Chapter 4: Data Collection & Methods

Our lives are not random. They certainly exhibit structure at all time-scales. How is this structure organized? What are its atomic elements? What is the network of dependencies connecting the past, present, and future moments? These are big questions, which we cannot address completely. However, through a few guiding principles (which we describe next) we can limit our analyses to an appropriate level-of-detail, thus enabling us to reasonably tackle the above questions. Since these questions need hard data to produce answers, we also address how to collect measurements of an individual's experiences and describe how we overcame this hurdle.

When the detective tries to understand the mind of the criminal, he attempts to place himself in the criminal's state of mind, duplicating the experiences and encounters that the criminal might have had up to and including the scene of the crime. This makes it possible for the detective to infer the missing pieces of evidence and perhaps predict the criminal's next move. Mapping this intuition to the case of a computational agent (the detective) and its user (the criminal), means that we should provide the computational agent the same inputs the user is receiving so as to allow the agent to understand or habituate to the experiences and thus perhaps predict future experiences of its user. This implies that we use wearable sensors that are unobtrusively integrated into the user's clothing. Furthermore, we will concentrate our efforts on sensors that parallel biological perception: vision, audition, and vestibular. Lastly, we must be able to capture the subject's experiences for as long as is reasonably possible. First-person, long-term sensor data is a guiding principle for our overall approach.

The second principle guiding this work is the use of peripheral or context-free perceptual methods. The definition of a context-free method is an algorithm or system that is effective in any context, thus independent of lighting, background auditory conditions, etc. Context-free methods generally rely on global features such as color histograms or optical flow in vision and spectrograms and peak tracking in the audio to provide useful descriptions of the raw sensor data. Thus if we have a speech detection module that operates robustly in most conditions, we can include it as a context-free perceptual method. Contrast this with the use of attentive or context-specific perceptual modules, such as today's state-of-the-art speech recognizers [55] [56] [7] [20] or face recognition systems [35] [53] that require knowledge of the current context in order to operate.

The third guiding principle of this work owes its inspiration to insect-level perception. It is inappropriate given the current state of the art to tackle the problem of how to give a machine human's level understanding of an individual's daily behavior without first granting it with an insect's level of understanding. Perhaps in certain cases, we can obtain near-human understanding by severely restricting the domain. However, in this work the completeness of the domain, that is an individual's day-to-day life, is a priority and hence we are guided to the more appropriate level of perception portrayed by insects. Similar to the representation-free approach of Rod Brooks, we avoid building complete models of the user's environment and instead rely on the redundancies in the raw sensor data to provide the structure. This philosophy implies the use of coarse level features and emphasizes robustness over detail (such as in the use of context-free methods over context-specific methods).

In this work we took a straightforward approach to addressing the issues of a similarity metric and temporal models of life patterns. We collected long-term sensor measurements of an individual's activity that enables the extraction of atomic elements of human behavior, and, the construction of classifiers and temporal models of an individual's day-to-day behavior. I will describe this data set and then describe in more detail methods for building coarse descriptions of the world, and thus a similarity metric. Last, we describe methods for extracting temporal models based on these coarse event descriptions.

4.1 The I Sensed Series: 100 days of experiences*

The first phase in statistically modeling life patterns is to accumulate measurements of events and situations experienced by one person over an extended period of time. The main requirement of learning predictive models from data is to have enough repeated trials of the experiment from which to estimate robust statistics. Experiential data recorded from an individual over a number of years would be ideal. However, other forces such as the computational and storage requirements needed for huge data sets force us to settle for something smaller. We chose 100 days (14.3 weeks) because, while it is a novel period for a data set of this sort, its size is still computationally tractable (approx. 500 gigabytes).

* The term "I Sensed" comes from a piece of historical conceptual art that has played a part in inspiring this thesis. In the 70's there was a Japanese conceptual artist named Kawara On [26] O. Kawara, *On Kawara: date paintings in 89 cities*, Museum Boymans-Van Beuningen, Rotterdam, 1992., who was in a way obsessed with time and the (usually) mundane events that mark its passage. His works such as the *I Met* and *I Went* series explored the kind of day-to-day events that tend to fall between the cracks of our memories. For years, everyday Mr. On would record the exact time he awoke on a postcard and send it to a friend or create lists of the people he met each day or trace on maps where he went each day. Other relevant works are his *I Got Up At*, *I Am Still Alive*, and the *I Read* series. His work raises a few interesting questions. If we had consistent records of some aspects of our day-to-day lives over a span of a lifetime, what trends could we find? What kinds of patterns or cycles would reveal themselves? Interestingly, we wouldn't need highly detailed memories to find these trends and patterns, just a consistent sampling in time. One of my dreams is to build a device that can capture these life patterns automatically and render them in a diary-like structure.

