# Duo: A Human/Wearable Hybrid for Learning About Common Manipulable Objects

Charles C. Kemp

MIT AI Lab, RM 935, 200 Technology Sq.
Cambridge, MA 02139
cckemp@ai.mit.edu
http://www.ai.mit.edu/projects/humans-as-robots/

**Abstract.** Humanoid robots would benefit from a better understanding of common manipulable objects and the human behaviors associated with them. Duo is a human/wearable hybrid that is designed to learn about this important domain of human intelligence by interacting with natural manipulable objects in unconstrained environments. Duo's wearable AI system measures the kinematic configuration of the human's head, torso and dominant arm, while watching the workspace of the human's hand through a head-mounted camera. Duo also requests helpful actions from the human through speech via headphones. This paper presents results on an initial set of behaviors for Duo which lead to high-quality segmentations of common manipulable objects in unconstrained human environments. In Duo, the wearable AI system essentially subsumes the abilities of its cooperative human partner by sharing the human's sensory input and directing a portion of the human's actions. Together, the cooperative human and the wearable AI system can be thought of as constituting a new kind of humanoid robot that complements more traditional, wholly synthetic humanoid robots by allowing researchers to circumvent some of the currently unsolved problems in the field, from dextrous object manipulation to unrestricted mobility.

## 1 Introduction

A great challenge in AI systems is the acquisition and use of a common sense understanding of the world [23]. As recognized in AI, and machine vision in particular, objects are a very powerful abstraction and serve as a useful level at which to represent common sense. Manipulable objects with which people regularly interact are an important class of objects for human intelligence, since they are used often and by many people. In order to achieve the long-term goals of artificial human intelligence, researchers must find ways to endow machines with common sense about objects and the ways in which people use them.

Humans acquire common sense about everyday objects through a lifetime of experience. Humanoid robots could serve as a direct approach to the acquisition of this type of competence, since a sufficiently sophisticated humanoid robot would be able to experience much of the world in the same way as humans.

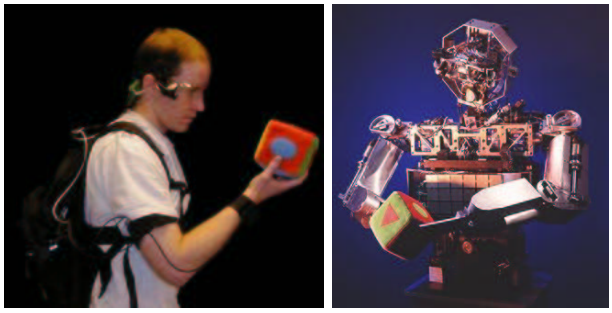

**Fig. 1.** Cog, a traditional, wholly synthetic humanoid robot is shown on the right. Duo, a human/wearable hybrid is shown on the left. The backpack holds batteries and a laptop that communicates wirelessly with a computer cluster. A glasses-mounted firewire camera captures video from the person's perspective. Kinematic measurements of the head and the dominant arm are performed by four Intersense devices with one worn on the head, one on the wrist, another on the upper arm, and the fourth on the torso.

Currently, however, humanoid robots have very limited experience with common manipulable objects in unconstrained environments due to obstacles ranging from mechanical design to social constraints on the use of autonomous robots.

A wearable system could serve as a good platform by which to learn about everyday objects. Duo is a platform that combines a wearable AI system with a cooperative human (see Figure 1) [17]. The wearable system captures video of objects as they are manipulated by the human, while simultaneously monitoring the kinematic configuration of the human's head, torso, and dominant arm (see Figure 2). Duo's name emphasizes that the system is composed of both the wearable AI system and a cooperative human working together as a unified entity. The wearable AI system both passively and actively observes the manipulation of objects in natural, unconstrained environments. When the wearable system is passively monitoring activity, the human contributes to the relationship by allowing the wearable system to very closely observe his activities. When the wearable system actively asserts itself, the human serves as an intelligent mechanical and computational infrastructure for the wearable system to control. The wearable AI controls this system in such a way as to make some problems much easier, such as visually segmenting a foreground object from the background.

In Duo, the wearable AI system essentially subsumes its human partner's abilities by sharing the human's sensory input and directing a portion of the human's actions. Together, the human and the wearable AI system constitute a complete platform whose control is well modeled as a subsumption architecture, with the wearable system as the top-layer of control (see Figure 3) [7]. We can helpfully think of this combined platform as a type of humanoid robot. From here on, this paper makes use of this perspective and considers human/wearable hybrids to be a new type of humanoid robot. In these terms, Duo is a humanoid

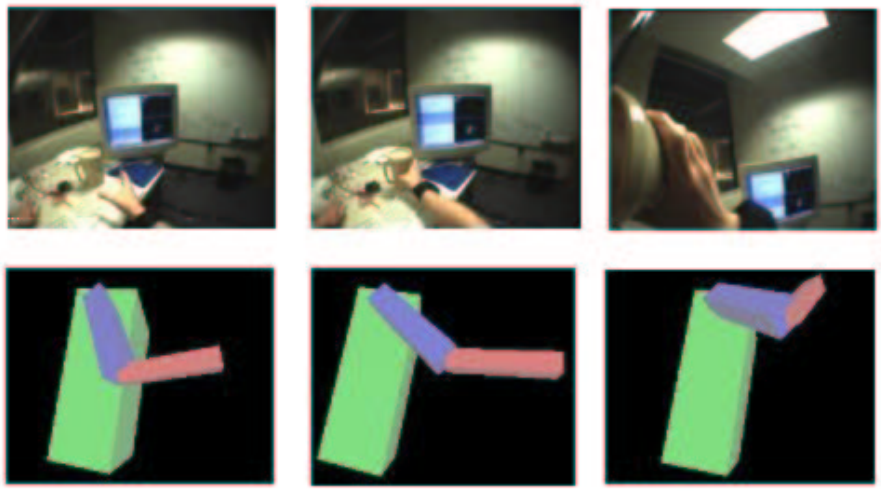

**Fig. 2.** This figure shows three snapshots of data from a sequence of activity monitored by Duo's wearable AI system. In the sequence, the human reaches for a cup and drinks from it. The top row consists of images from a head-mounted camera. The bottom row shows corresponding configurations of a kinematic model of the human based on approximate joint measurements and connections along with the orientation data provided by 3 Intersense devices, with the first on the torso, the second on the upper arm, and the third on the lower arm. These 3 devices provide 3 absolute orientations for a total of 9 angles.

robot that circumvents issues of detailed physical control by relying on innate abilities engendered by the cooperative human.

