The MERL Motion Detector Dataset

Chris Wren, Yuri Ivanov, Darren Leigh, Jonathan Westhues

TR2007-069 November 2007

Abstract

Looking into the future of residential and office building Mitsubishi Electric Research Labs (MERL) has been collecting motion sensor data from a network of over 200 sensors for a year. The data is the residual traces of year in the life of a research laboratory. It contains interesting spatio-temporal structure ranging all the way from the seconds of individuals walking down hallways, the minutes in the lobbies chatting with colleagues, the hours of dozens of people attending talks and meetings, the days and weeks that drive the patterns of life, to the months and seasons with their ebb and flow of visiting employees. This document describes that dataset, which contains well over 30 million raw motion records, spanning a calendar year and two floors of our research laboratory, as well as calendar, weather, and some intermediate analytic results. The dataset was originally released as part of the 2007 Workshop on Massive Datasets. The dataset can be obtained from http://www.merl.com/wmd.

MITSUBISHI ELECTRIC RESEARCH LABORATORIES
http://www.merl.com

The MERL Motion Detector Dataset

Christopher R. Wren Yuri A. Ivanov Darren Leigh
Jonathan Westhues

TR2007-069 November 2007

Abstract

Looking into the future of residential and office building Mitsubishi Electric Research Labs (MERL) has been collecting motion sensor data from a network of over 200 sensors for a year. The data is the residual traces of year in the life of a research laboratory. It contains interesting spatio-temporal structure ranging all the way from the seconds of individuals walking down hallways, the minutes in lobbies chatting with colleagues, the hours of dozens of people attending talks and meetings, the days and weeks that drive the patterns of life, to the months and seasons with their ebb and flow of visiting employees. This document describes that dataset, which contains well over 30 million raw motion records, spanning a calendar year and two floors of our research laboratory, as well as calender, weather, and some intermediate analytic results. The dataset was originally released as part of the 2007 Workshop on Massive Datasets. The dataset can be obtained from `http://www.merl.com/wmd`.

appears in 2007 ICMI Workshop on Massive Datasets

July 2, 2007: MERL released a year-long dataset from our lab-wide sensor network to provide the community with a rich benchmark for work on perceptual problems in sensor networks. The occasion for the release was the 2007 ICMI Workshop on Massive Datasets.

November 12, 2007: ICMI Workshop on Massive Datasets is held in Nagoya. Japan. A copy of this reference manual is published in the workshop proceedings in the ACM Digital Library

1 Purpose

Are the tools we use to understand our data scalable to the tens of millions of records, huge spans of time, minute details of behavior, and large geographic extent that future sensor networks will generate? In the future buildings will be studded with sensors. Every movement will generate a few bits of data. Every fluctuation in temperature will be recorded. Every deviation in lighting will be noticed. These large and complex datasets will challenge the tools we use today.

Looking into the future of residential and office building Mitsubishi Electric Research Labs (MERL) has been collecting motion sensor data from a network of over 200 sensors for a year. The data is the residual traces of year in the life of a research laboratory. It contains interesting spatio-temporal structure ranging all the way from the seconds of individuals walking down hallways, the minutes in lobbies chatting with colleagues, the hours of dozens of people attending talks and meetings, the days and weeks that drive the patterns of life, to the months and seasons with their ebb and flow of visiting employees.

The dataset contains well over 30 million raw motion records, spanning a calendar year and two floors of our research laboratory. As such it presents a significant challenge for behavior analysis, search, manipulation and visualization of the data. We have also prepared accompanying analytics such as partial tracks and behavior detections, as well as map data and anonymous calendar data marking the pattern of meetings, vacations and holidays.

MERL released this dataset to the community in July of 2007. The release was a central component of the 2007 Workshop on Massive Datasets, an attempt to focus the community on the important and interesting perceptual, visualization and data mining problems associated with these types of sensor networks. We hope the community will use the data to further research on analytic, visualization, and interface techniques in the context of the huge, detailed datasets of the near future.

2 Obtaining the Data

At the time of this writing, the dataset is hosted by the Mitsubishi Electric Research Laboratories in Cambridge, Massachusetts, USA. The entire dataset is approximately $600MB$ and is available via FTP. Details on data set access can be found on the 2007 Workshop for Massive Datasets web page: `http://www.merl.com/wmd`.

3 Requirements

In return for the right to use this resource in your research, we ask that you cite this report[23] in all publications that derive from your work on this data set.

