Human Torso Pose Forecasting in the Real World

Abhijat Biswas, Henny Admoni, Aaron Steinfeld
Robotics Institute
Carnegie Mellon University
Pittsburgh, Pennsylvania 15213
Email: {abhijat, henny, steinfeld}@cmu.edu

***Abstract*—In this paper, we describe a multi-modal approach to human torso pose estimation and forecasting. Our end-to-end system combines RGB images and point cloud information to reason about 3D human pose. We use a simple filter fit method to forecast torso pose. Further, we evaluate the forecasting performance quantitatively on the Human3.6M motion capture dataset and qualitatively on a furniture assembly task. Our simple forecasting algorithm outperforms complicated recurrent neural network methods, while being faster on the torso pose forecasting task.**

I. INTRODUCTION

Autonomous agents in dynamic environments must anticipate future events to mitigate latencies in cognitive processing [37]. This becomes more important as robots are increasingly deployed in highly dynamic and cluttered environments with imperfect information about their surroundings. For robots 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.

For example, consider a mobile, convention center robot that 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 [17, 4]. Finally, the robot needs to orient itself properly as it approaches the person [2]. Timely perception of torso pose is important for all of these steps.

In support of these, and similar interactions, we present a new human torso pose anticipatory model. We show that a simple filter and polynomial model outperforms deep neural networks for this task, while being much faster, hence allowing it to be deployed on low-cost mobile robots without necessitating significant and expensive computation.

Ideally, such a system is able to not only detect, but also track and forecast entire pose skeletons for multiple persons in 3D space and in real-time. Each of these problems are either challenging or well studied, but have computationally expensive solutions. As an illustration, consider that one of the best 2D multi-person pose detectors [3] can perform at around 18Hz [8] (using 2x Nvidia 1080 Ti GPUs). Furthermore, associated high dimensional non-linear dynamics as well as inherent stochasticity makes full body human pose forecasting inherently difficult.

To make this tractable, researchers have approached the forecasting problem by restricting the scope to a particular task, thereby reducing the dimensionality of the problem space. For example, some predictively model human reaching motions for a shared workspace assembly task [23], while others predict future hand locations in egocentric video to allow anticipatory motion planning and assistance [21].

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 orientation, an important cue correlated with motion intent and social engagement [35].

Our system uses multi-modal visual input data, namely RGB with scene depth data, and both estimates and forecasts a 3D torso plane. This in contrast to most previous body pose forecasting work [7, 6, 16, 24] that either use 2D or 3D articulated pose (with initial joint configurations obtained directly from a motion capture system). Such multi-modal sensing not only helps overcome depth ambiguity[15], 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.

Contributions: In this paper, we describe a novel end-to-end depth-based torso pose estimation and forecasting system combining both depth and RGB visual data. In particular, we show that a simple filtering and polynomial fitting algorithm outperforms more complicated recurrent neural network based forecasting approaches and is $45\times$ faster, giving up granularity. We evaluate the pose forecasting system quantitatively on the Human 3.6M dataset [14] and demonstrate the entire end-to-end system on a furniture assembly task.

II. RELATED WORK

Human modeling for robotics has taken various forms including estimating [3, 22, 30, 38] and forecasting [7, 16, 24] human pose from visual data, modeling human motion trajectories (both individual [39, 41] and group trajectories [36]), as well as human intent prediction [19].

A. Human intent prediction

Human intent prediction has emerged as an area of interest for effective human-robot collaboration. Previous work has utilized the intrinsic kinematics of the human anatomy [12], eye gaze [1, 13], and spatio-temporal structure of the task