Traditional humanoid robots, such as Cog, have detailed low-level control of a system with few inherent abilities, while a hybrid humanoid, such as Duo, has coarse high-level control of a system with true human-level abilities. For example, Duo's wearable AI system can easily request for an object to be picked up by the cooperative human. Yet through speech, Duo's wearable AI system would be hard pressed to control a human's joint torque with much resolution in time or space. In contrast, Cog allows for high resolution control of joint torques in time and space, but would require currently unknown and undoubtedly complex control techniques in order to command all of the joints to execute a grasping movement as versatile as a human's (see Figure 4).

These inherent and autonomous abilities of a human/wearable hybrid are in some sense analogous to the innate abilities present in many biological systems. Innate abilities, such as primitive face detection in infants, are often evident early in life, but get subsumed by more refined processing later in life. Initially from the wearable component's point of view, the innate human abilities are powerful, but enigmatic. Over time, however, the wearable component could potentially learn

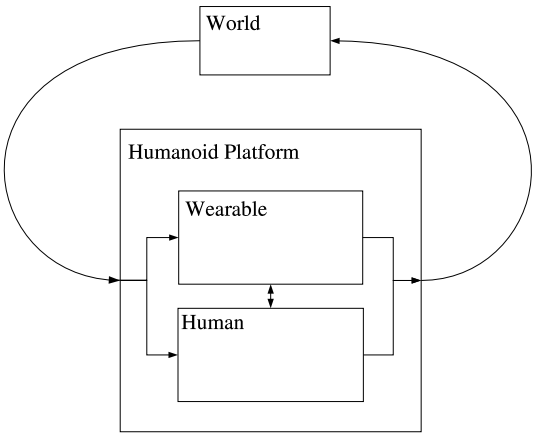


**Fig. 3.** This diagram shows the high-level architecture of the human/wearable hybrid. The wearable AI system and the cooperative human are both integral parts of the unified system. Together they can be usefully thought of as forming a new type of humanoid robot with a structure well modeled by Brook's subsumption architecture [7]. In the architecture the wearable AI system serves as the top-layer of control, since it subsumes the abilities of the cooperative human.

to better interpret and control these innate behaviors, much like the neocortex comes to dominate visual processing of faces as a child matures.

One additional pragmatic advantage of hybrid humanoids is that they tend to be substantially less costly to create and maintain. Unlike traditional humanoid robots, the requisite equipment is already widespread and construction primarily involves the integration of off-the-shelf products.

As the rest of this paper will show, human/wearable hybrids are a type of humanoid robot that is well suited for learning in natural, unconstrained environments about common manipulable objects and the human behaviors typically associated with them. Section 2 motivates the creation of Duo. Section 3 makes a broad survey of work related to Duo. Finally, Section 4 describes Duo and its current behaviors, after which Section 5 concludes the paper.

## 2 Motivations for Duo's Design

Dou is designed to be a learning system. This section starts by arguing for the significance and tractability of the specific learning domain for which Duo was designed. In general, learning is very hard, since it relates directly to search within spaces that can quickly explode exponentially. Consequently, the rest of

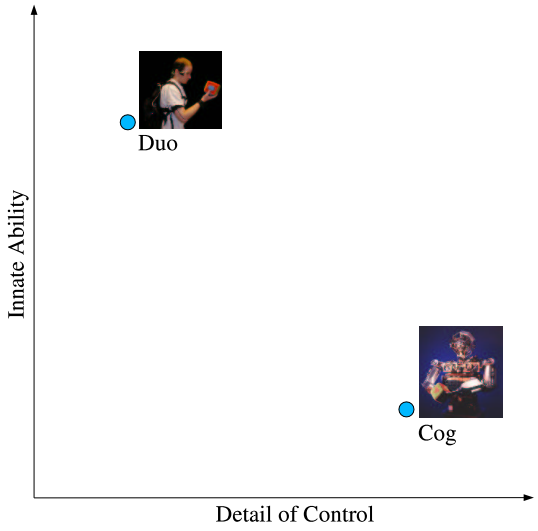


**Fig. 4.** Duo and Cog represent two extremes on this graph depicting the fundamental design tradeoff between the platform's innate abilities and the degree to which the system can be controlled. Duo represents a platform with exceptional innate abilities, but strong limits on the resolution at which the platform can be controlled. For example, Duo's wearable AI system can easily request for an object to be picked up by the cooperative human. Yet through speech, Duo's wearable AI system would be hard pressed to control a human's joint torque with much resolution in time or space. Cog represents the other extreme, since for example, Cog allows for high resolution control of joint torques in time and space, but would require currently unknown and undoubtedly complex control techniques in order to command all of the joints to execute a grasping movement as versatile as a human's.

this section looks at ways to simplify the learning problem, mostly by taking advantage of people.

### 2.1 Learning Systems Can Learn A Lot Through Common Manipulable Objects

By definition, manipulable objects are potentially useful components of the world and therefore worthy of special attention. Moreover, objects that people commonly manipulate are an especially important part of the world, since they are the objects to which people are most intimately tied. These objects are prevalent across natural human environments and they are the most generally useful objects for everyday activities. Most people only carry a small number of common objects with them and rely on the world to provide task relevant objects such as eating utensils, cups, white board erasers, tooth brushes, and pens. A humanoid robot working in natural human environments would be able to make wide use of

cognitive skills pertaining to common manipulable objects. Furthermore, people implicitly expect an intelligent entity to be familiar with these objects and their common uses and can become frustrated when this expected familiarity is not present. This implicit understanding about humans and their favorite objects suffuses human communication, interaction, and activities and is hence widely applicable.

