Predicting Arm Movements in Virtual Environments

Howon Lee, Jimmy Lee, Erik Brockbank

I. Introduction

Studies conducted at the Stanford Virtual Human Interaction Lab (VHIL) place subjects in virtual environments where they must converse and interact with other subjects as well as avatars and embodied agents. To do this, we place subjects in head-mounted displays, which block out other visual stimuli and allow the user to "enter" a virtual world. We also put infrared trackers on subjects, which allows the system to track the subject's movements. The goal of this project was to use this data to reduce the number of trackers which participants need to wear, by building a predictive model that would enable us to predict the location of a subject's elbows using head, ankle, and wrist trackers as input. This increase the immersiveness of our simulation, since fewer positioning trackers are needed, and allows us to more easily do experiments where the participants can see themselves in virtual reality. Many researchers have worked on the problem of creating a 3d model from motion tracking[1][2][3], but we believe that this is the first time anybody tried to infer the placement of motion tracking points from other motion tracking points.

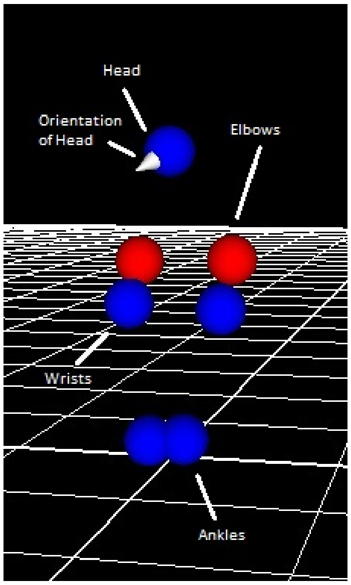
II. Data Collection

In order to gather an appropriate sample of labelled data, we had lab members do three tasks: sit down for five minutes while being recorded, stand for five minutes while being recorded, and move around the lab and tag virtual boxes placed at random locations by reaching out as if to touch them, again for five minutes. This is to get participants to walk around the lab and extend their arms at various heights and angles in order to gather a varied sample of typical, natural arm movements.

The participants wore infrared trackers on the wrists and elbows, as well as the ankles and one on the head for all of these tasks. The input features then were the position of each wrist and the ankles with respect to the head, as well as head location and orientation coordinates (pitch, yaw, and roll) of the head. The output features were the position of the elbows. Each of 8 participants performed the box task; the labelled data was collected 60 times per second and written to an external file, which we then used as our training and testing data. This gave us a total of around 120,000 data points.

During data collection, we identified many sources of possible error or misleading data, including the risk of technical failure (e.g. problems with infrared tracker operation or tracking and rendering during the box task). In order to accommodate this possibility, beyond collecting additional data, we wrote a script to "play back" each participant's session (see figure 1). For any given data file of input features, the script animates a basic stick figure to recreate the position and orientation values at each time t. This allowed us to review each session and see that the data had been properly collected and stored. Due to problems with the infrared trackers, we did in fact have to throw out 2 of the participants' data before beginning our training; these errors were only easily noticeable by running our replay script on the data we had collected.

Figure 1: *Playing back the data for each participant's session*



III. Learning Algorithms and Results

After collecting data from participants in the virtual box task and verifying their validity using our playback script, we used the data to train predictive algorithms on the input features. This is a supervised regression task. We started out with the typical linear regression and then took a number of the algorithms for supervised learning, modified to suit the task. We began by setting a basic linear regression to the data. For single participants, we were able to reach a Coefficient of Determination(R^2) value of 0.99 with 30% hold-out on elbow location prediction. However, on the aggregate of the data for all participants, R^2 was only 0.94. Unsatisfied with these numbers, we decided to try additional algorithms in the hopes that we might raise our quantitative accuracy measures before exploring subjective accuracy (described below). For a summary of all the results, see the table in the next page.

We followed linear regression with a ridge regression algorithm to see if added regularization might improve our results. This produced only a modest improvement, perhaps due to more systemic problems that were visible in the subjective testing phase. After the initial regressions, we implemented a random forest algorithm, which added a great deal of expressive power to our predictions; this was confirmed by the much higher coefficient of determination. We followed this with an artificial neural network, which regressed the datapoints closely, as reflected in the R^2 value of 0.988. This increased responsiveness with the more high level algorithms was exciting because it is perhaps reflective of a greater degree of complexity than we had initially anticipated in this problem, at least involving the features and training data that we chose to start with. One possible reason for the need for more expressive algorithms is that our inputs have fairly complicated interactions with respect to how they produce the output. We can imagine various ways in which the other input features might be nonlinear: when standing in a neutral position, ankle position and separation may be highly indicative of elbow position and when bending down (e.g. to touch a box low to the ground), head position and orientation might strongly reflect this. In any case, the ability to achieve high values for coefficient of determination or R2 with increasingly complex algorithms shows that these produce good regressions.

