Spatiotemporal Correspondence as a Metric for Human-like Robot Motion

Michael J. Gielniak
Dept. of Electrical & Computer Engineering
Georgia Institute of Technology
Atlanta, GA, 30332
mgielniak3@mail.gatech.edu

Andrea L. Thomaz
College of Computing
Georgia Institute of Technology
Atlanta, GA, 30332
athomaz@cc.gatech.edu

ABSTRACT

Coupled degrees-of-freedom exhibit correspondence, in that their trajectories influence each other. In this paper we add evidence to the hypothesis that spatiotemporal correspondence (STC) of distributed actuators is a component of human-like motion. We demonstrate a method for making robot motion more human-like, by optimizing with respect to a nonlinear STC metric. Quantitative evaluation of STC between coordinated robot motion, human motion capture data, and retargeted human motion capture data projected onto an anthropomorphic robot suggests that coordinating robot motion with respect to the STC metric makes the motion more human-like. A user study based on mimicking shows that STC-optimized motion is (1) more often recognized as a common human motion, (2) more accurately identified as the originally intended motion, and (3) mimicked more accurately than a non-optimized version. We conclude that coordinating robot motion with respect to the STC metric makes the motion more human-like. Finally, we present and discuss data on potential reasons why coordinating motion increases recognition and ability to mimic.

Categories and Subject Descriptors

1.2 [**Artificial Intelligence**]: Robotics—*Kinematics & dynamics, propelling mechanisms*; C.4 [**Performance of Systems**]: [Measurement techniques, performance attributes]

General Terms

Algorithms, measurement, performance

Keywords

Metrics, human-like motion, user study, mimicking

1. INTRODUCTION

When social robots interact with humans by communicating in a manner that is socially relevant and familiar to their human partners, the benefits are increased transparency; improved teammate synchronization; and reduced training cost for collaborative robots due to increased "user-friendliness." This has been called "natural human-robot interaction," and requires believable behavior to establish appropriate social expectations [6]. Furthermore, display of appropriate non-verbal communication behavior increases perception of agent realism [4]. Our work addresses this overall problem of how to generate believable or human-like motion for an anthropomorphic robot.

We hypothesize that spatiotemporal correspondence (STC) of distributed actuators (i.e. motor coordination) is a component of human-like motion. We present motion optimized with respect to a nonlinear STC metric based on Kolmogorov-Sinai entropy as both a synthesis and evaluation tool, and validate our results using a user study based on mimicking. The results show that STC-optimized motion is (1) more often recognized as a common human motion, (2) more accurately identified as the originally intended motion, and (3) mimicked more accurately than an unmodified version of human motion retargeted to the robot. We conclude that coordinating robot motion with respect to the STC metric makes it more human-like. Finally, we present and discuss potential reasons why coordinating motion increases recognition and the human ability to mimic.

2. RELATED WORK

2.1 Human-like Motion in Robotics

A fundamental problem with existing techniques to generate robot motion is data dependence. For example, a very common technique is to build a model for a particular motion from a large number of exemplars [13, 21, 15]. Ideally, the robot could observe one (potentially bad) exemplar of a motion and generalize it to a more human-like counterpart.

Dependence on large quantities of data is often an empirical substitute for a more principled approach. For example, rapidly exploring random tree (RRT) offers no guarantees of human-like motion, but relies upon a database to bias the solution towards realism. The database is a bottleneck for online planning, which can affect algorithm runtime [25]. Other techniques rely upon empirical relationships derived from the data to constrain robot motion to appear more human-like. This includes criteria such as joint comfort, movement time, jerk, [8, 5]; and human pose-to-target relationships [11, 23]. When motion capture data is used, often timing is neglected, causing robot manipulation to occur at unrealistic and non-human velocities [2, 1].

HRI '11 Lausanne, Switzerland

2.2 Existing Human-like Motion Metrics

Human perception is often the metric for quality in robot motion. By modulating physical quantities like gravity in dynamic simulations from normal values and measuring human perception sensitivity to error in motion, studies can yield a range of values for the physical variables that (according to the results of the study) are below the perceptible error threshold (i.e. effectively equivalent to the human eye) [18, 24]. These techniques are valuable as both synthesis and measurement tools. However, the primary problem with this type of metric is dependency on human input to judge acceptable ranges. Results may not be extensible to all motions without testing new motions with new user studies because these metrics depend upon quantifying the measurement device (i.e. human perception).

Classifiers have been used to distinguish between natural and unnatural movement based on human-labeled data. If a Gaussian mixture model, HMM, SLDS, naive Bayesian, or other statistical model can represent a database of motions, then by training one such model based on good motion-capture data and another based on edited or noise-corrupted motion capture data, the better predictive model for testing would have a higher log likelihood of the test data under the model. This approach is inspired by the theory that humans are good classifiers of motion because they have witnessed a lot of motion. However, data dependence is the problem, and retraining is necessary when significantly different exemplars are added [19].