The visual appearance of objects is strongly emphasized in the AI literature, much more so than the common human actions that are applied to objects. The ways in which an object is used is fundamental, as it relates to the object's function and the role it plays in satisfying a human's goals. In some sense an operational understanding of an object is more important than the specifics of its appearance. The answer to the question, "What can I do with object X?" directly communicates the behaviorally relevant value of an object by relating it to an intelligent system's objectives in the world. In contrast, the answer to the question, "What does object X look like?" directly gives information about how to visually locate the object in the world, but only indirectly suggests possible uses for the object and its corresponding behavioral value.

Duo's learning will focus on common manipulable objects as defined above above, by attending to objects that the human partner manipulates during everyday activities. These objects are important to the activities of the human partner and they are statistically more likely to be important to most people. Duo will also learn about the common actions applied to an object in parallel with the object's visual appearance by attending to the human partner's kinematic activity while working with an object. After learning about the appearance and use of objects, Duo will attempt to learn some of the effects that result from actions being applied to objects.

### 2.2 Learning Systems Should Exploit People!

Robotics researchers, inspired by human infant development, have convincingly argued that appropriately designed robots can make learning easier for themselves by taking advantage of human caregivers in ways analogous to human infants [12] [4]. Human/wearable hybrids also exploit cooperative people, albeit in ways that are less directly inspired by existing biological systems.

**People Are Beacons of Importance** People can serve as pointers to the important elements of the the vast streams of data to which Duo will be exposed.

This general theme shows up in several ways. By accompanying a variety of hosts over long periods of time, a system such as Duo can potentially perform a fairly unbiased sampling from the vast space of possible objects and actions. This is in contrast to humanoid robots that monitor a single environment, since human behavior will tend to be highly dependent on the environment's typical uses. People rarely eat dinner in the middle of the street, or play tennis in an office.

**People Are Active Perceptual Aids** Duo actively requests behaviors from the human that will help Duo solve perceptual problems. For example, when the wearable AI system requests that the cooperative human inspect an object more closely, the wearable system co-opts the person's natural inspection behavior so as to improve the view of the object. As another example, when Duo is segmenting an object, the wearable system asks the cooperative human to keep his head still if it is moving in a way that will degrade the segmentation. In the future, other simple predefined queries from the wearable to the human may help Duo usefully parse the kinematic activities of the human. For instance, Duo might request that the host "Do that again!".

**People Are Known Factors** Duo assumes that the same human will wear the wearable AI system for extended periods of time, which should allow for the creation of strong models of the person. For example, if Duo knows the dimensions of the human's hand as well as the appearance of the skin on the human's hand with respect to a variety of lighting conditions, Duo should be able to use these constants to better assess properties of various environments. For instance, if the hand suddenly appears blue and unadorned, the system could justifiably assume that the hand is under blue lighting. If an object held by the human's hand appears to be twice the size of the hand, Duo could make a good estimate of the object's actual size, especially when combined with the hand's 3D position as estimated by the arm orientations and the human's approximate body dimensions.

**People Are Annotators** The human can help Duo to both explicitly and implicitly annotate the experiences encountered. Future work may use text or speech based input, that would allow the human to provide names for common actions and common objects when Duo asks for them.

### 2.3 Learning Systems Should Study People Like An Ethologist

The majority of work that involves the observation of human behavior by a humanoid robot for the sake of learning takes place within a laboratory or an artificial environment. A hybrid platform such as Duo will be able to observe and learn about human behavior from a more ethological perspective, by following the person around in everyday life. This should lead to less bias in the observations, although the wearable system itself will have some impact on the human's behavior and the environment's responses [19]. Complex and natural environments are important for intelligent behavior and development. A better estimate of what objects and actions are common can be made from a wearable platform such as Duo. Similarly, testing algorithms on data acquired ethologically should help researchers to avoid using solutions that are overly specialized to a particular ecological niche. For example, image segmentation fundamentally depends on the background of the image - not just the foreground - but object segmentation research on humanoid robotics often uses images from a small set of natural

background environments, frequently with contrived object placement, that does not obey the natural statistics of object configurations within environments.

## 3 Work Related To Duo

Duo's design and objectives relate to a wide assortment of research specialties. This section attempts to touch on relevant work across these many disciplines in an effort to better understand the problem domain and Duo's approach.

### 3.1 A Brother In Arms

Taro Maeda's Parasitic Humanoid (PH) project at The University of Tokyo is another example of research on a human/wearable hybrid. Their independently developed work shares strong similarities to the work behind Duo, although their applications are quite different. Like Duo, the PH project makes use of absolute orientations measured by devices affixed to the human's body, although the PH system uses many more devices than Duo. The PH group has also developed a number of potentially useful sensors, including a small glasses-mounted eye tracking system, pressure sensors for the soles of shoes, and an interesting finger tip tactile sensor that unobtrusively estimates the force of contact between the human's finger and a surface.

The PH group's architecture can also be interpreted as a subsumption architecture. Their first application has apparently involved analyzing and modulating the walking rhythm of the human. Their work emphasizes the desire to create a parasitic system that will predict the behavior of a human given sensory input. In addition, they claim that the parasite will attempt to correct a human's behavior if it violates the parasite's prediction. [20]

### 3.2 Wearable Platforms

Within the wearable computer community designers typically strive to create devices that will help the person wearing the system [21] [35]. Thad Starner authoritatively states this common goal of wearable computing in the following quote from [35]:

> Wearable computing pursues an interface ideal of a continuously worn, intelligent assistant that augments memory, intellect, creativity, communication, and physical senses and abilities.

Wearable AI for a humanoid robot turns this goal around and makes the wearable AI's objectives paramount. An appropriate quote might paraphrase JFK and state "And so, my fellow researchers: Ask not what your wearable can do for you. Ask what you can do for your wearable." So although much of the work from the wearable computing community is related and applicable to Duo, the fundamental goals are distinct, which leads to different design criteria.

