In Search of a Natural Gesture

While computing has advanced exponentially, almost explosively, since the 1970s, input devices have only just begun to change. Why?

By Johnny Chung Lee

DOI: 10.1145/1764848.1764853

In many articles discussing the future of computing, you are very likely to find either a reference to, or a motivational assumption based on, a continued projection of Moore's law. This article will make no attempt to deviate from that steady tradition.

The regularity of references to "the law" probably extends from the fact that human behavior and clever uses for technology are notoriously difficult to predict, but the technology itself has demonstrated an unwavering trend over the past four decades. This trend is, of course, enabled only by the astonishing accomplishments by the engineering teams within the companies that manufacture computing equipment. Nevertheless, it doesn't take a huge amount of clairvoyance or risk-taking to claim that the trend will extend a bit further.

However, interface technology has not enjoyed the seven orders of magnitude in improvement of performance that core processors have achieved since 1970. In fact, aside from a slightly improved mechanical construction and visual polish, the input and output devices connected to the average desktop computer today are virtually identical to the ones used by Douglas Engelbart in his 1968 presentation, later referred to as "The Mother of All Demos." While there have certainly been several improvements along the way, such as the graphical user interface, trackpads, flat-panel displays, and touch screens, we still fundamentally operate our computers using a single 2D pointing device and a keyboard.

Yet in just the past two or three years, it is not too difficult to find articles proclaiming the "death" of the mouse and keyboard, or finding new product announcements promoting new methods of input and output as its key selling feature. Why has there been a recent spike in enthusiasm for new interface technology? In my opinion, it is because we've recently crossed the inflection point—an inflection point driven by Moore's Law and the limited growth of human attention.

CONSUMPTION-PRODUCTION IMBALANCE

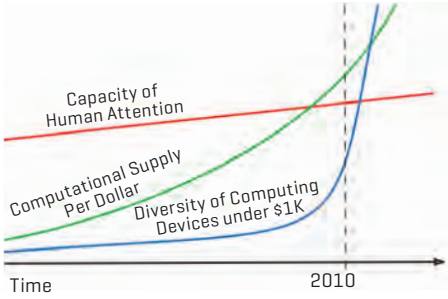
Over the past hundred years, the cognitive capacity for an individual to consume and produce information has stayed relatively constant or has increased only modestly. While technology has certainly made it much easier to saturate our input and output channels, the rate at which we can read, write, speak, listen, the density of pixels our visual system can resolve, the number of images we can meaningfully process per second, and the size of our fingers has not significantly changed. Yet the capacity for technology to supply us with content has grown in step with Moore's Law.

In the 1980s and 1990s, the consumer appetite for faster technology could not be satiated. But in recent years, the information supply has started to fundamentally surpass the ability of many people to absorb it. This creates a situation of computational surplus, which, economically, should dictate a substantial drop in cost.

Just a few years ago, a $100 laptop was considered a magnificent dream that would change the world. Now, it is possible to find a reasonably capable netbook for $100 on a good coupon day or even "free" with a network service contract. While it would have certainly been possible to manufacture $100 worth of computation in 1990, very few people would have found it satisfactory. That's not the case today. The average consumer's demand for more powerful technology has simply not kept up with the exponentially increasing supply.

Some have referred to this stall in performance demand as the era of "good enough computing." While "good enough" might suggest even further reduction in device cost, what's happening instead is it's becoming economically sensible to manufacture a wider variety of increasingly special purpose computers rather than expensive general purpose machines. For the price of a nice dinner, consumers can buy a computer that only plays music, only takes pictures, only shows maps, only plays games, only plays movies, or only lets you read the news. It's likely

Figure 1: The rise of diversification was a result of a computational surplus.



that we'll see a significant rise in the release of new form factors and targeted niche computing compared to what we have in the past. (**See Figure 1.**)

How is this relevant to interface technology?