In our literature search, we found no widely accepted metric for human-like motion in the fields of robotics, computer animation, and biomechanics. By far, the most common validation efforts rely upon subjective observation and are not quantitative. For example, the ground truth estimates produced by computer animation algorithms are evaluated and validated widely based on qualitative assessment and visual inspection. Other forms of validation include projection of motion onto a 2-D or 3-D virtual character to see if the movements *seem* human-like [14]. Our work presents a candidate metric for human-like motion that is quantitative.

3. APPROACH

Human motion is characterized by *interdependent* degrees-of-freedom (DOFs) [22]. This implies that the underlying movement data has a level of coordination—the phenomenon that motions of connected bodies or joints are related in terms of space and timing. Thus, in theory, if you increase the amount of spatial and temporal coordination for a given motion, it should become more human-like. Intuitively, coordination could be thought of as correlation or similarity in space and time of two trajectories, but correlation and similarity already have specific mathematical definitions. Bernstein theorizes that the amount of coordination in motion may only be limited by constraints (e.g. kinematic & dynamic limits, environmental, task) [3].

Spatial and temporal coordination of distributed actuators has long been an assumed part of human motion, and our unique mimicking experiment adds statistical evidence to this end. Humans have muscles, which inherently cause coupling, or coordination between DOFs. This basic insight is key to our metric. Since robots have motors (naturally uncoupled), we need to emulate human DOF coupling in robot motion to produce human-like motion. These observations inspired our experiment to optimize STC within motion and support STC as an element of human-like motion.

The concept that human muscles cause correspondence between DOFs may be obvious, but no prior work has extrapolated this to humanoid robot motion, noting the difference between human and robot actuators, in an effort to devise a method for synthesizing motion and evaluating the new hypothesis of whether robot motion is labeled "more natural" when coordinated spatially and temporally. We felt it is appropriate to verify the widely accepted truth about correspondence for human DOFs, which is why we analyze coordination in both human and robot motion in this work.

Humans' and robots' joints typically do not correspond in terms of location and degrees of freedom. Thus, information is lost when attempting to use human motion on a robot. For example, human translation at the shoulder will not appear on the robot, if the robot does not have a translatable shoulder DOF. Since coordination depends on the interaction between DOFs, then coordination is also lost along with the motion information lost when trying to apply human motion to robots.

In section 5.2, we describe a process called retargeting, which allows human motion to be applied to a robot. However, even the best retargeting algorithms still lose data in the transformation process. If sufficient interdependence between DOFs survives the retargeting process, then we hypothesize that optimizing the remaining trajectories with respect to STC will re-coordinate the motion given the constraints of the new, projected, robot kinematic chain (i.e. as much as possible given differences in DOF limits/locations).

We hypothesize that coordination will make motion appear more natural, for both human and robot motion. In other words, since human motion is spatially and temporally coordinated, then anthropomorphic robot motion will also appear more human-like if ST-coordinated. Therefore, we require an algorithm to synthesize an ST-coordinated (i.e. more natural) exemplar from a given input motion. Since our optimization depends on the amount of data surviving the retargeting process, which is dependent upon DOF similarity between the robot and human kinematics, humanoid robots are more likely to produce better results. Although in this paper, the optimized motion comes from motion-capture data, for our algorithm it does not need to come from motion capture or retargeting; it can come from anywhere (e.g. an animation of any quality, dynamic equations).

4. ALGORITHM

The spatiotemporal correspondence problem has already been heavily studied and analyzed mathematically for a pair of trajectory sets, where there is a one-to-one correspondence between trajectories in each set (e.g. two human bodies, both of which have completely defined and completely identical kinematic hierarchies and dynamic properties) [7, 10]. Given two trajectories x(t) and y(t), correspondence entails determining the combination of sets of spatial (a(t)) and temporal (b(t)) shifts that map two trajectories onto each other. In the absence of constraints, the temporal and spatial shifts satisfy the equations in 1, where reference trajectory x(t) is being mapped onto y(t).

$$\begin{aligned} y(t) &= x(t') + a(t) \\ t' &= t + b(t) \end{aligned} \tag{1}$$

where,
$t =$ time
$t' =$ temporally shifted time variable
$x(t) =$ first reference trajectory
$y(t) =$ second reference or output trajectory
$a(t) =$ set of time-dependent spatial shifts
$b(t) =$ set of time-dependent temporal shifts

The correspondence problem is ill-posed, meaning that the set of spatial and temporal shifts is not unique. Therefore, a metric is often used to define a unique set of shifts.

Spatial-only metrics, which constitute the majority of "distance" metrics, are insufficient when data includes spatial and temporal relationships. Spatiotemporal-Isomap (ST-Isomap) is a common algorithm that takes advantage of STC in data to reduce dimensionality. However, the geodesic distance-based algorithm at the core of ST-Isomap was not selected as the candidate metric due to manual tuning of thresholds and operator input required to cleanly establish correspondence [9]. Another critical requirement for a metric is nonlinearity, since human motion data is nonlinear.

Prokopenko et. al. used K_2 as a metric for STC for modular reconfigurable robots to identify optimal configurations that overcome environmental constraints on motion [17, 16]. We take advantage of their temporally extended version of the K_2 metric, which is nonlinear and an upper bound of Kolmogorov-Sinai entropy (KSE). In short, KSE describes rate of system state information loss as a function of time. Therefore, a lower value of K_2 is more optimal, since it indicates higher retention of system state information over time.