**Wearable Systems That Watch Hands And Objects** Within the wearable computing community many researchers have worked to interpret streams of data from wearable sensors such as cameras, microphones, and accelerometers. In the video diary work of Kawamura et al., [16], a wearable system records audio and video from a first person perspective and attempts to index the recordings for future recall. The system does pay attention to objects, but only if they have RFID (Radio Frequency Identification) tags attached to them.

A wearable system for real-time American Sign Language recognition by Starner et al. [36], used skin color to segment and track hands at 10Hz while viewed from a camera that faces down from the brim of a baseball cap. Sequences of feature vectors extracted from the segmented hand blobs were recognized as ASL words using pre-trained HMMs.

Many wearable computing systems process streams of data in an attempt to estimate the context of the situation within which a wearable system is being used, so that the system can behave in ways that match the current desires of the user [33]. Duo focuses on the human's hand and the objects with which it interacts, rather than broad contextual cues from the environment.

**Wearable Systems that Learn from People** Several researchers have recorded large data sets from wearable sensors worn by a person going through typical daily activities and then processed the recordings off-line with machine learning algorithms. Clarkson in [9] demonstrates that statistical learning methods applied to video, audio, and orientation data recorded from a backpack over 100 days can extract meaningful patterns of daily activity in terms of locations visited and sequences of locations visited at coarse and fine temporal resolutions.

Jebara and Pentland in [14] argue for statistical imitative learning of human behavior with a wearable platform that records Jebara's voice and video of his face while simultaneously recording audio and video from the immediate environment. Jebara and Pentland use statistical learning techniques to produce a generative model that attempts to synthesizes the appearance of Jebara's face and the sound of his voice when presented with video and audio from an environment. This prediction goal is similar to the stated goal of Maeda's Parasitic Humanoid.

The most fundamental distinction between our work and the work we cite here, is our use of real-time processing and control of a cooperative host. Processing pre-recorded sensory data does not allow the learning system to interact with the world. Requesting actions at opportune times will make learning easier for Duo. Duo also visually segments the hand and objects, which is significantly different from the approaches of Jebara, Pentland and Clarkson who use appearance based vision techniques applied to entire frames of video.

**Wearable Systems that Control People** Wearable applications often influence the behavior of the user. Some applications, such as instructive assistants that teach or guide the user, actually do attempt to directly control the behavior of the user through communications conveyed via speech or video displays [15]

[25] [32]. As an acknowledgement of this property of approximate cyborgs, Steve Mann states the following in his paper on humanistic intelligence (HI):

> When a wearable computer functions in a successful embodiment of HI, the computer uses the human's mind and body as one of its peripherals, just as the human uses the computer as a peripheral. This reciprocal relationship is at the heart of HI [21].

Clearly this notion strongly relates to Duo. However, the goals for these systems are very different from Duo, and still tend to emphasize the wearable system as a human helper.

### 3.3 Robotic Platforms

Robotics has served as the driving analogy for the design of Duo. This section briefly examines some related work on traditional robotic platforms.

**Robots that Socialize with People** Work associated with the social robot Kismet has demonstrated plausible mechanisms for learning through social interactions with people, which to some extent can be characterized as methods of controlling the caregiver [4]. Kismet has also used the caregiver to make some perceptual problems easier [5], which matches one of the motivations behind Duo. For example, through facial expressions and head posture, Kismet would influence the caregiver to stay at a more easily perceived distance from the robot.

Deb Roy's thesis work involved a robot that learned to visually recognize and associate an auditory name with simple objects that were placed before it [30]. Subsequent work within Deb Roy's group has involved a robot that asks people to describe an object that it selects from a small group of objects on a table using a laser pointer [31]. The way in which this system solicits help from a caregiver in order to learn about objects relates to Duo, although, as with most humanoid robotic systems, these robots have limited access to common manipulable objects in natural, unconstrained environments.

**Robots that Imitate People** Imitation has been proposed and pursued as a way for robots to learn useful skills [34]. Motion capture has been used for both teaching and teleoperating robots for various tasks [28]. Our work is not focused on imitation as such, although similar techniques and ideas will come into play if Duo is eventually able to request a behavior it has previously witnessed.

**Robots that Manipulate Objects** There exists a long history of work on AI for robots that manipulate objects within their environment [22] [10]. More recent work has used machine learning methods to perform tasks such as recognizing and then appropriately grasping or describing simple objects [26] [31]. This type of work is usually performed within a fixed environment into which researchers have brought a small set of objects and placed them before the robot.

The objects tend to be well matched with the perceptual capabilities of the robot, and the background is usually constant over time and simple in appearance. Due to mechanical and computational constraints the actions performed by the robots in manipulation tasks are usually either primitive, by human standards, or canned.

Although Duo only has high-level control of object manipulation, the actions will be complex everyday human actions. Likewise, the objects Duo encounters will be natural and complex.

**People that Manipulate Robots** Teleoperated humanoid robots are essentially Duo's subsumption architecture flipped upside down, with the human forming the top-layer of control. Although there are many interesting applications for teleoperated humanoids, their design has the effect of limiting the human's capabilities to those of the robot. Consequently, with today's robotics technology, a human controlling such a platform would have difficulty behaving naturally while manipulating common objects encountered in unconstrained environments. With respect to this learning problem, teleoperated humanoid robots would lead to more bias in the observations of human behavior, while being significantly more complex and costly than a system such as Duo.

The major advantage teleoperated robotics would offer is full knowledge of the information available to the human while manipulating objects. Given the mismatch between the resolution, frame-rate, and sensitivity of Duo's camera and the human's eyes, situations can occur where the wearable system will be unable to understand the human's behavior due to the discrepencies between the wearable's senses and the human's.

### 3.4 Machines that Watch from Afar

Observations made from a single camera that is stationary with respect to the world frame must contend with fluctuating resolution, occlusions, and widely varying perspectives of a person's natural object manipulation.

Hybrid systems have a distinct advantage because of their frame of reference, which significantly constrains the possible perspective variations. The wearable's frame is approximately fixed with respect to the hand's workspace in both position and orientation. Instead of simplifying the environment, the objects, or the actions, a hybrid humanoid simplifies its point of view on the activity and makes direct measurements with short range and contact based sensors. These advantages of a first person, situated, and embodied perspective on first person activities are true of traditional humanoid robots as well.

