Multi-person Motion Capture Dataset for Analyzing Human Interaction

David V. Lu*, Annamaria Pileggi* William D. Smart*†
* Washington University in St. Louis,
St. Louis, Missouri, 63108 USA
†Willow Garage
Menlo Park, CA 94025 USA
Email: davidlu@wustl.edu, wds@willowgarage.com

Abstract—**As robots become more articulated, the space of potential movements increases exponentially. The easiest-to-program and most efficient paths that robots move through are often not perceived by humans to be "natural." We present one potential source for information about the appropriate ways to move, in the form of motion capture data from actors. By leveraging the precise and specific nature of trained actors' movements, we can begin to see the qualities and relationships these motions have. In addition to the specifics that went into constructing this data set, we also present our initial principal component analysis of the motions.**

I. Motivation

For robots with some non-trivial number of degrees of freedom, the space of possible poses that a robot can take is quite large. Furthermore, the space of paths between two given poses is even larger. Some of the paths are infeasible due to resulting collisions or mechanical limitations. Obstacle avoidance and similar metrics slightly narrow the space. However, in the remaining space, there are still large numbers of paths that could be taken, even given a reasonable time constraint. Some of these paths are well-explored and/or well-defined, such as the most efficient one, minimizing some cost metric like energy used or distance traveled.

The most efficient path is not necessarily the right path for all contexts. While efficiency is a highly valued trait in industrial and other constrained applications, situations where robots must interact with humans often necessitate a different metric for judging potential paths. Efficient or easy-to-compute paths may be viewed as 'robotic', or at the very least, not very personable. Such paths are, by definition, not designed to carry any information that a person interacting with the robot could use to help facilitate the interaction. For situations that call for some measure of human robot interaction, a better method for planning paths is necessary.

Further complicating this issue is the fact that minor variations in the movements of robots can have profound effects on the way that a human will interpret it. It is not a matter of *what* the robots do, but *how* they do it, and how they should do it greatly depends on the context of the situation. What might be a perfectly acceptable movement in one context, may be grossly inappropriate in another. For humans, this behavior comes naturally, but for robots, this behavior must be constructed. Therefore, it is imperative for HRI researchers begin to examine the differences in movement in a multitude of contexts.

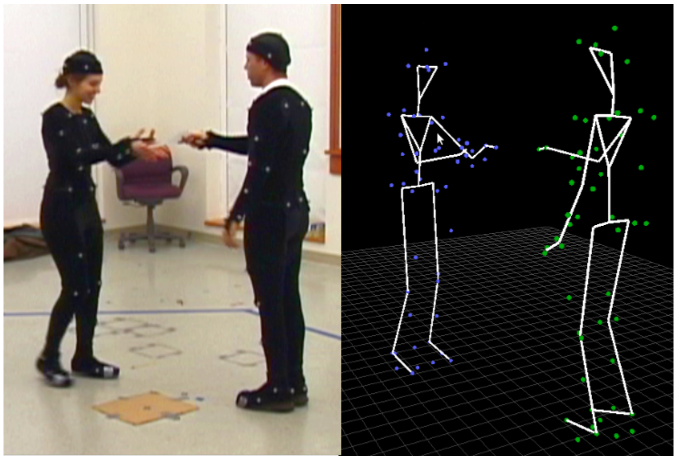

Fig. 1. The actors performing a scenario, as seen by a video camera (on the left) and the motion capture system (on the right).

One vital contextual factor is relationship. One of the primary influences of how humans interact with other humans is what the comparative relationship between them is. Do they both have equal status, as friends, colleagues or strangers, or does one person have a higher status, as an authority figure, boss or parent? There are many actions that most people would deem appropriate among friends, but when done in front of a superior would not be welcome. This is true even of subtler components of actions like posture, which can communicate a great deal of information about the person's general attitude toward the person with whom they are interacting.

In order to get an accurate sense of what sort of motions convey this information, we used classically trained stage actors in a motion capture system (see Busso et al. [2] for another use of actors as motion capture models). As we have discussed previously[5], the use of actors for improving human robot interactions is justified due to the actors' abilities to convey subtle clues about their internal state through their physical actions on stage. Furthermore, they are able to give repeatable performances when needed, and can perform naturally in constrained environments like motion capture studios, a trait not shared by the general population. Using actors as our models for motion allows us to tap into the wealth of knowledge provided by the human science of theatre.

In this paper, we present the dataset resulting from our efforts to learn about how actors move. We start by describing the conditions under which the series of motion capture trials were developed and captured. Then we show our initial results that came from performing principal component analysis on the motion capture data, and postulate how it may be used for developing more informative robot motions.

The combination of motion capture data with principal component analysis has been explored previously [1, 4, 3]. However, we believe this is the first dataset generated specifically for the analysis of two peoples' motions and how the relationship between them affects their motions.