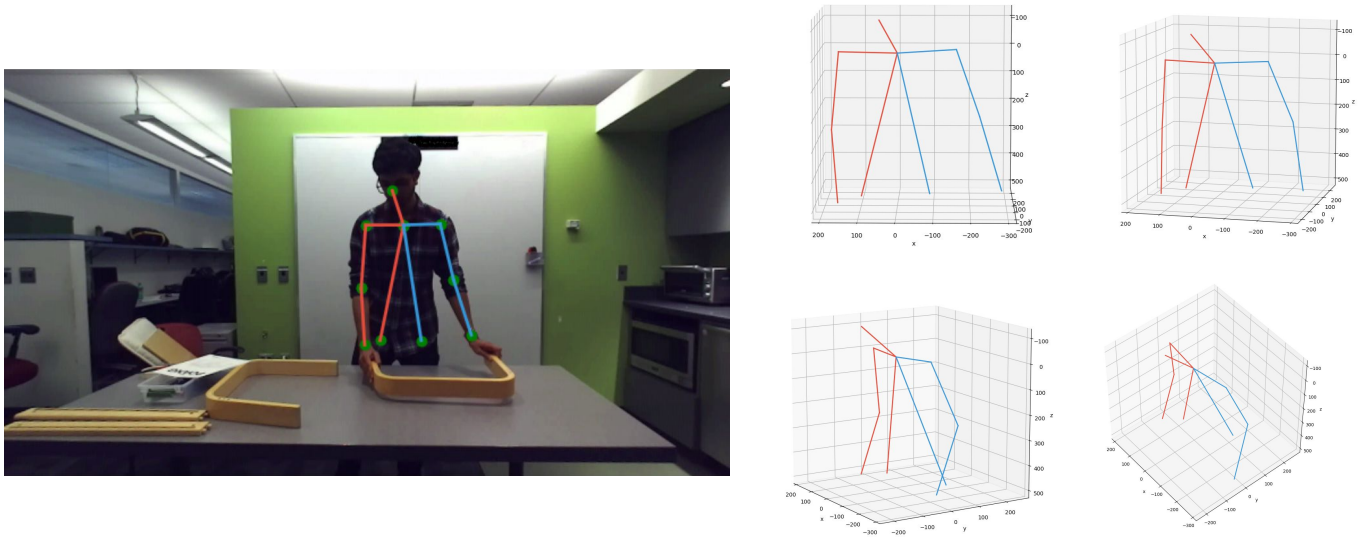

Fig. 1. Example input RGB image with overlayed with a 2D torso pose detection (left). Views of 3D pose extracted using the 2D detections (right). Points on the arm are shown for illustration, not included in torso pose calculations

space [19]. These methods have used graphical models such as MDPs or CRFs to encode constraints and spatio-temporal relationships.

Anticipatory modeling of human navigation trajectories has also received interest for mobile social robotics. In [34], the authors suggest the existence of biomechanical "turn indicators" of human motions that allow for human motion trajectory prediction. Maximal entropy inverse reinforcement learning approaches [18, 20, 41] have also been used to forecast pedestrian trajectories.

B. Human pose estimation and motion modeling

Human pose recovery from visual data is a challenging but incredibly useful task that has recently seen tremendous interest and success in the computer vision community. The focus has mostly been on joint keypoint localization using a single RGB camera in pixel space [22] of a single individual [28, 31, 38] and, more recently, multiple individuals [3, 5, 30]. The most successful models have employed graphical or neural network models trained on large datasets.

There has also been significant interest in 3D pose estimation from monocular images, despite the inherent depth ambiguity, that allows multiple plausible 3D pose hypotheses given a 2D pose estimate [15]. Several end-to-end models have been trained on this task that regress the individual skeletal keypoints [26, 27, 29, 32, 33, 40]. Even though these algorithms have the advantage of being able to work with inexpensive RGB cameras, these are single-person 3D pose estimators and, as such, suffer from depth ambiguity. In [25] the authors show that "lifting" the pose from 2D to 3D gives a surprisingly low error rate, suggesting that most of the error in these methods lies in the visual analysis of body pose rather than reasoning about the 3D structure of the 2D scene.

With human pose estimation a well-studied problem with accurate, real-time solutions in specialized use cases, modeling of human motion has received considerable interest in the literature. The vision community has approached this from the standpoint of forecasting articulated human poses without considering the acquisition of the pose skeletons themselves. For example, [7, 16, 24] all model and forecast human motion using recurrent neural networks. However, these methods do not model the global position of their subjects, instead choosing to focus 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 have selected the best performing method for comparison [24].

The graphics community also uses deep recurrent neural networks, primarily for character motion synthesis conditioned on human user input. For example, past work uses these methods to animate video game characters [9, 10, 11]. Even while generating realistic human motions, these methods fail to match ground truth human poses and suffer from discontinuity between the ground truth frames and first predicted frame.

In general, the intrinsic stochasticity of human motion does not allow for accurate forecasting of complete human poses over long horizons ($> 1s$) [24]. 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.

III. APPROACH

For a mobile robot's human perception system, we want the following characteristics:

- Real-time performance.
- A discriminative output signal that allows human attention/intent prediction.

- A smooth temporal model of human pose.

The last is non-trivial since previous supervised methods, such as [7, 16, 24], suffer from severe discontinuities at the beginning of the forecast, as shown in [24].

Our end-to-end system comprises a torso pose detection/estimation module followed by a forecasting module. The design choices we make throughout our system are influenced by the aforementioned desiderata.

Our desired pose detection system is required to reason about the world in 3 dimensions and hence monocular RGB approaches are not reliable. To achieve this, we currently use an off-the-shelf 2D human pose detection system [3] in conjunction with a stereo camera 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 we know the 2D joint locations, these are projected onto a registered point cloud obtained by triangulation in a separate step, giving us 3D joint locations. This depth data can come either from an IR-based time of flight camera or a stereo camera. To avoid restricting our domain to indoor use only, we use a stereo camera, the Stereolabs ZED for our system.