4.1 STC as a Synthesis Tool

In order to use K_2 to synthesize coordinated motion, two things are required: an optimization algorithm that can optimize with respect to a cost (or objective) function and one input motion to use as the basis (or reference) trajectory for the optimization. The K_2 metric presented in equation 3 becomes the cost function used in the optimization, with a K_2 value of zero being the target for the optimal solution.

Optimal control and dynamic time warping (DTW) are two examples of well-known approaches that can yield a solution for the optimal set of spatial and temporal shifts that solve the correspondence problem with respect to a cost or objective function, given the reference trajectory [20]. For example, if these were selected to generate optimized (i.e. coordinated) motion, the equation in 3 would be used as the cost function for either optimal control or DTW. Dynamic time warping is solved via dynamic programming [20].

During development, we optimized motion using both these algorithms, but in the end selected DTW as the algorithm to warp a trajectory with respect to our human-like metric (subject to constraints) because nonlinear optimal control formulations can be difficult to solve; however, the optimal control solution is known if the cost function is quadratic [22]. Given the constraints of robot actuators (e.g. finite space), the required convexity for an optimal control solution is handled by squaring the individual spatial and temporal terms on the right-hand side of our metric presented in equation 3. The new squared version still satisfies K_2 as an upper bound of KSE. However, the squared version is a less strict bound, which means that the minimum value of equation 3 is higher for optimal control due to the constraints of optimal control motion synthesis. In other words, an optimal control problem formulated by squaring individual K_2 terms has less potential to optimize coordination. Thus, DTW was selected over optimal control for synthesis.

The K_2 metric presented in equation 3 constrains the amount of warping in its three parameters: r, S, and T. The value r can be thought of as a resolution or similarity threshold. Every spatial or temporal pair below this threshold would be considered equivalent and everything above it, non-equivalent and subject to warping. We empirically determined a 0.1 N.m. threshold for r on our robot hardware.

To emulate the local coupling exhibited in human DOFs (e.g. ball-and-socket joints) on an anthropomorphic robot, which typically has serial DOFs, the spatial parameter, S, was selected to optimize based only on parent and children degrees-of-freedom, in the hierarchical anthropomorphic chain. Since our study, described in section 6, used predefined motions, temporal extent varied based on the sequence length for a given motion.

$$C_{(d_s,d_t)}(S,T,r) = \frac{\sum_{l=1}^{T}\sum_{j=1}^{T}\sum_{g=1}^{S}\sum_{h=1}^{S}\Theta(r-||V_l^g - V_j^h||)}{(T-1)T(S-1)S} \tag{2}$$

$$K_2(S,T,r) = ln(\frac{C_{d_s,d_t}(S,T,r)}{C_{d_s,d_t+1}(S,T,r)}) + ln(\frac{C_{d_s,d_t}(S,T,r)}{C_{d_s+1,d_t}(S,T,r)}) \tag{3}$$

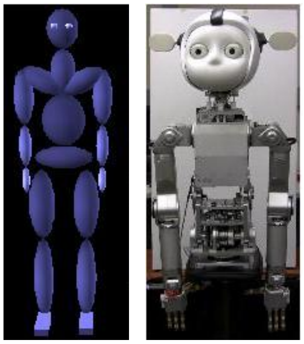
where,
$\Theta(...) =$ Heaviside step function
$V_i^k = [w_i^k, ..., w_i^{k+d_s-1}]$, spatiotemporal delay vectors
$w_i^k = [v_i^k, ..., v_{i+d_t-1}^k]$, time delay vectors
$v_i^k =$ element of time series trajectory for actuator k at time index i
$d_s =$ spatial embedding dimension
$d_t =$ temporal embedding dimension
$S =$ number of actuators
$T =$ number of motion time samples
$r =$ correspondence threshold

In our work, the term "spatial" warping is synonmyous with torque magnitude, since the given reference trajectories are torques for each DOF as a function of time.

4.2 STC as an Evaluation Metric

In order to use STC as a mechanism to evaluate motion quality with respect to human-likeness, we evaluate the spatial and temporal correspondence numbers on the right-hand side of equation 3 for a trajectory. Then, we follow the procedure outlined under 4.1 to optimize that trajectory with respect to spatial and temporal correspondence. Since this paper demonstrates that coordinated motion is more human-like, the difference between the optimal and original spatial and temporal numbers from equation 3 indicate "human-likeness" of the original trajectory. If the difference is small, the original trajectory is closer to being human-like.

5. IMPLEMENTATION



(a) Virtual Human. (b) Simon

Figure 1: The two platforms used in our experiment.

5.1 Research Platform

The platform for this research is an upper-torso humanoid robot we call Simon (Figure 1(b)). It has 16 controllable DOFs on the body and four on each hand. Each arm has seven DOFs (three at the shoulder, one at the elbow, and three at the wrist) and the torso has two DOFs, with one additional uncontrollable slave joint in the torso fore/aft direction. Simon has three DOFs for the eyes, two per ear, and four for the neck.