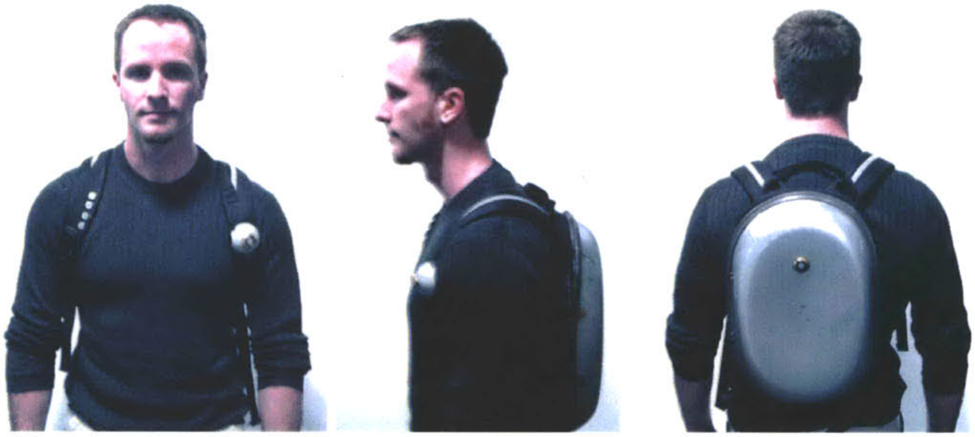

Figure 4-1: The Data Collection wearable when worn.

The wearable was worn from mid-April to mid-July of 2001 by the author. Refer to Figure 4-5 for actual excerpts from this data set during 4 example situations: eating lunch, walking up stairs, in a conversation, and rollerblading.

We designed and followed a consistent protocol during the data collection phase. Data collection commences each day from approx. 10am and continues until approx. 10pm. This varies based on the sleeping habits of the experimental subject. The times that the data collection system is not active or worn by the subject is logged and recorded. Such times are typically when: batteries fail, sleeping, showering, and working out.

In addition to the visual, aural, and orientation sensor data collected by the wearable, the subject is also required to keep a rough journal of his high-level activities to within the closest half hour. Examples of high-level activity are: "Working in the office", "Eating lunch", "Going to meet Michael", etc. while being specific about who, where, and why. Every 2 days the wearable is "emptied" of its data, by uploading to a secure server.

Persons who normally interact with the subject on a day-to-day basis and have a possibility of having a potentially private conversation recorded are asked to sign a consent form in which we formally agree to not disclose recordings of them in anyway without further consent. This way my data collection experiments were in full accordance with the Massachusetts state laws on recording audio & video in public.

