the person's torso being constrained in some fashion (e.g. by being placed in a rotating chair). The High Variance group comprises of all other activities such as Taking Photo, Posing. etc which have mostly stochastic motion where very little intent is telegraphed. In our opinion this is not really relevant to evaluate forecasting models since the premise of motion history based forecasting is that consecutive motions are correlated and motion intent is telegraphed. Nevertheless we perform evaluation for comparative purposes.

Once 3D torso pose is acquired, our analysis is local and does not consider inter-person effects, meaning single-person sequences are equivalent to multi-person scenarios for evaluation purposes. Prior pose forecasting work [19, 29] has also evaluated on this data. Consequently, an evaluation procedure exists for articulated pose forecasting which we adapted to the torso pose scenario. While this data seems like an ideal candidate for end-to-end system evaluation, lack of registered RGB and depth views render it unusable for that purpose.

Chapter 5

Results

5.1 Quantitative results

5.1.1 Torso pose recovery

Tables 5.1 and 5.2 show the results of our pose estimation method and a baseline using a state-of-the-art, learned, 3D pose predictor[30]. These results show that learning for 3D pose estimation may need more improvement before it can be used for accurate, real-time performance suitable for robot deployment.

We see comparable performance for our method and that of [30]. Since, the latter has access to information about ground truth rotation and translation (as a rigid transform), which

Table 5.1: Torso plane X,Y,Z estimation errors on Panoptic studio [20] data (centimetres/degrees). Note that a rigid ground-truth transform (GTT) is applied to [30] and these numbers are unattainable in real-world settings. For more details, see Chapter 3

Activity	Torso Center X (cm)		Torso Center Y (cm)		Torso Center Z (cm)	
	Ours	[30] + GTT	Ours	[30] + GTT	Ours	[30] + GTT
Range of Motion	7.64	4.46	3.91	12.26	15.59	4.73
Office	5.83	12.11	2.85	11.41	12.52	14.85

Table 5.2: Torso plane orientation estimation errors on Panoptic studio [20] data (centimetres/degrees). Note that a rigid ground-truth transform (GTT) is applied to [30] and these numbers are unattainable in real-world settings. For more details, see Chapter 3

Activity	Plane Azimuth (deg)		Plane Elevation (deg)	
	Ours	[30] + GTT	Ours	[30] + GTT
Range of Motion	30.25	33.98	14.44	8.24
Office	24.95	41.47	10.60	9.70

is not available in real world scenarios, we assert that our method enables 3D pose recovery with far less overhead.

In terms of computational performance, our bottleneck lies in the 2D pose estimation step. We are able to achieve a performance of 10 Hz using an Nvidia 1080Ti. However, our method is not tied to a particular type of 2D pose estimator. A faster pose estimator, such as the recent work in [39] (180 Hz on similar GPU with similar accuracy, real-time performance on CPU) can significantly speed up performance without sacrificing accuracy.

5.1.2 Torso pose forecasting

Table 5.3 shows the results of pose forecasting methods, including polynomials of degree $N = 1$ and 2, state-of-the-art human motion predictor (HMP, the quantitatively best performing method for on Human3.6M) [29], and a constant prediction baseline ($N = 0$) (where the last ground-truth torso plane orientation is predicted for the entire forecast window). The results also show the importance of the filtering step (see last row, where we omit the filter and directly fit an Nth order polynomial to unfiltered data.)

Tables 5.4 and 5.5 show the breakdown of torso plane orientation error by individual action type in the Human3.6M dataset.

A visual representation of the error as it evolves with the forecasting extent across time is shown in Figures 5.1, 5.2, and 5.3.

