

An Information Pipeline Model of Human-Robot Interaction

Kevin Gold
Department of Computer Science
Wellesley College
Wellesley, MA 02481
kgold@wellesley.edu

ABSTRACT

This paper investigates the potential usefulness of viewing the system of human, robot, and environment as an “information pipeline” from environment to user and back again. Information theory provides tools for analyzing and maximizing the information rate of each stage of this pipeline, and could thus encompass several common HRI goals: “situational awareness” [6], which can be seen as maximizing the information content of the human’s model of the situation; efficient robotic control, which can be seen as finding a good codebook and high throughput for the Human-Robot channel; and artificial intelligence, which can be assessed by how much it reduces the traffic on all four channels. Analysis of the information content of the four channels suggests that human to robot communication tends to be the bottleneck, suggesting the need for greater onboard intelligence and a command interface that can adapt to the situation.

Categories and Subject Descriptors

H.1.1 [Models and Principles]: Systems and Information Theory; H.1.2 [Models and Principles]: User/Machine Systems; I.2.9 [Artificial Intelligence]: Robotics

General Terms

Theory, Human Factors

Keywords

conceptual/foundational, information theory, human-robot interaction

1. INTRODUCTION

The three commonalities that human-robot interaction (HRI) researchers can agree on at the moment is that they study humans, robots, and interaction. This last necessitates, and indeed consists of, the transfer of information between human and robot. The robot may communicate

with the human through a GUI on a monitor, natural language, a literal wink and a nod, or even a physical design that conveys certain affordances; the human may communicate through a mouse, keyboard, natural language, facial gestures, or even by simply being passively observed. In all cases, human-robot interaction consists of the transfer of information. This paper is meant to explore the idea of treating HRI as the study of a loop of information between environment, robot, and human (Figure 1). Though several subfields that contribute to HRI, including psychology, HCI, machine learning, and signal processing, have previously examined the usefulness of information theory in their respective domains (as shall be discussed in more detail below), this paper is, to the author’s knowledge, the first exploration of the usefulness of information theory for understanding HRI.

Information theory is often used to understand the digital representation of sensory information, particularly in quantifying how much information is lost in its digitization, compression, and processing. The actions of the robot on the environment can themselves be characterized as containing a certain amount of information; a robot’s action is like a message from the robot to the physical world, with inevitable noise resulting from imperfect control. These additional channels of information between the robot and a physical environment distinguish HRI from human computer interaction (HCI). Though HCI, psychology, machine learning, and signal processing have all at one point or another used information theory as a tool of analysis, they have never all contributed to the same model. It may make sense, then, to consider whether information theory can act as a lingua franca for discussing findings from these disparate fields when tackling HRI problems.

By framing human-robot interaction and robot-environment interaction as channels of information, HRI researchers might quantitatively answer questions such as: How much information is being lost on the way to the human from the environment? How efficiently and robustly can the human act on the environment through the robot? How much processing must the robot itself provide to make up the difference between what is commanded and what must be physically performed? How should the interface be designed to maximally reduce the possibility of error, given a rate of commands from the user?

The current paper is not meant to be a complete theory of HRI from an information theoretic perspective. Rather, it is an attempt to gather various information theory findings in one place, and ask whether something might be gained

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

HRI’09, March 11–13, 2009, La Jolla, California, USA.
Copyright 2009 ACM 978-1-60558-404-1/09/03 ...\$5.00.

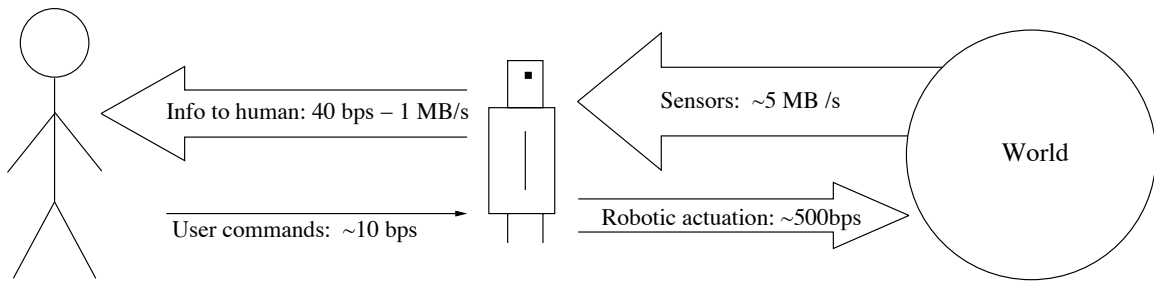


Figure 1: The information pipeline of HRI. Sensory throughput to the robot is typically higher than the bitrate of information conveyed to the human, and the bitrate of the human’s commands is smaller than the possibilities afforded by the robot. (Figures assume a color camera sensor, several degrees of freedom, and robot control via either speech, keyboard, mouse, or joystick.)

by considering these findings as parts of a whole. The analysis in this particular paper only goes as far as to consider which information channels tend to be larger than others, and the consequences of certain information channels being downstream from others. The hope is that other researchers might be interested in considering HRI from an information theoretic perspective, and do more with it.

