How a Smart Environment Can Use Perception

John Krumm Steve Shafer Andy Wilson
Ubiquitous Computing Group
Microsoft Research
Microsoft Corporation
One Microsoft Way
Redmond, WA 98052
{jckrumm | stevensh | awilson}@microsoft.com

Abstract

Perception is a vital part of a smart environment, both as a way to sense the state of the space and a way to detect commands from its users. This paper catalogs different sorts of perception for smart environments using the EasyLiving project at Microsoft Research as the main example.

1. Introduction

Without perception, ubiquitous computing would be cumbersome, and a smart environment would be impossible. With perception, an environment can come alive in its reactions to people and devices. All smart environments have some kind of sensing, be it cameras, microphones, active badges, pressure sensors in the floor, or other specialized sensors.

The EasyLiving project at Microsoft Research is a smart environment that demonstrates many different uses of perception in a living room. Part of our lab is shown in Figure 1. This paper catalogs different ways of using sensing in a smart environment using EasyLiving as the main example. Taken together, the list shows a surprising variety of compelling applications for perception as a part of ubiquitous computing.

We will organize the list of perception tasks by splitting the things that are perceived into people and objects. We will further split the tasks into those that sense context (state of the person or object) and those that sense intentional user interface commands. Table 1 shows the resulting four categories. The “Context” column concerns the perception of things in the environment that are not intentionally communicating with the system. This would include a motion detector watching for people in the room and a camera measuring the light level. The “UI” column concerns the perception of intentional commands to the system, such as a person making gestures or using a specialized UI device. The next two sections of the paper are organized along the columns and rows of the table as follows:

Figure 1: EasyLiving lab

2. Perception for Context
    2.1 Perceiving People in Context
    2.2 Perceiving Objects in Context
3. Perception for UI
    3.1 Perceiving People for UI
    3.2 Perceiving Objects for UI

2. Perception for Context

Perceiving context means sensing the state of the environment as it is without looking for any intentional commands. This allows the system to react automatically. Automatic behaviors based on context are usually very impressive in demonstrations

		Context vs. Intentional Communication	
		Context (Section 2)	UI (Section 3)
Type of Things Sensed	People	Sensing people as they go about normal activities (Section 2.1)	Sensing people who are giving specific commands to the intelligent environment (Section 3.1)
	Objects	Sensing normal objects in the environment (Section 2.2)	Sensing specialized UI devices (Section 3.2)

Table 1: Taxonomy of sensing tasks for an intelligent environment

of smart environments, but we have found we must be careful not to automate too much for fear of invoking an undesired behavior and frustrating the user.

2.1 Perceiving People in Context

Perceiving people in context lets the system react automatically to a person's identity, location, activity, and facial expression.

2.1.1 Person Recognition

Recognizing people in a space allows the environment to offer the right set of services for each person. In EasyLiving, this involves both setting a user's access privileges and customizing the environment to his or her preferences. For instance, a guest user in EasyLiving is only allowed to control the room's lights and run a web browser. An identified user is allowed to bring up his computing session. In addition, an identified user is given access to his own set of media files (*e.g.* MP3 music) to play on the room's media player. It is easy to imagine other preferences that might be invoked, like a person's preferred light settings, screensavers, and the alert signal for incoming messages.

In EasyLiving we have made a conscious decision not to automatically recognize people. Instead we offer two manual forms of recognition – a normal keyboard login and a fingerprint reader. We avoid automatic recognition for two reasons. The first is that people should be able to protect their own privacy and security by having the choice of whether or not the system knows who they are. The second reason is that people should not have to wonder whether or not the system has recognized them. An automatic recognition system may take some unknown amount of time to recognize a room's occupants, or it may fail altogether. User-invoked recognition, on the other hand, is more reliable and it can give immediate feedback when it works. Once the user has identified himself, he can instruct the system to forget his identity and thus regain his anonymity.

2.1.2 Person Location

Many behaviors are location specific, such as automatically turning on the lights near a person or routing telephone calls to the nearest telephone. In EasyLiving, we use the location of people measured from a vision system to control the lights, to have a Windows® session follow a user from screen to screen, to help pick audio speakers for music, and to play a game of "hotter/colder" where a person must walk around to find a certain spot in the room guided by hints from the computer.