**Machines that Spy on People with Objects** The research literature related to machine vision algorithms for surveillance from fixed perspective pre-recorded video of people is large [27], but the number of projects that detect arm activity in relation to objects appears to be small. The W4 real-time surveillance system has some methods of detecting whether or not a person is carrying an object,

based on symmetry and temporal analysis of silhouettes [13] [1]. They test their methods, which seem to require that the person be carrying a large object, on low resolution images of people cropped from wide area surveillance cameras.

**Machines that Watch Videos of People, Objects and Actions** Within machine vision and AI, a long ongoing thread of research has pursued the analysis of object manipulation in video [11]. However, all of the work cited below used video from fixed cameras with a good view of a well constrained, often canned, task, in a simplified environment.

In 1996 Brand created a blob-oriented 2D vision system that used 6 hand-coded networks similar to HMMs to recognize actions Brand called "touching", "putting", "getting", "adding", and "removing", in highly constrained video of human activity [2]. In 2000 Brand and Kettnaker described work on a system that automatically learned HMMs (the states, transitions, and parameter values) from a similar 2D blob oriented input from a well-positione stationary desk camera in an office [3]. Some of the actions could be loosely described as relating to object manipulation.

Duric et al. used optical flow, 3D object models, and a set of hand-coded categories to recognize a few types of tool use from 7 short videos of an arm using the tools in different ways, including a knife for sawing and stabbing. [11].

Darnell Moore's dissertation work is especially relevant to Duo [24]. His system tracked and recognized a variety of hand movement patterns as they related to a set of common objects, within a well specified task. He clearly demonstrated the value of associating actions with objects by using actions alone to recognize previously unseen objects. He also showed that action recognition can be aided by knowledge of the object to which the action is being directed. In addition, he used actions to help disambiguate the identity of objects when the visual evidence was weak. Moore also built higher level recognizers that could discern some larger scale tasks such as "washing dishes" and "cooking stir-fry".

However, his system was only tested on short scripted action sequences in carefully constructed environments with a well placed stationary ceiling camera. Like Brand's work, action recognition used simple HMMs applied to 2D trajectories of the person's hands. Object recognition was performed by simple template matching. The programmer had to first carefully specify the expected objects and their associated actions, and then train them on data prepared by hand. Moreover a GUI was used to segment the background scene into any pre-existing objects of interest by hand prior to running the system.

### 3.5 Systems that Capture Human Motions and Recognize Gestures

Since the 1980's motion capture technology has been successfully used to record and process the configuration of the human body, primarily for applications within animation and biomechanics [37].

Gesture recognition also has a long history that spans pen based gestures in Sketchpad in 1963 to more recent research in applications such as the automatic

recognition of American Sign Language (ASL). A large number of classification schemes have been explored in the literature although HMMs are the predominant classification method [24].

Rao et al. recently published work that clusters actions using points of high curvature along the trajectories of hands tracked in video [29]. The evidence they present for the utility and detectability of these points is compelling. Duo's simple detector for recognizing when the hand reaches for an object and grasps it was inspired by this work.

## 4 Duo: A Hybrid Humanoid

The value of hybrid humanoid robots is supported by the initial results from Duo. Duo was recently endowed with its first set of behaviors that relate to its long term objectives of learning about objects, actions, and effects. As described below, Duo currently acquires high quality segmentations of everyday manipulable objects in unconstrained natural environments. This initial application illustrates the benefits of this class of humanoid robot.

### 4.1 The Duo Platform, Described

The wearable side of Duo currently consists of a head-mounted firewire camera, 4 absolute orientation sensors, an LED array, and headphones. These sensory systems are connected to a laptop computer, which is placed in a backpack worn by the human. The backpack also contains rechargeable batteries that support full mobility by providing power to the camera, orientation sensors and LED array. The laptop wirelessly communicates with a dedicated cluster of computers via 802.11b. With the increasing availability of economical, broadband wireless connectivity this class of system will be able to function over an entire city in the relatively near future, although for now the system is primarily used inside the MIT AI Lab. Unlike systems that learn from prerecorded data captured by wearable systems [9], Duo must use real-time processing in order to make relevant requests for actions while exploring the world and testing hypotheses. The cluster provides processing power that will be necessary for computationally intensive real-time sensory processing. When unable to connect with the cluster the system could either shut down or perform more limited perceptual processing, such as recording significant events, as measured by the kinematic sensors to the hard disk for off-line processing. Duo uses custom clustering software running on top of the Debian distribution of the GNU/Linux operating system. The system also makes use of several open source libraries, including the Festival text to speech synthesizer and the Intel OpenCV computer vision library.

As shown in Figure 1, the initial system used a glasses-mounted camera. As depicted in Figure 5, the camera is now mounted on the brim of a hat. Testing indicated that this placement along with a wide angle lens gives Duo a better view of the dominant arm's workspace. The system uses Intersense Inertia Cube

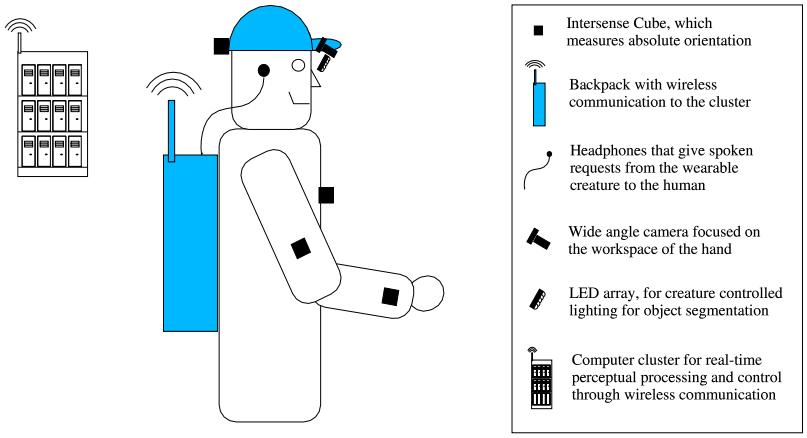


**Fig. 5.** This schematic depicts the most recent version of the human/wearable hybrid, which now uses a baseball cap, 4 Intersense Cubes, an LED array, a firewire camera, a laptop with wireless connectivity, headphones, a backpack with batteries, and a dedicated cluster of computers. The key shows the icons for these important components.