The next section will analyze the information content of the four channels of the loop in more detail, and will provide examples of how much information can be transferred by various input devices, sensors, and human cognitive abilities. Section 3 will present an analysis of the overall flow of information when the four channels are combined. Section 4 will consider what artificial intelligence can contribute to each of the channels. Section 5 will mention some possible uses of information theory in HRI.

2. THE FOUR CHANNELS

Figure 1 shows how information flows in a typical human-robot interaction. There are four main channels to consider in which information might be lost, distorted, or transformed.

2.1 Environment to Robot

The first channel of information is from the environment to the robot. A robot’s sensors and sensory programs limit the amount of information that the robot can pull from the environment; they determine the resolution, field-of-view, and quantization of visual information, the sampling rate, frequency response, number of channels of auditory information, and so on.

In choosing the type and placement of sensors on a robot, it is useful to consider the mutual information between their signals. High mutual information indicates that there is redundancy between sensor inputs, and that the sensors might be better placed in a different configuration to maximize the informational throughput. On the other hand, high mutual information can be desirable if the purpose of multiple sensors is redundancy. In this case, one might calculate the expected information conveyed by the sensors, taking into account the possibility of sensor failure. It is also possible to weight the value of bits from different sensors, if they are thought to be of different value to the human operator.

Example: Camera placement. Suppose a small robot can be equipped with two color cameras, which can either be placed binocularly in the front of the robot, or placed

one in the front, one rear-facing in the back. Placing the second camera in the front grants depth perception, which grants dR extra bits of information, where d is the quantized per-pixel depth information and R is the resolution in pixels; the rest of the information is essentially mutual information with the first camera. On the other hand, placing the camera in the back grants essentially $3cR$ additional bits of information, where c is the per-color bit depth (probably a byte) with essentially no mutual information between the two cameras. In general, $c \approx d$ and a rear-facing camera conveys more information. However, if bits of depth information are considered to be w_d times more valuable than bits of rear-facing camera information, then the designer must decide whether $w_d dR > 3cR$. Finally, if there is a P_f probability that the other camera fails during the mission, then the expected utility of the information from the front-facing second camera is $(1 - P_f)w_d dR + P_f 3cR$ versus $3cR$ for the rear-facing camera.

The value of some channels of sensory information relative to others need not be completely ad hoc; it is possible in some cases to directly calculate the utility of information. A discussion of calculating the value of side information in decision-making can be found in [5].

2.2 Robot to Human

Every robot must interact with a human in some way, be it through a GUI, speech, or gestures. Each of these channels presents limits on the maximum information that can be conveyed: a screen cannot be too cluttered, speech can only proceed at a certain rate, and gestures typically either only convey a bit of information (on/off) or rely on reference to the environment to convey information.

Example: Information rate of speech. Comfortable listening speed for speech is 150-160 words per minute [32], though this can be increased to 210 words per minute with no loss in comprehension [22]. Using Shannon’s entropy estimate of 11.82 bits per word in English [29], the information rate for this channel is 41 bits per second for rapidly spoken English.

Notice that because it is typical for remotely controlled robots to display their sensory information directly to the user, the bits used to convey or store this information may be much greater than their entropy relative to the task. For example, a .WAV file containing three spoken words may be roughly 1 second long, which at 44.1 KHz and 16 bit sampling requires 705,600 bits; yet by Shannon’s estimate,

each word should only require 11.82 bits in the most efficient encoding. Thus it is useful here to distinguish the bits the robot uses to encode the environment from the human’s encoding, which may be different. Generating more bits from the sensors is not necessarily more helpful if it does not increase the robot’s or the human’s understanding of the environment.

To that end, we can distinguish between *representation length* and *task entropy*. The representation length is the number of bits the robot uses to represent the environment; for example, the robot may encode a visual scene using three bytes per color pixel. The *task entropy* is a measure of the length of the most efficient encoding the robot could have of its environment, given its task. The latter depends only on the probabilities of the high-level events involved; it can be calculated using the standard entropy equation, $H = \sum_i -p_i \log p_i$, which is the average length of the most efficient encoding of the desired information.

Example: Coffee cam bot. Suppose there is a robot equipped with a video camera that exists solely to check whether the coffee in the next room is ready. The robot checks with its camera whether the coffee is ready, then announces its decision from the next room. Though the camera may encode $640 \times 480 \times 3 \times 10fps = 9216000$ bytes of information per second, the task entropy is only at most $2 * -0.5 \log 0.5 = 1$ bit of information, namely whether the coffee is actually ready or not; this is the only bit (literally) of task relevant information that the robot pulls from the environment. If the coffee is more likely to be ready than not, the task entropy may be even smaller; for instance, a 75% chance of coffee would result in a $-0.75 \log 0.75 - 0.25 \log 0.25 = 0.81$ bit entropy, such that the robot could record a coffee history over n checks using only about $0.81n$ bits using an efficient code. Increasing the resolution of the robot’s camera would increase the representation length, but would not affect the task entropy; rather, the visual image can be seen as a very redundant code for what is essentially a 1-bit message.

