Probable Path Inference for GPS Traces in Cities

Elena Agapie, Jason Ryder, Jeff Burke, Deborah Estrin

e.agapie@jacobs-university.de, jryder@cs.ucla.edu, jburke@remap.ucla.edu, destrin@cs.ucla.edu

Center for Embeded Networked Sensing, University of California, Los Angeles

ABSTRACT

This paper introduces a method for estimating the path of a person for Global Positioning System (GPS) traces. There are many sources of measurement error in location time-series data from GPS devices: undersampling, precision limitation, etc. Gaps in GPS data occur when the device loses reception or when it is turned off. In a technical pilot, 20 personal location traces were gathered during a four-day period using consumer GPS units; the individual traces obtained covered various periods of time, ranging from 6 to 12 hours per day. Several Geographic Information System (GIS) tools were used to process and visualize the collected data. Roadnav (described in section 2.5) was used as a basis for the map structure on which the algorithm for path prediction was built. All the data were related to the Topologically Integrated Geographic Encoding and Referencing (TIGER) map files that Roadnav uses to represent streets. The A* (A-star) algorithm, used for best path detection, was applied to estimate the path between extremity points of a GPS gap. For A* to use the data gathered with GPS units, preprocessing had to be run. Then we inferred values regarding traveled distances, average speed, and percentage of time spent in different zones of the city (residential, transportation, commercial). Significant issues about the accuracy of the information extracted from the GPS data are discussed. We made inferences about location during gaps that lead to more accurate calculation of other inferred properties such as the time spent in different zones.

Key Terms

GPS, Location computation, Path Profiling, Map-matching, A* algorithm

1. INTRODUCTION

The path inference problem has appeared in the context of the Personal Environmental Impact Report (PEIR) project. The PEIR project supports the assessment of personal and environmental impact and exposure. Data is collected from a person using the phone and the GPS unit with the purpose of determining a level of exposure to different types of pollution and a measurement of the impact an individual has in the local environment. This is done by building a location **trace** of the individual based on GPS location data, during a certain period of time and extracting information regarding his routes and the means of transportation used.

As was expected, while processing the acquired data, multiple **gaps** with no location information were discovered. Gaps appear either because the GPS unit loses reception (e.g. inside buildings or in hilly and downtown areas) or because the unit is turned off. Gaps can interfere with the inference process used to generate feedback. This applies to detecting the areas where the person is spending its time. The gaps in GPS data need to be resolved in order to acquire accurate data on the amount of *time spent within a zone* (e.g. commercial or transportation zones).

During a technical pilot, GPS units and Nokia N-series cell phones were used to gather data. The daily traces were differentiated according to the ID of the unit that gathered the data. Data was processed through several GIS tools described later in the paper. In order to handle periods for which no location information was obtained, we implemented an algorithm to reconstitute the trace in the period of the gap. The main technique used was based on **approximations of the speed** the subject had in periods before or after the gap, and in different zones on the overall trace. The processed data was input into an A* algorithm. This algorithm was implemented on the platform of the Roadnav software, which is able to relate the data to a representation of an associated road map.

2. METHODOLOGY

A *campaign for gathering data* was carried out to collect data from a group of subjects over a period of four days between the fixed times of 7am-10am and 4pm-7pm. This time interval was chosen to provide information on a person's location during rush-hours which can be applied to future studies in the PEIR project dealing with pollution issues.

Each person was equipped with a Nokia phone and a GPS unit, which the subjects were supposed to turn on during collection hours. All the data was uploaded to a single location in SensorBase and different traces were evaluated based on a daily period, on multiple days or am/pm intervals. Data was extracted from SensorBase, processed in ArcGIS and tagged with zoning information that was processed for location trace analysis. The obtained data was visually queried with GeoServer, and computations on different fields were done with Matlab. The algorithm for path profiling was built on the Roadnav structure.

2.1 Campaignr Processing

The data collected with the phone is transferred into a database through Campaignr, a software installed on the Nokia phones. To run a data collection campaign an XML file has to be added to the phone. This describes which phone sensors are used, the sampling interval and where to upload the acquired information. The data is uploaded to SensorBase in real time or when the user is connected to a network that allows access to the internet.

