Anticipation in Robot Motion

Michael J. Gielniak and Andrea L. Thomaz

Abstract— **Robots that display anticipatory motion provide their human partners with greater time to respond in interactive tasks because human partners are aware of robot intent earlier. We create anticipatory motion autonomously from a single motion exemplar by extracting hand and body symbols that communicate motion intent and moving them earlier in the motion. We validate that our algorithm extracts the most salient frame (i.e. the correct symbol) which is the most informative about motion intent to human observers. Furthermore, we show that anticipatory variants allow humans to discern motion intent sooner than motions without anticipation, and that humans are able to reliably predict motion intent prior to the symbol frame when motion is anticipatory. Finally, we quantified the time range for robot motion when humans can perceive intent more accurately and the collaborative social benefits of anticipatory motion are greatest.**

I. Introduction

Anticipatory motion is motion that prepares the viewer for a forthcoming action. Examples include the wind-up action of a pitcher or a large inhale of air before a strong exhale. Although anticipation is most commonly associated with preparation of momentum, in this work we show the communicative benefits of anticipation, and demonstrate its benefits for robots that interact with people. We show that anticipatory motion in gestures can be used to communicate motion intent earlier than motion without anticipation.

Advance knowledge of motion intent is useful to human partners in many different situations. For example, robot handoffs will be more fluid if the human partner has more time to prepare, since they will be aware of the robot's intent sooner.

We present an autonomous algorithm for the generation of anticipatory motion using one input motion. We demonstrate the benefits of adding the communication signal called anticipation to robot motion. We perform a three-part experiment to validate that anticipatory motion communicates intent to the human partner earlier than motion without anticipation. 1) We provide evidence that humans willingly identify anticipatory motion earlier and with higher accuracy than non-anticipatory versions. 2) We quantify the motion timing range when human partners can perceive anticipatory effects. 3) We validate our algorithm using still image frames selected uniformly from motions to prove that our technique extracts the pose that is most useful in helping humans identify a motion.

M.J. Gielniak is with Department of Electrical & Computer Engineering, Georgia Institute of Technology, 777 Atlantic Dr. NW, Atlanta, GA, 30332 USA (mgielniak3@mail.gatech.edu)

A. L. Thomaz is with the School of Interactive Computing, Georgia Institute of Technology, 801 Atlantic Drive, Room 256, Atlanta, GA, 30332 USA ((athomaz)@cc.gatech.edu)

II. Related Work

The concept of communicative anticipation is familiar in the domain of computer animation, as one of twelve principles of animation [1]. However, in this domain there are few autonomous algorithms to generate anticipatory motion. One example is based on principle component analysis and applies only to facial animation [2].

Another algorithm creates anticipatory motion that sets up momentum for a following motion, such as sports motions, like the retraction of an instrument (e.g. bat, racket, or club) before swinging or the drawback of an arm before a throw. This algorithm does not work well without an initial guess to the anticipatory pose before solving the optimization.

We focus on a different type of anticipatory motion that is beneficial to social robots: one that can be added to communicative gestures. These gestures usually do not require large momentum and do not exhibit a large change in the robot's center-of-mass over the duration of the motion.

In robotics, intentional action is widely achieved through human motion capture data retargeting [3], [4]. Our approach is complementary to this work, since motion capture data can be used as the input for our algorithm.

Other researchers have shown that human movement event sequences generate prior expectations regarding the occurrence of future events, and these expectations play a critical role in conveying expressive qualities and communicative intent through the movement [5]. However, our work is not in a musical control context and does not use a dynamic Bayesian framework for motion synthesis. Contrary to work that endows intent into motions through models derived from databases [6], we try to maximize the benefits of intent that exist within a gesture without modeling intent.

III. Algorithm

Our work is inspired by a concept from computer animation called a motion-graph [7], which identifies points of transition between frames in large databases of motion to create new concatenated motion sequences. We define one frame as $x(i) = \{x_1(i), x_2(i), \ldots, x_H(i)\}$ the set of all joint angles for a robot with H degrees-of-freedom (DOFs) at discrete time increment i. A trajectory, $x = \{x_1, x_2, \ldots, x_H\}, \forall i = 1, \ldots, T$, is defined as the set of all frames and all DOFs for all discrete time increments up to time T.