Having a larger than theoretically optimal representation length can be useful, however, because it can reduce human perceptual error. Though a single pixel on the screen is theoretically capable of sending 256 different messages, in practice this would be an absurd method of conveying information to a human, and some redundancy in the display is necessary to make a workable system. However, the original sensory signal is not necessarily the most error-free representation of the data, and the classic method of sending multiple independent messages can reduce the chance of human error. In general, in choosing a representation, the distribution of human errors should be taken into account as well.

Example: Measure twice, cut once. Suppose a surveyor robot takes a measurement of the distance to a distant hill, and sends this message to a human operator by means of a GUI. By displaying this information as a number, the human error is distributed by an order of magnitude because of the possibility of omitting or inserting a digit, or moving a decimal point. A bar graph display, on the other hand, limits the human error to a small normal distribution about the true measurement. Displaying both representations decreases the overall probability of error to the product of the two error probabilities, and reduces the variance accordingly as well. Multiple measurements from nearby locations, or

further alternative display methods with independent probabilities of error, can drive the probability of human error down to negligible levels, just as redundancy works in the classic noisy 1-bit channel [5].

2.2.1 Limitations of human short term memory

It is tempting to use the fact that humans can remember about 7 digits in the short term [20] to calculate a human short term memory capacity of roughly $7 \times \log_2 10 = 23$ bits, and go on to reason that it should be possible to remember 23 boolean values in short term memory. This is not correct; as Miller’s classic article points out [20], the number of items that people can remember does not appear to vary much with the information content of each item. Letters, digits, words, and binary values all appear to be remembered with roughly the same ease [11]. It has been suggested that this is because short-term memory is actually dictated by a short-term “phonological loop” in which humans can remember about 2 seconds of audio, regardless of content [2]. (This is possibly an argument for robots to report on their environments using “telegraphic” speech that omits articles and makes use of abbreviations.)

What *is* limited by information content is the number of different perceptual stimuli that humans can remember in short term memory with sufficient accuracy to classify a new stimulus. For example, it is possible to remember the positions of 9-10 points on a line distinctly enough to classify a new point correctly from among these alternatives; Miller calculates the human channel capacity of points presented on a line to be 3.25 bits, corresponding to 9.5 possible positions [20]. Some other interesting human channel capacities cited by Miller include 2.5 bits for pitch, 2.3 bits for loudness, 2.6 bits for area, 3 bits for angle, 3.1 bits for hue, and 2.2 bits for curvature [20]. Multidimensional stimuli increase the overall channel capacity, but with diminishing returns; for example, an early study showed that varying six different acoustic variables such as frequency, loudness, and duration resulted in a channel capacity of 7.2 bits, or about 150 different distinguishable sounds [24], though linearly adding the capacities would have resulted in at least 12 bits.

Example: Reporting battery levels via pitch. A group of mobile robots could report their battery levels to a human operator using pitch, where a high tone indicates a full battery and a low tone indicates low battery. Recalling the battery reports just delivered from a group of robots, a human operator will be able to distinguish between $2^{2.5} \approx 5$ qualitatively different battery levels from a maximum of about 7 robots, for a total of $2.5 \times 7 = 17.5$ bits of information.

Example: A faltering robotic smile. At 2.2 bits of memory for curvature, a human can remember about 5 different curves of a humanoid robot’s mouth. Thus, a human interacting with a robot can remember roughly two different levels of smile, two different levels of frown, and a neutral position, but any further subtlety in the robot’s smile will be lost on the user.

2.2.2 Situational Awareness

In introducing the notion of “situational awareness,” Endsley points out that situational awareness is not increased by merely providing more information to the human, but by

increasing the human’s ability to form a model of what is going on and predict what will happen [6]. Such a view would seem at first to reject the idea that information alone can measure situational awareness, or that aiming to increase information is a desirable goal.

Appealing to the idea of a mental model suggests that it is perhaps not the information content of the information delivered to the human that should be measured, but the information content of the human’s mental model. This could in theory be measured by asking human users to describe or draw their mental models of the robot’s situation, then calculate the number of bits implied by the users’ levels of detail and accuracy. The ability to predict future events can also be measured quantitatively in some cases, and the extent to which the human is correct can suggest a kind of Kullback-Liebler distance [5] between the human’s model and the true distribution of events.

Such measures would probably make the choice of how to represent task entropy much easier – it could be operationally defined as the entropy of the human’s description of the state of the environment, which is likely to be described in terms of the task.

2.3 Human to Robot

When considering various forms of robot control, it can be useful to consider the rate at which the human can output information. If the input device conveys less information content than the robot’s actions, the robot may need to perform additional onboard processing; but more information content can command a steep learning curve for the human. On the other hand, with the rise in popularity of video games has come a very large increase in the information capacity of input device channels.