2.1.3 Person Activity

People are always doing something. If a system is aware of their activity, even crudely, it can provide appropriate services. We have split activity awareness into detection, recognition, and learning.

2.1.3.1 Activity Detection

In activity detection, the goal is to simply determine whether or not there is activity in the space. This can be accomplished simply by looking for motion, either from a dedicated motion detector or a camera. If the space already has a person tracker, activity detection is free.

Besides the obvious security application, activity detection would be useful for creating a video history of what happened in a room. Activity would trigger the room's cameras to record video, which could later be browsed to answer questions like "Where did I leave my keys?" and "Who broke into our house?"

2.1.3.2 Activity Recognition

Activity recognition means recognizing some prespecified activity like cooking, sleeping, reading, or watching TV. Ideally, we would like to recognize what people are doing in order to assist them. For instance, reading the newspaper could trigger the right lighting. In EasyLiving our person tracker can distinguish standing from sitting, and we can trigger certain behaviors for each.

Activity recognition is not a mature research area, and nearly all activity recognition for smart environments is done with vision. General activity recognition is hard because a person's activity is defined by its evolution in time, location in space, and utilization of objects. In addition, there is no well-defined vocabulary of activity.

One of the most compelling forms of activity recognition is the monitoring of well-being, especially for the young and old. Audio and video baby monitors are available in stores. Elizabeth Mynatt of Georgia Tech has developed a system to monitor old people by measuring their activity levels and reporting to interested parties.

2.1.3.3 Activity Learning

Activity learning means discovering the normal patterns of activity in a space. If a system knows what activity to expect, it can use this information to improve its ability to perceive the environment. For instance, if people always reemerge from the bathroom within 30 minutes, then the system can devote extra attention to the bathroom door for this period to resume tracking. Or if the living room is typically empty in the middle of the night, the system could use this time to test the lights and reaquire a model of the room. In addition, a model of normal activity could be used to detect abnormal activity, which could trigger an alarm, such as no one reemerging from the bathroom after 30 minutes.

2.1.4 Expression Recognition

A person’s facial expression is a way of inferring what he wants. This would likely have to be accomplished with cameras in a smart environment, and such research is in its infancy. One exception is systems in automobiles that can tell if a driver is sleepy by looking at his eye blinks. If he blinks for too long, an alarm sounds.

2.2 Perceiving Objects in Context

Perceiving objects in context means that the system is measuring the state of normal objects in the room, like the lights and furniture. As with the perception of people in context, perceiving objects this way can be used to trigger automatic behaviors.

2.2.1 Object Tracking

Object tracking means continuously keeping track of an object’s location. In EasyLiving we track the location of a wireless keyboard on a coffee table with an overhead camera as shown in Figure 2. Combined with location of the room’s occupants, we use this capability to automatically direct the keyboard’s keystrokes to the typist’s computing session.

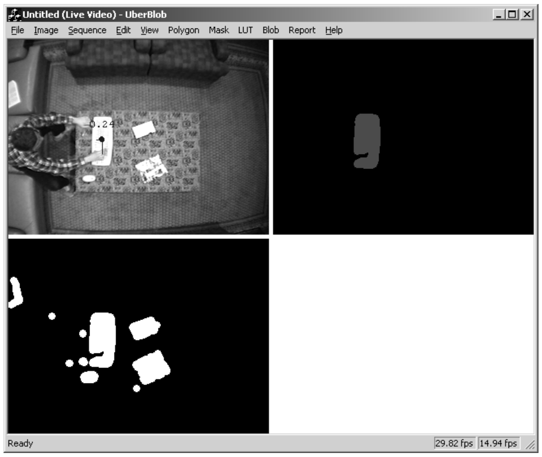
Untitled (Live Video) - UberBlob
File Image Sequence Edit View Polygon Mask LUT Blob Report Help
0.24
Ready
29.82 fps
14.94 fps

Figure 2: Tracking wireless keyboard with overhead camera

2.2.2 Object Recognition

Object recognition could be used to find all occurrences of an object in saved video. For instance, someone might ask to find all occurrences of a TV remote control or car keys over the past day. Object recognition would also be useful for maintaining an inventory, particularly of food. Many grocery store items are not bar coded, so a camera-based recognition system could automatically count, say, all the apples.