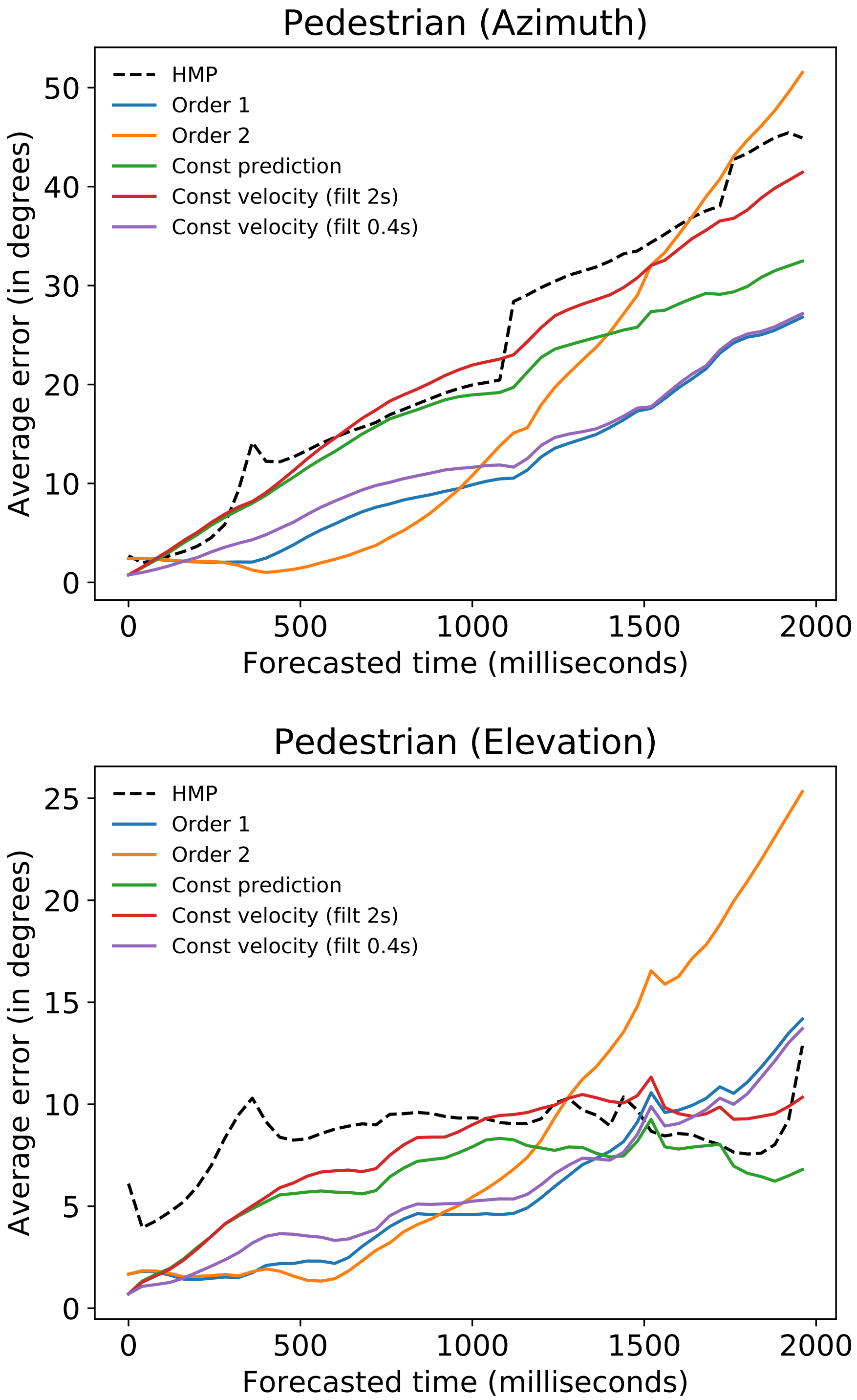

Figure 5.1: Average forecasting error (across ”Pedestrian” test sequences) vs. forecasting time extent for various categories of Human3.6M data (lower is better). The plots show our recommended method (Order 1), two baselines (Order 0 and 2), and the RNN-based method from HMP [29].

