2023 IEEE International Conference on Robotics and Automation (ICRA 2023)
May 29 - June 2, 2023. London, UK

Multi-modal Interactive Perception in Human Control of Complex Objects

Rashida Nayeem*, Salah Bazzi, Mohsen Sadeghi, Reza Sharif Razavian, and Dagmar Sternad

Abstract— **Tactile sensing has been increasingly utilized in robot control of unknown objects to infer physical properties and optimize manipulation. However, there is limited understanding about the contribution of different sensory modalities during interactive perception in complex interaction both in robots and in humans. This study investigated the effect of visual and haptic information on humans' exploratory interactions with a 'cup of coffee', an object with nonlinear internal dynamics. Subjects were instructed to rhythmically transport a virtual cup with a rolling ball inside between two targets at a specified frequency, using a robotic interface. The cup and targets were displayed on a screen, and force feedback from the cup-and-ball dynamics was provided via the robotic manipulandum. Subjects were encouraged to explore and prepare the dynamics by "shaking" the cup-and-ball system to find the best initial conditions prior to the task. Two groups of subjects received the full haptic feedback about the cup-and-ball movement during the task; however, for one group the ball movement was visually occluded. Visual information about the ball movement had two distinctive effects on the performance: it reduced preparation time needed to understand the dynamics and, importantly, it led to simpler, more linear input-output interactions between hand and object. The results highlight how visual and haptic information regarding nonlinear internal dynamics have distinct roles for the interactive perception of complex objects.**

I. Introduction

When a child shakes a present before opening it on Christmas morning, they can quickly guess what they received. Humans exhibit exquisite skill in perceiving objects through exploratory interactions [1]. This includes rattling boxes to gauge their contents, or squeezing fruits to feel their ripeness. This human ability of interactive perception, i.e., using forceful interactions with an object to gain information, has recently received substantial attention in the robotics community [2].

Interactive perception of non-rigid objects with internal degrees of freedom, such as sloshing liquids in containers, is of paramount interest to robotics [3]–[7]. Recent approaches in robotic manipulation have leveraged this interactive approach to both obtain information about the object and then to subsequently manipulate it [8]–[10]. However, visual information processing is extremely costly and the integration of different sensory information in robotic systems presents major computational challenges. Therefore, most control policies have relied exclusively on haptic or tactile signals to infer properties of the objects. For example, when grasping different rigid and non-rigid objects, tactile information was shown to enable successful manipulation [8]. Yet, we conjecture, integrating multiple streams of information could potentially lead to more informed control schemes.

Advances in robotics have been inspired by human research showing that information obtained through exploratory actions improves manipulation strategies [11], [12]. Humans routinely integrate haptic, acoustic and visual information for successful manipulation but each of these information sources may have differing impacts on behavior. Although it has been understood that humans are 'vision-dominant' [13], studies on manipulation have emphasized the intricate interplay of haptic and visual information [14]. However, these studies have focused on how humans reach to or handle rigid objects without complex internal dynamics. Only two previous studies on the manipulation of a linear mass spring examined the role of haptic information and reported that it is necessary for dexterous performance [15], [16]. How humans use both visual and haptic information to explore and manipulate objects with nonlinear internal dynamics, e.g., a cup of coffee, is still unknown. As robots aim to dexterously manipulate complex objects, it is useful to understand how humans interactively perceive and utilize different information channels for manipulation.

This experimental study is the first to investigate how different sensory modalities affect humans' ability to gain information about an object with nonlinear internal dynamics through interactive perception. In previous research, Sternad and colleagues have examined human control of an object with nonlinear internal dynamics inspired by 'carrying a cup of coffee' [17]–[22]. Using a virtual environment, subjects interacted with a cup-and-ball system, visualized on a screen, via a robotic manipulandum that moved the cup and also provided haptic feedback about the internal ball forces back to the user's hand. The dynamics of the cup-and-ball can evolve into complex and potentially chaotic behavior. Studies have found that during interaction humans aim to make cup-and-ball dynamics simpler, i.e., more predictable. A recent study by Nayeem et al. investigated how humans explored and prepared this system prior to a continuous rhythmic transport task [23], [24]. Results showed that subjects interactively prepared the object for the upcoming task: by 'jiggling' the system back and forth, they learned which initial states resulted in shorter transients to reach a more predictable steady state faster.

Using the same experimental paradigm, this paper explored how visual and haptic information about the internal dynamics affected subjects' exploration strategies, i.e., their interactive perceptual strategies. Two experimental conditions were implemented in a virtual environment: the first condition presented full haptic and visual feedback;

*Corresponding author: nayeem.r@northeastern.edu. Rashida Nayeem, Salah Bazzi, Mohsen Sadeghi, Reza Sharif Razavian, and Dagmar Sternad are in the Departments of Electrical and Computer Engineering, Biology, and the Institute for Experiential Robotics, Northeastern University, Boston, MA. This work was supported by NSF-NRI-1637854, NSF-M3X-1825942, NIH-R37-HD087089, awarded to Dagmar Sternad.