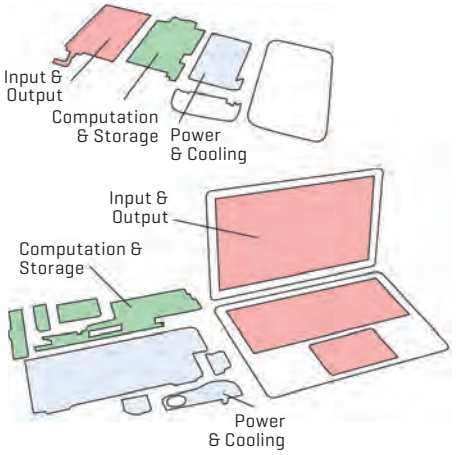
TASK-SPECIFIC DEVICES

As the diversity and specialization of devices increases, so does the diversity of interface technology. The best interfaces are typically task-specific. For example, the ideal interface for choosing a radio station while driving your car is not the same as the best interface for checking your email while sitting at your desk. As the distribution of computing shifts away from usage scenarios where a mouse and keyboard are acceptable, so does the adoption of alternative input methods, whether touch, motion control, location, gesture, voice, or some other physiological source.

If you look at the components within a modern laptop or mobile phone, you'll notice that there's actually very little "computer" in a computer today (**Figure 2**). The largest internal components of a modern laptop are already those dedicated to human input and output. As the physical space required for computation continues to fall or is even replaced with a high-speed network connection, the defining feature of the device and its suitable applications is the interface technology.

As a result, there is a very high demand in exploring novel ways of interacting with technology that permits alternative form factors, increases our capacity to express an idea, or improves our ability to absorb information. Computing will be defined by how we interact with the information rather than by the chipsets on the motherboard, or the operating system it runs. The quest to create new devices dedicated to solving each of our own specialized unsatisfied desires is largely lead by the search for better interface technology that can better understand what we want, when we want it, where we want it, in the way we want it.

Figure 2: The human interface hardware now dominates the form factor of many modern computing devices.



"Natural interaction is achieved through clever designs that constrain the problem in ways that are transparent to the user."

IN SEARCH OF NATURE

A phrase that has slowly received increasing traction, at least in the commercial exploration of alternative input technologies, is "natural user interface" (NUI). While there's no widespread consensus about the exact definition, NUI generally refers to an interface that is highly intuitive and effectively becomes invisible to the user when performing a task. It is an interface that can easily and efficiently transmit an idea from a user's mind into an action on the computer with little additional effort.

Don Norman described the philosophy well when he said, "The real problem with the interface is that it is an interface. Interfaces get in the way. I don't want to focus my energies on an interface. I want to focus on the job."

Unfortunately, the term NUI has also been coarsely applied to refer to anything that is not a typical keyboard and mouse. It's important to acknowledge that the philosophy behind a natural user interface is not conceptually incompatible with a mouse and keyboard. However, it has become much more popular to use the term "natural" when referring to multi-touch interaction, motion sensing, gesture input, and speech recognition.

These input techniques certainly offer a higher potential for expressing an idea to a computer with less distortion and rigid structure typically required by a mouse and keyboard. However, gesture and speech interfaces, in particular, have resonated well with the popular imagination. The allure of these methods of input is that they provide a glimpse into an easy-to-imagine vision of one day being able to communicate with a computer as easily and fluidly as we communicate with another human being using these skills we practice every day.

Now, it's debatable whether communicating with a computer in the same manner that we communicate with other humans is truly the most desirable interface for all tasks. To get a reasonably accurate picture of what a voice-and-gesture-only system might be like, imagine if the only input control to your computer were a video chat to a high school student sitting in a distant room, and all you could do is describe what you wanted. After a few minutes of saying, "Click on that. Move that ... no, not that one. The other window. I mean the browser window. Yeah. Make that bigger, I mean maximize it,"you will probably say, "where is my mouse and keyboard?"

An often unspecified detail of that vision is that the computer should be the embodiment of an exceptionally competent and omniscient human being with a reasonable personality that makes no recognition errors. But, for the sake of this article, I will concede there are components of that vision that are conceptually desirable enhancements to existing interface technologies and discuss some of its advantages and disadvantages. In particular, this article will discuss the gesture component in greater detail.

BODY MOVING

Body movements used to convey information from one person to another have been shown to be tightly coupled with simultaneous speech or co-verbal commands. According to researcher