We model the torso pose by the position and orientation 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,

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

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 $\tau \subset \mathcal{R}^3$ is the set of all torso joint locations. For annotated skeletons from the Human3.6M dataset, τ contains the shoulder joints, two hip joints, the mid-spine, and the tip of the tailbone.

In the forecasting sub-module, we predict elevation, azimuth, and absolute position of the torso plane over a 400ms time-window. Subsequently, we refer to the elevation and azimuth together as the plane orientation components. Together, the three components describe the plane uniquely. Once the torso plane is acquired from the pose detection module, we apply a low pass filter to the two orientation components. 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 HRI activities. This is followed by fitting an Nth order polynomial, which is then used to extrapolate a forecast for each individual component.

For the low pass filter, we use a second-order Butterworth filter. The cutoff frequency was empirically set to 5Hz.

IV. EXPERIMENTS

A. Datasets

We evaluated the system on two datasets:

- **Human 3.6M** [14]: This is currently the largest publicly available dataset of motion capture data, containing 7 actors performing 15 varied tasks such as walking, taking photos or giving directions. We used this dataset to perform quantitative analysis.
- **Furniture assembly**: For qualitative analysis of our entire system, we collected data in the real world on a furniture assembly task, in which a human subject followed print instructions to assemble an ottoman. This allowed us to demonstrate the performance of the entire end-to-end system qualitatively. We chose this task to illustrate the capabilities and limitations of the system.

B. Evaluation procedures

For quantitative evaluation on Human 3.6M [14], we used the same train-test split as [7, 16, 24] and compared against [24] since it is the quantitatively best performing model out of the three. The MoCap data was downsampled to 25Hz, as in previous work. During testing, skeletal poses over a 2 second sample (50 frames) were fed to a recurrent neural network, which then generated samples over a forecast window of 400 millisecond (10 frames) sample. The initial 50 frames are referred to as the conditioning ground truth.

Since our method focuses on torso planes rather than full body articulated pose, we must obtain ground truth planes from the MoCap data, by fitting a least squares plane to the hips, shoulders, and neck joints of an articulated pose obtained from the MoCap data, as described in Equation 1.

The aforementioned methods do not estimate the 3D pose of a human subject from visual data. Rather, they acquire the ground truth 3D poses directly obtained from the MoCap data accompanying Human3.6M. For evaluating our method in this experiment, our RGB and depth based pose estimation module was bypassed to keep the quantitative comparison of our forecasting system with [24] fair.

Instead of using the Euclidean distance in Euler angle space for all body joints (as in previous work), we computed the error of the plane orientation forecast. Specifically, we use the plane azimuth and elevation error over the forecast window. We chose this measure since it is most indicative of the macro-level expression of torso pose. See Table I for average azimuth and elevation angle error for each of the 15 Human3.6M activities across various methods.

V. RESULTS

A. Quantitative results

Table I shows the results of fitting polynomials of degree $N = 1$ to 4, the best performing quantitative method for human motion prediction (HMP) on Human3.6M [24], and a constant prediction baseline (where the last ground-truth torso plane orientation is predicted for the entire forecast window). The results also demonstrate the importance of low pass filtering in the last line, where we omit the filtering step and directly fit an Nth order polynomial to the unfiltered data.

B. Qualitative results

The furniture assembly task was recorded using a commercially available ZED stereo camera and processed offline. See Figure 1 for a sample visualization of the estimated 3D pose. We see realistic 3d pose estimates notwithstanding occlusions.

TABLE I
TORSO PLANE ORIENTATION FORECASTING (400MS) ERRORS ON H3.6M DATA (DEGREES)