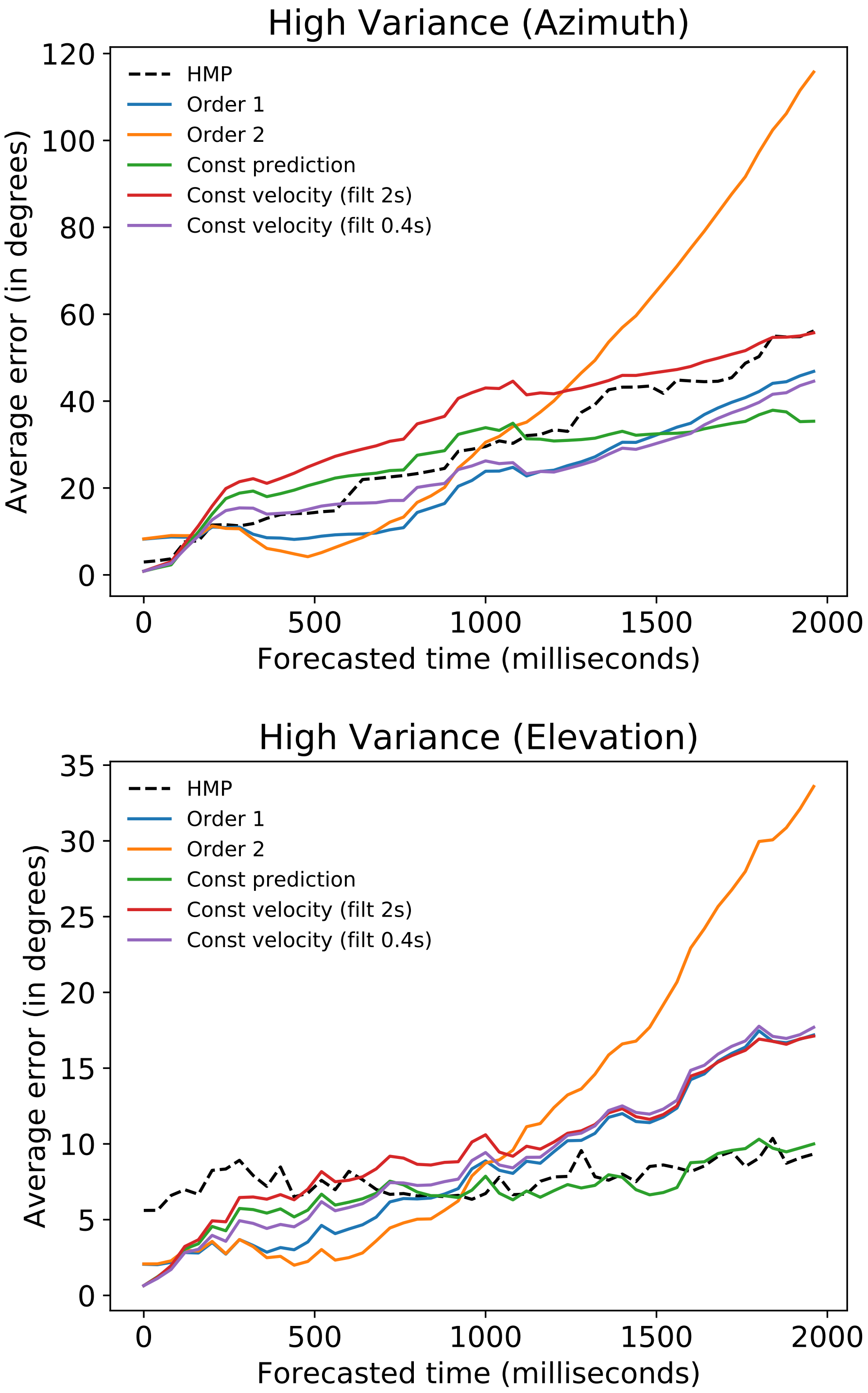

Figure 5.2: Average forecasting error (across ”High Variance” test sequences) vs. forecasting time extent for various categories of Human3.6M data (lower is better). The plots show our recommended method (Order 1), two baselines (Order 0 and 2), and the RNN-based method from HMP [29].

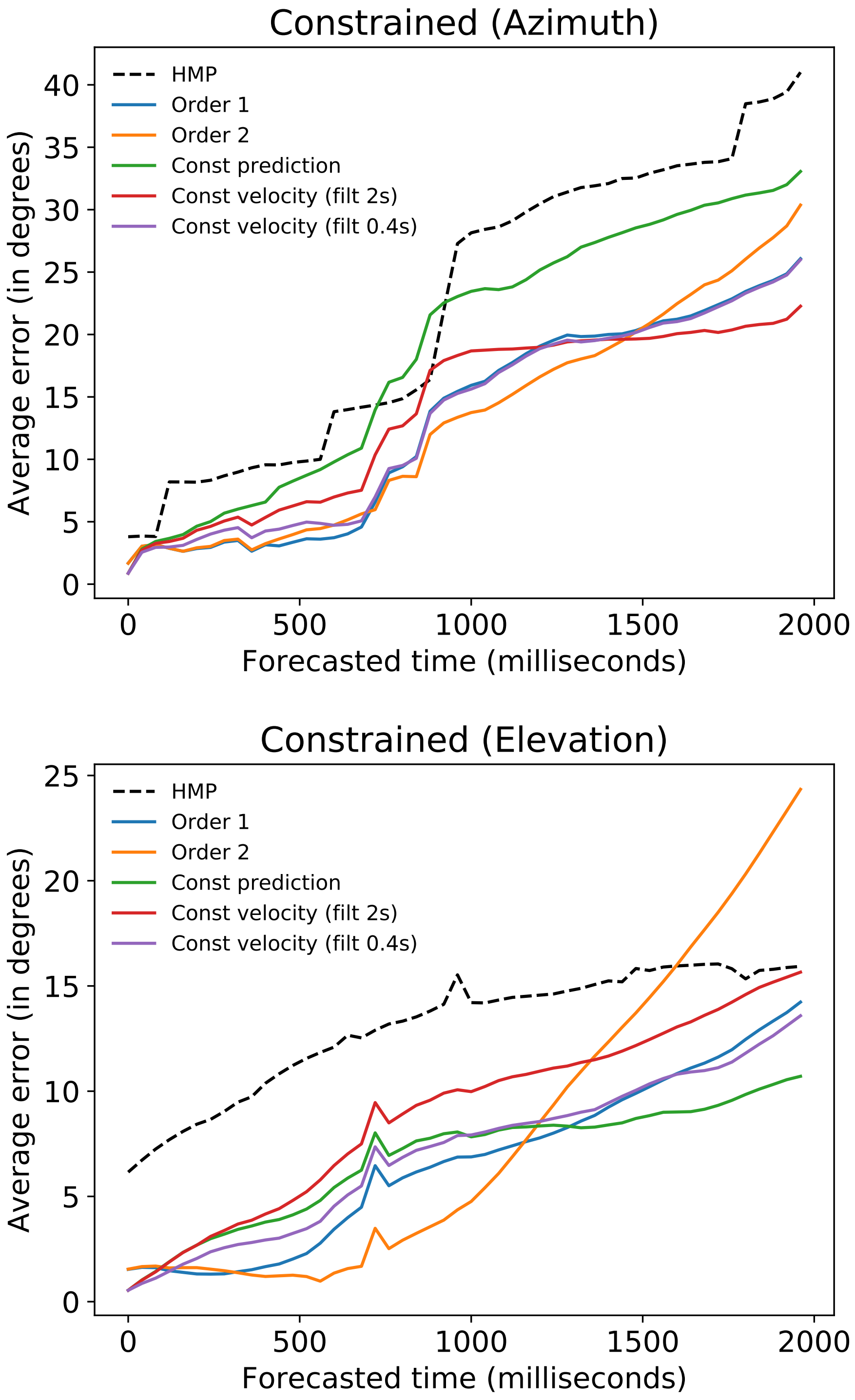

Figure 5.3: Average forecasting error (across ”Constrained” test sequences) vs. forecasting time extent for various categories of Human3.6M data (lower is better). The plots show our recommended method (Order 1), two baselines (Order 0 and 2), and the RNN-based method from HMP [29].