Example: Atari joystick to NES controller to WASD keys and mouse. Before the rise of the Nintendo NES in the mid 80’s, joysticks typically contained four digital sensors corresponding to the up, down, left, or right directions, so that moving a joystick in a diagonal direction activated at most two of these sensors. With a single fire button and at most 8 directions, these controllers could convey at most $\log 8 + 1 = 4$ bits per sample cycle. The NES kept the 4 digital directional sensors, but placed them under a directional pad to be operated by the left thumb, allowing the right thumb to operate two buttons that could be pressed simultaneously or one of two buttons in the center of the controller that for practical purposes had to be hit separately, for a total of $3 + \log(2^2 + 2) = 5.6$ bits per sample. Today, “first-person shooter” games assign the 4 directional buttons to the middle three fingers of the left hand operating the W, A, S, and D keys of the keyboard, leaving the pinky to hold Shift or Control for special actions, the thumb able to hit the space bar, the unused directional finger free to hit one of about 6 number keys for weapon changes, and the right hand free to manipulate the mouse for aiming and the two mouse buttons for firing, for a total of $3 + \log 3 + 1 + \log 6 + 2q + 2 = 10.2 + M$ bits per sample, where M is the per-sample information conveyed by mouse movement (see below).

Because processor speeds have far exceeded human reaction time, it makes more sense to calculate the number of samples using reaction time instead of the actual sampling rate of the input device to determine the true number of bits conveyed by the user. Reaction time is roughly 200 ms

for a basic response to a visual stimulus [33], but may be more to formulate a complex multi-button reaction of the kind described here.

Human factors play a large role in the information throughput of devices that require pointing, such as mice, trackballs, and analog joysticks; the rate at which targets can accurately be selected is based on their distance from each other, D , and their size W . Fitts’s law [7] is a mathematical model of pointing that can be used to calculate this throughput; ISO Standard 9241-9 [12] gives

$$Throughput = \frac{ID_e}{MT} \quad (1)$$

where

$$ID_e = \log_2\left(\frac{D}{W} + 1\right) \quad (2)$$

and MT is the movement time; the quantity is given in bits. One study reported a throughput of 4.9 bits per second (bps) for a mouse, 3.0 bps for a trackball, 1.8 bps for a joystick, and 2.9 bps for a touchpad [18]. These figures are much smaller than the limits imposed by the hardware because of the error introduced by reaching movements, which increases according to Fitts’s law with the size of the movement.

Spoken language is another possible means of delivering commands to a robot. A typical dictation rate is about 105 words per minute [17], which with no errors would result in about $1.75 \log V$ bits per second, where V is the vocabulary size. However, speech recognition can have a high error rate, and the actual throughput of information can drop by a factor of 5 when phrase repetition is taken into account [14].

Various groups have also experimented with adding recognition for prosody, or the pitch contours of speech, an information signal that people produce naturally to convey affect, mood, and relative interest [21]. According to the TOBI model of prosody [3], words can be stressed with either high (H^*) or low (L^*) tones, or can contain two-part changing tones that can be stressed either on the low or high tone (L^*+H versus $L+H^*$, etc.; asterisk indicates stress). Allowing a word to contain one of these six patterns, or no detectable stress at all, increases the information content by up to 2.8 bits per word, or 7 bits per second at a natural speaking rate. However, the more complex tones of TOBI are still controversial and difficult to detect automatically [15], and so the bit rate for prosody may be closer to 4 bits per second.

2.4 Robot to Environment

Finally, when acting on user commands, the robot performs some kind of physical action in the environment, which can be seen as either a path through physical space or a path through multidimensional joint space. The information content of this path is the information content of the action, which is determined in a large part by the number of degrees of freedom (DOF) of the robot.

Example: Analog servos. The throughput for an individual DC motor or servo varies, but an analog servo typically updates its position at 30Hz, while a digital servo updates its position at 300 Hz. Though servo angle can vary continuously, transit time to the target position effectively limits the bandwidth of the information conveyed by each servo motor. A typical transit time for an analog servo is

roughly 0.2 sec/60 degrees, meaning that an analog servo can only effectively update its position by about 10 degrees in either direction per signal. Since noise can result in servo positions being off by a few degrees, a quantization of the servo position might assign about three degrees to each position, resulting in 3 bits per signal or a maximum bitrate of around 90 bps.

For many robots, the number of DOF in the robot’s end-effectors is larger than the number of DOF in the user’s input device, or the number of DOF the user can effectively control. In this case, there are three options: the subspace of available actions can be reduced to match the DOF of the input device; more information can be acquired from the user over time by making the commands less than real-time; or the missing information can be supplied by the environment itself, with the robot’s onboard intelligence converting the sensory information into path information.