5.2 STC Application

Any input robot motion can be spatially and temporally coordinated using the algorithm described above, but in our evaluation we use trajectories collected from human motion-capture equipment and then retargeted to the Simon robot. The motion-capture trajectories include position data from the 28 upper-body markers on the suit. A similar set of constraints or handles, which are positioned on the target kinematic heirarchy (i.e. the robot model), serve to sufficiently constrain the problem so the mapping between human motion capture markers and robot markers is one-to-one. An optimization problem is solved, iterating over the human motion capture data as a function of time, aligning human and robot markers. The optimal mapping allows for scaling the overall size of Simon based on human participant's size, given that the proportions of Simon's parts remain constant with respect to each other. This ensures maximum amount of information preservation over the retargeting process. Upon termination of the optimization, a set of 28 time-varying point constraints exist on the robot body that align optimally with the human constraint markers from the motion-capture data. The time-varying point constraints on the robot create a motion trajectory, in Simon joint angles, that can be executed on the robot [26]. Then this retargeted motion is optimized with respect to STC as described in Sec. 4.1. Our hypothesis is that this makes the motion more human-like on the robot hardware.

6. EVALUATION

The purpose of this evaluation is to quantitatively support that spatiotemporal correspondence of distributed actuators is a good metric for human-like motion.

6.1 Hypotheses

Since human motion exhibits spatial and temporal correspondence, robot motion that is more coordinated with respect to space and timing should be more natural. Thus, we hypothesize that STC is a metric for human-like motion.

In order to test this hypothesis, we require a quantitative way to measure naturalness. We cannot optimize motion with respect to our metric and use distance measures between human and robot motion variables (e.g. torques, joint angles, joint velocities) due to the DOF correspondence problem. Given that no naturalness measure exists, we designed a user-study based on mimicking. In short, we have people mimic robot motions created by different motion synthesis techniques and the one that humans can mimic "best" (to be defined later) is the technique that generates the most human-like motion. Thus, our experiment relies on the idea that a human-like motion will be easier for people to mimic accurately. And, since practice makes perfect, we look only at people's first mimicking attempt. We theorize that awkward, less natural motions will be harder to mimic, people will be less coordinated with the seen motion.

6.2 Experimental Design

We look for differences in mimicking performances across the three experimental groups of stimulus motion:

- **OH**: Twenty motions were captured by a single male (one of the authors), we call this data "original human".
- **OR**: The human motions were retargeted to the Simon hardware using the position constraint process described in section 5.2. We call this the "original retargeted" data set.
- **OC**: The retargeted motion was then coordinated using the algorithm and metric described in section 4. This set of motions is called "original coordinated."

The twenty motions included common social robot gestures with and without constraints such as waving and object-moving, but also nonsense motions like air-guitar.[1]

In the experiment, participants are shown a video of a motion and then asked to mimic it. The OR and OC motion trajectories were videotaped on the Simon hardware from multiple angles for the study. Similarly, the original human motion was visualized on a simplified virtual human character (Fig. 1(a)); also recorded from multiple angles.

Forty-one participants (17 women and 24 men), ranging in ages from 20-26 were recruited for the study. Each participant saw a set of twelve motions from the possible set of twenty, randomly selected for each participant in such a way that each participant received four OH, four OR, and four OC motions each. This provided us with a set of 492 mimicked motions total (i.e. 164 motions from each of three groups, with 8-9 mimicked examples for each of 20 motions).

6.2.1 Part One - Mo-cap Data Collection

Each participant was equipped with a motion capture suit and told to observe videos projected onto the wall in the mo-cap lab. They were instructed to observe each motion as

[1] The accompanying video shows several examples. The full set of the motions used is: shrug, one-hand bow, two-hand bow, scan the distance, worship, presentation, air-guitar, shucks, bird, stick 'em up, cradle, take cover, throw, clap, look around, wave, beckon, move object, and call/yell.

long as necessary (without moving) until they thought they could mimic it exactly. The videos looped on the screen showing each motion from different view angles so participants could view each DOF with clarity. Unbeknownst to them, the number of views before starting was recorded as a measure for the study. When the participant indicated they could mimic the motion exactly, the video was turned off and the motion capture equipment was turned on, and they mimicked one motion. Since there is a documented effect of practice on coordination [12], we capture only their initial performance. This process was repeated for all twelve motions. Every person's first motion was treated as a practice motion and is not included in our analyses. Only if a participant was grossly off with respect to timing or some other anomaly occurred, were suggestions made about their performance before continuing. This happened in two cases.

Constraints accompanied some motions, such as objects or eye gaze direction. These constraints were given to the participants when necessary to facilitate ability to mimic accurately. For example, a cardboard box and two stools were given for the object moving motion. For all participants, the constraint locations and the standing position of the participant were identical. When constraints were given, they were given in all cases (i.e. OH, OR, and OC).