2.2 SensorBase Processing

The PEIR project uses the following data gathered from phones:

1) phone ID,
2) date and time stamp,
3) latitude and longitude at every moment of time,
4) speed,
5) accuracy on location coordinates and speed,
6) number of satellites used.

The data is kept in the phone and then uploaded to a database (SensorBase) where it can be queried through a web interface but not visualized.

All data concerning location is stored in SensorBase in a JavaScript Object Notation (JSON) format that can easily be parsed to obtain the location and speed values. The data is extracted from SensorBase and parsed into CSV (comma separated values) files for processing with ArcGIS.

2.3 ArcGIS Processing

ArcGIS is a software tool that allows one to create maps, perform some spatial analysis, and view and query maps. It also allows the use of predefined layers of maps (such as highways, zone types and pollution related maps) that one can intersect with traces from the data.

The CSV files where data is transferred for processing are input to ArcGIS to tag data with different location zones (such as residential, transportation, commercial, etc) according to the coordinates of the point on the map. Different queries are also made in ArcGIS based on speeds during travel and different map layers on which the GPS information is overlaid. For the purpose of the project, the files that were processed in ArcGIS were further used for evaluating the data.

2.4 GeoServer and Geospatial Queries

GeoServer is an open source server that connects location data with a variety of map formats, images or geospatial data. The primary use of GeoServer is to overlay the traces with the Google Maps application and perform different types of queries on the data.

The data is uploaded into a PostgreSQL database. This is done either by using a script that uploads data from SensorBase automatically or at periodic time intervals, or by uploading the data from the processed CSV files. Both options will allow queries on time periods (data from a certain day, from certain time intervals of the day, multiple time intervals of multiple days). Furthermore, spatial queries can be made to intersect the trace with a given zip code or to overlay multiple traces distinctly. The data from the CSV files also permits the addition of zone tags into the query. As a result, any of the

time queries above can be intersected with one or more given location zones (see Figure 2.4a and 2.4b).

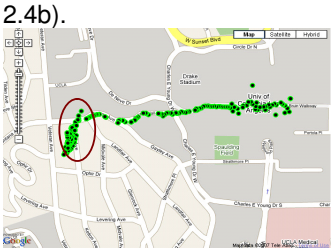


Figure 2.4a A trace with multiple points in residential area

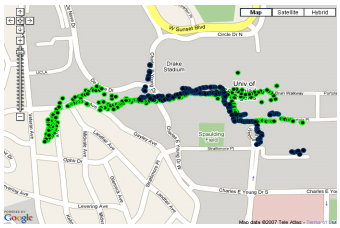


Figure 2.4b Multiple traces overlaid

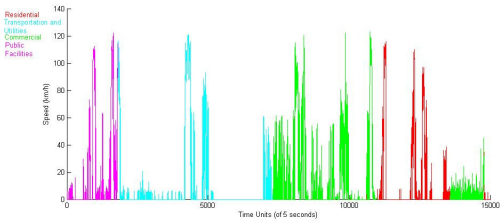
2.5 Roadnav and the Road Network

Roadnav is anopen source software that deals with path representations and shortest distance directions. The road map used to map the GPS data comes from Roadnav. The software uses TIGER maps for the representation of the US road system. It also allows the user to set preferences for finding the shortest path through different types of roads. The software was used to implement the A* algorithm for finding the most probable path based on the GPS information. The road map representation provided by the program is used as a support for the GPS data.

Roadnav caption

3. LOCATION TRACE ANALYSIS

The first purpose of trace analysis is to obtain information related to the **time spent in different zones** of the city. The **zones** that we are dealing with are Commercial, Public Facilities, Transportation and Utilities, Residential, Vacant, Open Space and Recreation, Under Construction, Agriculture, and Industrial.



Variation of speed through zones

A problem arises when there are GPS data with periods of time that do not have location information or no data is present for some interval of time. We classify a set of temporarily consecutive samples (and/or absence of samples) of this nature as a **gap period**. Some assumptions have to be made when dealing with this data. This will be discussed in the following paragraphs. An alternative to using only basic statistics on the available data is to use path inference as an additional method to find zoning information during gap periods.

3.1 Inferences about percentage of time spent in different zones

The first method for inferring the percentage of time in each zone is to look at the data samples that are tagged within the zone and compute an average of **time spent in the zone**.

If two consecutive readings are tagged with different zones, the difference of time between the samples is split between the two zones. The impact of this approximation depends on the frequency of position sampling.