Example: Robonaut. Robonaut is a humanoid robot built for remote teleoperation in space, consisting of two seven degree of freedom arms, two five finger hands with a total of 12 joints, a two degree of freedom neck with multiple cameras, and a three degree of freedom waist [4]. Though Robonaut’s input devices include wearable gloves and position trackers for the arms, neck, and waist of the operator, in practice the human operator cannot effectively control such a high-information signal in real-time. As a result, all three strategies mentioned above are used to augment the information content of the human’s signal: primitives for hand motions reduce the space of possible motor actions of the robot’s hand; the human can issue verbal “freeze” and “thaw” commands to various joints so that the robot parts can be controlled in serial over time; and work continues on finding ways to extract the necessary information for motor tasks from the sensory and motor inputs [23].

3. THE INFORMATION PIPELINE

When a human-controlled robot must respond to an event in the environment, the information travels through a loop from environment to robot to human and back again (Figure 1). The information from the environment first passes to the robot through the sensors, then from the robot to the human via an interface. The human then reacts by sending some command (or no command at all, which is still informative) to the robot, and the robot performs some action on the environment. The final action of the robot is ultimately some noisy function of the event that prompted the human to act.

With this loop and the preceding analysis of the four channels in mind, there are a few observations that can be made.

3.1 The Human to Robot Bottleneck

If at any point in the loop, one of the channels has a limited throughput, then the information content of all the subsequent channels is limited as well. For example, if the robot cannot perceive more than 3MB/s of information from the environment, it cannot possibly pass on more information than this to the human. If the human’s commands have a throughput of no more than 4 bps, then the robot’s actions are similarly constrained to no more than $2^4 = 16$ possibilities for every second the user spends sending commands. Which channel is the usual bottleneck?

Modern robots meant for human control typically use video input, which in color at 320×240 resolution is about 230 KB

per frame, or on the order of 2-6 MB per second depending on frame rate. The user’s input devices possess a throughput that is magnitudes smaller, on the order of bits per second once user accuracy is taken into account. Thus, under almost all circumstances, the human’s commands will contain less information than this sensory input.

The robot to human channel lies somewhere between these, depending on how one wishes to quantify the task relevant information. The raw video data can be displayed on a GUI, but the genuinely new information that updates the human’s mental model of the situation is undoubtedly more compressible than this. However, the updates to the human’s mental model of the situation should generally contain more information than the human’s signal to the robot, because the human’s commands typically only manipulate the robot, which is only one element of the environment among many which the human must represent. Thus, even if the human possesses a highly compressed model of the situation, the human to robot commands will contain less information than the channel to the human.

Given that even a single analog servo can sample information at a rate much faster than most human input devices can accurately supply it (see above), it is probably fair to say that the transmission bottleneck will usually be in the Human to Robot channel. Though human control may appear to be able to control servos with a high degree of accuracy, such control generally requires feedback from the robot, which reduces the human’s throughput relative to the amount of information actually being provided.

This is not to say that the Human to Robot channel provides the least useful information, or that it is most in need of augmentation – but it is the speed bottleneck in the information transmission loop, and therefore most in need of well-chosen codes. These commands must somehow span the space of the robot’s actions using as few bits as possible. As a result, efficiency can be improved by using commands that access multiple motors and initiate sequences of actions that tend to co-occur, instead of relying on low-level commands that are essentially “uncompressed.”

3.2 Processing Cannot Increase Information

The data processing theorem [5] states that any deterministic function of a random variable cannot contain more information than the random variable itself. This has important implications for the HRI information pipeline, because the robot is essentially only performing deterministic functions on the input from the environment and the user, and the human user’s responses are presumably also more or less deterministic functions of the communications from the robot and the human’s goal. That is to say, the robot’s final action a is a function of the human’s command c and the robot’s sensor input s , while the human’s command is itself a function of the information i the human received, which in turn was a function of the robot’s sensor input, and the human’s goal g : $a = a(s, h(i(s), g))$. There are some interesting implications that follow from the fact that there are only two real random variables here, s and g , but several composed deterministic functions.

First, no amount of processing of the robot’s sensor signals can increase the information available to the human, beyond what was in the signals themselves. At best, the robot might change information from a modality that is difficult to process to one that is more lucid (for example, display-

ing range-finding information as a bar instead of a number), but no amount of “data mining” can present a pattern to the human that was not there to be seen in the first place.

Second, the amount of information in the signal from the environment strictly decreases as it passes through the loop. If the robot has access to the human’s goal ahead of time, or knows it from a previous communication, the human essentially contributes processing power and algorithms, but not new information.

Third, a robot that acts on its sensory information autonomously is inherently more expressive than one that is restricted to human control, assuming it already has access to the human’s goal. In terms of information content, $H(a(s, g)) \geq H(a(s, h(i(s), g)))$. Greater information content of the robot’s actions would mean that it would be able to use its native motor speeds and resolutions faster and more accurately, allowing it to follow paths of motion that would be impossible under human control.