Our anticipatory motion algorithm begins with the assumption that a trajectory exists to which anticipation will be added. This original motion can be observed, come from a database, can be learned (through demonstration or

otherwise), or can be provided by any standard means that trajectories are generated for robot actuators.

Creating a motion graph is like clustering, where one cluster or one node is defined as $C = \{x(i) : dist(x(i), x(g)) \leq dist_{threshold}, \forall x(g) \in C\}$ the set of all frames with some distance less than or equal to some given threshold with respect to all other frames in cluster C. The distance measure need not be calculated in joint-space, and will be discussed in detail in section III-C.

The key insight of our algorithm is that gestures used in social communication have a hand or body configuration that represents a *symbol*, which has a commonly accepted meaning. If it is possible to extract that symbol and create a variant of the same motion which displays that symbol sooner, the motion becomes *anticipatory*, in that the human partner has advance knowledge of what motion the robot is performing. We believe this will improve interactions (e.g., allowing the human partner to better coordinate with the robot in collaborative tasks [8]). For this work, we exclude facial gestures and motions for which anticipation is used for the sense of building momentum.

A. Determine Gesture Handedness: One or both hands?

Non-facial gestures for anthropomorphic robots are either one-handed or two-handed. Two-handed gestures represent a more constrained system, and they are commonly associated with a body posture that is part of the symbol. For one-handed gestures, usually the corresponding arm configuration supersedes importance of the torso posture. Waving and pointing are examples of one-handed gestures, whereas, shrugging ('I don't know') and bowing are two-handed.

Logically, in one-handed gestures, the DOFs for one arm move much more than the DOFs of the other arm. Additionally, anthropomorphic robots usually have symmetric arms, with DOFs in the same locations relative to the end-effector. Therefore, pairwise comparisons in variance (equation 1) can be made for each DOF between both arm chains to determine if a particular DOF on one arm is moving significantly more than the corresponding DOF on the opposite arm.

$$v(x_m) = \sum_{i=0}^{N} \frac{(x_m(i) - \mu_{x_m})^2}{N-1} \quad (1)$$

where,
v(x_m) = joint angle variance of arm DOF m
x_m = DOF m original joint angle trajectory
μ_{x_m} = mean joint angle for trajectory of DOF m
N = number of samples in arm DOF m trajectory

Under the similar arms assumption, 'handedness' of the gesture reduces to a linear regression of the variances. The least-squares minimization in equation 2 is solved using pairwise left arm and right arm DOF data.

$$\hat{\beta} = \arg\min_{\beta_1,\beta_0} \sum_{m=0}^{M} (v_l(x_m) - (\beta_1 * v_r(x_m) + \beta_0))^2 \quad (2)$$

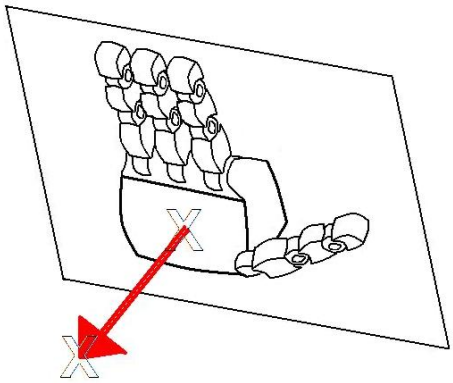

Fig. 1. The hand normal vector extends outward from the palm perpendicular to the plane of the hand.

where,
$v_l(x_m)$ = joint angle variance of left arm DOF m
$v_r(x_m)$ = joint angle variance of right arm DOF m
β_0, β_1 = regression parameters
M = number of DOFs in one arm

If the correlation coefficient from the regression term in equation 2 approaches 1.0, then we classify the gesture as two-handed. No two-handed motion in our experiments had a correlation coefficient (R^2 value) below 0.998.

B. Find & Extract the Symbol

Since our work is focused on hand and body gestures, the *symbol* is a unique hand configuration that holds a social meaning. Thus, we search the input motion for a representative hand configuration.