For gap periods in traces, which could be due to loss of satellite signal, communication error, or device being switched off, we apply the following rules:

- if the samples at the start and at the end of the gap are in the same zone, we assume that the device was in that zone during the gap;
- if the samples are in different zones, we consider it in the starting zone half of the time and in the ending zone half of the time (as an example see Figure 3.1a).

Different values for the gap are evaluated. Regardless of the length, we associate a **tolerance** of 5, 10 or 30 minutes to gaps and consider this tolerance as part of a certain zone while ignoring the rest of the gap. For a gap of size *x,* and a tolerance of size *t* we will consider:

size(x) = x, if x < t
t, else

This associates periods of the gaps to certain zones. It gives an estimate of how to associate the gap periods to zones. Since it is not a very robust method, the results are different for each tolerance level If the gap is longer, the difference in result obtained for each tolerance will be greater (see Figure 3.1b).

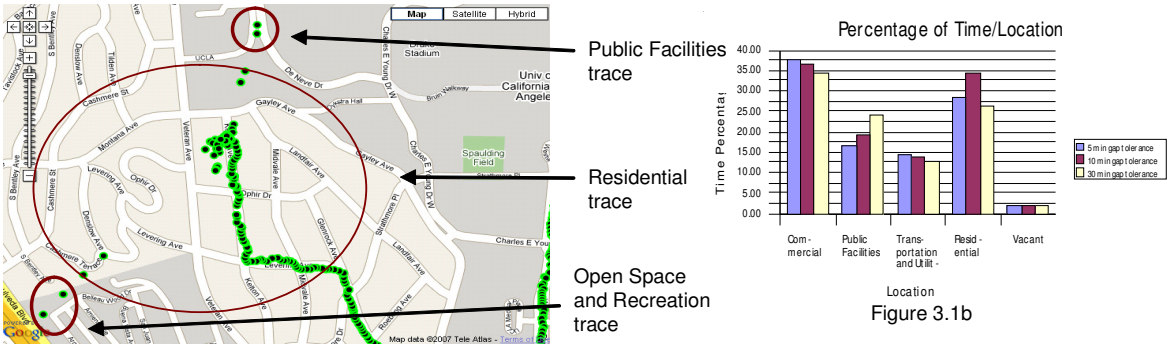


Figure 3.1a Points and zones they belong to

Figure 3.1b

3.2 Discussion

Since there are a significant number of gaps in the GPS data, we explore a more precise means of estimation. Inferring the most probable path during these periods seems to be an alternative that would provide enough information. We claim that determining the precise path of the subject will give more information on the exact zone that was visited. The path inference algorithm is presented in the next section.

4. PATH INFERENCE

Considering the above results, an algorithm is implemented for path inference. The GPS data collected was related to an existing representation of a road map implemented through the Roadnav open source software. The A* algorithm was used to estimate the missing path within a gap. In order to apply the algorithm to the GPS data **speed analysis** was done.

4.1 A* algorithm

The A* algorithm is used to find the **best path** between two nodes of a graph. The algorithm uses a heuristic to determine the best solution at every point. The priority of a path is determined by a cost function

$$f(x) = g(x) + h(x),$$

where

- *f* is the cost function at point *x,*
- *g* is the cost of the path from the starting point to *x* and
- *h* is the heuristic used to estimate the cost from *x* to the destination point.

The smallest cost path is always chosen. A* runs in polynomial time if the heuristic satisfies

$$| h(x) - h^*(x) | \leq O(\log h^*(x))$$

where h^* is the optimal heuristic. The optimal heuristic is the real cost to get from point *x* to the ending point [8].

4.2 Heuristic for Path Estimation

The algorithm finds the **shortest path** between two points by choosing the most probable segment of road followed at every step. The heuristic uses speed constraints from the GPS data and from the road map that is available in Roadnav. We define the following variables that are used in the heuristic:

Average speed in zone is the average speed that the person has in the different zones previously mentioned. It has been computed based on partial traces during four days.

Expected traveling speed is the estimated speed that a person was having during the period of the GPS gap. This is obtained by looking at the average speed on a time segment before and after entering the GPS gap and at the average speed in the zone corresponding to the gap. The zone of the gap is defined by the zone where the extremity points are located.