Table 5.3: Torso plane orientation forecasting errors for various forecasting windows on H3.6M [17] data (degrees)

Forecast time→	400 ms				1 s				2 s			
Activity ↓	Const	$N = 1$	$N = 2$	HMP[29]	Const	$N = 1$	$N = 2$	HMP[29]	Const	$N = 1$	$N = 2$	HMP[29]
					Plane Azimuth							
Constrained	4.24	**2.87**	2.92	7.13	9.83	5.49	**5.37**	11.48	18.98	**13.11**	**13.11**	22.32
HV	10.87	9.38	9.1	**8.45**	19.13	**11.01**	11.03	15.72	26.32	**21.98**	38.92	28.98
Pedestrian	4.4	2.17	**2.09**	5.02	10.46	4.82	**3.22**	11.45	18.23	**11.32**	16.69	22.83
All	7.36	5.77	**5.64**	7.33	14.3	7.93	**7.58**	13.45	22.26	**16.89**	25.87	25.53
All (no filter)	7.36	**5.4**	13.56	7.33	14.3	14.15	50.81	**13.45**	**22.26**	27.2	155.87	25.53
					Plane Elevation							
Constrained	2.32	**1.45**	1.54	8.13	4.62	3.24	**1.99**	10.83	6.77	**6.63**	8	13.06
HV	3.59	**2.79**	2.81	7.21	5.3	4.32	**3.53**	7.09	**6.68**	8.48	11.65	7.69
Pedestrian	2.81	**1.6**	1.68	6.53	4.86	2.6	**2.34**	8.02	6.25	**5.61**	8.4	8.55
All	3.01	**2.11**	2.16	7.38	4.98	3.62	**2.78**	8.52	**6.63**	7.29	9.79	9.65
All (no filter)	**3.01**	3.26	16.35	7.38	**4.98**	8.84	67.58	8.52	**6.63**	17.33	202.66	9.65

Table 5.4: Torso plane azimuth forecasting (400ms) errors on H3.6M [17] data (degrees)

Action	Plane Azimuth (Degrees)						
	HMP[29]	Const	$N = 1$	$N = 2$	$N = 3$	CVel (2s)	CVel (0.4s)
walking	4.44	2.39	2.45	2.76	2.59	3.45	2.78
eating	4.48	0.97	0.99	1.11	2.94	2.21	1.49
smoking	1.43	0.89	0.74	0.83	17.49	1.53	0.96
discussion	3.13	3.79	4.7	5.53	4.43	3.99	5.92
directions	2.66	1.62	1.3	1.34	4.45	2.64	2.22
greeting	12.68	11.24	10.14	10.24	4.18	17.1	10.93
phoning	7.44	7.38	7.53	8.46	6.54	7.06	8.04
posing	9.65	6.81	7.22	7.69	13.25	10.24	6.16
purchases	18.85	14.4	13.25	14.33	22.44	21.63	13.57
sitting	2.49	1.79	2.01	2.28	2.93	2.42	2.3
sittingdown	5.36	3.32	3.33	3.62	5.76	5.82	3.51
takingphoto	2.14	4.13	5.24	5.5	5.88	3.8	2.73
waiting	26.96	23.68	21.87	22.97	4.509	27.96	23.16
walkingdog	4.89	3	2.6	2.97	9.34	7.43	2.69
walkingtogether	3.88	1.13	1.21	1.28	3.13	2.93	1.82
All (average)	7.36	5.77	**5.64**	6.06	7.32	8.01	5.88

Table 5.5: Torso plane elevation forecasting (400ms) errors on H3.6M [17] data (degrees)

Action	Plane Elevation (Degrees)						
	HMP[29]	Const	$N = 1$	$N = 2$	$N = 3$	CVel (2s)	CVel (0.4s)
walking	1.91	1.32	1.25	1.28	4.29	1.91	1.76
eating	1.13	1.01	1.05	1.09	3.78	1.21	1
smoking	1.77	1.19	1.19	1.25	7.77	2.02	1.42
discussion	1.42	0.83	0.79	0.84	4.22	1.64	1.31
directions	2.12	1.89	2.04	2.1	5.97	2.16	1.75
greeting	3.35	3.41	3.29	3.34	4.23	3.43	3.29
phoning	1.63	1.41	1.6	1.66	8.37	1.49	1.61
posing	5.46	3.61	4.16	4.38	8.33	5.32	4.3
purchases	6.54	4.17	4.08	4.26	9.81	7.6	5.21
sitting	3.21	1.87	1.98	2.1	10.79	3.12	2.44
sittingdown	3.85	1.79	1.88	1.98	9.91	4.12	2.67
takingphoto	3.17	2.89	2.76	2.92	10.73	3.78	2.37
waiting	3.05	2.77	2.52	2.58	7.16	3.92	3.46
walkingdog	4.75	2.21	2.57	2.75	12.46	4.66	1.91
walkingtogether	1.77	1.25	1.22	1.25	2.84	1.82	1.67
All (average)	3.00	**2.11**	2.16	2.25	7.38	3.21	2.41