Our insight is that gestures have a direction constraint, since one or both hands during gestures are typically directed toward something; for example, consider stop gestures, waving, beckoning, or pointing, all of which make no sense if the hand changes orientation relative to some world constraint. Thus, we use the hand normal vector (HNV) which is directed outward from the plane that is parallel to the palm of the robot's hand. This unit vector (palm normal) is calculated for all discrete time increments in the trajectory, and it is represented in world coordinates, not a local vector relative to the hand orientation. As shown in Figure 1, the hand normal vector is easiest calculated from two world points: the hand centroid $[x_{hc}, y_{hc}, z_{hc}]$ and a point one unit in the direction perpendicular outward from the palm $[x_{hnv}, y_{hnv}, z_{hnv}]$ by subtraction. We modify the original motion graph implementation to cluster all frames of the motion based upon the hand normal vector using a corpus of frames from a single motion rather than a set of motions [7]. Multiple features in addition to the HNV could be combined for posture extraction with our approach, but we demonstrate the power of this technique to extract the symbol frame using a single feature. As you add more features for posture extraction, accurate symbol extraction is likely to increase.

Given two hand normal vectors from frames f_1 and f_2

respectively, f_1 and f_2 belong to the same motion graph cluster if criteria in equation 4 is satisfied for appropriately sized increment thresholds for the two rotation angles, $\Delta\theta$ and $\Delta\phi$.

$$\begin{aligned} \theta &= \cos^{-1}(z_{hnv} - z_{hc}) \\ \phi &= \tan^{-1}\frac{(y_{hnv} - y_{hc})}{(x_{hnv} - x_{hc})} \end{aligned} \tag{3}$$

$$\Delta\theta > |\theta_{f1} - \theta_{f2}| \quad and \quad \Delta\phi > |\phi_{f1} - \phi_{f2}| \tag{4}$$

where,
hnv = coordinate of HNV endpoint in world coordinates
hc = coordinate of hand centroid in world coordinates
ϕ_{f1} = ϕ formed from f_1 coordinates
θ_{f2} = θ formed from f_2 coordinates

A single representative frame for each cluster is created by joint-wise average of euler angles (equation 5) for each DOF, $p_{cluster} = \{p_{cluster_m}\}, m = 1, \ldots, M$, which creates frames in the anticipatory motion which do not occur in the original motion. If the HNV angular thresholds for clusters are set too high then the graph will have few clusters. If the threshold is too low, then the motion graph will simply devolve into the original motion.

$$p_{cluster_m} = \sum_{x(i) \in cluster} \frac{x(i)}{Y} \tag{5}$$

where, Y = number of frames in the cluster.

To identify the symbol cluster in the motion graph, we assume that the gesture contains a set of hand poses that ensure that the expressive message is received. Thus, a large number of frames will contain the symbol hand configuration. We define the symbol cluster as the cluster in the motion graph with the largest number of frames.

C. Find Cluster-to-Cluster Transitions

We extend our previous cluster metric in equation 3 to incorporate derivative information by using windows of frames to become the distance metric (equation 6) for determining cluster-to-cluster transitions from the motion graph.

$$\begin{aligned} D &= dist(p_u(i), p_w(i)) \\ &= \sum_{i=1}^{K} |\theta_{p_u}(i) - \theta_{p_w}(i)| + |\phi_{p_u}(i) - \phi_{p_w}(i)| \end{aligned} \tag{6}$$

where,
$p_u(i)$ = i^{th} frame in window beginning at frame $p_u \in C_u$
$p_w(i)$ = i^{th} frame in window beginning at frame $p_w \in C_w$
D = HNV angular distance metric
w, u = cluster indices being checked for transition $w \neq u$
K = number of samples in transition window

Smaller values of D identify cluster pairs that are candidates for transition points. Low thresholds on D will create lower graph connectivity. Longer original trajectories have higher potential for creating motion graphs which have more node transitions. Higher graph connectivity allows the symbol to occur sooner in the anticipatory motion, as compared to a graph with lower connectivity.

D. Compose the Anticipatory Motion

Anticipatory motion is extracted from the motion graph by beginning at the cluster that contains the initial frame from the original motion and following the path with fewest number of transitions to the "symbol" cluster, so that the symbol will occur as soon as possible in the anticipatory motion. For motions with cyclic components, e.g. waving, the symbol cluster in the motion graph may be passed more than once. For cyclic motions, we constrain our resultant anticipatory motion to exhibit the same number of cycles as the original motion, which is easily accomplished by observing the number of temporal discontinuities for frames in the symbol cluster. This is possible because our original motion is produced from a continuous trajectory for each DOF which is discretely sampled. After passing the symbol cluster the same number of times as in the original motion, the anticipatory motion can take any path to conclude at the cluster that contains the final frame from the original motion.