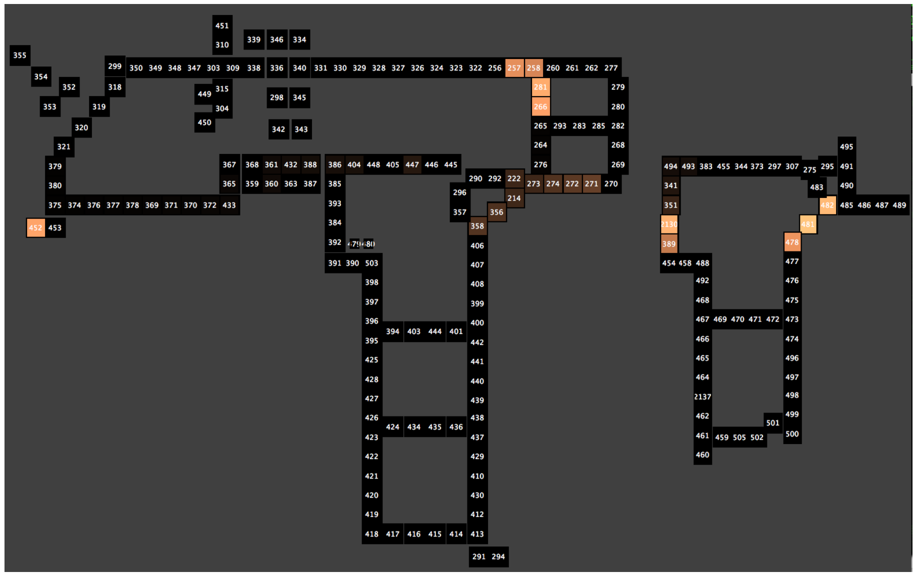

Figure 1: A snapshot of the motion data, bright blocks are sensors that have fired recently, with fading indicating older activations.

4 Data Description

The dataset is comprised of several segments:

- raw motion data
- calibration data
- calendar data
- solar and weather data
- intermediate track analytics

The section contains detailed descriptions of each of these subsets.

4.1 Raw Motion Data

The primary data stream is the output of the motion detector network. The sensors are ceiling mounted at approximately two meter intervals along hallways and in grids covering public spaces such as lobbies and meeting rooms. There are no sensors in individual offices. The sensors are installed with the intention of covering the floor area completely with little or no overlap between sensor fields of view. The ceiling height varies, but is approximately three meters in most areas. Figure 1 shows the tiled arrangement of the sensors and a snapshot of motion activity.

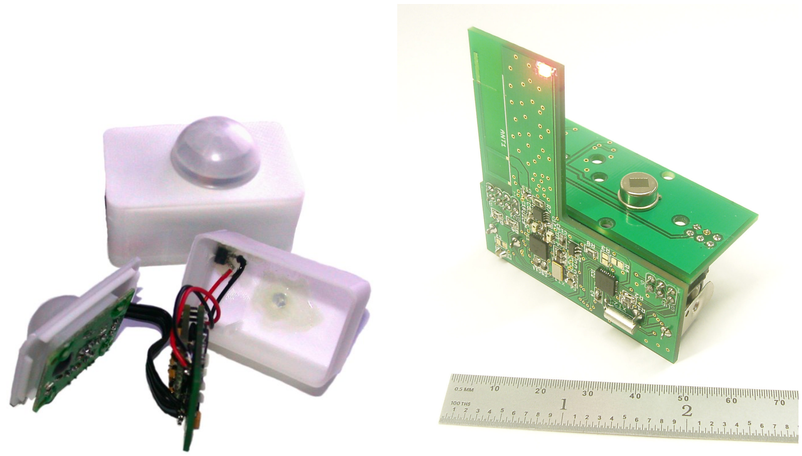

Figure 2: The hardware implementation of a motion detector node. Left: early MITes-based version, Right: An ultra-low-power, standards-based prototype.

The sensors use the MITes platform[15] for processing and communication coupled to a modified KC7783R sensor board. Two versions of the hardware are shown in Figure 2. The MITes-based node used to collect the data in this release is pictured on the left. The re-engineered node on the right is used in new installations and it features an 802.15.4 radio, a modular design, and ultra-low power consumption for drastically extended battery life.

The MITes employ an unreliable network protocol that does not contain checksum information. As a result packets are sometimes lost, duplicated, or garbled. An attempt has been made to filter out the duplicated and garbled packets. However it is certain that there are packets missing from in the data and any analysis should take that into account. The loss rate is low, but is difficult to measure since it depends on the network load.

The sensor boards are the type commonly found in motion activated lighting fixtures and security sensors. They are passive infrared motion detectors. They work by sensing light emitted in the far-infrared by warm objects and signal on high-frequency changes in the scene at those frequencies. They were modified to reduce their adaption rate from minutes to seconds. Since the timing of the detections depends on the analog characteristics of the circuit, the minimum inter-detection time varies, but is observed to typically be around 1.5 seconds.