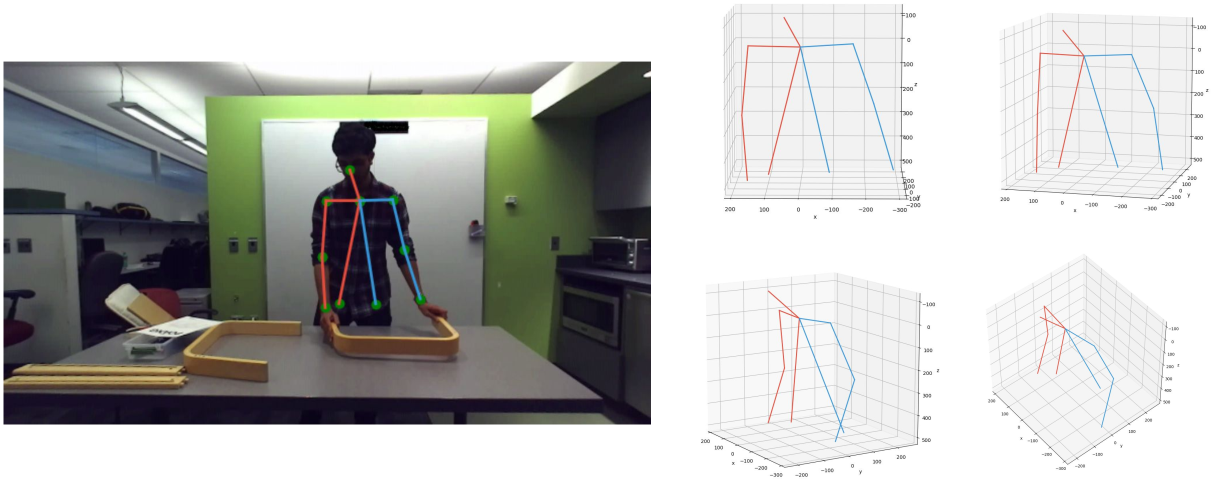

Figure 5.4: Qualitative torso pose recovery results on a furniture assembly task. The input sensor is a consumer stereo (ZED) camera

5.2 Qualitative results

For qualitative analysis of our pose recovery system, we collected data in the real world on a furniture assembly task, in which a human subject followed print instructions to assemble an IKEA ottoman. We chose this task to illustrate the capabilities and limitations of the system. We saw realistic 3D pose estimates in the output and included the results in a video demonstration. Some sample poses are shown in Figure 5.4 and the video demonstration can be viewed at this link.

5.3 Discussion

A few trends can be seen in Table 5.3 and the error graphs in Figs. 5.1, 5.2, and 5.3.

First, the plane azimuth is harder to predict than the elevation, given the higher error rates across all 15 activity sequences and various methods. However, the best average error for both torso orientation components is under 5 degrees. This is small enough to not cause ambiguity in most real-world activities.

Second, the filtering step is essential. Without it, we see larger errors in the polynomial fitting and the errors tend to explode in the larger forecasting windows (Table 5.3). This suggests that the forecast becomes unstable for higher order approximations due to susceptibility to high frequency components in the pose variation.

Third, the recurrent neural network model from HMP[29] tends to make much larger errors than our simple 1st degree (linear) polynomial fit, especially over the short-to-medium term (i.e., $400\ ms$-$1s$) and over all windows for the Pedestrian group of activities. The HMP errors also show higher variability across tasks than our method.

This suggests that such models are either over-fitting or that the error they are trained to minimize is unsuitable for our task. That is, recurrent neural network based methods try to minimize a quantitative loss without reasoning about the temporal smoothness of human motion. Thus, these methods can suffer from unrealistic discontinuities. This is reinforced by the observation that these errors are larger in the High Variance tasks such as “Taking Photo” (both upright and kneeling poses) or “Directions” (high variance poses).

Fourth, the constant velocity baselines described in Section 4.1.2 and seen in Figs. 5.1, 5.2, and 5.3 come very close to the first order spline. In particular, the need for the low pass filter is further emphasized by the exploding error on averaging

Our forecasting algorithm is inexpensive to compute while being faster and more accurate (for short horizons) than previous work. The method described in [29] (HMP) takes about $35\ ms$ for one forward pass on a dedicated Nvidia Titan X GPU. This translates to a maximum sampling rate of 28 Hz, assuming desktop-level hardware is available on-board. Note that this is the computational cost of just the HMP forecasting method. This is significant since it my be an additional bottleneck if used in conjunction with another 2D/3D pose recovery method for end-to-end pose recovery and forecasting over our forecasting method. Our method takes approximately $0.715\ ms$ on an Intel i7-6700HQ CPU (laptop processor). This makes our method about $45\times$ faster on cheaper and more accessible hardware.