2023 IEEE International Conference on Robotics and Automation (ICRA) | 979-8-3503-2365-8/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICRA48891.2023.10160375

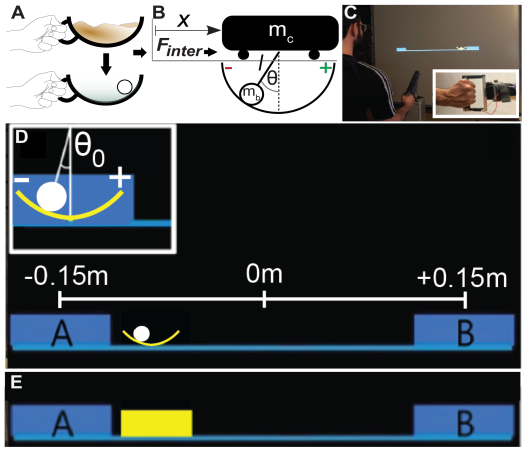

Fig. 1. **A.** Experimental task inspired by transporting a cup of coffee, simplified to a cup with a ball sliding inside. **B.** Mechanical model of cart-pendulum system. **C.** Participant holding the robot handle to move the cup while viewing the system on a screen. Inset shows the subject's grip of the robot handle. **D.** Display in the *Full Information* condition, inset shows definition of the ball angle. **E.** Display for the *Hidden Dynamics* condition.

the second condition occluded visual information about the ball dynamics. Subjects' interactive perceptual actions were quantified by their ability to converge to a control strategy that minimized transients and reached a steady state faster. Results showed that without visual information about internal dynamics, subjects required more time for exploration and excited the system with a wider range of frequencies, yet were less likely to find the optimal solution.

II. Methods

A. Experimental Task, Apparatus, and Data Acquisition

Subjects interacted with a 'cup of coffee' simulated in a virtual environment. Simulating a 3D cup with sloshing coffee would be computationally expensive and was not a viable option for real-time virtual rendering. Therefore, the task was simplified to transporting a 2D semicircular cup moving on a horizontal line with a ball sliding inside the cup (Fig.1). Since the ball was sliding instead of rolling, the system was mechanically equivalent to a cart sliding on frictionless line with a suspended frictionless pendulum. The pendulum bob was represented by the ball, and the cup position corresponded to the cart position. The arc of the cup corresponded to the rotational path of the pendular bob, i.e., ball (Fig.1). While simplified, the task retained the basic challenges of transporting a cup of coffee: underactuation and nonlinear internal dynamics. The equations of motion for the system are:

$$(m_c + m_b)\ddot{X} = \underbrace{m_b l[\dot{\theta}^2 sin\theta - \ddot{\theta} cos\theta]}_{F_{ball}} + F_{inter} \quad (1)$$

$$\ddot{\theta} = -\frac{\ddot{X}}{l} cos\theta - \frac{g}{l} sin\theta. \quad (2)$$

F_{inter} is the force applied by the human hand on the cup. X and θ are cup position and ball angle, respectively. The ball angle when at the bottom of the cup defined 0deg; clockwise direction was negative. F_{ball} denotes the force that the ball exerts on the cup. Parameters used to simulate the system were: cup mass m_c=2.4kg, ball mass m_b=1.0kg, pendulum length l=0.45m, and gravitational acceleration g. The values were chosen to be heavy enough for the subjects to feel F_{ball} upon their hand, but light enough to avoid fatigue.

Subjects interacted with the virtual cup-and-ball via a robotic manipulandum capable of haptic force feedback (HapticMaster, Fig.1C) [25]. The cup was shown as a yellow arc and a small white circle rolling inside represented the ball. The subject grasped the robot handle to control the displacement of the cup, X, shown on a screen 2m in front of them. The robot was admittance-controlled: the subject's force, F_{inter}, on the manipulandum resulted in cup displacement, according to Eqs. (1-2). A custom-written C++ program based on the HapticAPI computed the cup and ball kinematics that then controlled the cup and ball on the visual display. The ball force F_{ball} was haptically fed back to the subject's hand at 120Hz update rate. Corresponding to the horizontal cup movement, the movement of the robot handle was also restricted to a horizontal line. Two blue rectangular target boxes delimited the cup's peak-to-peak amplitude for the instructed back-and-forth movements (Fig.1D). The cup's rim was at $\pm$50deg; the ball could not 'escape' from the cup, but if it exceeded $\pm$50deg, it would rotate above the cup rim.

B. Experimental Conditions and Task Instructions

Two groups of 9 healthy college-aged subjects each performed one of the following two conditions: in the *Full Information* condition subjects interacted with the simulated cup and ball, while receiving full visual and haptic feedback (Fig.1D). In the *Hidden Dynamics* condition a yellow rectangle covered the system. Subjects were unaware they were manipulating a cup with a ball rolling inside (Fig.1E). Only the F_{ball} acting upon their hand via the manipulandum would provide haptic information to infer the object's dynamics.

At the start of each trial, the cup was positioned in Box A with the ball at rest (0deg). Subjects were instructed to move the cup in rhythmic fashion between Box A and Box B (0.3m apart) for 15s paced by an auditory metronome that was set to 0.60Hz (Fig.1C). Prior to starting the prescribed rhythmic movement, subjects were encouraged to explore and prepare the cup-and-ball dynamics by 'jiggling' the cup. This interactive preparation interval was not limited in time, but the cup motions were constrained to the left half of the screen (Fig.1D). They were told find a preparation strategy that would allow them to complete the prescribed rhythmic task to the best of their ability. Once subjects felt ready, they moved the cup towards Box B (0.15m) and continued moving rhythmically between the two boxes (Fig.2A) at a pace of 0.60Hz. The metronome began when the participant reached Box B for the first time. The experiment consisted of 120 trials for each condition, which lasted 40 minutes. All subjects gave written informed consent, as approved by the Institutional Review Board of Northeastern University. De-indentified data is publicly available online [26].