The **average traveling speed** is computed by taking the average of the three aforementioned speeds. We define a **margin** of three minutes before and after the gap for computing the speed at the extremities. The average speed in this time interval is computed by considering the speed from any GPS recording, including the ones that are static. The average speed per zone also includes gaps of less than two minutes. If the extremity points of the gap are in different zones the speed corresponding to the zone of the gap is averaged from the speeds of the two zones.

Straight line speed is the speed computed from the information regarding distance between the extremity points of the gap and the length of time of the gap. If the value of this speed is too small, we can deduce that either the period of time was too long or the distance was too small. If the distance is small, it could be inferred that the user did not travel from the initial location. If the time is too long compared to the distance, the information regarding the path may be insignificant. The subject could have been relatively static, or could have traveled and returned.

Roadnav defines different types of roads based on their coding from the TIGER files. These are defined as small roads, large roads or highways (each can be one-way or two-way). Some default **road speeds** are defined for every type of road as follows: 25 mph for small roads, 45 mph for large roads and 65 mph for highways.

The heuristic function *h* is defined as

$$h(x) = l * \mathrm{abs}(rSp - trSp)$$

where

- *rSp* is the road speed and
- *trSp* is the average traveling speed, based on the type of road where the gap extremity points are located and *l* is the length of the current segment of road.

This makes the heuristic function return a smaller value as the traveling speed is closer to the road speed.

4.3 Distance and speed analysis

4.3.1 Distance analysis

Distance computation is necessary for determining **straight line speed** between two points. Given the distance between the extremities of a gap and the time difference, the estimated speed between the points is computed.

The data was processed to give decimal coordinates of latitude and longitude for every point. From these coordinates the distance was computed using the **Haversine Formula** that is based on the great-circle distance. This is assumed to give an error of about 0.5% when used with the sphere radius of 6372.795 km (Earth's radius). The distance is defined as

$$\Delta\sigma = 2\sin^{-1}\left(\sqrt{\sin^2\left(\frac{\phi_f - \phi_s}{2}\right) + \cos\phi_s \cos\phi_f \sin^2\left(\frac{\lambda_f - \lambda_i}{2}\right)}\right)$$

where $\phi_s, \lambda_s, \phi_f, \lambda_f$ are geographical latitude and longitude points.

4.3.2 Speed analysis

Speed computations were made for the purpose of determining the average speed in a certain zone or on a certain segment of the travel. Problems appeared when taking into consideration the speed through the entire

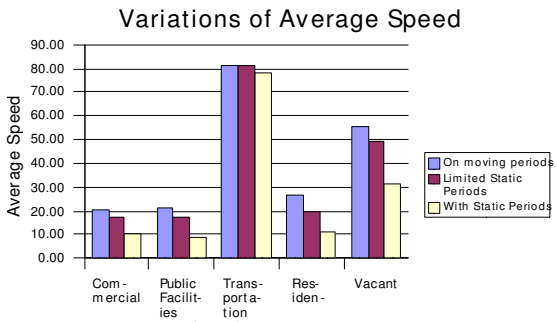


Figure 4.3.2a Average speed computed with different conditions

segment or only considering the non-stationary speed.

There may be segments of time when speed is zero because the person is intentionally residing at a location and should not be added to the average speed on the segment of time. When this occurs that should be considered the end of a segment. However, there may also be periods when the person stops for red lights and this significantly influences the average speed.

Figure 4.3.2a shows the variations of the speed when *static speed* is taken into consideration.

- The highest average speed was when the static periods were ignored, and it decreased as speed was computed using limited static periods (these had a range of less than two minutes; if the stop was longer then it was discarded).
- The lowest average speed occurred when the static periods were entirely taken into consideration.

Another observation on the graph is that the difference between the variations of speed in the three categories can give information on how much time a person spends moving through a certain zone. For example the variation in the Transportation and Utilities area is very small which leads to the conclusion that the person had limited static periods in those zones. In the Vacant zone the variation is significantly higher which means that more stationary periods were encountered.

4.4 Examination and behavior modeling of the algorithm

The algorithm was tested on a set of traces from one of the subjects. The data contained information from all types of zones. Processing of the data was conducted and the following variables were made available: the coordinates of every GPS gap, the average speed at the extremities and the duration of the gap. Only the gaps greater than one minute in length were tested. Gaps were also simulated by deleting samples of data and simulating gaps.