After each motion the participant was asked if they recognized the motion, and if so, what name they would give it (e.g. wave, beckon). Participants did *not* select motion names from a list. After mimicking all twelve motions, the participant was told the original intent (i.e. name) for all 12 motions in their set. They were then asked to perform each motion unconstrained, as they would normally perform it. This data was recorded with the motion captured equipment and is labeled the "participant unconstrained" (PU) set.

While the participants removed the motion capture suit, they were asked which motions were easiest and hardest to mimic and which motions were easiest and hardest to recognize. They gave their reasoning behind all of these choices.

Thus, at the conclusion of part one of the experiment, the following data had been collected for each participant:

- Motion capture data from 12 mimicked motions:
  - 4 "mimicking human" (MH) motions
  - 4 "mimicking retargeted" (MR) motions
  - 4 "mimicking coordinated" (MC) motions
- Number of views before mimicking for each of the 12 motions above
- Recognition (yes/no) for each of the 12 motions
- For all recognizable motions, a name for that motion
- Motion capture data from 12 "participant unconstrained" (PU) performances of the 12 motions above.
- Participant's selection of:
  - Easiest motion to mimic, and why
  - Hardest motion to mimic, and why
  - Easiest motion to recognize, and why
  - Hardest motion to recognize, and why

6.2.2 *Part Two - Video Comparison*

After finishing part one, participants watched pairs of videos for all twelve motions that they had just mimicked. The participant would watch the OR and OC versions of the robot motion one after the other, but projected in different spatial location on the screen to facilitate mental distinction. The order of the two versions was randomized. The videos were shown once each and the participants were asked if they perceived a difference. Single viewing was chosen because it leads to a stronger claim if difference is noted after only one comparison viewing. Then, the videos were allowed to loop serially and the participants were asked to watch the two videos and tell which they thought looked "better" and which they thought looked more natural. The participants were also asked to give reasons for their choices. Unbeknownst to them, the number of views of each version before deciding was also collected.

Thus, at the conclusion of part two of the experiment, the following data had been collected for each participant:

- Recognized a difference between OR and OC motion after one viewing (yes/no); for each of 12 motions mimicked in Part One (section 6.2.1)
- For motions where a difference was acknowledged,
  - Selection of whether OR or OC is "better"
  - Selection of whether OR or OC is more natural
- Rationale for "better" and more natural selections
- Number of views before better/natural decisions

Table 1: Percent of motion recognized correctly, incorrectly, and not recognized by participants.

	Human (OH)	Retarg (OR)	Coord (OC)
% correct	72.1	46.6	87.2
% incorrect	19.4	42.3	9.1
% not recog.	8.5	11.0	3.7

7. RESULTS

7.1 Coordination Increases Recognition

The results from our study allow us to conclude that STC optimized motion makes robot motion easier to recognize. The data in table 1 represents the percentage of participants who named a motion correctly and incorrectly, as well as those who opted not to try to identify the motion (i.e. not recognized). This data is accumulated over all 20 motions and sorted according to the three categories of stimulus video: OH, OR, and OC. Coordinated robot motion is correctly recognized 87.2% of the time, and is mistakenly named only 9.1% of the time. These are better results than either human or retargeted motion. Additionally, coordinating motion leads human observers to try to identify motions more frequently than human or retargeted motion (i.e. not recognized = 3.7% for OC). This suggests that coordinating motion makes the motion more familiar or common.

Considering the data from table 1 on a motion-by-motion basis, percent correct is highest for 16/20 coordinated motions and lowest for 17/20 retargeted motions. In 17/20 motions % incorrect is lowest for coordinated motions, and in a different set of 17/20 possible motions, % incorrect is highest for retargeted motion. These numbers support the aggregate data presented in table 1 suggesting that naming accuracy, in general, is higher for coordinated motion,

Table 2: Percent of responses selecting types of motions as easiest and hardest motion to recognize.

	Human (OH)	Retarg (OR)	Coord (OC)
Easiest	14.8	9.9	75.3
Hardest	11.5	78.3	10.2

and lower for retargeted motion. Comparing only coordinated and retargeted motion, % correct is highest for 19/20 possible motions, and in a different set of 19/20, % incorrect is highest for retargeted motion. This data implies that relationships for recognition comparing retargeted and coordinated robot motion are maintained, in general, regardless of the particular motion performed. For reference, overall recognition of a particular motion (aggregate percentage) is a function of the motion performed. For example, waving was correctly recognized 91.7% of the all occurrences (OH, OR, and OC), whereas imitating a bird was correctly recognized overall only 40.2% of the time.

Our subjective data also supports the conclusion that coordinated motion is easier to recognize. Participants were asked which of the 12 motions that they mimicked was the easiest and hardest to recognize. Table 2 shows the percentage of participants that chose an OH, OR, or OC motion, indicating that 75.3% of participants chose a coordinated motion as the easiest motion to recognize, while only 10.2% chose a coordinated motion as the hardest motion to recognize. A large majority of participants (78.3%) selected a retargeted motion as the hardest motion to recognize.

When asked, participants claimed that coordinated motion was easiest to recognize because it looked better, more natural, and was a more complete and detailed motion. On the other hand, retargeted motion was hardest because it looked "artificial" or "strange" to participants.