C. Performance Metrics

The effects of interactive perception were first assessed by the initial ball states that participants adopted before starting the rhythmic task. The duration of the transients in each trial measured how quickly subjects reached a steady state,

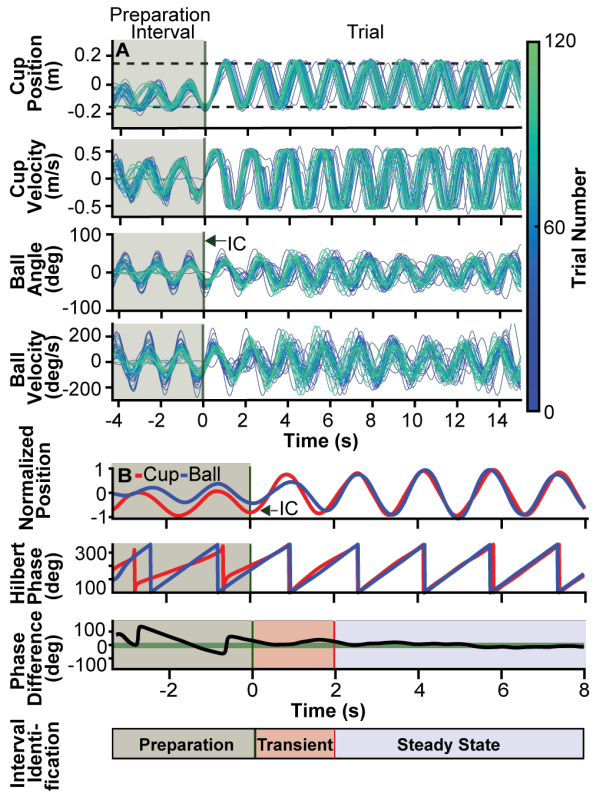

Fig. 2. **A.** Time series of cup and ball position and velocity for all trials of an example subject in *Full Information* condition. Early trials are in dark blue, later trials are in green. IC indicates initial conditions, defined as the state variables when cup velocity was zero prior to reaching Box B for the first time. Trials are aligned by this point, defined as time 0. **B.** Calculation of transient duration. When relative phase between cup and ball phase was less than 27deg, the end of the transient and start of steady stat was defined.

and therefore more predictable dynamics [23], [24]. These two metrics served as an indicator of how well subjects had inferred the object's dynamics via preparatory activity. To further examine whether the two experimental conditions (Full Information and Hidden Dynamics) elicited different preparatory activity, the frequency components in the preparation interval were also analyzed. In addition, system identification was applied to characterize the activity serving interactive perception. All calculations were performed in Matlab (Mathworks, v.29b, Natick MA).

1) Initial Conditions: Initial conditions θ_0 and $\dot{\theta}_0$ were determined at the instance when the subject started the cup movement towards Box B (0.15m), i.e., the final zero cup velocity before reaching Box B (Fig.2A). The states that defined the system's initial conditions were ball angle θ_0 and ball velocity $\dot{\theta}_0$; cup position X_0 was not included in further analyses as it was the at the center of Box A. Movements before this time point were considered preparation.

2) Transient Duration: As the rhythmic cup movement began, the cup-and-ball system exhibited a transient prior to reaching a steady state. To calculate the duration of this transient, the steady state for the system had to be defined first. For rhythmic cup movements at the metronome frequency of 0.60Hz, the cup and ball position were in phase (Fig.2B). To compute the end of the transient and start of the steady state, the instantaneous phase of the cup and ball position were calculated using Hilbert transform [27]. Relative phase between cup and ball, the difference between the two phase signals, served to indicate when the system entered a steady state. A relative phase less than 27deg (15% percent of $\pm$180deg) for the rest of the trial marked the end of the transient and the start of steady state [23], [24]. The time between the initial conditions and start of steady state defined the transient duration.

Given the known dynamics of the system, forward dynamic simulations were run to evaluate which initial conditions led to shorter transients. These simulations required an input force to the cup-and-ball system, i.e., a controller. As a simple choice, the control input was the desired rhythmic cup trajectory $X_{des}(t) = (A/2)sin(2\pi ft + \pi/2))$ coupled to a hand impedance, i.e., a linear spring K in parallel with a damper B [23], [24], [28]. The equations of motion of the coupled model include (1)-(2) with F_{inter} expressed as:

$$F_{inter} = -K(X - X_{des}) - B(\dot{X} - \dot{X}_{des}) \qquad (3)$$

The hand impedance acted as a proportional derivative controller that reduced any divergence of $X(t)$ from $X_{des}(t)$ due to ball forces F_{ball} [19]. Stiffness K and damping B were constants; their respective values were estimated from each experimental trial using an optimization method [19], [23], [24]. For the forward simulations, the mean constant values were used: K= 40N/m and B= 20Ns/m. To evaluate the effect of initial ball states on the transient duration, the cup-and-ball system was forward-simulated for different θ_0 ($\pm$90deg, 1deg step size) and $\dot{\theta}_0$ ($\pm$150deg, 1deg step size) to produce a heat map of transient durations.

3) Preparation Interval: Participants had no time limit for their preparatory interactions, hence the duration of this interval was also informative. The preparation interval was the time between the start of the trial to the point where initial conditions were determined. To characterize preparatory activity, cup position was parsed into cycles and their frequencies determined. All frequencies of a trial were pooled across subjects and binned into 0.02Hz bins. The frequencies were binned into 10 trial intervals and summarized in a time-frequency plot which showed the changes in preparation cup frequencies across practice.