$$x_{new_d}(t) = slerp(p_u(t), p_w(t), \alpha(t)) \tag{7}$$

where,
$\alpha(t)$ = weight function at index t, designed for continuity
d = d^{th} DOF in full body posture
t = frame index during transition, -1<t<K

Anticipatory motion is synthesized in one of two ways: (1) When few or no candidate transitions exist that will allow an anticipatory variant of the motion to be extracted from the motion graph, splines are used between frames that represent the clusters' joint-space averages to generate the anticipatory variant to guarantee continuity of posture, velocity, acceleration, and other higher order derivatives. (2) When motion needs to be generated using more of the frames from the original motion (e.g. when higher frequency infomation would be lost in joint-space blending), the transition window from Section III-C can be utilized for spherical linear interpolation (equation 7) with properly designed blending weight function for continuity (see reference [7] for examples of weighting functions that offer C^1 continuity).

Regardless of the choice of (1) or (2), the anticipatory motion is reproduced using the joint angle data from all the DOFs. Since one frame of joint-angle data consists of all DOFs needed to generate motion, during motion synthesis redundancy is not an issue for our approach, as it might be if we were representing trajectories for reproduction as sequences of HNVs, thereby creating a many-to-one mapping from Cartesian space to joint space.

IV. Hardware Platform

The platform for this research is an upper-torso humanoid robot we call Simon (Figure 2). It has 16 controllable DOFs

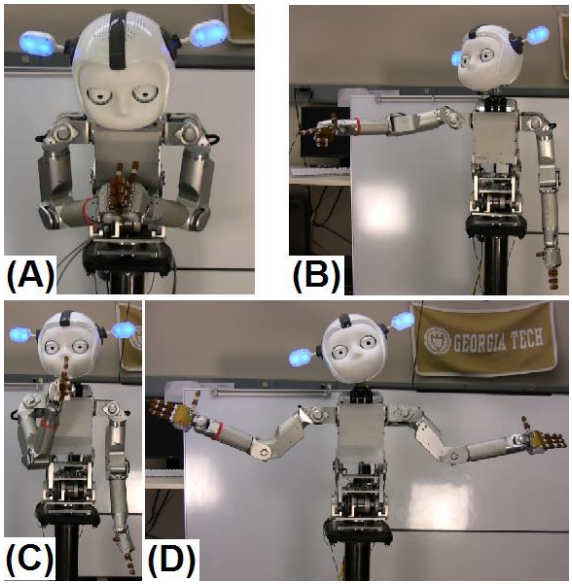

Fig. 2. Example symbols on the Simon hardware extracted using our algorithm. (A) Bow (B) Point (C) Shhh... (D) I Don't Know

on the body and 4 on each hand. Each arm has 7 DOFs (3 at the shoulder, 1 at the elbow, and 3 at the wrist) and the torso has 2 DOFs, with 1 additional uncontrollable slave joint in the torso fore/aft direction. Simon has 3 DOFs for the eyes, 2 per ear, and 4 for the neck.

V. Hypotheses

We have three hypotheses about the anticipatory motion generated using our algorithm:

- H1: The symbol extracted using our procedure will yield the frame from the motion with the highest recognition accuracy of any frame in the trajectory.
- H2: The anticipatory motions generated by using our algorithm will allow human observers to label motion intent sooner compared to how fast they can label the intent of the original motion.
- H3: Anticipatory motion is beneficial in helping observers predict motion intent only during a specific range in timing relative to the symbol. If an observer watches robot motion beyond the symbol frame, they will be able to predict motion intent equally well for anticipatory motion and the original counterpart.

We designed three experiments to test our hypotheses, which were conducted on separate days using different sets of human subjects.

VI. Experiment 1: Symbol Validation

A. Experimental Design

To test H1, we have a set of thirteen original motions that were either animated with Maya 3-D animation software or retargeted from human motion capture data. These original motions are executed on the hardware and videotaped to access individual frames. Our algorithm was used to extract the symbol from each of these original gestures[1]. Then we find the frame in the original motion that is nearest to the symbol, using a Euclidean distance metric in torque space. This becomes the representative symbol frame in the original motion. Some examples of symbols from our motions are shown in Figure 2.