II. Motion Capture Data

For our initial round of performing motion capture on actors, we focused on a number of general scenarios for a robot inhabiting a typical office environment. These included simple behaviors like passing someone in the hallway, giving someone an object or stopping someone to talk. The actors were directed to perform these tasks with a variety of given circumstances. The primary variable element of the circumstances that changed was the relationship between the two actors in the scene, putting them into a colleague/colleague relationship, or a boss/subordinate relationship.

Two actors were selected from the Washington University community by Pileggi, who also filled the role of director for these interactions. Over the course of three months, the actors rehearsed the interactions and found the best way for them to authentically create the scenarios in the motion capture studio. After the conclusion of the rehearsal process, the actors performed the interactions in Vanderbilt University's motion capture studio manufactured by Vicon. Each actor wore fifty-three markers on their body (five on the head, fourteen on the torso, eight on each arm, and nine on each leg). Using its array of cameras, the Vicon system recorded the x, y, and z coordinates of each marker, which it output into a C3d file. The resulting files were parsed, labeled and analyzed using custom-built packages made with the ROS framework [1].

Each of the interactions centered around one of five different scenarios: Passing, Impasse, Stopping, Exchange and Follow. The **Passing** scenario is the simplest, involving the two actors passing each other in a hallway. The **Impasse** scenario adds a twist in that the hallway is too narrow for both to pass each other at the same time. Instead, one must give way and let the other pass. The **Stopping** scenario involves the two actors meeting in a hallway and one actor, called the initiator, stops the other, the recipient, to talk. The **Exchange** scenario was built around the exchange of an object, with the two actors meeting, and the initiator gives the recipient a folder with papers in it. Finally, the **Follow** scenario involved the initiator getting the recipient to follow them somewhere else.

These scenarios were selected due to the varying types and degrees of interaction between the two actors that they necessitated. The Passing scenario has little to no interaction between the two actors. The Impasse scenario adds an obstacle that forces the two to interact in some way in order to obtain their objective of getting down the hallway. The Stopping scenario adds a slightly simpler interaction to the Passing scenario, in that the initiator aims to change the recipient's actions by making him or her stop and talk. Similarly, the Exchange and Follow scenarios are other variations that force interaction between the two.

Within the different scenarios, there were a variety of specific circumstances which the actors used to inform their actions. Sometimes the Follow scenario entailed two colleagues meeting and deciding to go to lunch, while other times it necessitated one person asking the other to get to a meeting quickly. These variations served two purposes. For one, they kept the actors' motions specific and organic, not allowing them to generalize a useless, generic version of the scenario. Second, the differing circumstances introduced a degree of variance, ensuring that any algorithms developed from this training data did not become overfit.

III. Principal Component Analysis

There are two problems we had to tackle to analyze this data. First, we had to determine what the essential ways the actors moved were. Secondly, once we had extracted the types of movement that occurred, we had to determine how those movements related to our original semantic labels for the motions.

A. Setup

We chose Principal Component Analysis (PCA) as our initial foray on the motion capture data for a number of reasons. First, we were looking for commonalities among different trials, but were unsure of what those commonalities would look like. Hence, a machine learning method that explore essentially unlabeled data seemed appropriate. Second, given the large numbers of markers for each actor and the dependence of markers on each other, a dimensionality reduction was needed to make working with the data more manageable.

We represented the data on a time-frame by time-frame basis. Although the native representation of the motion capture data uses a global XYZ coordinate frame, we chose to use a local coordinate frame for each actor when performing PCA on the data. Using the positions of the six waist markers at each time frame, a six dimensional central reference frame was calculated. For each frame, a data vector was composed using the pose of the central reference frame relative to the global reference frame, combined with the pose of each marker relative to the central reference frame. The pose for the central reference frame is specified using six numbers (x,y,z and $roll$, $pitch$, and yaw), while each marker is represented by three (x,y and z). Hence, each time frame is represented by a K degree vector, where $K = 6 + 3m$, and m is the number