System identification methods were used to characterize input-output behavior during preparation. Linearizing Eqs.1-2 around the pendulum's downward position, a 4^{rd}-order transfer function described the system dynamics with interaction force as an input and cup position as an output. Therefore, a 4^{rd}-order linear transfer function was fit between interaction force (input) and cup position (output), using tfest.m (Mathworks, v.29b; [29]–[31]). 30 trials were used as the system identification needed to be adequately trained. This tested if subjects linearized dynamic behavior during preparation. To check sensitivity to parameters, functions of 3^{rd}- and 5^{th}-order were also fitted. Specifically, each transfer function was fit between the timeseries of the interaction force and cup position for the first and last 30 trials of each individual to assess if there was a change in preparation activity with practice. The fitting error was calculated as the

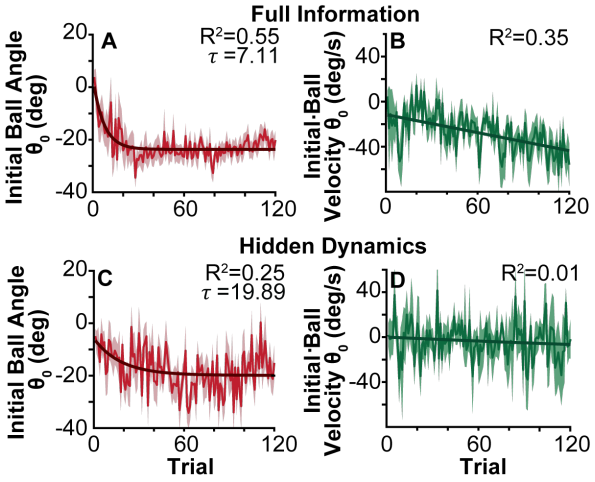

Fig. 3. **A.** Initial ball angles for *Full Information* averaged across subjects for each trial; shading indicates one standard deviation. **B.** Initial ball velocity for *Full Information*. **C.** Initial ball angle for *Hidden Dynamics*. **D.** Initial ball velocity for *Hidden Dynamics*.

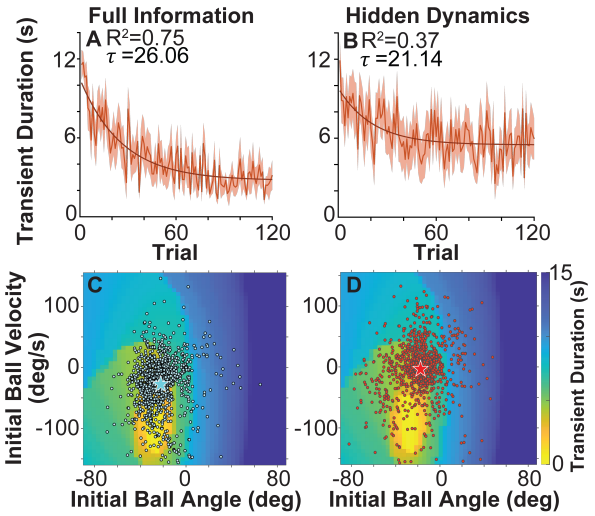

Fig. 4. **A.** Average transient duration across subjects over trials in *Full Information*; shading indicates one standard deviation. **B.** Average transient duration across subjects over trials in *Hidden Dynamics* condition. **C** Heat map of simulated transient durations for different initial ball angles and velocities for 0.6 Hz frequency in *Full Information*. Experimental data are overlaid, with center denoted by the star. **D.** Same heat map as in C. with experimental data from *Hidden Dynamics* overlaid; center of data denoted by the star.

mean squared error (MSE) between experimental output (cup position) and the model estimation for the corresponding input (interaction force).

4) Statistical Analyses: The performance metrics (initial conditions, transient duration, duration of the preparation interval) showed approximately exponential trends across trials. Therefore, the averaged data was fit with exponential functions, and the time constant τ indicated the rate of convergence to the final value; R^2 specified the goodness of fit. As initial ball velocity exhibited poor fits to exponential functions ($R^2 < 0.10$), linear functions were used. For each metric, the difference between early (first 5 trials) and late (last 5 trials) performance within groups was compared using paired t-tests. Unpaired t-tests quantified the differences across groups. For all tests $p < 0.05$ was considered significant. Finally, to identify differences in the quality of fit from the system identification procedure, a regression analysis (general linear model) was used with model order, early-late, feedback conditions and subject as fixed factors [32]. All individual trials were fed into the regression model.

III. Results

Subjects in both *Full Information* and *Hidden Dynamics* conditions performed the rhythmic task as instructed at the metronome frequency of 0.60Hz and with an amplitude conform with the distance between the target boxes (0.30m). There was no apparent change in amplitude or frequency over the 120 trials. The average frequencies and standard deviations across trials and subjects were 0.59 ± 0.001Hz and 0.60 ± 0.012Hz in the two conditions. The average movement amplitudes were: 0.32 ± 0.003m, and 0.31 ± 0.003m.

A. Initial Conditions

In the *Full Information* condition, the initial ball angle θ_0 clearly converged to preferred values with practice, while initial ball velocity decreased linearly $\dot{\theta}_0$ (Fig. 3A,B). The exponential and linear fits are shown by solid black lines. The colored data denote the mean and one standard deviation for each trial number across 9 subjects. The average ball angle θ_0 decreased with a time constant τ=7.11 trials to an asymptote of -23deg. Average values in the first 5 trials dropped from -6.64 ± 13.27deg to -21.59 ± 7.17deg in the last 5 trials, ($t(8)$=2.8, p=0.02). Ball velocity $\dot{\theta}_0$ declined linearly with a slope -0.29 deg/s·trial and an intercept -10deg/s; there was also a significant difference between the first 5 trials (-13.49 ± 28.64deg/s) to the last 5 trials (-44.57 ± 33.64deg/s); $t(8)$=2.75, p=0.02; Fig. 3.