We wish to penalize our distance metric for pose variations that appear significantly different when viewed in Cartesian space. We rejected a joint-angle-space metric because this treats all DOFs similarly when viewing poses in Cartesian space. This is inappropriate because moving the wrist a small amount does not make the robot configuration appear as different as moving the torso by the same amount. Furthermore, since gestures are predominantly free-space motions, payloads are irrelevant, and we avoid tuning the weights for a weighted joint-space metric by using a torque-space metric. A torque-space metric more consistently gives a 'weighted' distance metric, which yields greater penalty for deviations in DOFs closer to the root (center-of-mass) in the chain, and produces better pose-dependent penalties in Cartesian space for gestures.

Once the symbol frame in the original motion is determined, the same distance metric is used to calculate the maximum composite torque-space distance from all other frames in the motion with respect to the symbol frame. This quantifies the range of poses for a given original motion. Our experiment cannot use all frames from all motions because some of original motions have over 300 frames. Thus, we sample this space uniformly to select six other uniformly distanced frames. A true uniform sampling is not possible since we have a finite number of frames from which to select. Therefore, the selection of the frames is as close to uniform as possible. For all thirteen motions, selections included included frames from before and after the symbol. All frames came from the original motion, since we are testing whether the extracted frame is the frame of the original motion that produces the highest motion recognition accuracy from participants (i.e. the true symbol).

In this experiment, participants view one frame from each of the 13 motions in a random order, and are asked to label it with their best guess. Subjects are given the option to abstain from guessing, if they have no label for the motion. As a practice example, participants view one of seven possible frames from one of the 13 motions (randomly selected). After the practice example, the subjects view only one of the seven possible frames from each of the remaining twelve other motions. Motion order and frame are randomized. They are not allowed to go back review a previous motion or change a label once they have guessed.

This experiment contains one independent variable, which is distance from symbol frame. Since seven still frames were used for each of the thirteen motions, 224 participants were recruited to participate, yielding a sample size of 32 per still image. All participants saw one of seven possible random

[1]The gestures are: bow, beckon, shrug, point, stop, wave, fist bump, yes!, Mmm...tasty, cuckoo, knock, Shhh..., reach

TABLE I

COEFFICIENT OF DETERMINATION FROM MONOTONICALLY DECREASING POWER SERIES FITS OF PERCENTAGE CORRECTLY LABELED VERSUS DISTANCE FROM SYMBOL FRAME. MAXIMUM PERCENTAGE CORRECTLY LABELED (PCL) FOR ANY STILL IMAGE FOR EACH MOTION FROM EXPERIMENT ONE.

Motion	Coeff. of Det.	Max. PCL
Bow	0.9025	93.8
Beckon	0.9804	59.4
I Don't Know	0.9336	62.5
Point	0.9269	87.5
Stop	0.8668	53.1
Wave	0.9934	81.3
Fist Bump	0.9836	56.3
Mmm...Tasty	0.8767	9.4
Knock	0.9966	40.7
Yes!	0.9634	40.7
Cuckoo	0.9734	15.6
Shhh...	0.8352	81.3
Reach	0.8825	25.0
Composite	0.9944	54.3

frames from all thirteen motions.

In experiment one it is important to realize that overall gesture recognition accuracy is irrelevant for this specific analysis. We are testing our algorithm to determine if it can pick out the "best" frame from the given set of all possible frames in a motion. Thus, we only care about the relative recognition accuracy between frames of the same gesture. "Best" is the frame with highest recognition accuracy relative to all other frames.

B. Results

In order to demonstrate the relationship between frames in the motion relative to the symbol frame, we present results that depict the correlation between our distance metric (relative to the symbol) and percentage of participants who correctly labeled each still image. These results are shown for all thirteen motions from our study in Table I, where the numbers presented are the coefficients of determination from a monotonically decreasing power series fit of "percent correctly labeled" versus "distance from the symbol" ordered so that the symbol frame is far left and the frame furthest from the symbol is far right. For the results presented in Table I, any still images that had 0% correct recognition for any motion are excluded from the analysis.

A coefficient of determination of 1.0 means that the percent of participants who correctly label motion intent is perfectly correlated to distance from the symbol frame. Thus, the composite statistic (using the data from all thirteen motions) of 0.9944 indicates a strong correlation between torque-space distance from the symbol frame and ability of participants to accurately predict motion intent from still images. In short, when frames further from the symbol are shown to participants, they are less likely to predict the motion intent accurately.