The majority of participants agree that coordinated motion is "better" and more natural. In 98.98% of the trials, participants recognized a difference between retargeted and coordinated motion after only one viewing. When difference was noted, 56.1% claimed that coordinated motion looked more natural (27.1% chose retargeted), and 57.9% said that coordinated motion looked "better" (compared with 25.3% for retargeted). In the remaining 16.8%, participants (unsolicited) said that "better"/more natural depends on context, and therefore they abstained from making a selection. Participants who selected coordinated motion indicated they did so because it was a "more detailed" or "more complete" motion, closer to the "expectation" of human motion.

Statistical significance tests for the results in tables 1 and 2 were not performed due to the nature of the data. Each number is an accumulation expressed as a percentage. The data is not forced choice; all participants are trying to correctly recognize the motion; some attempt and fail, and some do not attempt because they cannot recognize the motion.

7.2 Coordination Makes Motion Human-Like

Four sets of motion-capture data exist from the first part of the experiment (section 6.2.1): mimicking human (MH), mimicking retargeted (MR), mimicking coordinated (MC), and participant unconstrained (PU) motion. Analysis must occur on a motion-by-motion basis. Thus, for each of the 20 motions, there is a distribution of data that captures how well participants mimicked each motion. For each participant, we calculate the spatial (SC) and temporal (TC) correspondence according to equation 3, which resolves each motion into two numbers, one for each term on the right-hand side of the equation. For each motion, 8-9 participants mimicked OH, OR, and OC. There is three times the data for the unconstrained version because regardless which constrained version a participant mimicked, they were still asked to perform the motion unconstrained. Thus for analysis we resolve MH, MR, MC, and PU into distributions for SC and TC across all participants. There are separate distributions for each of the 20 motions, yielding 4 x 2 x 20 unique distributions. Now, our goal is to analyze each of the SC and TC results independently on a motion-by-motion basis, in order to draw conclusions about MH, MR, MC, and PU. We use ANOVAs to test the following hypotheses:

- **H1:** Human motion is not independent of constraint. In other words, all the human motion capture data, (MH, MR, MC, and PU) does not come from the same distribution. The F values, for all twenty motions, ranged from 7.2 to 10.8 (spatial) and 6.9 to 7.6 (temporal) which is greater than $F_{crit} = 2.8$. Therefore, we conclude at least one of these distributions is different from the others with respect to SC and TC.

- **H2:** Mimicked motion is not independent of constraint. In other words, all mimicked (i.e. constrained) data, (MH, MR, and MC) come from the same distribution. In these ANOVA tests, values for all 20 motions ranged between 6.1-8.6 (spatial) and 5.3-6.6 (temporal), which are greater than $F_{crit} = 3.4$-3.5. Therefore, we conclude that at least one of these distributions for mimicked motion is statistically different.

- **H3:** Coordinated motion is indistinguishable from human motion in terms of spatial and temporal coordination. MH, MC, and PU motion all come from the same distribution. $F_{observed}$ of 0.6-1.1 (spatial) and 0.9-1.9 (temporal), which are less than F_{crit} of 3.2-3.3, meaning that with this data there is insufficient evidence to reject this hypothesis for all twenty motions.

Since we are able to isolate that retargeted motion is different from the other spatial and temporal correspondence distributions in mimicked motion, at this point, we perform pairwise t-tests to determine the difference between data sets on a motion-by-motion basis. Table 3 shows the number of motions for which there is a significant difference in spatial correspondence (the table on temporal correspondence is identical but not shown). For example, when participants mimic retargeted motion, twenty motions are statistically different than the original retargeted performance. However, for mimicking human or coordinated motion, the distributions fail to be different from their original performance for both spatial and temporal coordination (H3). From this, we conclude that humans are not able to mimic retargeted motion as well as the coordinated or human motion.

Since the above statistical tests do not allow us to conclude that distributions are identical (H3), we present a regression analysis of the data across all twenty motions to determine how correlated any two variables are in our study. For the purpose of this regression analysis the variables are either the mean or the standard deviation of SC, TC, or STC, for each of the distributions (OH, OR, OC, MH, MR, MC, PU).

Table 3: Number of motions with p<0.05 for pairwise spatial correspondence comparison t-tests for the indicated study variables. Note: Table is identical for temporal correspondence.

	OH	OR	OC	MH	MR	MC	PU
OH	X	20	0	0	20	0	0
OR	X	X	20	20	20	20	20
OC	X	X	X	0	20	0	0
MH	X	X	X	X	20	0	0
MR	X	X	X	X	X	20	20
MC	X	X	X	X	X	X	0
PU	X	X	X	X	X	X	X

Table 4: R^2 value from linear regression analysis on spatial (SC), temporal (TC) and composite mean coorespondence (STC) for pairs of variables. R^2 = 1 (perfectly correlated); 0 = (uncorrelated). Note the high correlation between mimicked human and coordinated motions seen in row 14.