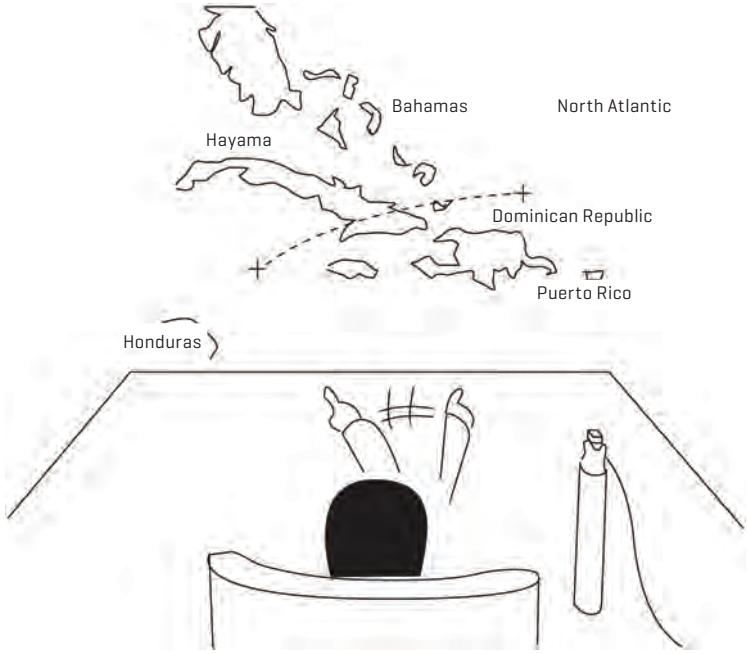
David McNeill's 1992 study, 90 percent of these types of communicative gestures are found to be associated with spoken language. Additionally, gestures often identify underlying reasoning processes that the speaker did not or could not articulate providing a complementary data source for interpreting a set of utterances.

Thus, gesture and speech go hand-in-hand in daily human-to-human communication, and it would be appropriate for any interactive system that attempts to provide a similar level of fluidity to be designed with that in mind.

Such systems that combine more than one mode of input are often called multimodal interfaces. A well known example demonstrating the power of combining speech and gesture is the Put-That-There system created by Richard A. Bolt in 1980 shown in **Figure 3**. This system allowed the operator to sit comfortably in a chair, point his arm at a distant location on a large display wall and issue verbal commands such as "move that" and then pointing at a different location, continue the command "... there."

The gesture provided object focus and task parameters, and the speech component provided the task selection and event execution. These modalities complement each other's strengths, combining the physical specificity of pointing with the random access temporal nature of speech.

Figure 3: Richard Bolt's Put-That-There system combined speech and gesture input.



GESTURE CHALLENGES

While a number of prototype systems that used gesture alone have been demonstrated to be reasonably functional, many of these systems typically relied on a unique set of hand or body poses that must be learned by the user, which trigger a small number of pre-assigned actions.

In 1986, Jean-Luc Nespoulous identified three classes of communicative gestures that have come into common use: mimetic, deictic, and arbitrary. Mimetic gestures are motions that are intended to be representative of an object's shape or behavior. For example, indicating the shape of a person's beard, the size of a box, or the act of tumbling. Deictic gestures are used to provide context or explanatory information such as pointing at the object of conversation, or indicating the direction for an action to be taken. Arbitrary gestures are learned motions typically used in specific communication settings, such as the hand signals used in airplane guidance, baseball pitches, or infantry coordination.

In the context of a gestural interface prototype, arbitrary gestures are highly popular choices in research systems because they can be easily designed to be distinctive for the sake of recognition and segmentation. But, these gesture sets tend to require significant user training, and they map to a rigid set of commands.

In general, performing complex communicative or manipulation tasks using free-air gestures alone without tactile feedback or co-verbal commands is actually quite unnatural. However, there may be opportunities to take advantage of the expressive power of deictic and mimetic gestures to augment or supplement interaction tasks because users will have a tendency to produce these gestures without additional prompting or training. Unfortunately, these gestures are not typically easy to segment and are subject to high variability between individuals.

In 1989, Alex G. Hauptmann attempted to study the degree of consistency and variability in unprompted gestures when users attempted to perform a three-dimensional spatial manipulation task. The users were asked to try to perform a translation, rotation, or scaling operation on a virtual wireframe cube rendered on a computer display. Upon completion, a human operator observing the hand gesture would attempt to simulate the resulting output. While there were coarse similarities in the type of gesture performed for each of the three tasks, individuals varied significantly in the number of fingers used, the position and orientation of their hands, the number of hands used, and the alignment of the hand movements.

Hauptmann made no attempt to make the computing system recognize and track these gestures as real interactive controls, which weakens the validity of certain conclusions as interactive feedback would certainly impact user behavior. However, the findings do indicate that a fully functional system would have to accommodate a high degree of variability between users.