Furthermore, in 12 of the 13 motions, the highest percent labeling accuracy occurred at the symbol frame. The exception was the 'reach' motion, where the labeling accuracy was 3% higher for the frame closest to the symbol. Reaching is a strong function of directionality. A reaching motion played forward looks like a 'placing' motion (without context) and a reaching motion executed backward is easily mistaken for a 'picking' motion. This directionality is absent in still frames, which suggests that prediction of intent for reaching depends more on context than the other motions in our study.

Given the high correlation between recognition accuracy and distance from symbol frame across motions and the fact that 12 of 13 motions had highest concentration of recognition accuracy near the symbol frame, we conclude that H1 holds true for our symbol extraction method.

VII. EXPERIMENT 2: COMMUNICATION OF INTENT

A. Experimental Design

Experiment two is designed to test whether humans can perceive motion intent sooner in anticipatory motion. We also test whether humans are confident enough to consistently guess a motion's intent prior to the symbol frame. We selected six motions from the previous experiment, and designed html and javascript code that would progress through videos at random. To measure accurate data, all participants accessed the code through the same computer, running the files locally on the computer, rather than over the internet. Only six motions were selected for this experiment because we wanted common gestures and communicative motions that would be familiar to the largest number of participants. Therefore, we selected six of the eight motions from experiment one for which any frame was correctly labeled by greater than 53.0% of participants, as shown in Table I. We believe it is valid to eliminate less common motions such as 'Mmmm...tasty' which had a maximum of 9.4% correct recognition for any still image frame in experiment one because for this experiment overall gesture recognition accuracy matters.

Participants viewed each of the six motions only once, and for each motion, the motion version (anticipatory or original) was randomly selected. The participants were instructed to click the "stop video" button immediately, when they thought that they could label the motion. The stop video button was very large to minimize cursor localization time lag. After clicking, the screen would change to a blank screen with an empty prompt for typing a label. The code logged the time from start of video playing to click of the stop button. If the user didn't click the stop button and the end of the video was reached, the screen would automatically transition to the page prompting for the motion label.

Videos of the robot hardware were used instead of the actual hardware for two reasons: safety and data integrity. It is safer to stop a video and have it disappear, than to have the real hardware freeze and hold position upon press of a button. Second, we wanted the motion and all poses to disappear from view to ensure that our participants were not relying upon the final keyframe in decision making.

To encourage participants to watch as much motion as they needed, "no label" was not an option. If a participant left the box blank or input characters that were not a label, this data

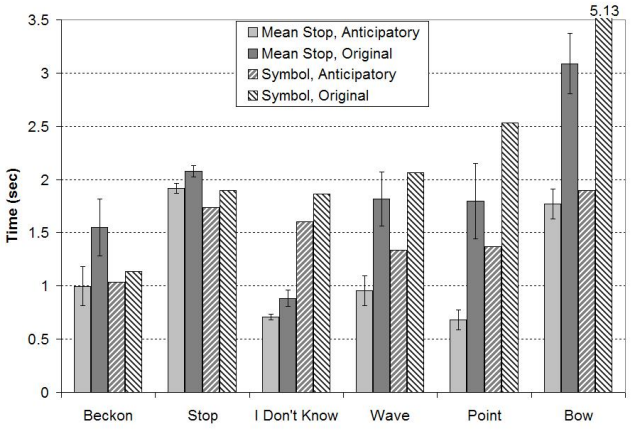

Fig. 3. Average stop times and symbol times for all six motions, both anticipatory and original versions. From left to right, difference in original and anticipatory symbol times increase. Error bars show that average stop time between original and anticipatory motions is significant for all motions.

was excluded from the experimental results. Since we are testing for the time when a human can first label a motion, if they cannot label the motion, then their answer provides no data for this question. Ideally, we want all participants to label motions correctly and wait only the minimum time necessary before pressing the stop button. Subjects were not allowed to re-watch any videos.

To provide extra incentive, participants were told that only the two participants with the fastest times to correctly label all six motions would receive $10.00 each for their participation in the experiment. The instructions were clear that only times when videos were actually playing counted toward their cumulative time total. They could spend as much time as they wanted typing in their labels for the motions.