Subjects in the *Hidden Dynamics* group were not aware that they were moving a cup with a ball inside, as it was occluded by a solid rectangle (see Fig.1E). However, F_{ball} acted on the hand via the robot handle. Subjects converged to a preferred θ_0, despite being deprived of visual information about the ball. The average θ_0 values declined exponentially, with $\tau = 19.89$; convergence to the final value of -20deg was slower in this group (Fig. 3C). The initial ball angle averaged across the first 5 trials changed from -7.45 ± 10.59deg to -20.66 ± 9.66deg in the last 5 trials ($t(8)$=2.47, $p = 0.04$). The final θ_0 in the *Full Information* and *Hidden Dynamics* were not significantly different ($t(16)$=-0.28, p=0.79). For the *Hidden Dynamics* group, $\dot{\theta}_0$ was highly variable and showed no significant trend across the experiment (Fig. 3D). Values in the first 5 trials were 7.29 ± 20.51deg/s and 0.18 ± 42.66deg/s in the last 5 trials ($t(8)$=0.40, p=0.69). Initial ball velocity in the last 5 trials between *Full Information* and *Hidden Dynamics* were significantly different ($t(16)$=-2.53, p=0.02).

B. Transients

Following the convergence to preferred initial states, it was expected that transient durations would decrease with practice. The average transient durations across all subjects in the *Full Information* condition declined with a decay constant of $\tau = 26.06$ trials to an asymptote of 2.76s, Fig.4A. In the first 5 trials the average duration was 12.04 ± 2.38s decreasing

to 3.62 ± 1.43s by the last 5 trials ($t(8)$=9.57, p=$1.73 \cdot 10^{-08}$). Subjects with *Hidden Dynamics* also shortened their transients with a similar time constant of $\tau = 21.14$ trials to a final value of 5.49s (Fig.4B). The average duration in the first 5 trials, 11.24 ± 2.55s, decreased to 6.88 ± 3.67s by the last 5 trials ($t(8)$=3.07, p=0.006). However, transient durations with *Hidden Dynamics* were not shortened to the same degree as with *Full Information*. The values achieved in the last 5 trials were significantly different in the two conditions ($t(16)$=2.94, p=0.009).

To evaluate the effect of the initial conditions of cup and ball on the transients, we used the simple control model to simulate the cup-and-ball dynamics. The objective was to compare which of the two perceptual conditions were closer to achieving optimal preparation. Fig.4C,D both show the same heat map of simulated transient duration for a range of initial ball angles θ_0 and initial ball velocities $\dot{\theta}_0$. Yellow areas indicate initial ball states that produced the shortest transients in simulation. As illustrated, the range of initial ball states that produced transient durations < 0.1s were between -26.97deg and -18.38deg (θ_0) and between -157.13deg/s and -65.46deg/s ($\dot{\theta}_0$). The center was at θ_0=-22.67deg and $\dot{\theta}_0$=-111.3deg/s.

All experimental data from the *Full Information* and *Hidden Dynamics* conditions were overlaid onto the same simulated landscape (Fig.4C,D respectively). Subjects with visual and haptic feedback chose θ_0, $\dot{\theta}_0$ that produced transients close to the optimum in simulation. The cyan star shows the center of the data at θ_0=-22.17deg and $\dot{\theta}_0$=-27.68deg/s. Subjects without visual feedback were further away from the optimal states. The center of the data, shown by the red star, was at θ_0=-17.57deg and $\dot{\theta}_0$=-3.43deg. These results indicate that if only provided haptic information about the internal dynamics, subjects could find a mapping between initial states and simplified dynamic behavior. However, for convergence to the global solution, subjects also required visual information about the internal dynamics. To understand why the preferred initial states differed between the two groups, the preparatory activity was analyzed further.

C. Characterization of Preparatory Activity

One indicator of whether the preparatory activity changed with practice was the duration of the preparation interval (Fig.5A,B). The duration in the *Full Information* condition was 17.69 ± 10.61s on average in the first 5 trials and decreased to 5.76 ± 5.04s in the last 5 trials ($t(8)$=3.34, p=0.008; Fig.5A). The time constant of an exponential function fit was τ=16.09. Subjects in the *Hidden Dynamics* condition required more trials to converge to a final value τ=48.41 (Fig.5B), but their preparation interval duration also decreased significantly from 23.42 ± 13.46s to 5.40 ± 1.80s ($t(8)$=3.88, p=0.004).

Cycle-by-cycle frequencies during the preparation interval, summarized as histograms in Fig.5C,D, revealed that in both conditions, subjects initially explored a wide array of frequencies from 0.50-0.80Hz. Across trials, subjects narrowed this range to frequencies around 0.60Hz, coincident with the subsequent metronome frequency. It is notable that

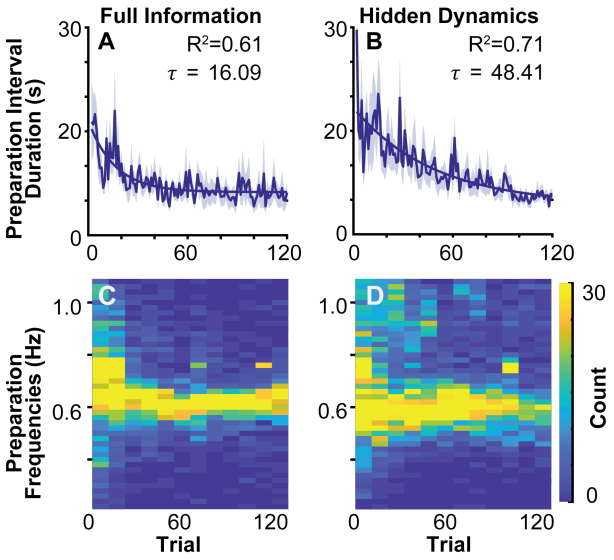

Fig. 5. **A.** Duration of preparation interval over trials in the *Full Information* condition, averaged across subjects per trial. **B.** Duration of preparation interval over trials in the *Hidden Dynamics* condition, averaged across subjects per trial.**C.** Histograms of individual cycle frequencies visited during the preparation interval in *Full Information*. Yellow indicates high occurrence of a frequency. **D.** Histograms of individual cycle frequencies visited during the preparation interval in *Hidden Dynamics*.

this distribution was slightly wider in *Hidden Dynamics*.