Bolt attempted a partial implementation in 1992, but this system only pro-

vided constrained object rotations and single-handed object placement in a limited application scenario using two six-degree-of-freedom-tracked gloves with articulated fingers and heavily relied on co-verbal input for action selection and control. Furthermore, the variations Hauptmann observed occurred in the constrained scenario where people were seated within 1 meter of a computer display and were prompted to perform a simple spatial operation on a single object. As the assumptions are pulled back on this problem, the opportunity for variation goes up exponentially, such as allowing multiple objects on the screen simultaneously, using non-spatial actions such as changing the object color, varying the seated posture relative to the screen or standing at different distances, allowing multiple users to attempt simultaneous control, and even choosing to perform other peripheral tasks within the tracking volume without activating the system. Variations in body shape, size, and cultural background only exacerbate the difficultly in interpreting a given gesture, or finding a common gesture for a given desired action.

The complexity of a gesture recognition system is roughly proportional to the complexity of the input vocabulary. For example, if all that is desired is either motion or non-motion, there are a variety of sensors such as accelerometers that can provide a simple data stream that is relatively easy to examine to obtain this signal. In the case of an accelerometer, other data such as the magnitude and direction of motion, or the static orientation relative to gravity are moderately easy to extract. However, as the desired expressiveness of the input system goes up, so must the complexity of the gesture system.

In effect, the device must have an understanding of the world that is not only capable of distinguishing the target input set, but all other similar gestures, in order to recognize that they are not part of the input set. Otherwise, the number of false positives may be unacceptable. If the goal is to recognize a "jump," simply looking for vertical movement would be insufficient if a "squat" should not be considered a "jump." Should sitting down and then standing up be considered a jump? What about walking? Is a one legged-kick a jump? What about individuals who jump at different heights?

"Computing will be defined by how we interact with the information rather than by the chipsets on the motherboard, or the OS."

GUIDING GESTURES

In this respect, freeform gesture recognition shares many of the difficulties of unstructured speech recognition. Many spoken words have very similar acoustic properties. People speak with different accents and dialects. There are multiple ways of expressing the same thought. Recognizing isolated words without additional context information is generally unreliable. Recognizing speech in the presence of other noise can significantly reduce accuracy. Identifying when the user wants to engage and disengage with the system can be challenging. Without the speech equivalent of push-to-talk, prompted input, or escape keywords, gesture interaction suffers from the "Midas touch" problem of excessive accidental or false activations.

Alternatively, naively applying rigid structure and increasing recognition requirements would negate any potential benefit from user intuition and simply replace it with frustration from excessive false negatives. Understanding the strengths and weakness of a particular input method is fundamental to understanding what combination of tools will make for a successful user experience. The design should provide just enough guidance using other techniques to prevent the user from falling into the poor performing areas of gesture and speech recognition. If done correctly, a relatively small amount of recognition work can provide a delightful experience giving the illusion that the technology has merely understood their intention.

For guidance toward solutions to this problem, it's helpful to revisit the philosophies behind "direct manipulation" that made the graphical user interface successful, as described by Ben Shneiderman in 1983. One of the tenants of the design was that displaying immediate visual feedback for input is essential. This is the communicative common ground between yourself and the device that indicates it has an understanding of what you are doing, and that you understand what it is doing in response to your actions. The number of degrees of freedom that can be captured in the feedback should be as high as reasonably possible, giving users the information they need to quickly adjust their input to accomplish the desired output.

Interactive objects should have a visual representation with understandable metaphors. The visual appearance of an interactive object should provide some affordance as to the actions or gestures to which it can respond. The interface should clearly provide rapidly accessible, complimentary, and reversible commands.

NATURAL IS IN THE DESIGN

Regardless of the technology being used, a good interface experience is one that is able to capture the intent of a user's behavior with as little distortion as possible. While gesture and speech technologies offer greater potential for us to express our thoughts and ideas without thinking about the constraints of the interface, accurately reconstructing those ideas within the computer does not come from technology for free. Natural interaction is achieved through clever designs that constrain the problem in ways that are transparent to the user but fall within the capabilities of technology. A good user experience is achieved only through the hard work of individuals with domain expertise and empathy for those who are not like themselves.

Biography

Johnny Chung Lee is a researcher in Microsoft's Applied Sciences Group exploring interface novel software and hardware technology. In 2008, he graduated from Carnegie Mellon University with a PhD in human-computer interaction and was named into *MIT Technology Review*'s TR35. He is known for his video demonstrations of alternative applications for the Nintendo Wii remote that have received more than 10 million views.