Action	Plane Azimuth						Plane Elevation					
	HMP[24]	Const	$N=1$	$N=2$	$N=3$	$N=4$	HMP[24]	Const	$N=1$	$N=2$	$N=3$	$N=4$
Walking	2.15	4.44	2.94	**1.87**	2.35	3.16	3.52	**1.91**	2.18	2.45	3.36	6.67
Eating	1.99	4.48	1.14	**0.71**	1.20	1.76	3.84	**1.13**	2.18	1.86	2.24	4.73
Smoking	1.63	1.43	**0.71**	1.24	1.18	0.93	3.57	1.77	**1.26**	2.45	2.25	2.87
Discussion	13.89	3.13	**1.06**	1.83	3.17	3.21	3.93	1.42	**1.11**	2.27	1.45	9.80
Directions	30.40	2.66	2.02	**1.82**	2.61	6.30	6.01	2.12	**1.88**	2.61	5.36	8.41
Greeting	27.11	3.75	2.70	**2.69**	3.49	5.97	5.39	3.35	3.47	4.90	**2.98**	8.14
Phoning	**3.69**	7.44	4.05	7.30	10.09	21.06	4.06	**1.63**	1.84	1.92	3.50	5.45
Posing	7.96	9.65	**4.30**	7.24	7.99	13.11	6.82	5.46	3.67	**3.48**	4.94	5.73
Purchases	14.85	**5.64**	9.12	6.55	21.35	26.77	8.97	6.54	**5.45**	11.17	13.7	10.35
Sitting	2.55	2.49	2.55	**1.64**	7.22	9.58	4.52	3.21	**1.46**	1.99	4.34	3.32
Sitting Down	**4.30**	5.36	7.11	7.02	5.03	10.41	4.80	3.85	2.23	2.40	**1.90**	4.52
Taking Photo	32.63	2.14	**1.46**	1.61	3.21	3.82	8.63	**3.17**	5.56	4.77	22.71	28.37
Waiting	11.58	4.28	2.10	**1.58**	2.45	4.23	5.34	**3.05**	3.27	3.89	6.27	22.07
Walking Dog	4.75	4.89	**2.39**	2.73	4.42	6.71	7.64	4.75	**3.64**	4.45	7.33	8.6
Walking Together	1.31	3.88	1.46	**1.23**	1.24	3.4	4.53	**1.77**	2.04	2.72	4.40	8.52
All (average)	10.72	4.38	**3.01**	3.14	5.13	8.02	5.44	3.01	**2.74**	3.55	5.78	9.17
All (average, no filter)	10.72	4.38	**3.33**	10.54	96.34	657.73	5.44	**3.01**	3.26	16.35	187.08	1628.74

C. Analysis

Some trends can be seen in Table I. First, the plane azimuth is harder to predict than the elevation, given the higher error rates across all 15 actions 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 the filtering step, we not only see larger errors in the polynomial fitting but larger standard deviation among errors corresponding to various values of N. This suggests that the estimate is unstable. (See last two rows of Table I)

Third, the recurrent neural network model tends to make much larger errors than our simple 1st degree (linear) polynomial fit. 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 the 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 highly stochastic tasks such as "Taking Photo" (both upright and kneeling poses) or "Directions" (high variance poses).

Fourth, our forecasting algorithm is inexpensive to compute while being faster and more accurate than previous work. The method described in [24] (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 onboard. Our forecasting 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. However, note that HMP forecasts full body articulated 3D pose while we only model the torso plane.

From qualitative evaluation, it is clear that using a point cloud constructed from a stereo camera is vulnerable to errors by occlusion while projecting the 2D points onto 3D, even if the pose detector can circumnavigate occlusions. This is apparent when the nearly assembled ottoman occludes the person and 3D projection is inaccurate.

VI. CONCLUSION

We propose a novel end-to-end torso pose estimation and forecasting system. In addition to torso pose orientation, we predictively model absolute torso position, which is ignored in previous work. We present quantitative results to evaluate the forecasting system and show that our simple filter and fit method outperforms complex recurrent neural network methods while being approximately $45\times$ faster on the torso plane forecasting task.

A. Limitations & Future Work

We imposed several constraints on this work, including a focus solely on the torso plane. We also only address the single human case, but this approach could easily be extended to multiple people in a top-down fashion.

In future work, we would like to extend this to a more full-fledged perception system and apply it to tasks like social navigation, that will require multi-person perception. We would also like to collect data from the perspective of a mobile robot for evaluation closer to likely real-world applications.

ACKNOWLEDGMENTS

This work was funded by a grant (IIS-1734361) from the National Science Foundation.

References

[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] 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.

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

[4] 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.

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

[6] 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.

[7] 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.

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

[9] Daniel Holden, Jun Saito, Taku Komura, and Thomas Joyce. Learning motion manifolds with convolutional autoencoders. In *SIGGRAPH Asia 2015 Technical Briefs*, page 18. ACM, 2015.

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

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

[12] 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.

[13] 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.

[14] 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.

[15] 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.

[16] 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.

[17] 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.

[18] 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.

[19] 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.

[20] 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.

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

[22] 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.

[23] 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.

[24] 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.

[25] 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.

[26] 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.

[27] 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.

[28] 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.

[29] 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.

[30] 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.

[31] 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.

[32] 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.

[33] 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.

[34] Vaibhav V Unhelkar, Claudia Pérez-D'Arpino, Leia Stirling, and Julie A Shah. Human-robot co-navigation using anticipatory indicators of human walking motion. In *Robotics and Automation (ICRA), 2015 IEEE International Conference on*, pages 6183–6190. IEEE, 2015.

[35] 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.

[36] Anirudh Vemula, Katharina Muelling, and Jean Oh. Social attention: Modeling attention in human crowds. *arXiv preprint arXiv:1710.04689*, 2017.

[37] David Vernon, Claes Von Hofsten, and Luciano Fadiga. *A roadmap for cognitive development in humanoid robots*, volume 11. Springer Science & Business Media, 2011.

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

[39] 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.

[40] 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.

[41] 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.