This small but visible difference in 'jiggling' frequencies motivated further analysis using system identification. Fig.6A,B illustrates timeseries of interaction forces and the resulting cup positions for all trials in the preparation interval for two example subjects. The subject in the *Hidden Dynamics* condition exhibited more variability in preparation frequencies than the subject in *Full Information*. To capture the specific dynamic behavior, system identification methods were applied to fit 3rd-, 4th- and 5th-order linear transfer functions to the data. Goodness of fit quantified by the MSE values are summarized in Fig.6C. The averaged MSE fits to training data across subjects from the first 30 and last 30 trials for the three transfer functions are shown. Regression analyses found a significant difference between the two experimental conditions ($p < 0.001$), and between early and late practice ($p < 0.001$). In *Full Information*, the decrease in MSE from early to late training was significant ($t(16) > 2.5$, $p < 0.03$, for all orders of linear fits), demonstrating that subjects learned to simplify or linearize their dynamic behavior in the preparation interval (Fig.6C). In contrast, in the *Hidden Dynamics* condition, the change in MSE values from early to late practice was not significant and indicated that subjects did not learn to produce more linearized input-output behavior ($t(16) < 1.49$, $p > 0.17$, for all orders of linear fits). Comparison between the MSE values in early stage of training between the two conditions showed no significant difference ($t(16) < 0.86$, $p > 0.4$, for all orders of linear fits). However, comparison of values in later trials showed that MSE values in *Full Information* were significantly lower than those in *Hidden Dynamics* ($t(16) > 2.1$, $p < 0.05$ for all orders of linear fits). Subjects in the *Hidden Dynamics* condition excited complex dynamic

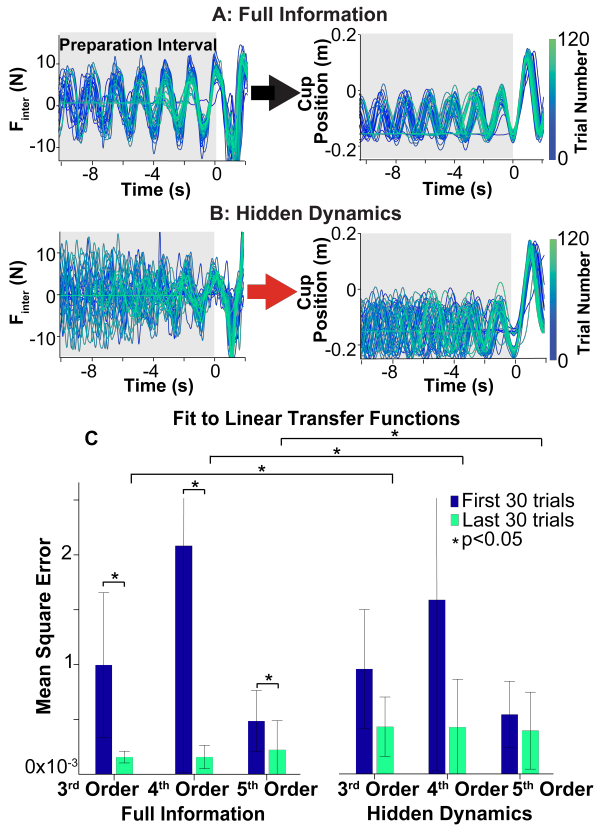

Fig. 6. **A.** Input-output behavior (interaction force to cup position) of all trials of one subject during the preparation interval with *Full Information*. Early trials are in dark blue, late trials are in green. **B.**. Input-output behavior during the preparation interval of one subject in the *Hidden Dynamics condition*. **C.** Mean square error (MSE) of early and late trials in fitting 3 different linear functions for *Full Information* (left) and *Hidden Dynamics* (right).

modes and could not linearize dynamics, and therefore could not simplify behavior.

IV. Discussion

This paper investigated the effect of visual and haptic information in humans when exploring and preparing a complex object for a transport task. Examples for the challenges that manipulation of complex objects pose abound, both for humans and robotic systems, ranging from opening a box with unknown contents to carrying a cup filled with liquid without spilling [33], [34]. This study compared interactive strategies to identify the effect of visual and haptic feedback about the internal dynamics. With full visual and haptic information, subjects successfully explored the system's dynamics to converge relatively quickly to optimal initial conditions for the task. Transients significantly decreased as a consequence and subjects reached a steady state faster. Without visual information, convergence to initial conditions was less optimal and transients did not decrease to the same degree. Analysis of preparation activity revealed that with unrestricted information, subjects achieved linear mappings between interaction forces and the resulting cup position.

Perhaps the most striking difference between the two conditions was their transient durations (Fig.4A,B). Shorter transients and, hence, longer steady state behavior is desirable, as in steady state the system exhibits predictable dynamics. This is likely the result of different preparatory actions as the length of the preparation interval differed accordingly. Multi-modal sensory feedback not only aided the identification of object dynamics, but also facilitated learning of simpler control strategies. Subjects with unrestricted information were able to simplify preparatory behavior of the cup-and-ball by eliciting linear dynamics. This finding is in line with existing human movement studies, which showed that augmenting visual feedback with haptic information in a ball bouncing task enhanced subjects' learning of open-loop stable control strategies [35], [36]. Furthermore, we believe that the dual role of haptic feedback, in carrying both information and mechanical power, may have enhanced the perception and learning of interaction dynamics. Leveraging this feature of haptic feedback may be useful for robotic applications and warrants further investigation.