2's to measure the absolute orientation of 4 body parts in a common world coordinate system. These devices combine inertial measurements with gravimetric and magnetic measurements that correct for drift using the direction of gravity and the earth's magnetic field. As shown in diagram 5, these 4 absolute orientation sensors are affixed to the lower arm, upper arm, torso and head of the human. In order to estimate the configuration of the person's head and dominant arm, these orientations can be used with a simple kinematic model based on approximate dimensions and connectivity of the body parts. The wearable system makes spoken requests through the headphones and uses the LED array to aid vision.

### 4.2 Duo's First Behavior System

Duo's current set of behaviors are designed to acquire high quality segmentations of the hand-held objects a person works with during the day. While the human goes about his daily activities, Duo attempts to detect when the person is working with a new hand-held object. If Duo detects that the arm has reached for an object and picked the object up, it asks to see the object better. When the cooperative person brings the object close to his head for inspection, Duo recognizes the proximity of the object to the head using the kinematic model, and turns on a flashing array of white LEDs, which allows Duo to easily segment

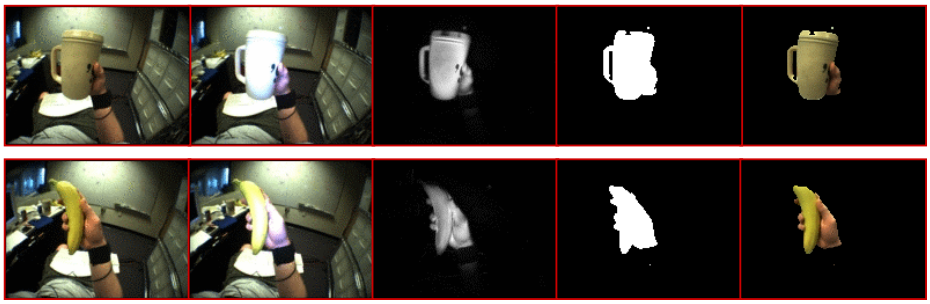

**Fig. 6.** The system currently achieves segmentation by active sensing. When the wearer brings an object up into view (first column), an oscillating light source is activated (second column). The difference between images (third column) is used to compute a mask (fourth column) and segment out the grasped object and the hand from the background via a simple threshold.(fifth column).

the object and the hand (see Figure 6). This method has the additional benefit that the effective resolution of the object in the image is increased. While the human is holding the object close to his head, Duo also kinematically monitors head motion. If the human's head motion goes above a threshold, Duo requests that the person keep his head still in order to improve the segmentation. (see Figure 7 for a detailed explanation of the behavior system.)

Currently, Duo uses a simple hand-coded matched filter to detect when a person is likely to be grabbing a new object. The filter operates on measurements derived from the kinematic model and its estimated configuration based on the measured orientations from the human body. Specifically, the filter is run on the results of projecting the estimated velocity of the wrist, with respect to the world's coordinate system, onto a unit vector extending from the center of the torso to the wrist. The resulting measurement indicates the velocity at which the hand is moving toward or away from the center of the human's torso. The matched filter simply detects when the wrist moves away from the torso for an extended period at a relatively high velocity, slows down to a stop, and then moves toward the torso at a relatively high velocity for an extended period. This method is related to work done on invariant action recognition through the detection of points of high curvature in the path of the hand [29]. Although simple and imperfect, Duo's detector is sufficient to acquire excellent segmentations of everyday objects and should help the system bootstrap more refined methods of detection.

The array of white LEDs provide active illumination that clearly differentiates between foreground and background since the illumination rapidly declines as a function of depth. By simply subtracting the illuminated and non-illuminated images from one another and applying a constant threshold, Duo is able to segment the object of interest and the hand (see Figure 6). By keeping his head still, a cooperative human minimizes image motion, which improves the success of the simple segmentation algorithm and reduces the need for motion

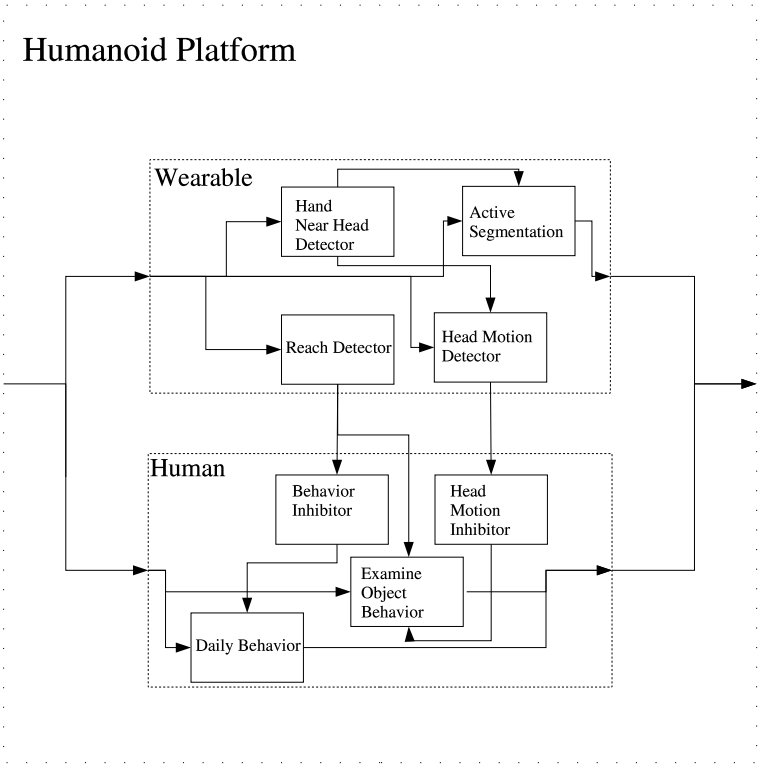


**Fig. 7.**

**Duo's First Behavior System:** This block diagram shows the detailed architecture for the first behaviors implemented on Duo. Although simple, these behaviors effectively demonstrate the value of human/wearable hybrids for learning about common manipulable objects in unconstrained environments. This block diagram clearly shows the subsumption architecture with the wearable AI system serving as the top-layer control. This diagram also includes a conceptual block diagram of the behaviors at work within the human and their relationship to the wearable's inner workings.