The raw data is available as compressed ASCII text files in five parts:

`0114.txt.gz` Mar 21 23:00:25 2006 – Jun 11 00:26:40 2006

`0115.txt.gz` Jun 11 00:26:40 2006 – Oct 4 18:13:20 2006

`0116.txt.gz` Oct 4 18:13:20 2006 – Jan 28 11:00:00 2007

`0117.txt.gz` Jan 28 11:00:00 2007 – May 24 05:46:40 2007

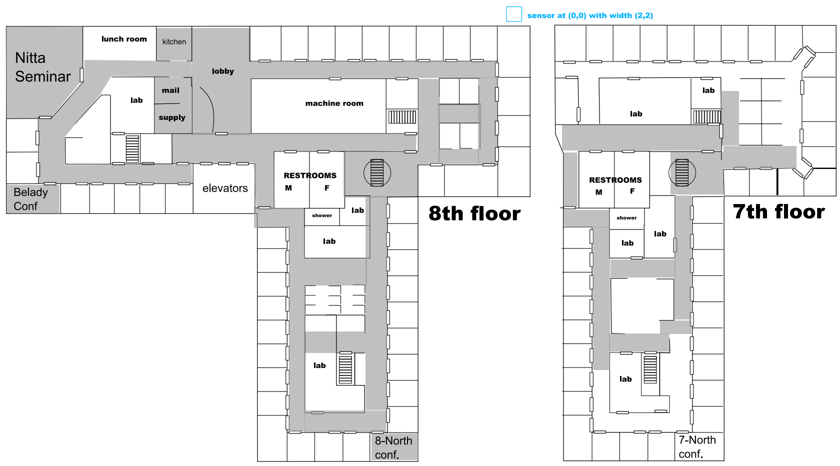

Figure 3: The floor plan of the experimental area. Offices are mostly on the outside of the building. The areas observed by sensors (shaded) are hallways, lobbies, and meeting rooms.

`0118.txt.gz` May 24 05:46:40 2007 – Jul 2 15:41:50 2007

The filename refers to the high-order bits of the timestamps on the data contained in each file.

The files contain data like this:

```
470 01179980510828 01179980511853 1.0
469 01179980512169 01179980513193 1.0
467 01179980513580 01179980514609 1.0
468 01179980514573 01179980515598 1.0
```

The first element is the sensor identification number. The second and third numbers are the timestamps of the beginning of the event. The fourth number is a meaningless place holder value.

The map in Figure 3 depicts the test area. Executives and administrators occupy the wing on the right right side of the eighth floor map. Researchers occupy the bottom and left wings, and most of the 7th floor. The central core of the building contains restrooms, lobbies, elevators, and on the eighth floor, the mail room and the kitchen. There are several stairwells that connect the floors.

We have been collecting data at this facility since October of 2005. Data from the entire area depicted on that map has been continuously recorded since March 2006. The system generates approximately two million motion detections per month.

4.1.1 Time

The data is timestamped with the number of milliseconds since the epoch: January 1, 1970 UTC. Like the windows system clock, this number becomes larger than 2^{32} after 50 days, on February 19, 1970, so you must take care to use 64-bit integer representations when manipulating timestamps:

- `__int64` or `long long` in C
- `use bignum` in PERL
- `java.math.BigInteger` in Java

4.1.2 Senor Geometry and the Map

Each sensor ID corresponds to a unique sensor. The sensors IDs are associated with physical space in the file `map.txt`. This file contains one row per sensors with the sensor ID followed by eight coordinates that specify the four corners of a quadrilateral in meters:

```
sid,x1,y1,x2,y2,x3,y3,x4,y4
214,-13.3,23.1,-13.3,25.3,-15.5,25.3,-15.5,23.1
222,-13.3,20.9,-13.3,23.1,-15.5,23.1,-15.5,20.9
256,-15.5,8.3,-15.5,10.5,-17.7,10.5,-17.7,8.3
257,-13.3,8.3,-13.3,10.5,-15.5,10.5,-15.5,8.3
```

The resulting spatial distribution of sensors is illustrated in the image `map.png`.

A map of the space that provides context for the sensors can be found in `floor_plan.pdf`. The exact relationship between the sensors and this floor plan is encoded in the Scalable Vector Graphics file `floor_plan.svg` The object named "globalframe" is a one meter square located at the origin in sensor coordinates, and as such it may be used to derive the coordinate transform between map coordinates and sensor coordinates.