B. Results

Eighty-two participants contributed to our second experiment. Using only the correctly labeled responses, Figure 3 clearly shows that average participant stop time is statistically significant for all six motions. Since there are two versions of each motion and we used only the correct number of responses, the average stop times in Figure 3 are determined from 37-41 correct responses. The motions in Figure 3 are ordered from left to right in order of increasing symbol time difference between the original and anticipatory motions to demonstrate that predictability of motion intent from the symbol is not a strong function of how much the symbol moves relative to its timing in the original motion. Even with as little as 100 milliseconds symbol timing difference in the beckon motion, intent is still more easily predicted with anticipatory motion. One possible reason is that motion at one time frame is dependent upon the previous frames. Thus, moving the symbol affects the previous frames to varying extents. In doing so, human observers have more information about intent even before the symbol frame, and the symbol frame is not the sole means by which human perceive intent.

The majority of participants watched motions less than the symbol time before stopping to label the motion. Across all six motions, 74% of correct responses from anticipatory motions were labeled before the symbol, and 65.9% of original motions were labeled before the symbol. From this, we conclude that people are developing a mental model of motion intent while viewing motion. This prediction of intent via the motion communication channel could explain why turns can overlap in turn taking activities, or humans can react preemptively to partner motions in collaborative tasks.

On average, with the six anticipatory motions in our experiment, participants reacted 697 milliseconds sooner to anticipatory motion with correctly labeled responses for motion intent. This finding supports H2, and provides evidence that (1) when motions are familiar, humans can discern intent from motion, (2) the social cue for turns of dynamic collaboration is not restricted to action completion, (3) anticipatory motion leads to earlier correct labeling of motion intent than motions that are not anticipatory.

Finally, it is interesting to compare the results from our first two experiments. Overall response rates for correct labeling are higher for motion than for static images. This suggests that motion conveys more information about intent than a single frame. Even though both experiments are largely devoid of context, ambiguities in motion intent are better resolved by viewing more frames.

VIII. Experiment 3: Quantifying the Time Range When Anticipatory Motion is More Beneficial

A. Experimental Design

In our final experiment we determine the time range over which anticipatory motion is beneficial for human-robot communication. We are attempting to make generalizations across motions in this experiment, and therefore we needed a variable that is not specific to a particular motion. H3 is based upon the logic that if the symbol is the most important frame in the entire motion, then once an observer has seen it, they will gain very little (in terms of determining intent) from watching the rest of the motion. To generalize across motions, we use percent of symbol time as the variable by which we divided group categories in this experiment. For example, 10% means that the [test video length] divided by the [time at which the symbol occurs] equals 0.1.

There are two independent variables for this experiment: motion end time and motion type. Motion end time has one of seven values: videos that ended at 20% increments with respect to symbol time, up to 120%, and the entire motion. With each of these there are two possible motion types that a participant could see: anticipatory or original motion.

Using web-based code running on a single computer (similar to the setup in experiment two), we serially randomly displayed one of the two possible versions for each of six different motion videos, each of which ended at one of seven possible (randomly-selected) end times. Subjects watched one video to the predetermined concluding point, then the screen would blank and prompt them for a label. We encouraged participants to label the motions, even if they were uncertain. Only correctly labeled data was included in our analysis.

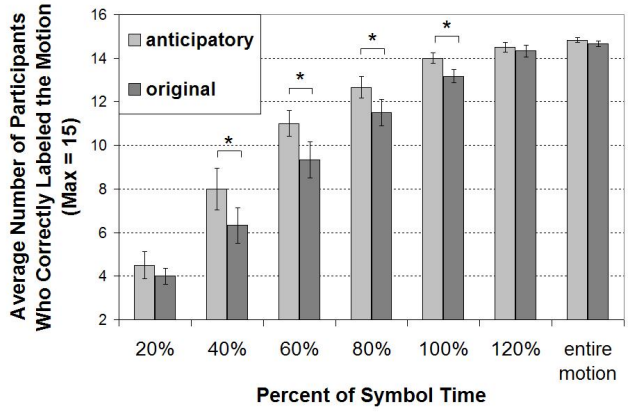

Fig. 4. Average number of participants who correctly labeled the motions based on motion end time relative to symbol frame. Comparisons in each division for original and anticipatory motion. 6 motions total, 15 participants per motion. * = statistically significant.