2.2.3 World Model Acquisition

A world model is a database that stores the location and type of objects and people in the environment. In EasyLiving, it is vital that we know the locations of computer monitors, audio speakers, and furniture. We have developed an automatic way of finding computer monitors, called AirNet, using cameras and an attention-getting pattern flashed on the monitors. We have a semi-automated way of placing the furniture in our model using images from the room’s cameras. The result is shown graphically in Figure 3.

2.2.4 Device Control

Continuous monitoring of the environment means a system can contain feedback control loops to maintain certain conditions. A simple example is a thermostat for HVAC control. With cameras it would be possible to continuously control the amount of light in a room, dimming the lights when sunlight is present and gradually brightening them in the

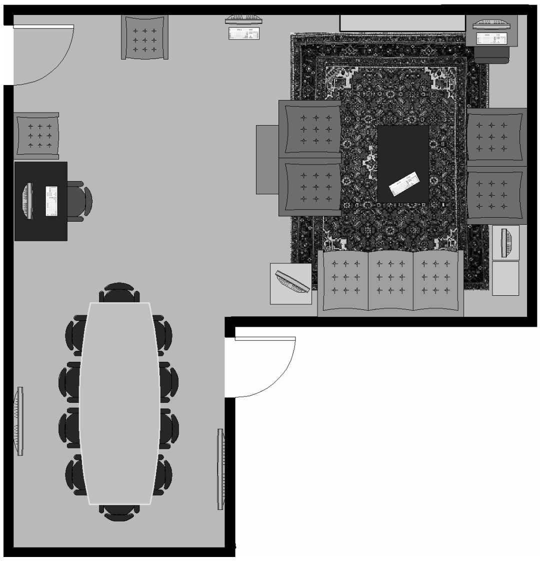

Figure 3: World model shows location of furniture in EasyLiving

evening. The cameras would be used to measure the illumination.

2.2.5 Background Maintenance

The background is, in general, the signal against which new measurements are compared. In EasyLiving, as well as in many other person-tracking systems, we keep background images for all the cameras in order to detect people in the environment. It would also be advantageous to model background statistics for audio to aid speech recognition. These background measurements must be maintained so they will reflect any changes in the background such as a moved piece of furniture or a new audio noise source. Background maintenance differs from the world model in that the background is usually represented at the signal level, while things in the world model are usually represented at the object level.

3. Perception for UI

This section on Perception for UI covers our other major category of perception, the first one being Perception of Context. Perception for UI concerns the perception of things that are meant as intentional commands to the smart environment, such as a gesture from a person or a signal from a special UI device.

3.1 Perceiving People for UI

In Section 2.1, we discussed the perception of people going about their normal activities. This could be used to trigger automatic behaviors. A perceptive environment can also look and listen for certain actions from people that are specifically intended as communication with the environment, which we discuss in this section.

3.1.1 Gestures

User interfaces with gestures have been studied extensively for the case of a user sitting in front of a monitor. These techniques can be generalized for a user in a room being observed by cameras. Except for pointing, it is an open question whether users would rather give commands with gestures or speech.

3.1.2 Pointing

Pointing has been documented as users' preferred method of giving "locatives" to a computer, such as points on a map. It is in general easier to point at something than describe its location with typed or spoken words. In EasyLiving, a user can position the cursor on a large display by pointing with his arm as shown in Figure 4. We sense the pointing direction with the same stereo cameras that we use to track people.

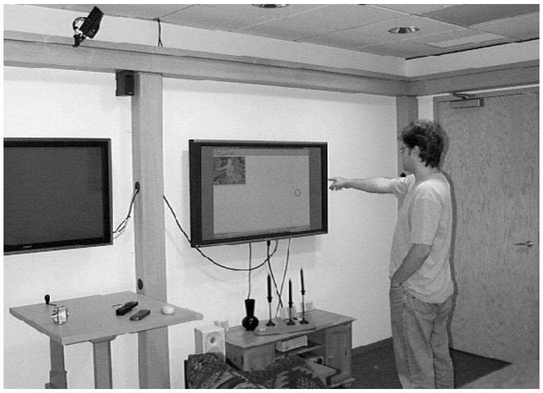

Figure 4: EasyLiving measures where a person is pointing to position a cursor

3.1.3 Gaze

People often look toward the object they are talking about, such as, "Close the shades on that window." Like pointing, gaze is a way to communicate a location. Gaze measurement is an ongoing research area, mostly for users sitting directly in front of a monitor.