	Variables	SC	TC	STC
1	OH v. MH	0.9783	0.9756	0.9914
2	OH v. MR	0.6339	0.0427	0.5483
3	OH v. MC	0.9792	0.965	0.9933
4	OH v. PU	0.9859	0.9378	0.9843
5	OR v. MH	0.0103	0.0009	0.0022
6	OR v. MR	0.0915	0.008	0.0526
7	OR v. MC	0.0001	0.0002	0.0004
8	OR v. PU	0.0011	0.0003	0.0001
9	OC v. MH	0.9494	0.9626	0.9819
10	OC v. MR	0.6084	0.0491	0.5176
11	OC v. MC	0.9834	0.962	0.9918
12	OC v. PU	0.9836	0.9414	0.9795
13	MH v. MR	0.6412	0.0421	0.5612
14	MH v. MC	0.9531	0.9749	0.9809
15	MH v. PU	0.969	0.9271	0.9756
16	MR v. MC	0.6728	0.0516	0.5365
17	MR v. PU	0.6414	0.017	0.5076
18	MC v. PU	0.9881	0.9144	0.9822

However, OH, OR, and OC are only one number (not a distribution) so they are not included in the standard deviation analysis. The intuition for this analysis is that if two variables are highly correlated with respect to both mean and variance, then it is further evidence that their distributions are similar. Specifically, we seek results showing high correlation between the human and coordinated motions.

These results, the R^2 values from the linear data fits, are shown in tables 4 and 5. This shows that participants mimicking coordinated and human motion are highly correlated (line 14 in table 4 and line 2 in table 5, lightly shaded), whereas participants' mimicking retargeted motion is less correlated to all other data including the original human performance (lines 2, 6, 10, 13, and 16 in table 4). In short, this means that any data with high correlation would be a excellent linear predictor of the other variable in the pair. These higher correlations between human and coordinated motion are further evidence that coordinated motion is more human-like than retargeted motion.

Table 5: R^2 value from linear regression analysis on standard deviation of spatial, temporal and composite correspondence for pairs of study variables. R^2 = 1 (perfectly correlated); 0 = (uncorrelated). Variables not shown have a standard deviation of 0. Note the high correlation between mimicked human and coordinated motions seen in row 2.

	Variables	SC	TC	STC
1	MH v. MR	0.1005	0.1231	0.3507
2	MH v. MC	0.8847	0.7435	0.9842
3	MH v. PU	0.0674	0.0906	0.8348
4	MR v. MC	0.0746	0.1749	0.346
5	MR v. PU	0.5002	0.0002	0.2239
6	MC v. PU	0.0986	0.096	0.8537

Table 6: Percent of responses selecting types of motions as easiest and hardest motion to mimic.

	Human (OH)	Retarg (OR)	Coord (OC)
Easiest	14.6	9.8	75.6
Hardest	31.7	56.1	12.2

Furthermore, the standard deviation correlation on line 3 in table 5 is low for the spatial and temporal components, which shows that mimicking does in fact constrain people's motion. Variance increases for the PU distribution because humans are free to perform the motion as they please. This validates our premise in this study that mimicking performance is a method by which to compare motion.

The data taken for number of views before mimicking (NVBM) also supports the claim that coordinated motion is more human-like. On average, humans view a retargeted motion more before they are able to mimic (3.7 times) as compared to coordinated motion (2.7 times) or human motion (2.4 times). Pairwise t-tests between these, on a motion-by-motion basis for NVBM, show that 19 of 20 retargeted motions exhibit statistical significance ($p<0.05$) when compared with human NVBM whereas only 3 of 20 coordinated motions NVBM are statically different ($p<0.05$) from human NVBM. This suggests coordinated motion is more similar to human motion in terms of preparation for mimicking.

Of the 12 mimicked motions, each participant was asked which motion was easiest and hardest to mimic. Of all participant responses, 75.6% of motions chosen as easiest were coordinated motions, and only 12.2% of participant responses chose a coordinated motion as hardest to mimic (table 6). We return to our assertion stated earlier where we claimed that a human would be able to more easily mimic something common and familiar to them. Our results suggest that coordination adds this quality to robot motion, which improves not only ability to mimic, as presented earlier, but also perception of difficulty in mimicking (table 6).

During questioning we gained insight into people's choices of easier and harder to mimic. Participants felt that human and coordinated motion were "more natural" or "more comfortable." Participants also indicated that human and coordinated motion are easier to mimic because the motion is "more familiar," "more common," and "more distinctive." In comparison, some people selected retargeted motion as being easier to mimic because fewer parts are moving in the motion. Others said retargeted motion is hardest to mimic because the motion felt “artificial” and “more unnatural.”

7.3 SC and TC are better than STC

In equation 3, the individual terms (spatial and temporal) on the right-hand side can be evaluated separately, rather than summing to form a composite STC. In our analysis, when the components are evaluated individually on a motion-by-motion basis, 20 of 20 retargeted motions exhibit statistical difference ($p<0.05$) from the human mimicked data and 0 of 20 coordinated motions exhibit correspondence that is not statistically different ($p>0.05$) than human data distribution (table 3). However, with the composite STC used as the metric, only 16 of 20 retargeted motions are statistically different than the original human performance ($p<0.05$). Thus, we recommend the SC/TC individual components be used independently as a metric for human-likeness.