B. Results

Two-hundred ten participants were recruited, which yields a sample size of fifteen per time division. For each of the seven stop time divisions (20%, 40%,...,120%, and entire motion) and two motion type groups (original and anticipatory), we tallied statistics across all motions for the average number of participants who correctly labeled the motion video. Each group in each division is an average of six motions[2], and all participants observed all six motions, where each motion was randomly selected as original or anticipatory.

The results in Figure 4 show that in the time range between 40% to 100% of the symbol, anticipatory motion is recognized more accurately. After the symbol time and too early in the motion the differences between anticipatory motion and the original motion are not statistically significant. We speculate that for 20% of the symbol time, too little motion is seen to yield accurate guessing. As the motions progress beyond the symbol time in each video the anticipatory effects are not beneficial with respect to predicting intent because both the original and anticipatory motions have both shown enough representative motion. The data in Figure 4 supports H3. Moreover, it quantifies the time range relative to the symbol when benefits of including an early symbol in motion are gained: 40% to 100% of symbol time.

Recognition accuracy is much higher for motions than for static images. For example, using the six motions from experiment three compared to same six static symbol frames from experiment one, we notice that when participants watch all motion up to the symbol frame, recognition accuracy is 89.4% (average for all six motions). However, using only the static symbol frames, average recognition accuracy for all six motions drops to 72.4%. Motion is easier to correctly label as more of it is seen.

IX. Discussion

The value of anticipatory motion is derived from the benefits of knowing motion intent sooner. In our experiment, we show a result of 697 milliseconds earlier average reaction time with anticipatory motion. There are many examples of how earlier reaction time can be beneficial during interaction with humans. For example, in coupled interaction tasks, 697 extra milliseconds to respond can make the difference between not dropping the object that the human and robot are carrying together. Or, for robots directing traffic, 697 extra milliseconds can make the difference between life and death. Even in ordinary interactions, 697 extra milliseconds can possibly alter perceived responsiveness of the agent and ultimately be the difference between frustration and pleasure for the human partner in the interaction. The instrumental task utility of anticipatory motion is an important element of future work in this domain.

The main limitation of our algorithm is that it depends on the motion having variance in the hand normal vector throughout the duration, as would be expected from human motion. For extremely simple motions, such as those where very few DOFs are moving, no anticipatory motion can be produced using the current formulation of our algorithm.

X. Conclusion

We presented an autonomous algorithm that creates anticipatory motion variants from a single motion exemplar that has hand and body symbols as a part of its communicative intent. We validated that our algorithm extracts the most salient frame (i.e. the true symbol) which is most informative about motion intent to human observers. Furthermore, we showed that anticipatory variants allow humans to discern motion intent sooner than motions without anticipation, and that humans are able to reliably predict motion intent prior to the symbol frame when motion is anticipatory. Finally, we quantified the time range for robot motion when the collaborative social benefits of anticipatory motion are greatest.

[2]bow, beckon,'I don't know', point, stop, wave

References

[1] O. Johnston and F. Thomas, *The illusion of life: Disney animation*. Disney, 1995.

[2] J. Choi *et al.*, "Anticipation for facial animation," *Proceedings of 17th International Conference on Computer Animation and Social Agents*, 2004.

[3] N. Pollard *et al.*, "Adapting human motion for the control of a humanoid robot." in *Proceedings of International Conference on Robotics and Automation*, 2002, pp. 1390–1397.

[4] O. Jenkins and M. Mataric, "Deriving action and behavior primitives from human motion data." *Intelligent Robots and Systems.*, vol. 3, pp. 2551–2556, 2002.

[5] D. Swaminathan *et al.*, *Computer Music Modeling and Retrieval. Sense of Sounds*. Springer-Verlag, 2008.

[6] N. Rezzoug, "Virtual humanoids endowed with expressive communication gestures : the hugex project," *Systems, Man, and Cybernetics*, vol. 5, pp. 4445–4450, 2006.

[7] J. Lee *et al.*, "Motion graphs." in *Proceedings of the 29th Annual Conference on Computer Graphics and Interactive Techniques*, July 2002, pp. 473–482.

[8] A. Nijholt *et al.*, *Verbal and Nonverbal Features of Human-Human and Human-Machine Interaction*. Springer-Verlag, 2008.