4.2 Calendar Data

The calendar data is extracted from the MERL corporate calendar and provides both temporal and spatial context for the sensor data. We have included data on holidays, the number of people officially out of the office on a given day, times and locations of meetings, dates of fire alarms in the building, and the daylight savings time shifts. The file `CALENDAR.txt` contains more detailed information as well as pointers to individual data files. The datafile are typically two columns: start and end timestamps for each event, one event per row. If the least significant digits of a timestamp are zero, that indicates that the event is only specified down to the granularity of the non-zero portion of the timestamp, so 1143824400000 indicates that the timestamp is only know to the nearest $100000ms$, or approximately to the minute.

4.3 Solar and Weather Data

The file WEATHER.txt contains pointers to sunrise and sunset times as well as daily weather summaries from the National Climactic Data Center.

The file sun.txt is list of the sun rise and sun set time on the given date. These times are given in standard time, so the daylight savings time shift is not represented in this data.

The file clouds.txt contains data on temperatures, precipitation type,barometric pressure, visibility, and wind speeds, It is taken directly from the NCDC, and the file format is explained in detail in the file WEATHER.txt.

4.4 Intermediate Track Analytics

It is not possible to recover complete tracks of people moving around the space using only motion detectors due to the data association problem. The sensors do not provide enough information to recover from ambiguities. It *is* possible to flawlessly track the movements of an individual *if* they are the *only* person in the space. In this case there will never be an ambiguity about which person is causing a particular sensor activation, because there is only a single person: the data association problem thus becomes trivially degenerate. As soon as a two individuals pass near each other an ambiguity arises: two individuals enter an ambiguity and two leave, but there is no way to distinguish one from the other.

Instead of representing each possible hypothetical labeling as a distinct element in our database we instead record the transitions that we are sure of and link them together at points of ambiguities to form a directed graph. This graph will contain the true track as a subgraph, but also contains all alternative, consistent hypothetical tracks as well.

The tracklets could be regenerated from the raw motion data, but we include our interpretation of the tracklet graph for convenience. The representation of tracklet graph is explained in detail in the file TRACKLETS.txt.

5 Data Management

We store our data in a database to facilitate efficient access and search. Most of the data files are formatted to be read directly into a database, specifically MySQL version 5. The *.sql files are examples of creating tables, loading data, and performing example queries. It is not necessary to use a database, most of the files will also load into Matlab or can be read easily into custom software. Beware however, that the dataset is sizable.

6 Bibliography

Below is a list of relevant references into the liturature. This is not intended as an exhaustive bibliography, but containes references to our work with this dataset, and pointers into the liturature to related work on motion sensors, cameras networks, sensor networks generally, privacy, social networks, and the perception of activity.

7 Acknowledgments

The authors would like to express their gratitude for the undying support and unrelenting encouragement of Dr. Joseph Marks, Director of of the Research Laboratory of the Mitsubishi Electric Research Laboratories when this sensor network was conceived, designed, built, and installed. We also wish to thank the current management of MERL for their continued support of this project. And finally we must thank the employees, interns, delivery people, cleaning crew, and visitors who contributed data to this set by moving around 201 Broadway between March 2006 and March 2007.

References

[1] Gregory Abowd, Aaron Bobick, Irfan Essa, Elizabeth Mynatt, and Wendy Rogers. The aware home: Developing technologies for successful aging. In *Proceedings of AAAI Workshop on Automation as a Care Giver*, 2002.

[2] Ryan Aipperspach, Elliot Cohen, and John Canny. Modeling human behavior from simple sensors in the home. In *Proceedings Of The IEEE Conference On Pervasive Computing*, 2006.

[3] A. Baumberg and D. Hogg. An efficient method for contour tracking using active shape models. In *Proceeding of the Workshop on Motion of Nonrigid and Articulated Objects*. IEEE Computer Society, 1994.

[4] Aaron F. Bobick. Movement, activity and action: the role of knowledge in the perception of motion. *Philosophical Transactions: Biological Sciences*, 352(1358):1257–1265, 1997.

[5] T. Choudhury and A. Pentland. Characterizing social networks using the sociometer. In *Proceedings of the North American Association of Computational Social and Organizational Science (NAACSOS)*, 2004.

[6] Gamhewage C. de Silva, Toshihiko Yamasaki, and Kiyoharu Aizawa. An interactive multimedia diary for the home. *Computer*, 40(5):52–59, 2007.

[7] N. Eagle and A. Pentland. Reality mining: Sensing complex social systems. *Personal and Ubiquitous Computing*, 10(4):255–268, 2006.

[8] Irfan Essa. Ubiquitous sensing for smart and aware environments. *IEEE Personal Communications*, October 2000. Special Issue on Networking the Physical World.