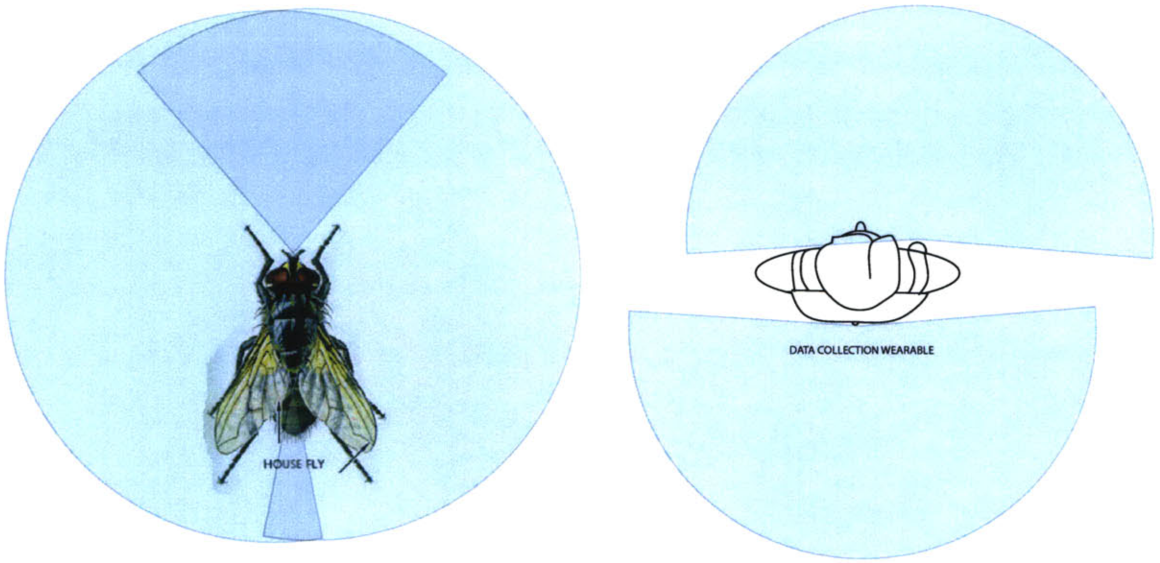

Figure 4-2: Comparison of the field-of-view for the common household fly and the data collection wearable used in the I Sensed series.

4.1.1 The Data Collection Wearable

The sensors chosen for this data set are meant to mimic insect senses. They include visual (2 camera, front and back), auditory (1 microphone), and gyros (for 3 degrees of orientation: yaw, pitch and roll). These match up with the eyes, ears, and inner ear (vestibular), while taste and smell are not covered because the technology is not available yet. The left-right eye unit placement on insects differs from that front-back placement of the cameras in our system. However, they are qualitatively similar in terms of overall resolution and field-of-view (see Figure 4-2). Other possibilities for sensors that have no good reason for being excluded are temperature, humidity, accelerometers, and bio-sensors (e.g. heart-rate, galvanic skin response, glucose levels). The properties of the 3 sensor modalities are as follows (see Figure 4-4):

Audio: 16kHz, 16bits/sample (normal speech is generally only understandable for persons in direct conversation with the subject.)
Front Facing Video: 320x240 pixels, 10Hz frame rate (faces are generally only recognizable under bright lighting conditions and from less than 10ft away.)
Back Facing Video: 320x240 pixels, 10Hz frame rate (faces are generally only recognizable under bright lighting conditions and from less than 10ft away.)

Orientation: Yaw, roll, and pitch are sampled at 60Hz. A zeroing switch is installed beneath the left strap that is meant to trigger whenever the subject puts on the wearable. Drift is only reasonable for periods of less than a few hours.

The wearable is based on a backpack design for comfort and wardrobe flexibility. The visual component of the wearable consists of 2 Logitech Quickcam USB cameras (front- and rear-facing) modified to be optically compatible with 200° field-of-view lenses (adapted from door viewers). This means that we are recording light from every direction in a full sphere around the user (but not with even sampling of course). The front-facing camera is sewn to the front strap of the wearable and the rear-facing camera is contained inside the main shell-like compartment. The microphone is attached directly below the front-facing camera on the strap. The orientation sensor is housed inside the main compartment. Also in the main compartment are computer (PIII 400Mhz Cell Computer) with a 10GB hard drive (enough storage for 2 days) and batteries (operating time: ~10 hrs.). The polystyrene shell (see Figure 4-1) was designed and vacuum-formed to fit the components as snuggly as possible while being aesthetically pleasing, presenting no sharp corners for snagging, and allowing the person reasonable comfort while sitting down.

Since this wearable is only meant for data collection, its input and display requirements are minimal. For basic on/off, pause, record functionality there are click buttons attached to the right-hand strap (easily accessible by the left-hand by reaching across the chest). These buttons are chorded for protection against accidental triggering. All triggering of the buttons (intentional or otherwise) is recorded along with the sensor data. Other than the administrative functions, the buttons also provide a way for the subject to mark salient points in the sensor data. The only display provided by the wearable is 2 LEDs, one for power and the other for recording.

4.1.2 The Data Journal

Organizing, accessing, and browsing such a large amount of video, audio, and gyro data is a non-trivial engineering task. So far we have a system that allows us to fully transcribe the "I Sensed" series and to access it arbitrarily in a multi-resolution and efficient manner. This ability is essential for learning and feature extraction techniques talked about later in this paper. All data (images, frames of audio, button presses, orientation vectors, etc.) are combined and time synchronized in our data journaling system to millisecond accuracy (see Figure 4-3).