[1] http://www.ros.org/wiki/motion_capture

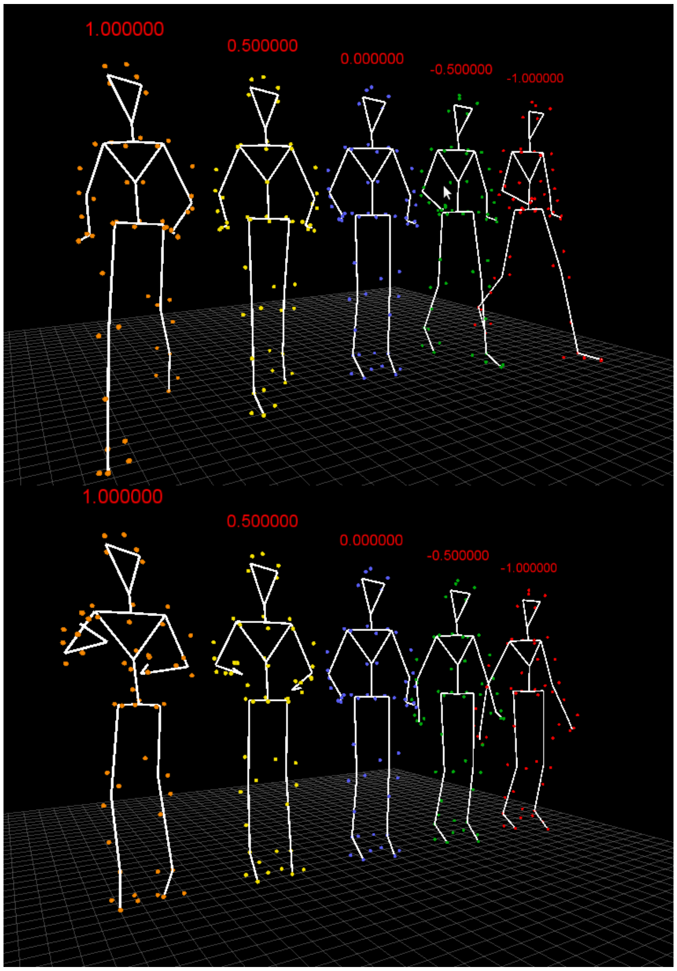

Fig. 2. A visualization of the second and third principal components, showing the markers and a simple skeleton interpolated from those markers. The five figures in each image show the resulting position of the markers when the eigenvalue for the particular component is set to -1, $-.5$, 0, .5 or 1. Note: For this visualization, the central reference frame of each figure was set manually.

of markers on each actor. In this work, $m = 53$, making $K = 165$. With T data points, we end up with one $K \times T$ matrix for each actor. Using the central reference frame for the data massively decreased the variance of each data point.

PCA was performed using the set of transformed data vectors for each actor in each trial, giving us the set of principal components and and a set of coefficients for each time frame. The sets of coefficients were then matched back with the actor and trial associated with them.

B. Analysis

Performing PCA in the manner described in the previous section, we found a number of interesting components. The most significant component moved the central reference frame for the figure across the stage, which was the primary direction which the actors traveled. The second most significant component, seen in the top of Figure 2, shows the basic mechanics of a walking motion. As the component rises and lowers in value, both legs move back and forth, matched with alternating arm motions.

The discovery of this as the largest component related to body pose was not unexpected, since walking played a role in each of the scenarios. However, the fact that such a clear walking cycle was observed validates, at least initially, our approach to this type of movement analysis. The component managed to encapsulate a complex movement using the relatively primitive representation described above. It may have been easier to recreate such a motion in the joint angle space, but the fact that it was extractable in this form lead us to believe that other less obvious motions will also be easily extracted.

The rest of the components express other portions of the actors' movements. For example, the third most significant component, seen in the bottom of Figure 2 shows the raising and lowering of the arms. How each of these components specifically relates to the semantic labels is explored in the following section.

C. Labeling the Components

While many of the components derived through PCA are well-formed and interesting looking, the main objective in exploring them was to see how they related to the original semantic labels. Using the components, we would like to be able to derive two main labels. First, there is the scenario for the particular trial. While this mainly becomes a gesture-recognition-like problem, finding the components that relate to the specific role within the scenario presents an interesting problem. Secondly, and more importantly, we have the relationship label. This is a subtle problem since the overall motions of the interactions stays constant even when the relationship changes (i.e. an item is still exchanged), so the motions become even more nuanced.

Our initial attempts to learn these labels used the average eigen-value for each component over all the time points for each trial, giving us a representation of how much any particular component was used in any of the trials. We have used this to find some initial correlations between the components and labels using a simple entropy-based method. These result remain unpublished until we can tweak our learning algorithm to better suit the dataset.

D. The Space Between

We also have some ongoing work to more explicitly model the two actors together, rather than separately, as all the previously mentioned work has done. Instead of one motion capture trial resulting in two $K \times T$ matrices, we instead concatenate them for a single $2K \times T$ matrix. This gains us the insight of how the actors move in relation to each other.

In addition to seeing how the new components form and relate the one actor to the other, we also use this as a way to investigate how the two actors relate to each other temporally. We can measure the synchronicity of the two actors in the scene by seeing how many components express movement in both actors, and how many deal only with one or the other. Furthermore, by introducing an artifical offset between the two actors' motions, we can see if there is a causal relationship between the two, where one actor moves, and the other reacts and follows with their own movement. If this is the case, we can then start to hypothesize about how the synchronicity can be used to help infer the relationship.