It is important to note that HMP forecasts full body articulated 3D pose using about $34000\times$ as many learned parameters, while we only model the torso plane. This makes the $45\times$ running time speed-up less surprising. However, this difference in pose information demonstrates the power of our forecasting technique. Since HMP models the torso with more granularity and parameters than our method, it is reasonable to expect much better performance in the medium to long term. This is however not always the case, as can be seen from Table 5.3

Chapter 6

Conclusion

6.1 Summary

We propose a novel end-to-end torso pose estimation and forecasting system which is relevant for rapid perception and re-planning loops of robot decision making in highly dynamic environments, such as the case of social navigation in an autonomous mobile, service robot.

We parameterized torso pose uniquely by the position and orientation of a torso plane (Equation 3.1). We evaluated the pose estimation quantitatively and compare against a state-of-the-art monocular approach, showing comparable results against a strong baseline. The evaluation was performed in a replicable manner using a publicly available dataset while also simulating the single viewpoint sensing of a mobile robot, thus allowing fair and easy benchmarking in the future.

In addition to torso pose estimation, our approach predictively models absolute torso position. We present a comparative quantitative evaluation and show that our simple filter and fit method outperforms complex recurrent neural network methods for the short-to-medium horizon case while being competitive over the long horizon case. For walking motions, it also accurately predicts the torso facing direction (plane azimuth) which is an important predictive cue of pedestrian trajectory intent. In this context, larger models or more complex modeling has not lead to better forecasting performance. Further, our method is approximately $45\times$ faster on the torso plane forecasting task, implying suitability to navigation in human environments. We also identified a need for more realistic pedestrian datasets to evaluate pedestrian pose and trajectory detection and forecasting systems.

6.2 Future work

In future work, we would like to apply our method to tasks that require multi-person pose perception, like social navigation, to measure the intent prediction capability of torso pose. We imposed several constraints on this work, including a focus solely on the torso plane (as opposed to full-body), which we will provide more validation for, via a downstream task such as planning for social navigation, in future work. Comparing the two pose representations in this manner would quantify the difference in intent-prediction capability, allowing an informed trade-off of computational load vs performance. This requires the collection of a dataset with naturalistic pedestrian motions, on-board robot point-of-view sensing which is calibrated to instrumentation around the robot to generate pose annotations for pedestrians around the robot in a global frame. We plan to collect such a dataset.

Appendix A

Appendix

A.1 Human 3.6M

A.1.1 Categorization of Human 3.6M Activities

The categories used in Chapter 3 are composed of the following actions:

Pedestrian	**Constrained**	**High Variance**
Walking	Eating	Posing
Walking Together	Smoking	Purchases
Walking Dog	Phoning	Discussion
	Sitting	Directions
	Sitting Down	Greeting
		Taking Photo
		Waiting

Table A.1: Categorization of Human 3.6M dataset

Bibliography

[1] Henny Admoni and Siddhartha Srinivasa. Predicting user intent through eye gaze for shared autonomy. In *Proceedings of the AAAI Fall Symposium Series: Shared Autonomy in Research and Practice (AAAI Fall Symposium)*, pages 298–303, 2016. 2

[2] Alexandre Alahi, Kratarth Goel, Vignesh Ramanathan, Alexandre Robicquet, Li Fei-Fei, and Silvio Savarese. Social lstm: Human trajectory prediction in crowded spaces. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 961–971, 2016. 2.1

[3] E. Avrunin and R. Simmons. Socially-appropriate approach paths using human data. In *The 23rd IEEE International Symposium on Robot and Human Interactive Communication*, pages 1037–1042, Aug 2014. doi: 10.1109/ROMAN.2014.6926389. 1

[4] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. In *CVPR*, 2017. 1, 2, 2.3, 3.2, 4.1.1

[5] I. Chatterjee and A. Steinfeld. Performance of a low-cost, human-inspired perception approach for dense moving crowd navigation. In *2016 25th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*, pages 578–585, Aug 2016. doi: 10.1109/ROMAN.2016.7745176. 1

[6] Paul Dierckx. *Curve and surface fitting with splines*. Oxford University Press, 1995. 3.3.2, 3.3.2, 3.3.2

[7] Hao-Shu Fang, Shuqin Xie, Yu-Wing Tai, and Cewu Lu. RMPE: Regional multi-person pose estimation. In *ICCV*, 2017. 2.3

[8] Hao-Shu Fang, Yuanlu Xu, Wenguan Wang, Xiaobai Liu, and Song-Chun Zhu. Learning pose grammar to encode human body configuration for 3d pose estimation. 2018. 1

[9] Zhijie Fang and Antonio M López. Is the pedestrian going to cross? answering by 2d pose estimation. *arXiv preprint arXiv:1807.10580*, 2018. 2.4

[10] Zhijie Fang, David Vázquez, and Antonio M López. On-board detection of pedestrian intentions. *Sensors*, 17(10):2193, 2017. 2.4