4. THE ROLE OF ROBOTIC INTELLIGENCE

As we have seen, human attention and short term memory place a bottleneck on the amount of information that can be successfully transmitted to the human, while input devices and the speed of speech typically limit the rate at which information can be transmitted to the robot. Obviously, any interaction that can be handled autonomously by the robot, without consulting the human, reduces the reliance on these information bottlenecks. Thus, while Figure 1 may appear to assume a paradigm of teleoperation, it in fact is an argument against teleoperation and for increased autonomy, given the small throughput of the human-to-robot channel. However, even when a human is involved in the loop, robotic intelligence can serve to improve the effectiveness of all four information channels.

Maximizing environment-to-robot throughput: Exploration. The environment-to-robot channel typically pulls sensory signals from the environment at a steady rate; however, the information content of these signals can be quite small if the robot is not doing anything on its own. Video and other sensor signals from a motionless robot are highly correlated from one moment to the next, whereas a robot that is actively exploring its environment produces more information. In the presence of a task, information can be defined relative to the task’s goals (the aforementioned “task entropy”), and the robot can act to maximize information gain relative to the task. In the absence of a task, the robot can still act to increase its overall sensory information by exploring areas of the environment with high uncertainty. The use of information content to guide exploration is explored in [30].

Maximizing robot-to-human throughput: Finding efficient representations of sensor information. To the extent that the robot can combine sensory information over time and find interesting patterns, it can reduce the attentional and short term memory load on the human. Examples include creating mosaics of the environment from individual video frames [27], Simultaneous Localization and Mapping [30], and identifying and storing key frames of human actions in the video sequence [1].

Maximizing human-to-robot throughput: Learning and offering efficient commands. Commands can be delivered much more quickly when they are delivered at a high level of abstraction; “get the chair and bring it here” is

much more succinct than specifying the sequence of movements necessary to carry out this plan in detail. To that end, robotic intelligence can make human-robot communication more efficient by providing methods for on-the-fly vocabulary acquisition [10], learning to imitate actions [28], and interpreting natural language [13]. Another possibility for non-verbal interfaces is for the robot to offer high-level commands based on predictions of the user’s intentions – for instance, offering a limited-time hotkey to zoom in on a particular object that the robot judges to be novel.

Maximizing robot-to-environment throughput: Motor control. The traditional scope of robotics still applies to human-robot interaction, and efficient path planning, inverse kinematics, and control are necessary to handle the extra degrees of freedom beyond what the human operator can efficiently control.

5. WHAT MIGHT INFORMATION THEORY GET US?

So far, a case has been presented for the idea that information theory can describe many aspects of HRI and the various roles of AI in HRI, but it has not yet been shown that information theory itself is useful. Here are some ways in which the information theory framework may prove to be useful.

Metrics. Calculating the information throughput of a particular sensor configuration, algorithm, or interface gives a way of evaluating different methods or algorithms before moving to the human studies phase.

Theoretical bounds. Calculating the theoretical capacity of a particular channel can inspire more creative ways to make full use of it.

Finding the bottleneck. When building a complete HRI system, finding the channel with the smallest information rate can help focus on the weak points of a system, while avoiding overengineering of others. High resolution sensors and extra DOF on a robot may go to waste if the human in the equation cannot take advantage of them.

Applicability of AI results. Several popular artificial intelligence algorithms, such as decision trees [26] and expectation maximization [19], operate by calculating the representation length (or equivalently, the posterior probability) of the data under different models of the world, and choosing the model that minimizes representation length (maximizes posterior probability). Being explicit about the relevant information in a robotics problem makes it easier for AI specialists to apply these abstract techniques to real-world robotics problems, and analyze the effects of choices elsewhere in the pipeline.

Mathematical theorem mining. Information theory has been around longer than HRI, and has a much deeper trove of theorems than can be explored here.

6. CONCLUSIONS

This paper has argued that many problems of HRI can be viewed as problems of sending efficient and error-resistant signals between environment, robot, and human. Information theory provides a way to quantify how changes to a robot’s configuration, algorithms, or interface can affect this flow of information. For designers of whole systems, thinking about HRI as an information pipeline can focus design attention on the parts of the system where bottlenecks arise,

and allow informed decisions about which methods to use in combination with each other. For designers of particular parts of the HRI interaction, information theory provides a way of quantifying the success of one’s own piece of the puzzle, and communicating that success to people in other subfields. For the AI researcher, the information pipeline model encourages the application of information theoretic machine learning techniques on not just the data from the environment, but all four channels of information. For the theoretician, information theory opens up HRI as a fertile ground for the application of theorems hitherto unknown to HRI researchers, with the potential for surprising empirical predictions.