[9] Deborah Estrin, Ramesh Govindan, John Heidemann, and Satish Kumar. Next century challenges: scalable coordination in sensor networks. In *MobiCom '99: Proceedings of the 5th annual ACM/IEEE international conference on Mobile computing and networking*, pages 263–270, New York, NY, USA, 1999. ACM, ACM Press.

[10] Jason Hill, Robert Szewczyk, Alec Woo, Seth Hollar, David E. Culler, and Kristofer S. J. Pister. System architecture directions for networked sensors. In *Architectural Support for Programming Languages and Operating Systems*, pages 93–104, 2000.

[11] Yuri A. Ivanov, Christopher R. Wren, Alexander Sorokin, and Ishwinder Kaur. Visualizing the history of living spaces. *Transactions on Visualization and Computer Graphics*, 13(6):1153–1160, 2007.

[12] Ishwinder Kaur. Openspace: Enhancing social awareness at the workplace. Master's thesis, Massachusetts Institute of Technology, 2007.

[13] Ser-Nam Lim, Larry S. Davis, and Ahmed Elgammal. A scalable image-based multi-camera visual surveillance system. In *IEEE AVSS*, Miami, Florida, July 2003.

[14] T. Mori, K. Asaki, H. Noguchi, and T. Sato. Accumulation and summarization of human daily action data in one-room-type sensing system. *Intelligent Robots and Systems, 2001. Proceedings. 2001 IEEE/RSJ International Conference on*, 4:2349–2354, 2001.

[15] E. Munguia Tapia, S. S. Intille, L. Lopez, and K. Larson. The design of a portable kit of wireless sensors for naturalistic data collection. In *Proceedings of PERVASIVE 2006*, Dublin, Ireland, 2006. Springer-Verlag.

[16] Ali Rahimi, Brian Dunagan, and Trevor Darrell. Simultaneous calibration and tracking with a network of non-overlapping sensors. In *Computer Vision and Pattern Recognition*, pages 187–194. IEEE Computer Society, June 2004.

[17] Carson J. Reynolds and Christopher R. Wren. Worse is better for ambient sensing. In *Pervasive: Workshop on Privacy, Trust and Identity Issues for Ambient Intelligence*, May 2006.

[18] H. S. Sawhney, A. Arpa, R. Kumar, S. Samarasekera, M. Aggarwal, S. Hsu, D. Nister, and K. Hanna. Video flashlights: real time rendering of multiple videos for immersive model visualization. In *Proceedings of the 13th Eurographics workshop on Rendering*, pages 157–168, 2002.

[19] Chris Stauffer and Kinh Tieu. Automated multi-camera planar tracking correspondence modeling. In *Computer Vision and Pattern Recognition*, pages 259–266. IEEE, July 2003.

[20] Daniel H. Wilson and Chris Atkeson. Simultaneous tracking & activity recognition (star) using many anonymous, binary sensors. In *The Third International Conference on Pervasive Computing*, pages 62–79, 2005.

[21] Christopher R. Wren, U. Murat Erdem, and Ali J. Azarbayejani. Functional calibration for pan-tilt-zoom cameras in hybrid sensor networks. *ACM Multimedia Systems Journal*, 12(3):255–268, December 2006.

[22] Christopher R. Wren, Yuri A. Ivanov, Darren Leigh, and Joanathan Westhues. Buzz: measuring and visualizing conference crowds. In *SIGGRAPH '07: ACM SIGGRAPH 2007 emerging technologies*, page 25, New York, NY, USA, 2007. ACM.

[23] Christopher R. Wren, Yuri A. Ivanov, Darren Leigh, and Jonathan Westhues. The merl motion detector dataset: 2007 workshop on massive datasets. Technical Report TR2007-069, Mitsubishi Electric Research Laboratories, Cambridge, MA, USA, August 2007.

[24] Christopher R. Wren, Yuri A. Ivanov, Darren Leigh, and Jonathan Westhues. The merl motion detector dataset: 2007 workshop on massive datasets. Technical Report 069, Mitsubishi Electric Research Laboratories, 2007.

[25] Christopher R. Wren, David Minnen, and Srinivas G. Rao. Similarity-based analysis for large networks of ultra-low resolution sensors. *Pattern Recognition*, 39(10):1918–1931, October 2006. Special Issue on Similarity-Based Pattern Recognition.

[26] Christopher R. Wren and Emmanuel Munguia Tapia. Toward scalable activity recognition for sensor networks. In *Lecture Notes in Computer Science, Volume 3987*, pages 168–185, Dublin, Ireland, 2006. 2nd International Workshop on Location- and Context-Awareness.