8. CONCLUSION

We have demonstrated a metric that can be used to synthesize and evaluate motion. If the goal of anthropomorphic robots is to communicate with humans as humans communicate with each other, then robot motion can be improved with this metric to create more human-like motion.

We presented spatial and temporal correspondence (i.e motor coordination) as an quantitative metric for human-like motion. We also presented objective and subjective data to support our claims that motion optimized with respect to STC is more human-like and easier to recognize, and therefore has benefits for human-robot interaction. Our metric is useful as both a tool for human-like motion synthesis for anthropomorphic robots and as a measurement instrument for the evaluation of human-like motion.

9. REFERENCES

[1] T. Asfour et al. The humanoid robot ARMAR: Design and control. *1st IEEE-RAS Intl. Conference on Humanoid Robots*, September 2000.

[2] T. Asfour et al. Control of ARMAR for the realization of anthropomorphic motion patterns. *Second Intl. Conf. on Humanoid Robots*, pages 22–24, 2001.

[3] N. A. Bernstein. *The Coordination and Regulation of Movements*. Oxford, UK: Pergamon Press, 1967.

[4] J. Cassell. Embodied conversational agents: Representation and intelligence in user interfaces. *AI Magazine.*, 22(4):67–84, 2001.

[5] T. Flash and N. Hogan. The coordination of arm movements: An experimentally confirmed mathematical model. *The Journal of Neuroscience*, 5(7):1688–1703, July 1985.

[6] T. Fong et al. A survey of socially interactive robots. *Robotics and Autonomous Systems*, 42, 2003.

[7] M. Giese and T. Poggio. Morphable models for the analysis and synthesis of complex motion pattern. *Intl. Journal of Computer Vision*, February 2000.

[8] K. Harada et al. Natural notion generation for humanoid robots. In *Proceedings of the 2006 IEEE/RSJ Intl. Conference on Intelligent Robots and Systems*, pages 833–839, October 2006.

[9] O. Jenkins and M. Mataric. A spatiotemporal extension to isomap nonlinear dimension reduction. Technical Report CRES-04-003, University of Southern California Center for Robotics and Embedded Systems, 2004.

[10] A. Kolmogorov. Entropy per unit time as a metric invariant of automorphisms. *Doklady Akademii Nauk SSSR*, 124:754–755, 1959.

[11] K. Kondo. Inverse kinematics of a human arm. Technical Report CS-TR-94-1508, Stanford University, 1994.

[12] B. Lay et al. Practice effects on coordination and control, metabolic energy expenditure, and muscle activation. *Human Movement Science*, 21:807–830, 2002.

[13] J. Lee et al. Interactive control of avatars animated with human motion data. In *Proceedings of the 29th Annual Conference on Computer Graphics and Interactive Techniques*, pages 491–500, 2002.

[14] T. Moeslund and E. Granum. A survey of computer vision-based human motion capture. *CVIU*, pages 231–268, 2001.

[15] D. Oziem et al. Combining sampling and autoregression for motion synthesis. In *Proceedings of the Computer Graphics Intl. Conference*, pages 510–513, June 2004.

[16] M. Prokopenko et al. Evolving spatiotemporal coordination in a modular robotic system. In S. Nolfi et al., editors, *From animals to animats 9: 9th Intl. Conference on the Simulation of Adaptive Behavior (SAB 2006), Rome, Italy*, volume 4095 of Lecture Notes in Computer Science, pages 558–569. Springer Verlag, September 2006.

[17] M. Prokopenko et al. Measuring spatiotemporal coordination in a modular robotic system. In L. Rocha et al., editors, *Proceedings of Artificial Life X*, 2006.

[18] P. Reitsma and N. Pollard. Perceptual metrics for character animation: Sensitivity to errors in ballistic motion. *ACM Trans. Graph.*, 22(3):537–542, 2003.

[19] L. Ren et al. A data-driven approach to quantifying natural human motion. *ACM Trans. Graph.*, 24(3):1090–1097, 2005.

[20] S. Salvador and P. Chan. Toward accurate dynamic time warping in linear time and space. *Intelligent Data Analysis*, 2007.

[21] Y. Song et al. Unsupervised learning of human motion. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 25(7):814–827, July 2003.

[22] E. Todorov and M. Jordan. Optimal feedback control as a theory of motor coordination. *Nature Neuroscience*, 5(11):1226–1235, October 2002.

[23] R. Tomovic et al. A strategy for grasp synthesis with multifingered robot hands. In *Proceedings IEEE Intl. Conf. on Robotics and Automation*, pages 83–89, 1987.

[24] D. Wu. Ad hoc meta-analysis of perceived error in physically simulated characters. Game Developers Convention. Microsoft Game Studios, 2009.

[25] K. Yamane et al. Synthesizing animations of human manipulation tasks. *ACM Transactions on Graphics*, 23(3):532–539, 2004.

[26] J. Yang et al. Capturing and analyzing of human motion for designing humanoid motion. In *Intl. Conference on Information Acquisition*, pages 332–337, June/July 2005.