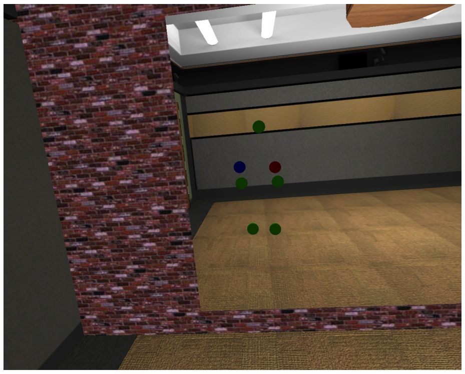
We used Scikit-Learn for training[5], except for the neural networks, for which we used PyBrain[4].

Algorithm	R^2 with 30% holdout
Linear Regression	0.940
Ridge Regression	0.943
Random Forest	0.998
Artificial Neural Networks (1 hidden layer of 10 neurons)	0.988

3.1 Subjective Results

An interesting property of our particular learning problem is that we can talk not only about objective accuracy in the quantitative means described above, but also subjective accuracy, related to whether the algorithm provides a realistic experience for participants. If we were to use an algorithm to render elbow location in real time based only on wrist location, it is imperative that subjects not find their elbows to be in a different location in the virtual world from where they would expect them to be. This reduces "presence" - the feeling for a participant that she is actually in the virtual world - an important feature in obtaining accurate results from experiments at VHIL.

Figure 2: *Real-time virtual environment with elbow prediction*



In order to confirm whether or not our quantitative accuracy corresponds to this more subjective accuracy, we wrote a script that is able to run our learning algorithms and render their predictions in real time. We place a subject in a simple virtual environment that includes a virtual mirror, allowing the subject to see all of their movements. We then feed the subject's wrist, ankle, and head location, as well as head orientation to the algorithm in real time and render the algorithm's projected elbow location for the participant to see in the mirror (see figure 2). This is vital because it allows us to test the true accuracy of our algorithms: whether or not they are able to render elbow location in a way that feels realistic to the participant. Further, it enables us to ask questions that would be very difficult without this feature, such as how our algorithms are able to handle novel or very fast movements and difficult arm positions in addition to the standard movements provided in our training data. Additionally, the real-time validation provides an early answer to a speed/accuracy tradeoff in how we render predicted arm movements. While our quantitative results above focused on accuracy, speed is a significant factor for the learning task, since presence is reduced when a virtual reality environment renders at less than 60 frames per second. Without some way of testing the performance of our algorithms on the spot, we have no way of knowing where the most optimal region lies between speed and accuracy, or whether there is one at all.

After obtaining the quantitative accuracy results described above, we ran each of our predictive algorithms in the real-time validation script to examine whether good machine learning could be aligned with the goals of the project which have to do with the subject's experience in the virtual world. Our results surprised us and were very informative in thinking about future directions. The first observation was that the linear and ridge regression, despite their relatively low accuracy rates, offer great advantages in terms of speed; they allowed for very natural renderings of elbow location that felt accurate for basic motions such as walking and generic arm movements. However, the consequences of their low coefficients of determination became apparent when we attempted novel or even slightly unique body movements; for example, we found that raising either ankle off the ground typically lifts the corresponding elbow up towards the head, which is of course not representative of normal ankle movements. While this was the most egregious failure of the algorithm, similar patterns emerged for other movements that were not explicitly included in the training data set. As described in section IV below, this problem, a sort of over-fitting, may be remedied by broadening our training data to include a more diverse set of body and arm motions.

Another pattern that emerged, though less conclusively, was the possibility of a handedness bias. Certain positions and movements of either hand similar to those involved in the cube task produced differential responses in the left and right elbows. Since the majority of our subjects were (evidentially based on what we saw during the data collection and what would be expected of a normal sample set) right handed, we suspect that there may be a scarcity in the data of left arm movements, which could be causing higher error rates and some of the asymmetrical results witnessed during real-time validation. Alongside these patterns, there were no observable differences between the linear and ridge regression.