For each channel in Figure 1, quantitative results can be obtained about the total information throughput, and these suggest ways to improve the interaction. “Situational awareness” can be increased by increasing the information obtained from the sensors, either by adding more sensors or by reducing mutual information between existing sensors, but also by finding concise and task-relevant representations of the data, then communicating these in a manner that takes into account human perceptual error and short term memory constraints. Information throughput from the human to the robot can be increased by choosing interfaces with high information throughput, and using AI techniques to generate efficient, high-level commands that are specific to the situation. Information throughput of the robot to environment channel can be maximized by the application of AI and vision techniques that use information from the environment to expand high-level human commands into complex activities in the real world, allowing the use of more DOF than would be possible if the human had to control the robot directly. Artificial intelligence, which subsumes machine vision, machine learning, control, and other subfields, can play a role in increasing the amount of task-relevant information that passes over each channel while reducing the overall load on the human-robot channels.

The success of this formulation of HRI will depend on whether task entropy can be quantified and agreed upon for particular tasks. Increasing information without regard to task is unlikely to be useful; random numbers, for instance, possess a large amount of information, but randomly injecting them into communication will not achieve anything. Instead, information flowing to the human should be quantified relative to the events of interest to the operator – the locations of obstacles and goals, the state of the robot, and so on – and it is these high-level events, rather than the low-level sensory information, that should be the focus of information theoretic analysis. Nor is this to say that robots should only communicate to their users in abstract paraphrases of the situation. A direct video feed is a useful way of including the redundancy necessary for the robot-human channel to be error resistant. But it is just one way of introducing such redundancy, and it may not be the best way. Separation of source and channel is potentially a useful way of thinking about the distinction between the information the robot is conveying to the human, versus the means by which it is presented.

This formulation of HRI does not say much about important work on the psychological perception of robots [9, 25, 31], nor anthropological studies of how robots are used in the home [8]. It is therefore not a complete description of HRI, nor is it intended to be. However, these studies can inform

the information pipeline model by revealing what information naive users may be getting from or giving to the robot regardless of the designer’s intentions. For example, prosody may exist as a side channel in natural language interactions regardless of whether the robot designer makes use of it. A field study may reveal that operators are actually using the sound of the robot’s motors to make certain decisions, a side channel of information that may make some GUI elements redundant. An impression of a robot as dangerous or unapproachable may kill certain channels of communication. Theoretical analysis cannot replace field studies, but it can make use of them and, in some cases, motivate or justify them.

One weakness of an information theoretic model is that the analysis gets somewhat murky on the human end of the pipeline. Several psychological effects, such as the processing cost of goal switching, the improved performance that results from practice, the results of comfort and stress, and even the varying levels of situational awareness produced by different systems, are not quantitatively understood well enough to make theoretical predictions about how they impact information throughput. However, information theory does provide another means of measuring these effects, by focusing on the very quantifiable information throughput of the human rather than the debatable mental models that produce it. In some cases, information throughput may be more relevant than simple response time, since response time cannot be fairly compared across input and output methods; in other cases, it may be more accurate than Likert scales, since human subjective judgments can be very unreliable when made about the subject’s own performance [16].

Several subfields of HRI have at one point had a love affair with information theory: psychology, machine learning, HCI. In each case, the fields at some point moved on to a framework that was more specific to their field. But it may be worth reconsidering these older viewpoints when analyzing how different parts of the HRI framework interact, since information theory and HRI are both, at their cores, about effective communication.

Acknowledgments

This work was generously supported by the Norma Wilentz Hess Faculty and Program Fund in Computer Science at Wellesley College.

7. REFERENCES

- [1] D. Ayers and M. Shah. Monitoring human behavior from video taken in an office environment. *Image and Vision Computing*, 19(12), 2001.
- [2] A. D. Baddeley and G. Hitch. Working memory. In G. A. Bower, editor, *The psychology of learning and motivation: advances in research and theory*, pages 47–89. Academic Press, New York, 1974.
- [3] M. E. Beckman, J. Hirschberg, and S. Shattuck-Hufnagel. The original tobi system and the evolution of the tobi framework. In S.-A. Jun, editor, *Prosodic Typology: The Phonology of Intonation and Phrasing*, chapter 2. Oxford University Press, Oxford, England, 2005.
- [4] W. Bluethmann, R. Ambrose, M. Diftler, S. Askew, E. Huber, M. Goza, F. Rehnmark, C. Lovchik, and D. Magruder. Robonaut: A robot designed to work