[11] Efstathios P Fotiadis, Mario Garzón, and Antonio Barrientos. Human detection from a mobile robot using fusion of laser and vision information. *Sensors*, 13(9):11603–11635, 2013. 2.2

[12] Katerina Fragkiadaki, Sergey Levine, Panna Felsen, and Jitendra Malik. Recurrent network models for human dynamics. In *Computer Vision (ICCV), 2015 IEEE International Conference on*, pages 4346–4354. IEEE, 2015. 1, 2, 2.4, 3, 4.1.2

[13] Hidalgo, Gines. Openpose performance benchmark, 2017. `https://docs.google.com/spreadsheets/d/1-DynFGvoScvfWDA1P4jDInCkbD4lg0IKOYbXgEq0sK0/edit#gid=0`, Accessed: 05-16-2018. 1

[14] Daniel Holden, Taku Komura, and Jun Saito. Phase-functioned neural networks for character control. *ACM Transactions on Graphics (TOG)*, 36(4):42, 2017. 2.4

[15] Ninghang Hu, Aaron Bestick, Gwenn Englebienne, Ruzena Bajscy, and Ben Kröse. Human intent forecasting using intrinsic kinematic constraints. In *Intelligent Robots and Systems (IROS), 2016 IEEE/RSJ International Conference on*, pages 787–793. IEEE, 2016. 2

[16] Chien-Ming Huang and Bilge Mutlu. Anticipatory robot control for efficient human-robot collaboration. In *The Eleventh ACM/IEEE International Conference on Human Robot Interaction*, pages 83–90. IEEE Press, 2016. 2

[17] Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. Human3.6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(7):1325–1339, jul 2014. (document), 1.1, 2.3, 4.1.1, 4.2, 5.3, 5.4, 5.5

[18] Ehsan Jahangiri and Alan L Yuille. Generating multiple diverse hypotheses for human 3d pose consistent with 2d joint detections. *arXiv preprint arXiv:1702.02258*, 2017. 1, 2.3

[19] Ashesh Jain, Amir R Zamir, Silvio Savarese, and Ashutosh Saxena. Structural-rnn: Deep learning on spatio-temporal graphs. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5308–5317, 2016. 1, 2, 2.4, 3, 4.1.2, 4.2

[20] Hanbyul Joo, Tomas Simon, Xulong Li, Hao Liu, Lei Tan, Lin Gui, Sean Banerjee, Timothy Scott Godisart, Bart Nabbe, Iain Matthews, et al. Panoptic studio: A massively multiview system for social interaction capture. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017. (document), 1, 2.3, 4.1.1, 4.2, 5.1, 5.2

[21] Rachel Kirby, Reid Simmons, and Jodi Forlizzi. Companion: A constraint optimizing method for person-acceptable navigation. In *IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*, pages 607–612, September 2009. 1

[22] Kris M Kitani, Brian D Ziebart, James Andrew Bagnell, and Martial Hebert. Activity forecasting. In *European Conference on Computer Vision*, pages 201–214. Springer, 2012. 2, 2.1

[23] Hema S Koppula and Ashutosh Saxena. Anticipating human activities using object affordances for reactive robotic response. *IEEE transactions on pattern analysis and machine intelligence*, 38(1):14–29, 2016. 2

[24] Markus Kuderer, Henrik Kretzschmar, Christoph Sprunk, and Wolfram Burgard. Feature-based prediction of trajectories for socially compliant navigation. In *Robotics: science and systems*. Citeseer, 2012. 2.1

[25] Jangwon Lee and Michael S Ryoo. Learning robot activities from first-person human videos using convolutional future regression. *Image*, 500:500, 2017. 1

[26] Alon Lerner, Yiorgos Chrysanthou, and Dani Lischinski. Crowds by example. In *Computer Graphics Forum*, volume 26, pages 655–664. Wiley Online Library, 2007. 2.1

[27] Zhao Liu, Jianke Zhu, Jiajun Bu, and Chun Chen. A survey of human pose estimation: the body parts parsing based methods. *Journal of Visual Communication and Image Representation*, 32:10–19, 2015. 2, 2.3

[28] Jim Mainprice, Rafi Hayne, and Dmitry Berenson. Predicting human reaching motion in collaborative tasks using inverse optimal control and iterative re-planning. In *Robotics and Automation (ICRA), 2015 IEEE International Conference on*, pages 885–892. IEEE, 2015. 1

[29] Julieta Martinez, Michael J Black, and Javier Romero. On human motion prediction using recurrent neural networks. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4674–4683. IEEE, 2017. (document), 1, 2, 2.4, 3, 4.1.1, 4.1.2, 1, 4.2, 5.1.2, 5.1, 5.2, 5.3, 5.3, 5.4, 5.5, 5.3

[30] Julieta Martinez, Rayat Hossain, Javier Romero, and James J Little. A simple yet effective baseline for 3d human pose estimation. In *IEEE International Conference on Computer Vision*, volume 206, page 3, 2017. (document), 2.3, 4.1.1, 5.1.1, 5.1, 5.2