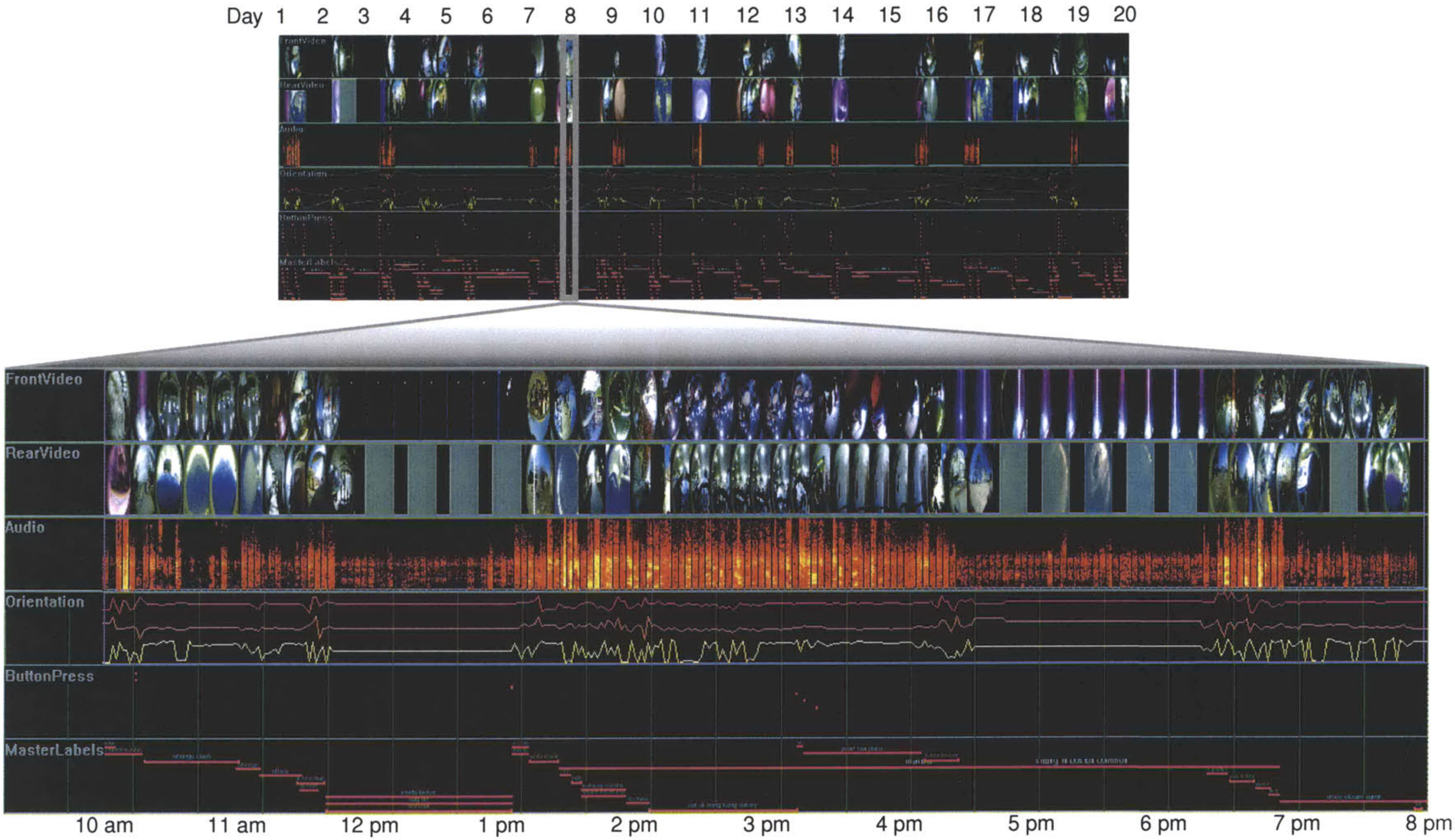

Figure 4-3: The Data Journal System: provides a multiresolution representation of the time-synchronized sensor data.