- with humans in space. *Autonomous robots*, 14(2-3):179-197, 2003.
- [5] T. M. Cover and J. A. Thomas. *Elements of Information Theory*. Wiley-Interscience, 2nd edition, 2006.
- [6] M. R. Endsley. Design and evaluation for situation awareness enhancement. In *Proceedings of the Human Factors Society 32nd Annual Meeting*, Santa Monica, CA, 1988. Human Factors Society.
- [7] P. M. Fitts. The information capacity of the human motor system in controlling the amplitude of movement. *Journal of Experimental Psychology*, 47(6):381-391, 1954.
- [8] J. Forlizzi and C. DiSalvo. Service robots in the domestic environment: A study of the roomba vacuum in the home. In *Proceedings of the 2006 ACM Conference on Human-Robot Interaction*, pages 258-265, Salt Lake City, UT, 2006. ACM Press.
- [9] R. Gockley, J. Forlizzi, and R. Simmons. Interactions with a moody robot. In *Proceedings of the 2006 ACM Conference on Human-Robot Interaction*, pages 186-193, Salt Lake City, UT, 2006. ACM Press.
- [10] K. Gold, M. Doniec, and B. Scassellati. Learning grounded semantics with word trees: Prepositions and pronouns. In *Proceedings of the 6th International Conference on Development and Learning*, London, UK, 2007.
- [11] J. Hayes. Memory span for several vocabularies as a function of vocabulary size. Technical report, MIT Acoustics Laboratory, 1952. Cited in (Miller 1956).
- [12] ISO. *ISO 9241-9: Ergonomics of Human System Interaction Part 9: Requirements for non-keyboard input devices*, 2000.
- [13] D. Jurafsky and J. H. Martin. *Speech and Natural Language Processing*. Prentice Hall, Upper Saddle River, NJ, 2000.
- [14] C. Karat, C. Halverson, D. Horn, and J. Karat. Patterns of entry and correction in large vocabulary continuous speech recognition systems. In *Proceedings of the ACM Conference on Human Factors in Computing Systems (CHI)*, pages 568-575, New York, 1999. ACM.
- [15] E. S. Kim, K. Gold, and B. Scassellati. What prosody tells infants to believe. In *Proceedings of the International Conference on Development and Learning*. IEEE/Cognitive Science Society, 2008.
- [16] J. Kruger and D. Dunning. Unskilled and unaware of it: How difficulties in recognizing one's own incompetence lead to inflated self-assessments. *Journal of Personality and Social Psychology*, 77(6):1121-1134, 1999.
- [17] J. R. Lewis. Effect of error correction strategy on speech dictation throughput. In *Proceedings of the Human Factors and Ergonomics Society*, Santa Monica, 1999. HFES.
- [18] I. S. MacKenzie, T. Kauppinen, and M. Silberberg. Accuracy measures for evaluating computer pointing devices. In *Proceedings of the ACM Conference on Human Factors in Computing Systems (CHI)*, pages 9-16, New York, 2001. ACM.
- [19] G. J. McLachlan and T. Krishnan. *The EM algorithm and extensions*. Wiley, New York, 1997.
- [20] G. A. Miller. The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological Review*, 63:81-97, 1956.
- [21] H. Mixdorf. Speech technology, TOBI and making sense of prosody. In *Speech Prosody*, volume 3, pages 31-38, 2002.
- [22] N. Omoigui, L. He, A. Gupta, J. Grudin, and E. Sanocki. Time-compression: Systems concerns, usage, and benefits. In *Proceedings of the ACM Conference on Human Factors in Computing Systems (CHI)*, pages 136-143, New York, 1999. ACM.
- [23] R. A. Peters, R. E. Bodenheimer, and O. C. Jenkins. Sensory-motor manifold structure induced by task outcome: Experiments with Robonaut. In *Proceedings of the IEEE-RAS International Conference on Humanoid Robots*, pages 484-489. IEEE, 2006.
- [24] I. Pollack and L. Ficks. Information of elementary multidimensional auditory displays. *J. Acoustic Society of America*, 26:155-158, 1954. Cited in (Miller 1956).
- [25] A. Powers and S. Kiesler. The advisor robot: Tracing people's mental model from a robot's physical attributes. In *Proceedings of the 2006 ACM Conference on Human-Robot Interaction*, pages 218-225, Salt Lake City, UT, 2006. ACM Press.
- [26] J. R. Quinlan. Induction of decision trees. *Machine Learning*, 1:81-106, 1986.
- [27] H. S. Sawhney, S. Hsu, and R. Kumar. Robust video mosaicing through topology inference and local to global alignment. In *Computer Vision - ECCV '98*. Springer, Berlin, 1998.
- [28] S. Schaal. Is imitation learning the route to humanoid robots? *Trends in Cognitive Sciences*, 3(6), 1999.
- [29] C. E. Shannon. Prediction and entropy of printed english. *Bell Systems Technical Journal*, 30:50-64, 1951.
- [30] S. Thrun, W. Burgard, and D. Fox. *Probabilistic Robotics*. MIT Press, Cambridge, MA, 2006.
- [31] C. Torrey, A. Powers, M. Marge, S. R. Fussell, and S. Kiesler. Effects of adaptive robot dialogue on information exchange and social relations. In *Proceedings of the 2006 ACM Conference on Human-Robot Interaction*, pages 126-133, Salt Lake City, UT, 2006. ACM Press.
- [32] J. R. Williams. Guidelines for the use of multimedia in instruction. In *Proceedings of the Human Factors and Ergonomics Society 42nd Annual Meeting*, pages 1447-1451, 1998.
- [33] R. Woodworth. *Experimental Psychology*. Holt and Co., New York, revised edition, 1954.