Subjects without visual information about the internal dynamics did not converge to a preferred initial ball velocity (Fig.3D). While not a singular contributor to performance, this absence of convergence to a specific value indicated that subjects were not able to estimate ball velocity solely using force feedback. This result showed that different sensors can observe different states. For a robotic task, when full state observability is crucial, a multi-modal sensory stream would be advantageous.

In the absence of vision, the decline in the duration of the preparation interval was also much slower (5A,B). This was accompanied by a similarly slow convergence of the range of frequencies employed in the preparation interval (5C,D) implying that without visual information, subjects demanded more interactions with the object to identify its properties and dynamics. With an eye to robotics, a multi-modal sensory stream could potentially facilitate more efficient object property estimation and learning of a manipulation skill. This could lead to more agile robots in manufacturing, military and healthcare settings.

In robotics the vast majority of approaches that model learning through object interaction have only used one mode of data: either visual or tactile feedback [37]–[43]. However, the results presented here highlight the importance of equipping a learning system with multiple sensory modalities. There is no one sensor, whether visual or haptic, that alone is adequate for learning the dynamics of an object. The integration of multi-modal data facilitates more efficient and robust interactive perception. From a robot control perspective, this has its own challenges given the heterogeneous nature of the data and their different dimensions, frequencies, and characteristics. Therefore, more investigation is warranted.

This study only investigated the visual and haptic sensory modalities. Yet, humans incorporate an even richer array of sensory modalities including proprioception, vestibular, auditory and olfactory feedback. In robotics, it would be useful to explore the value of adding a wider set of modalities to go beyond vision and tactile sensing, such as auditory and pressure sensors. For future work, we would like to expand our set of experimental conditions to investigate the effects of a wider range of sensory modalities and their integration.

REFERENCES

[1] L. A. Jones and S. J. Lederman, *Human hand function*. New York: Oxford University Press, 2006.

[2] J. Bohg, K. Hausman, B. Sankaran, O. Brock, D. Kragic, S. Schaal, and G. S. Sukhatme, "Interactive perception: Leveraging action in perception and perception in action," *IEEE Transactions on Robotics (T-RO)*, vol. 33, no. 6, pp. 1273–1291, 2017.

[3] C. L. Chen, J. O. Snyder, and P. J. Ramadge, "Learning to identify container contents through tactile vibration signatures," *IEEE International Conference on Simulation, Modeling, and Programming for Autonomous Robots (SIMPAR)*, pp. 43–48, 2016.

[4] H.-J. Huang, X. Guo, and W. Yuan, "Understanding dynamic tactile sensing for liquid property estimation," *Robotics: Science and Systems (RSS) 2022*, pp. 72–82, 2022.

[5] M. Eppe, M. Kerzel, E. Strahl, and S. Wermter, "Deep neural object analysis by interactive auditory exploration with a humanoid robot," *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 284–289, 2018.

[6] H. P. Saal, J.-A. Ting, and S. Vijayakumar, "Active estimation of object dynamics parameters with tactile sensors," *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 916–921, 2010.

[7] C. Matl, R. Matthew, and R. Bajcsy, "Haptic perception of liquids enclosed in containers," *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 7142–7149, 2019.

[8] R. Calandra, A. Owens, D. Jayaraman, J. Lin, W. Yuan, J. Malik, E. H. Adelson, and S. Levine, "More than a feeling: Learning to grasp and regrasp using vision and touch," *IEEE Robotics and Automation Letters (RA-L)*, vol. 3, no. 4, pp. 3300–3307, 2018.

[9] C. Wang, S. Wang, B. Romero, F. Veiga, and E. Adelson, "Swingbot: Learning physical features from in-hand tactile exploration for dynamic swing-up manipulation," *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 5633–5640, 2020.

[10] S. Tian, F. Ebert, D. Jayaraman, M. Mudigonda, C. Finn, R. Calandra, and S. Levine, "Manipulation by feel: Touch-based control with deep predictive models," *International Conference on Robotics and Automation (ICRA)*, pp. 818–824, 2019.

[11] S. J. Lederman and R. L. Klatzky, "Extracting object properties through haptic exploration," *Acta Psychologica*, vol. 84, no. 1, pp. 29–40, 1993.

[12] M. Turvey and C. Carello, "Chapter 11 - Dynamic Touch," in *Perception of Space and Motion*, ser. Handbook of Perception and Cognition, W. Epstein and S. Rogers, Eds. San Diego: Academic Press, 1995, pp. 401–490.

[13] I. Rock and C. S. Harris, "Vision and touch," *Scientific American*, vol. 216, no. 5, pp. 96–107, 1967.

[14] R. L. Klatzky, S. J. Lederman, and D. E. Matula, "Haptic exploration in the presence of vision." *Journal of Experimental Psychology: Human Perception and Performance*, vol. 19, no. 4, p. 726, 1993.

[15] F. C. Huang, R. B. Gillespie, and A. D. Kuo, "Visual and haptic feedback contribute to tuning and online control during object manipulation," *Journal of Motor Behavior*, vol. 39, no. 3, pp. 179–193, 2007.

[16] F. Danion, J. S. Diamond, and J. R. Flanagan, "The role of haptic feedback when manipulating nonrigid objects," *Journal of Neurophysiology*, vol. 107, no. 1, pp. 433–441, 2012.

[17] S. Bazzi, J. Ebert, N. Hogan, and D. Sternad, "Stability and predictability in dynamically complex physical interactions," *IEEE International Conference on Robotics and Automation (ICRA)*, pp. 5540–5545, 2018.