[31] Dushyant Mehta, Helge Rhodin, Dan Casas, Pascal Fua, Oleksandr Sotnychenko, Weipeng Xu, and Christian Theobalt. Monocular 3d human pose estimation in the wild using improved cnn supervision. In *3D Vision (3DV), 2017 Fifth International Conference on*, 2017. 2.3

[32] Dushyant Mehta, Srinath Sridhar, Oleksandr Sotnychenko, Helge Rhodin, Mohammad Shafiei, Hans-Peter Seidel, Weipeng Xu, Dan Casas, and Christian Theobalt. Vnect: Real-time 3d human pose estimation with a single rgb camera. *ACM Transactions on Graphics (TOG)*, 36(4):44, 2017. 2.3

[33] Justin Miller, Andres Hasfura, Shih-Yuan Liu, and Jonathan P How. Dynamic arrival rate estimation for campus mobility on demand network graphs. In *Intelligent Robots and Systems (IROS), 2016 IEEE/RSJ International Conference on*, pages 2285–2292. IEEE, 2016. 2.2

[34] Alejandro Newell, Kaiyu Yang, and Jia Deng. Stacked hourglass networks for human pose estimation. In *European Conference on Computer Vision*, pages 483–499. Springer, 2016. 2.3

[35] Georgios Pavlakos, Xiaowei Zhou, Konstantinos G Derpanis, and Kostas Daniilidis. Coarse-to-fine volumetric prediction for single-image 3d human pose. In *Computer Vision and Pattern Recognition (CVPR), 2017 IEEE Conference on*, pages 1263–1272. IEEE, 2017. 2.3

[36] Stefano Pellegrini, Andreas Ess, Konrad Schindler, and Luc Van Gool. You'll never walk alone: Modeling social behavior for multi-target tracking. In *Computer Vision, 2009 IEEE 12th International Conference on*, pages 261–268. IEEE, 2009. 2.1

[37] Leonid Pishchulin, Eldar Insafutdinov, Siyu Tang, Bjoern Andres, Mykhaylo Andriluka, Peter V Gehler, and Bernt Schiele. Deepcut: Joint subset partition and labeling for multi person pose estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4929–4937, 2016. 2, 2.3

[38] Varun Ramakrishna, Daniel Munoz, Martial Hebert, James Andrew Bagnell, and Yaser Sheikh. Pose machines: Articulated pose estimation via inference machines. In *European Conference on Computer Vision*, pages 33–47. Springer, 2014. 2.3

[39] Taiki Sekii. Pose proposal networks. In *The European Conference on Computer Vision (ECCV)*, September 2018. 5.1.1

[40] Xiao Sun, Jiaxiang Shang, Shuang Liang, and Yichen Wei. Compositional human pose regression. In *The IEEE International Conference on Computer Vision (ICCV)*, volume 2, 2017. 2.3

[41] Denis Tome, Christopher Russell, and Lourdes Agapito. Lifting from the deep: Convolutional 3d pose estimation from a single image. *CVPR 2017 Proceedings*, pages 2500–2509, 2017. 2.3

[42] Marynel Vázquez, Aaron Steinfeld, and Scott E Hudson. Parallel detection of conversational groups of free-standing people and tracking of their lower-body orientation. In *Intelligent Robots and Systems (IROS), 2015 IEEE/RSJ International Conference on*, pages 3010–3017. IEEE, 2015. 3

[43] Marynel Vázquez, Elizabeth J Carter, Braden McDorman, Jodi Forlizzi, Aaron Steinfeld, and Scott E Hudson. Towards robot autonomy in group conversations: Understanding the effects of body orientation and gaze. In *Proceedings of the 2017 ACM/IEEE International Conference on Human-Robot Interaction*, pages 42–52. ACM, 2017. 1

[44] Anirudh Vemula, Katharina Muelling, and Jean Oh. Social attention: Modeling attention in human crowds. In *2018 IEEE International Conference on Robotics and Automation (ICRA)*, pages 1–7. IEEE, 2018. 2, 2.1

[45] Shih-En Wei, Varun Ramakrishna, Takeo Kanade, and Yaser Sheikh. Convolutional pose machines. In *CVPR*, 2016. 2, 2.3

[46] Kota Yamaguchi, Alexander C Berg, Luis E Ortiz, and Tamara L Berg. Who are you with and where are you going? In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pages 1345–1352. IEEE, 2011. 2

[47] Xingyi Zhou, Xiao Sun, Wei Zhang, Shuang Liang, and Yichen Wei. Deep kinematic pose regression. In *European Conference on Computer Vision*, pages 186–201. Springer, 2016. 2.3

[48] Brian D Ziebart, Nathan Ratliff, Garratt Gallagher, Christoph Mertz, Kevin Peterson, J Andrew Bagnell, Martial Hebert, Anind K Dey, and Siddhartha Srinivasa. Planning-based prediction for pedestrians. In *Intelligent Robots and Systems, 2009. IROS 2009. IEEE/RSJ International Conference on*, pages 3931–3936. IEEE, 2009. 2, 2.1

[49] Christian Zimmermann, Tim Welschehold, Christian Dornhege, Wolfram Burgard, and Thomas Brox. 3d human pose estimation in rgbd images for robotic task learning. *arXiv preprint arXiv:1803.02622*, 2018. 2.3