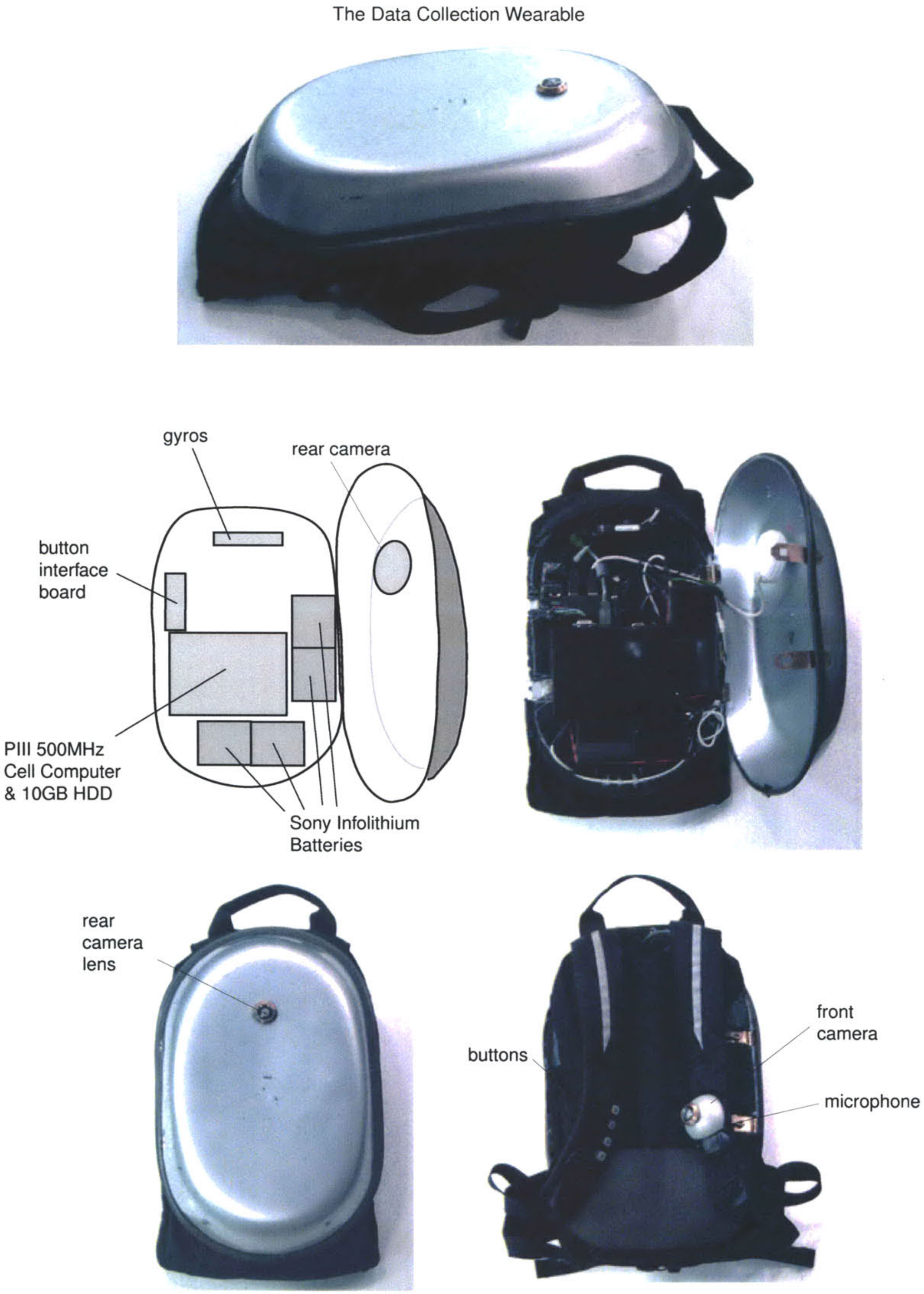
The Data Collection Wearable
gyros
rear camera
button interface board
PIII 500MHz Cell Computer & 10GB HDD
Sony Infolithium Batteries
rear camera lens
front camera
buttons
microphone