[18] ——, "Stability and predictability in human control of complex objects," *Chaos*, vol. 28, no. 10, p. 103103, 2018.

[19] P. Maurice, N. Hogan, and D. Sternad, "Predictability, force, and (anti)resonance in complex object control," *Journal of Neurophysiology*, vol. 120, no. 2, pp. 765–780, 2018.

[20] B. Nasseroleslami, C. J. Hasson, and D. Sternad, "Rhythmic manipulation of objects with complex dynamics: Predictability over chaos," *PLoS Computational Biology*, vol. 10, no. 10, p. e1003900, 2014.

[21] H. Guang, S. Bazzi, D. Sternad, and N. Hogan, "Dynamic primitives in human manipulation of non-rigid objects," *IEEE International Conference on Robotics and Automation (ICRA)*, pp. 3783–3789, 5 2019.

[22] W. J. Sohn, R. Nayeem, I. Zuzarte, N. Hogan, and D. Sternad, "Control of Complex Objects: Challenges of Linear Internal Dynamics," *IEEE/RAS-EMBS International Conference on Biomedical Robotics and Biomechatronics (BioRob)*, pp. 1229–1235, 2020.

[23] R. Nayeem, S. Bazzi, N. Hogan, and D. Sternad, "Transient behavior and predictability in manipulating complex objects," *IEEE International Conference on Robotics and Automation (ICRA)*, pp. 10 155–10 161, 9 2020.

[24] R. Nayeem, S. Bazzi, M. Sadeghi, N. Hogan, and D. Sternad, "Preparing to move: Setting initial conditions to simplify interactions with complex objects," *PLOS Computational Biology*, vol. 17, no. 12, p. e1009597, 2021.

[25] R. Q. Van Der Linde and P. Lammertse, "HapticMaster - A generic force controlled robot for human interaction," *Industrial Robot*, vol. 30, no. 6, pp. 515–524, 2003.

[26] R. Nayeem, "Role of Visual and Haptic Feedback in Complex Object Exploration," 2022. [Online]. Available: https://doi.org/10.7910/DVN/G2OZZE

[27] S. L. Hahn, *Hilbert transforms in signal processing*. Boston: Artech House, 1996.

[28] R. S. Razavian, S. Bazzi, R. Nayeem, M. Sadeghi, and D. Sternad, "Dynamic primitives and optimal feedback control for the manipulation of complex objects," *IEEE International Conference on Robotics and Automation (ICRA)*, pp. 7555–7562, 2021.

[29] H. Garnier, M. Mensler, and A. Richard, "Continuous-time model identification from sampled data: implementation issues and performance evaluation," *International Journal of Control*, vol. 76, no. 13, pp. 1337–1357, 2003.

[30] L. Ljung, "Experiments with identification of continuous time models," *IFAC Proceedings Volumes*, vol. 42, no. 10, pp. 1175–1180, 2009.

[31] P. Young and A. Jakeman, "Refined instrumental variable methods of recursive time-series analysis Part III. Extensions," *International Journal of Control*, vol. 31, no. 4, pp. 741–764, 1980.

[32] J. Brüderl and V. Ludwig, "Fixed-effects panel regression," in *The SAGE Handbook of Regression Analysis and Causal Inference*, H. Best and C. Wolf, Eds. London: Sage Publications, 2014, vol. 327, p. 357.

[33] R. I. C. Muchacho, R. Laha, L. F. Figueredo, and S. Haddadin, "A solution to slosh-free robot trajectory optimization," pp. 223–230, 2022.

[34] H. C. Mayer and R. Krechetnikov, "Walking with coffee: Why does it spill?" *Physical Review E*, vol. 85, no. 4, p. 046117, 2012.

[35] D. Sternad, H. Katsumata, M. Duarte, and S. Schaal, "Bouncing a ball: Tuning into dynamic stability," *Journal of Experimental Psychology: Human Perception and Performance*, vol. 27, no. 5, pp. 1163–1184, 2001.

[36] D. Sternad, M. Duarte, H. Katsumata, and S. Schaal, "Dynamics of a bouncing ball in human performance," *Physical Review E*, vol. 63, no. 1, p. 011902, 2000.

[37] J. Kenney, T. Buckley, and O. Brock, "Interactive segmentation for manipulation in unstructured environments," *IEEE International Conference on Robotics and Automation (ICRA)*, pp. 1377–1382, 2009.

[38] M. Krainin, B. Curless, and D. Fox, "Autonomous generation of complete 3d object models using next best view manipulation planning," *IEEE International Conference on Robotics and Automation (ICRA)*, pp. 5031–5037, 2011.

[39] S. Dragiev, M. Toussaint, and M. Gienger, "Gaussian process implicit surface for object estimation and grasping," *IEEE International Conference on Robotics and Automation (ICRA)*, pp. 2845–2850, 2011.

[40] J. Bohg, M. Johnson-Roberson, M. Björkman, and D. Kragic, "Strategies for multi-modal scene exploration," *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 4509–4515, 2010.

[41] M. Björkman, Y. Bekiroglu, V. Högman, and D. Kragic, "Enhancing visual perception of shape through tactile glances," *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 3180–3186, 2013.

[42] J. Ilonen, J. Bohg, and V. Kyrki, "Three-dimensional object reconstruction of symmetric objects by fusing visual and tactile sensing," *The International Journal of Robotics Research*, vol. 33, no. 2, pp. 321–341, 2014.

[43] P. K. Allen and R. Bajcsy, "Two sensors are better than one: example of integration of vision and touch," in *3rd International Foundation of Robotics Research (ISRR)*, France, 1985.