A small number of cases were tested for the moment. Further analysis has to be made on the available data. What has been visually observed is that in cases when the subject had a high speed, the path inferred used mainly highways, when they were available. When the speed was low, the path mainly included small roads.

The algorithm does not consider situations like the occurrence of consecutive gaps with a very short interval between their extremities. In this case, the information received from the speed analysis is not accurate, since the speed around the extremities of one gap will include a distinct gap from which no relevant information can be extracted. Instead, both gaps should be merged into single one that can use information from the extremities.

The results obtained for gaps over long distances might not be very precise compared with the real path. The average speed of the extremities could change when the person is traveling over long distances and on various types of roads. The end-point information is not sufficient for predicting a long path. The algorithm only deals with car traces. If a subject is walking and taking paths that are not available on the road map, the path cannot be computed.

4.5 Future work

Path prediction between zones can undertake the following refinements by using the boundary between the zones. First, a route can be chosen from the first extremity of the gap to the boundary and then another route is chosen from the boundary to the second extremity point based on each zone's information. Second, the zone speed on this transition section can be computed by taking the distance between the extremity points as a straight line and then weighting the average speed according to the percentage of the line that lies in each zone.

The maximum speed of the subject on different types of roads should also be taken into consideration. The subject might travel with a speed of 50 mph on highways, but that would not mean that he will actually prefer large roads. Road speeds can be set for every particular user according to his speed pattern.

The future step for choosing the route is to make use of the speed limit of every road as well as analyzing the variation between a person’s average speed and the speed limit. Using this approach a better prediction can be made on how the speed will vary on a specific road. Also, we can assume that

red lights or stop signs are found at regular intervals. As a result we can add stationary speed on a periodic basis.

The distance between the extremity points of the gap can also be used when making a decision about a path. Besides considering speed as a constraint for the algorithm, we can also compute the path that is closest in distance with the one we are expecting. We can assume the expected distance is computed by considering the shortest distance between the two extremities, the average road speed on this distance and the time length of the gap. A length of the path would be approximated and this can be used to choosing the probable path by selecting the one with the length closest to the above estimated distance.

On large sets of data some filtering techniques should be used. If sampling would be done at very short periods of time, like one second, there would be a large set of data that could contain redundant information. Instead, the data should be reduced by considering some algorithm on the information received from the data.

Additional information that can be extracted from the GPS data concerns the heading of the subject. Looking at the direction at the extremity points can lead to an inference on the possible heading of the subject in between the points and several paths can be excluded along the path.

5. RELATED WORK

Many studies have been made on GPS systems and mobile services. Most of the work focuses on methods for building accurate road networks, road mapping techniques, shortest path algorithms and sensor errors. Looking into the accuracy of both the data coming from the GPS units, the road map and the way they relate with each other was very helpful for our work.

Some important issues in map matching are determining the exact location of a point with respect to the map and correctly identifying the road of travel (see [3]). Previous information from the GPS-estimated position is frequently used. Probabilistic and geometric approaches are used to determine the likelihood that a point is situated on a certain road.

Furthermore, algorithms such as variations of Dijkstra, lane clustering or road segmentation are used for determining the shortest path between two points or for map-matching. This comes up either in the context of finding direction between two locations or in finding a satisfying road to map points from the GPS data. The GPS routing problem has been approached by geometrical or statistical means (see [9]). All the methods have to deal with noisy data.

Techniques for tracing moving objects are used to predict future position of a mobile device. User profiles have to be built regarding speed and acceleration or other known data. Many algorithms deal with inferring and learning transportation routines. Algorithms are often based on clustering and need historical data that would be used for setting up user patterns on the paths.

Different types of sensor errors are closely related to our work. Estimating the position reported by the GPS unit can be affected by map errors that occur from an incomplete or erroneous representation of roads (intersections, one-way streets, legal turnings). Although GPS has a high accuracy, the coverage is reduced as a result of weak satellite geometry, narrow streets, hilly or downtown areas. In large buildings no receptions might be acquired. Distance measurement errors can affect the location of a subject on a street or a lane, especially in dense urban areas. Furthermore, distance measurement yields unclear results when dealing with roads versus GPS traces. Since the road network is usually based on segments of roads and a vehicle is usually correlated to the center of the lane, a standard error can be computed with respect of the GPS deviation from the real road.