**Within the Wearable AI System:** The *Reach Detector* activates when it kinematically detects that the human has reached for and acquired an object. The *Hand Near Head Detector* activates when it kinematically detects that the human's hand is close to his head. The *Head Motion Detector* activates when it detects head motion and the *Hand Near Head* detector is active. The *Active Segmentation* block turns on the flashing LEDs when the *Hand Near Head Detector* is active. It also performs image differencing and thresholding to segment actively illuminated foreground objects from the background.

**Within the Cooperative Human:** The *Behavior Inhibitor* and the *Examine Object Behavior* are both activated when the human hears the wearable ask to see the object better. The *Behavior Inhibitor* subsequently inhibits the *Daily Behavior* that the human was in the midst of performing. The *Examine Object Behavior* causes the human to bring the object up to a natural location in front of the eyes in order to better observe it, which in turn triggers the wearable's the *Hand Near Head Detector*. The *Head Motion Inhibitor* activates when the human hears the wearable request that he keep his head still. It subsequently inhibits the head motion performed by the *Examine Object Behavior*.

compensation prior to subtracting the images. The location at which the LED array is most effective sits about 25cm from the face, centered on the eyes. Humans also get a strong sense of depth around this location through stereopsis. Duo could feasibly use a stereo camera configuration to get a similar segmentation, but the computational cost and additional hardware complexity would not be justified for this application. Also, less obtrusive infrared LEDs could probably be substituted for the white LEDs, but debugging would be more difficult and less feedback would be provided to the human about optimal object placement and system activity.

### 4.3 Duo's Simple Example

In contrast to a traditional humanoid robot, a few simple behaviors allowed Duo to jump directly to interacting with common manipulable objects in natural, unconstrained environments from which it acquires useful, high-quality segmentations. Duo only has coarse high-level control of its body, yet this control was sufficient to move interesting objects to a location ideal for segmentation. Furthermore, requesting other object directed actions, such as picking up an object of Duo's choosing, are entirely feasible using current AI technologies.

Results from Duo should be complementary with results from traditional humanoid robots. For example, vision systems trained on the first person manipulation of everyday objects should be useful to these robots. Likewise, the kinematic recordings of object related actions could potentially serve as useful hints when traditional humanoids attempt to work with everyday objects through methods similar to those that have been used when combining standard motion capture data with humanoid robots.

### 4.4 Duo's Future

This simple set of behaviors for segmenting everyday objects should serve as a useful foundation from which Duo can bootstrap its learning. First, the segmentation results will be used as a type of ground truth for training and evaluating a more general statistical segmentation algorithm which is being researched [18]. Second, the segmentation results and subsequent tracking of the segmented object will be used to help the system detect and recognize that particular object during future activities without the aid of the LED array and using the same statistical framework as the segmentation algorithm [18]. Third, the hand-coded action recognizer that detects when the human has reached and grabbed an object, will serve as the starting point for learning a more accurate, empirically based recognizer for object related actions. The human's actions will help to interpret whether or not an object is in hand, by either ignoring requests stemming from false positives, or by showing Duo an empty hand. Likewise, when an object is successfully segmented, the system will associate the observed kinematic data with the segmented object, in order to begin learning the common human behaviors that relate to particular objects.

## 5 Discussion and Conclusions

Hybrid humanoids have the ability to encounter diverse human-relevant stimuli that are currently inaccessible to humanoid robots for now and the near future. They circumvent many of the very difficult unsolved problems of humanoid robots, such as low-level motor control, allowing researchers to attack interesting problems that can be beneficially investigated in parallel with unsolved low-level problems. As such, human/wearable hybrids may serve as a bridge to more functional humanoid robots.

Many of the arguments for applying traditional, wholly synthetic humanoid robotics to the study of artificial human intelligence apply well to human/wearable hybrids [6]. As sensory technology improves, the wearable component of hybrid systems will be able to experience the environment in a way that is increasingly similar to the human's experience. Likewise, opportunities for more direct methods of influencing human behavior will become available as technologies for integrating flesh and machines advance [8].

Technological innovation continues to diversify the methods that are available for human and machine cooperation. The opportunity to be the bottom layer in a subsumption architecture may seem like a poor deal for the cooperative human. But care in system design could potentially make the experience of helping a more primitive system learn enjoyable, as it often is when playing the role of caregiver for infants and children, who are frequently demanding and manipulative in their own way. Similarly, although this paper has focused on wearable AI systems that would direct human action for their own benefit, these types of systems are not incompatible with the typical goals of wearable computing. For example, the flow of assistance could gradually shift from benefitting the wearable AI system to benefitting the cooperative human, as the wearable AI learns enough to be helpful. In the far future, even if the process of helping a hybrid system learn is not fully entertaining, the creation of intelligent systems that understand us and our everyday lives should lead to helpful applications that would make the investment worthwhile.

## 6 Acknowledgements

Rod Brooks's support is greatly appreciated. Una-May O'Reilly and Paul Fitzpatrick provided very helpful comments on this manuscript. Eduardo Torres-Jara helped significantly with the design of an analog circuit for the system. Funds for this project were provided by DARPA as part of the "Natural Tasking of Robots Based on Human Interaction Cues" project under contract number DABT 63-00-C-10102, and by the Nippon Telegraph and Telephone Corporation as part of the NTT/MIT Collaboration Agreement.

## References

1. Chiraz BenAbdelkader and Larry Davis. Detection of people carrying objects: a motion-based recognition approach. In *Proceedings of FGR'02*, 2002.