IV. Discussion

This paper presents our initial exploration of the movements of actors using motion capture technology. It by no means represents a complete analysis of the data. In keeping with the open source philosophy, all of the raw and processed motion capture data has been posted online (http://www.cse.wustl.edu/~dvl1/motion_capture/) in hope that others will find the information informative.

There is the further question of how best to apply these insights to robots. Our plan is to create a heuristic algorithm for translating human movement onto the morphology of robots. This is a multifaceted problem, which forces us to consider differences in morphologies, possible reach spaces, hardware limits and human reactions. However, once the most effective way to translate the motions is found, then we end up with a more robust way to have robots interact with people. After the motions are transferred onto the robot, we intend to test the effects of the different components on human reactions to the robot through live demonstrations, which we hypothesize will match our initial semantic labels.

We also plan on performing these same analyses using different representations of the same data. Instead of using the pose at each time frame, we have also performed some initial investigations of using a representation that utilizes not the pose but the *change in pose* at each time frame. Furthermore, instead of using the marker positions as the raw data, we have also considered fitting a skeleton to the points and using the relative joint angles as the input to PCA.

As mentioned earlier, many hours went into the process of creating this data set even before stepping foot into the motion capture studio. The end result is presented in this paper, but it is also worth noting that the scenarios ultimately decided on were the result of a long negotiation process. There was a constant need to balance the scientific requirements of the data gathering and needs of the actors, satisfying their need to create authentic scenes with our need to collect data in a systematic manner. This underlines the value in having a long collaborative relationship, as it was much easier to resolve the differences as part of an ongoing discussion. Furthermore, it was essential to have someone who was both versed with the technical and artistic side of things to help translate the ideas from one group to the other.

One of the more interesting results of this particular episode in our collaboration was the exploration of the space between art and science. There are two beliefs, fundamental to each side, that are at odds when applying science to art and vice versa. Scientists tend to believe that all phenomenon can be measured, modeled and ultimately reproduced. Artists tend to believe that there is a unique human element to art, that humans alone are capable of recognizing and producing. We found that the best way to reconcile these beliefs was to work toward the goal of developing models that aspire to the human performance, which still leaves a lot of room for improvement before the models are even close enough to the humans for comparison. This produced that data needed for our modeling, while also giving the actors an opportunity to hone their craft. According to the acting methodology offered by Stanislavski[6], an important skill for actors is the ability to embue each physical action with more and more specificty. Exploring this level of precision made the motion capture trials a worthwhile exercise for the actors.

Finally, we have found that this investigation of relationship to be a key stepping off point for discussions of human robot interactions. Actors often focus their scene work by examining their character's relationship to other characters in the play. In HRI, we are lead to the question: what role do we want the robot to play? Our impression is that people desire robots that are generally subservient, and sometimes, friendly. Ergo, a robot that orders you around is less acceptable, at least given current societal prejudices. This makes finding what sorts of motions are used in each situation even more pressing.

V. Thanks

Thanks to Bobby Bodenheimer and his students for the use of and assistance with their motion capture studio, and very special thanks to Justin Rincker and Amanda Spector, our wonderfully talented and patient actors. This work was partially supported by the National Science Foundation under award IIS-0917199 and by a research gift from Willow Garage, Inc. At the time of writing, Smart is on a paid sabbatical at Willow Garage, Inc.

References

[1] J. Barbič, A. Safonova, J.Y. Pan, C. Faloutsos, J.K. Hodgins, and N.S. Pollard. Segmenting motion capture data into distinct behaviors. In *Proceedings of Graphics Interface 2004*, pages 185–194. Canadian Human-Computer Communications Society, 2004. ISBN 1568812272.

[2] C. Busso, M. Bulut, C.C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J.N. Chang, S. Lee, and S.S. Narayanan. IEMOCAP: Interactive emotional dyadic motion capture database. *Language Resources and Evaluation*, 42(4):335–359, 2008.

[3] A. Fod, M.J. Matarić, and O.C. Jenkins. Automated derivation of primitives for movement classification. *Autonomous robots*, 12(1):39–54, 2002. ISSN 0929-5593.

[4] O.C. Jenkins and M.J. Mataric. Deriving action and behavior primitives from human motion data. In *Intelligent Robots and Systems, 2002. IEEE/RSJ International Conference on*, volume 3, pages 2551–2556. IEEE, 2002. ISBN 0780373987.

[5] David V. Lu and William D. Smart. Human-robot interactions as theatre. In *RO-MAN 2011*. IEEE, In preparation.

[6] Constantin Stanislavski. *An Actor Prepares*. Theatre Arts Books, 1989.