Our work needs accurate information on all the above issues. Our study does not focus on fixing any map-matching or map representation errors. Instead, all the map-matching techniques are used as Roadnav implements them. The main focus is to find the best heuristic of choosing the path based on the available data. The data is not gathered on a long period of time so there is no significant user history with regards to the places he is visiting. However we are trying to use as much of the information we have, resulting in a speed profile set for the subject and the zones he is visiting. GPS

errors are one of the most important factors in our results and that is where the concern for our main problem comes from. The gaps in GPS data resulted from different type of sensors presented above are the main focus of the study. The information from GPS data is processed such that time, distance and speed constraints can be used on inferring a probable path during the gap. The A* algorithm is used for computing the best distance between two points. It is applied on the extremities of the gap with speed constraints coming from the user and from the current road such that it produces a probable path using much of the information available.

6. CONCLUSION

This paper studies methods and problems that appear when dealing with data from GPS units in periods when no location information is available. Although most of the data gathered by GPS units contains relevant information on the location of the subject gathering the traces, there is still a large set of gaps in this data that make a difference on the overall analysis of the set of data. The traces obtained from the GPS are used for a further analysis. Our main application is looking at the percentage of time a person spends in different zones of the city. Because it was observed that a significant amount of time had no information regarding the zone where the subject was located, the gaps were analyzed. The main focus was to locate the subject on a path during the gap periods. The A* algorithm was used to determine the best path between the extremities of a gap. The heuristic of the algorithm was based on speed constraints. This implied a set of analysis on speed during different segments of time, zones and speed from the perspective of the road map we are using.

Several issues have appeared during the analysis. There are numerous situations that a subject can face while traveling and have to be considered (like the random occurrence of static periods). We are focusing the analysis on the speed constraints and getting accurate estimations that A* can use to decide the best route. There are several solutions proposed, e.g. considering heading, speed limits of roads, boundaries of zones or maximum speed. Further experiments should also be conducted on a larger set of data. Also, we believe that a data collection campaign should be run for gathering data on a longer period of time and analyzing patterns.

7. ACKNOWLEDGEMENTS

We would like to thank NSF and the Center for Embedded Networked Sensing at UCLA for supporting this research. Also we want to thank all the researchers from the Urban Sensing team and to Ryan Rosario from the UCLA-Statistics Department.

8. WORKS CITED

[1] Wu Chen, Zhilin Li, Meng Yu, and Yongqi Chen, Effects of Sensor Errors on the Performance of Map Matching, *The Journal of Navigation*(2005), 58, 273–282.
[2] Alminas Civilis, Christian S. Jensen, Techniques for Efficient Road-Network-Based Tracking of Moving Objects, *IEEE Transactions on Knowledge and Data Engineering*, VOL. 17, NO. 5, MAY 2005
[3] G. R. Jagadeesh, T. Srikanthan and X. D. Zhang, A Map Matching Method for GPS Based Real-Time Vehicle Location, *The Journal of Navigation*(2004), 57, 429–440.
[4] Cesar A. Quiroga, Darcy Bullock, Travel time studies with global positioning and geographic information systems: an integrated methodologyTransportation Research Part C: Emerging Technologies, Volume 6, Number 1, February 1998
[5] Changqing Zhou, Dan Frankowski, Pamela Ludford, Shashi Shekhar, Loren Terveen, Discovering personal gazetteers: an interactive clustering approach, *Proceedings of the 12th annual ACM international workshop on Geographic information systems*
[6] Daniel Ashbrook, Thad Starner, Using GPS to learn significant locations and predict movement across multiple users, *Springer London* , Volume 7, Number 5 / October, 2003
[7] Lin Liao, Donald J. Petterson, Dieter Fox, Henry Kautz. Learning and inferring transportation routines ,*Artificial Intelligence*, Volume 171, Issues 5-6, April 2007, Pages 311-331
[8] Russell, S. J.; Norvig, P. (2003). Artificial Intelligence: A Modern Approach
[9] Stefan Edelkamp, Stefan Schroedl, Route planning and map inference with global positioning traces, *Computer science in perspective*, Pages: 128 – 151, 2003

[10] Donald Patterson, Lin Liao, D Fox, H Kautz, Inferring High-Level Behavior from Low-Level Sensors, *UbiComp* 2003