2. Matthew Brand. Understanding manipulation in video. In *Proceedings of FGR'96*, pages 94–99, 1996.
3. Matthew Brand and Vera Kettnaker. Discovery and segmentation of activities in video. *PAMI*, 22(8):844–851, August 2000.
4. C. Breazeal. *Sociable Machines: Expressive Social Exchange Between Humans and Robots.* PhD thesis, MIT, 2000.
5. C. Breazeal, A. Edsinger, P. Fitzpatrick, and B. Scassellati. Active vision systems for sociable robots. *K. Dautenhahn (ed.) IEEE Transactions on Systems, Man, and Cybernetics*, 31(5), 2001.
6. R. Brooks, C. Breazeal(Ferrell), C. Kemp R. Irie, M. Marjanovic, B. Scassellati, and M. Williamson. Alternative essences of intelligence. In *AAAI98*, pages 961–967, Madison, WI, 1998.
7. R. A. Brooks. A robust layered control system for a mobile robot. *IEEE Journal of Robotics and Automation*, 2(1):14–23, March 1986.
8. Rodney A. Brooks. *Flesh and Machines.* Pantheon Books, 2002.
9. Brian Patrick Clarkson. *Life Patterns: structure from wearable sensors.* PhD thesis, MIT Media Laboratory, September 2002.
10. J. Connell. *A Colony Architecture for an Artificial Creature.* PhD thesis, MIT AI Lab, 1989.
11. Zoran Duric, Jeffrey A. Fayman, and Ehud Rivlin. Function from motion. *PAMI*, 18(6):579–591, June 1996.
12. C. Ferrell and C. Kemp. Presentation slides for the talk associated with this paper : An ontogenetic perspective to scaling sensorimotor intelligence. In *Embodied Cognition and Action: Papers from the 1996 AAAI Fall Symposium*, pages 45–49. AAAI Press, 1996.
13. Ismail Haritaoglu, David Harwood, and Larry S. Davis. W4: Real-time surveillance of people and their activities. *PAMI*, 22(8):809–830, 2000.
14. Tony Jebara and Alex Pentland. Statistical imitative learning from perceptual data. In *2nd International Conference on Development and Learning, ICDL'02*, June 2002.
15. Ashish Kapoor, Selene Mota, and Rosalind W. Picard. Towards a learning companion that recognizes affect. In *Proceedings from Emotional and Intelligent II: The Tangled Knot of Social Cognition, AAAI Fall Symposium*, November 2001.
16. Tatsuyuki Kawamura, Yasuyuki Kono, and Msatsugu Kidode. Wearable interfaces for a video diary: Towards memory retrieval, exchange, and transportation. In *Proceedings of the International Symposium on Wearable Computers 2002*, 2002.
17. Charles C. Kemp. Humans as robots. In *MIT Artificial Intelligence Laboratory Research Abstracts*, chapter Robotics, pages 332–334. MIT AI Lab, September 2002.
18. Charles C. Kemp. *Thesis Proposal: Humans as Robots.* PhD thesis, MIT, December 2002.
19. Peter Lynch. Cyberman, 2001. A film documenting some of the adventures of Steve Mann.
20. Taro Maeda, Hideyuki Ando, Maki Sugimoto, Junji Watanabe, and Takeshi Miki. Wearable robotics as a behavioral interface - the study of the parasitic humanoid -. In *Proceedings of the 6th International Symposium on Wearable Computers (ISWC'02)*, October 2002.
21. Steve Mann. Wearable computing: Toward humanistic intelligence. *IEEE Intelligent Systems*, pages 10–14, May-June 2001.
22. M. Minsky. Serpentine hydraulic robot arm. a patent apparently exists and the robotic arm currently resides within the Boston Museum of Science, 1967.

23. Marvin Minsky. *The Emotion Machine*, chapter Part VI, Common Sense of The Emotion Machine. http://web.media.mit.edu/ minsky/E6/eb6.html, 2002.
24. Darnell Janssen Moore. *Vision-Based Recognition of Actions Using Context*. PhD thesis, Georgia Institute of Technology, April 2000.
25. Jennifer J. Ockerman and Amy R. Pritchett. Preliminary investigation of wearable computers for task guidance in aircraft inspection. In *Proceedings of the 2nd International Symposium on Wearable Computers 1998*, October 1998.
26. Josef Pauli. Learning to recognize and grasp objects. *Machine Learning*, (31):239–259, 1998.
27. Alex Pentland. Looking at people: Sensing for ubiquitous and wearable computing. *PAMI*, 22(1):107–119, January 2000.
28. Nancy S. Pollard, Jessica K. Hodgins, Marcia J. Riley, and Christopher G. Atkeson. Adapting human motion for the control of a humanoid robot. In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA2002)*, May 2002.
29. Cen Rao, Alper Yilmaz, and Mubarak Shah. View-invariant representation and recognition of actions. *IJCV*, 50(2):203–226, 2002.
30. Deb Roy. *Learning from Sights and Sounds: A Computational Model*. PhD thesis, MIT Media Laboratory, 1999.
31. Deb Roy, Peter Gorniak, Niloy Mukherjee, and Josh Juster. A trainable spoken language understanding system for visual object selection. Submitted to the International Conference of Spoken Language Processing, 2002. (in review).
32. Feiner S., MacIntyre B., and Seligmann D. Knowledge-based augmented reality. *Communications of the ACM*, 36(7):52–62, July 1993.
33. Daniel Salber, Anind K. Dey, and Gregory D. Abowd. The context toolkit: Aiding the development of context-enabled applications. In *Proceedings of CHI'99*. ACM Press, May 1999.
34. Stefan Schaal. Is imitation learning the route to humanoid robots? *Trends in Cognitive Sciences*, pages 233–242, 1999.
35. Thad Starner. The challenges of wearable computing: Part 1. *IEEE Micro*, pages 44–52, July-August 2001.
36. Thad Starner and Alex Pentland. Real-time american sign language recognition using desk and wearable computer based video. *PAMI*, 20(12):1371–1375, 1998.
37. David J. Sturman. A brief history of motion capture for computer character animation. This paper was published on the web, you can find it by searching for the title on Google!, 1994.