3.1.4 Speech

From a user's point of view, speech is one of the easiest and potentially most effective ways of intentionally communicating with a smart environment. In EasyLiving we have shown how a user, wearing a microphone, can give commands to the room such as turning on the lights. One problem is finding a way to tell which utterances are intended as commands to the room and which are regular speech. In our system, the user has to first address the room by name, then give the command.

Speech input would be much more convenient if the user did not have to wear a microphone. Arrays of microphones placed around the room are close to solving this problem.

3.2 Perceiving Objects for UI

There may be specialized objects in the room that can be used for UI. These are different from the room's "normal" objects which can be used independently of the room's perceptive abilities. The most basic of the specialized objects is a simple remote control for a television or audio system.

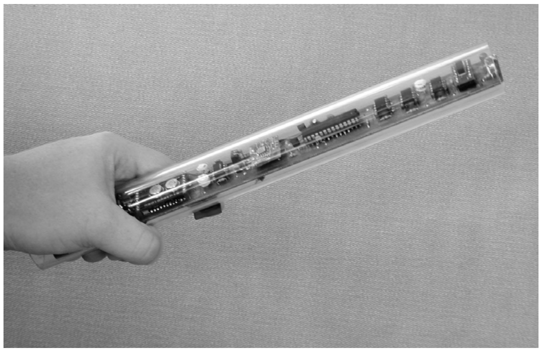

Figure 5: Prototype XWand used for controlling devices in smart room.

3.2.1 XWand

EasyLiving's XWand is a novel wireless sensor package in the shape of a wand that senses its own orientation with respect to the room. A prototype is shown in Figure 5. The idea is to use a very simple UI to control multiple devices, and rely on the intelligence of the room to determine what the user is trying to do. One application is in device control. A user can select a device by pointing at it with the wand, and perform an action on the device by making a gesture with the wand, or potentially speak a command in combination with the pointing gesture. In this case, a host computer with an RF receiver performs gesture recognition and decides on an appropriate course of action given that the wand is pointing at a particular device.

3.2.2 Phicons

Pioneered by Hiroshi Ishii of MIT's Media Lab, phicons are physical icons – physical objects that represent data or commands. In EasyLiving we have used phicons to give commands in two different ways. One allowed a person to rotate a colored cube in view of a camera. Rotating the cube would allow the user to select commands for a media player, such as "stop", "start", and "fast forward". Another phicon is a virtual mouse shown in Figure 6. This technology lets a user control the position of the cursor on the screen by moving his finger on a pad of paper. The finger and pad are tracked by a camera.

4. Privacy

More perception increases the potential usefulness of a smart environment, but it also increases concerns about privacy. There are things that are useful to perceive but that should not be made accessible to outside parties. Even something as simple as activity detection could indicate when a house is safe to burglarize. Solutions include encrypting data and blocking access to the environment's network from outside.

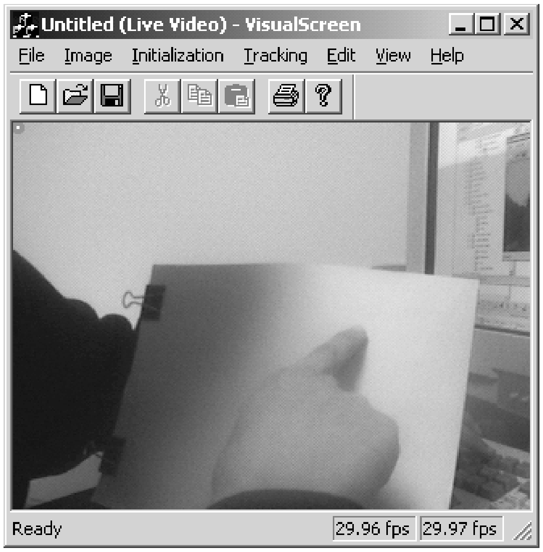
Untitled (Live Video) - VisualScreen
File Image Initialization Tracking Edit View Help
Ready
29.96 fps
29.97 fps

Figure 6: The virtual mouse lets a user control the position of the cursor on the screen by moving his finger on a pad of paper.

5. Conclusion

Perception is clearly a vital part of smart environments. With perception an environment can respond intelligently to its occupants and let them access computing in novel ways. This paper cataloged different types of perception by splitting the tasks along the lines of perceiving people and objects for both context and UI.