Figure 4-4: The Data Collection Wearable Schematic

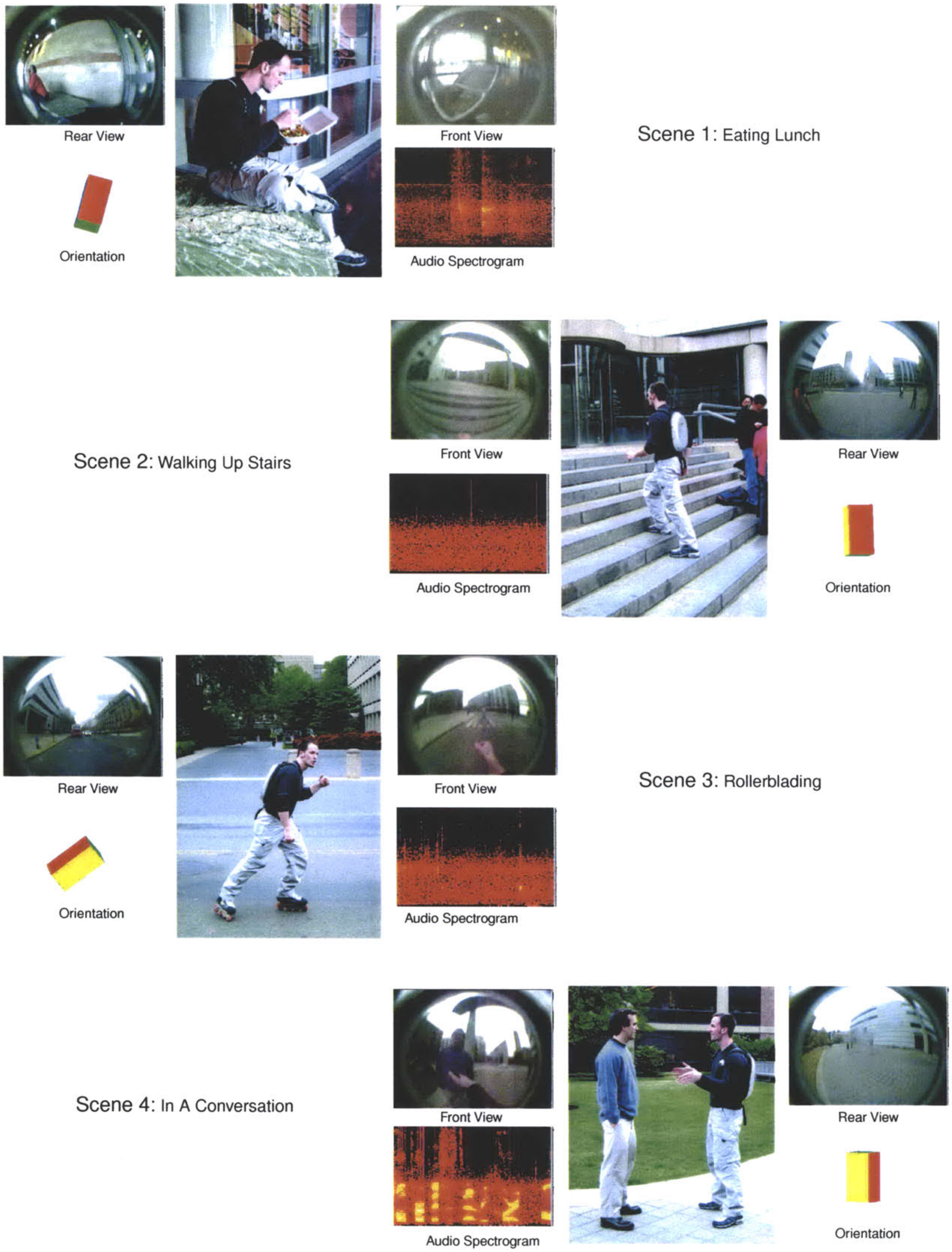

Figure 4-5: Some excerpts from the "I Sensed" series

Chapter 5: The Similarity Measure

Before we can answer any of the questions about classification, prediction or clustering, we first need to determine an appropriate distance metric with which to compare moments in the past. We will look at how to determine what are the appropriate intervals to be comparing and how to quantify their similarity. While doing so we present new methods for data-driven scene segmentation. We will then present methods for determining the similarity of pairs of moments that span time-scales from seconds to weeks. The tools we build up in this chapter provide the foundations for classification and prediction.

5.1 The Features

The first step in aligning sensor data is to decide on an appropriate distance metric on the sensor output. Possibly the simplest similarity measure on images is the L_1 norm on the vectorized image. Computer vision researchers typically avoid using such a simple metric because of its vulnerability to differences in camera position and orientation and opt instead for orientation-invariant representations, such as color histograms or image moments. However, as mentioned before there is clear evidence [54] that insects (and in many cases humans) store view-dependent representations of their surroundings for later recall and matching. In this case the dependency of the image and the camera position and orientation is an advantageous one. Throwing away the information that links an image to the state of the camera at the moment of capture doesn't make sense when the task is to situate the camera wearer.

There is an interesting side-note on the choice of the exponent in the Minkowski metric. Researchers in biological perception have noticed repeatedly that simple creatures such as insects (particularly bees) appear to use an L_1 norm on visual discrimination tasks, but as the creature gets more complex (say humans) they discretely switch between the L_1