When we initially ran our random forest algorithm, we found that its predictions were typically fairly accurate as measured by R^2, much more so than the linear and ridge regressions for more unusual arm movements, which might have been expected from the increased robustness and higher coefficient of determination for the random forest model. However, we observed that these gains were offset by a problem of jittery rendering, in which the predicted elbow locations, even while fairly accurate, seemed to jump around and were somewhat delayed behind actual movements. When we ran the artificial neural network, its predicted arm positions seemed to be excessively dominated by the rotation of the subject's body, and was also jittery as well. Although we used these classification algorithms in a regression mode, they didn't seem to behave as smoothly as we might have hoped they would.

In summary, we were thrilled to discover that our learning algorithms could be translated to what often amounted to very accurate real-time elbow location predictions, although there is certainly room for improvement in the regression models. The below table lists links to videos of the virtual mirrors for each of the regressions which we ran.

Algorithm	Link
Linear Regression	`http://youtu.be/tsDJ9x6cPJI`
Ridge Regression	`http://youtu.be/rx3U3_7bJ-0`
Random Forest	`http://youtu.be/ALF-tUbB8ag`
Artificial Neural Networks (1 hidden layer of 10 neurons)	`http://youtu.be/hLk8RGxADlE`

IV. Future Research and Directions

This project offers opportunities for further work in several directions. There are of course several steps we might take with respect of improving our machine learning on the project.

The first is to incorporate a more diverse set of training data. What we discovered from the subjective evaluation of our algorithms was that the ones that have the most optimal prediction speed (and consequently the most fluid rendering), namely the linear regression and ridge regression, are highly susceptible to overfitting in a way that might be improved through a more diverse set of training data. For example, one problem we saw with both algorithms was that raising one ankle or the other often causes the corresponding elbow to raise as well, presumably due to a lack of training data in which the ankles are raised but not the elbows. Imagine touching a virtual box hanging in the air: your ankle moves along with your arm, because your whole body reaches out to touch it. Improving our training set in order to diversify body movements might help avoid such problems with the linear and ridge regression models.

One additional pattern we observed in the data is that there is evidence (described above in Results) of a handedness bias where people completing the box task were more likely to use their right hands and therefore, right elbow location prediction is sometimes asymmetric to left elbow prediction. Future expansions of our training data set might look at further examining and correcting for this feature.

Finally, assuming high degrees of accuracy are attainable with the improvements above and looking beyond the immediate future, we hope this research contributes to the growing field of human motion reconstruction. It is worth acknowledging that in our particular case it may not in fact be of high practical value to render elbow location based on our given input features, as long as it remains fairly easy to simply track and render elbow movements of our participants without causing excessive discomfort or inaccuracy using the same trackers already placed on the wrists and ankles. However, as a matter of academic inquiry in the field of motion reconstruction, it may be an open question whether elbow location is highly predictable using wrist location; if, as our research seems to suggest, that is indeed the case, this leads to a number of questions about which joints have primacy in the sense of predictive power over the rest of the body's movements. In the field of social sciences, we may well ask how much about non-verbal cues can be inferred from predictive models that are able to map whole arm or body movements using only a few key tracking locations.

4.1 Acknowledgements

We would like to thank Prof. Jeremy Bailenson of the Communications department and our labmates at the Virtual Human Interaction Lab. We would also like to thank the course staff of CS229, especially Richard Chen and Mahdi Soltanolkotabi, for advice with regards to the project.

IV.References

[1] Lenarcic, J., and Philippe Wenger. "Human Motion Reconstruction by Direct Control of Marker Trajectories." *Advances in Robot Kinematics: Analysis and Design.* New York: Springer, 2008.

[2] Khatib, O., E. Demircan, V. De Sapio, L. Sentis, T. Besier, and S. Delp. "Robotics-based Synthesis of Human Motion." *Journal of Physiology-Paris* 103.3-5 (2009): 211-19.

[3] Lai, Ying-Xun et al. "3D Adaptive Reconstruction of Human Motion From Multi-Sensors." *The Third International Workshop on Wireless Sensor, Actuator and Robot Networks* (2011)

[4] Tom Schaul, Justin Bayer, Daan Wierstra, Sun Yi, Martin Felder, Frank Sehnke, Thomas Ruckstiess, Jurgen Schmidhuber. PyBrain.

[5] Pedregosa et al., Scikit-learn: Machine Learning in Python, JMLR 12, pp. 2825-2830, 2011.