InSense: Interest-Based Life Logging

Mark Blum, Alex (Sandy) Pentland, and Gehrard Tröster
Massachusetts Institute of Technology and Swiss Federal Institute of Technology in Zürich

Our wearable data collection system lets users collect their experiences into a continually growing and adapting multimedia diary. The system—called inSense—uses the patterns in sensor readings from a camera, microphone, and accelerometers to classify the user's activities and automatically collect multimedia clips when the user is in an "interesting" situation.

In 1945 Vannevar Bush proposed the Memex (short for memory extender) as a device for storing first-person information that's automatically linked to a library, able to display books and films from the library, and automatically follow cross-references from one work to another.[1] This "enlarged intimate supplement to memory"[1] has spawned a variety of modern projects such as the Remembrance Agent,[2] the Familiar,[3,4] myLifeBits,[5] Memories for Life,[6] and What Was I Thinking.[7]

Each of these more recent projects focuses on organizing, categorizing, and searching a massive store of relatively unedited personal data, such as recorded video and audio. The techniques employed for finding relevant items are mostly speech and image recognition, sometimes in combination with machine learning for data mining. The problem is conceived as first recording everything, then filtering the information to find items relevant and interesting to the user.

In contrast, with our system inSense, we've shifted the problem from the offline analysis of collected data to the online evaluation of a user's current situation. We evaluate the user's context in real time and then use variables like current location, activity, and social interaction to predict moments of interest. Audio and video recordings using a wearable device can then be triggered specifically at those times, resulting in more interest per recording. Some previous examples of this approach are the Familiar and iSensed systems,[3,4,8] which structure multimedia on the fly; the eyeBlog system,[9] which records video each time eye contact is established; and the SenseCam,[10] which records images and sound whenever there's a significant change in the user's environment or the user's movement.

Several reasons exist for making the change from record-and-analyze to annotate-on-the-fly. First, real-time annotation of multimedia allows real-time sharing between users: for example, "Here, take look at this, it's interesting!" Second, online annotation means we don't have to physically remove the data from a body to use it. This is an important privacy consideration, especially when systems such as these are to be used when traveling or on vacation.

In this novel approach, we use a wearable system with acceleration and audio sensing to perform real-time context recognition. Based on the current context classification, inSense uses an interest prediction algorithm to assess the current situation. If the system detects a moment of interest, it takes a picture and stores a short audio clip.

Hardware platform

The hardware platform used is based on low-cost sensors and leverages off commodity hardware. It consists of a personal digital assistant (a Sharp Zaurus SL6000L PDA), two wireless accelerometers, and the matching receiver.[11] This provides the following sensing layout:

- a triaxial accelerometer on the left side of the hip (~90 Hz, 10 bit);
- a triaxial accelerometer worn on the wrist of the dominant hand (~90 Hz, 10 bit);
- audio recorded from the wearer's chest (8 KHz, 16 bit);
- images or video taken from the wearer's chest (for example, one image per minute at 480 × 480 pixels); and
- a wireless frequency (WiFi) access point sniffing with the PDA (every 100 seconds, for location).

We believe that this minimal set of sensors (see Figure 1) is sufficient to classify many interesting dimensions of context. This assumption is supported by previous work in wearable computing.[12,13]

Data collection and annotation

We chose four concurrent categories—location, speech, posture, and activities—to represent many diverse aspects of a user's context (see Table 1). The labels within each category are

1070-986X/06/$20.00 © 2006 IEEE

Published by the IEEE Computer Society

mutually exclusive and represent situations in everyday life.

Subjects wear the system for several hours without interacting with it. Audio and acceleration signals are recorded continuously. The camera takes pictures once a minute and WiFi access points are logged to establish location. After the recording session, the user employs an offline annotation tool, which presents an image at a time, the corresponding sound clip, and a list of labels from which to choose. This naturalistic approach reflects the statistics of a user's everyday life and, apart from the annotated data, also lets us establish the conditional probabilities between the subject's activities. That is, this experience-sampling approach lets us learn, for instance, that users never type while bicycling.

While annotating the user's minute-by-minute activities and context, we also asked each user to rate how interesting the collected images and audio were. These ratings help us learn an *interest operator*, relating the user's context and activity to how interesting the collected images and sound are. For instance, using this approach we can learn that images and sound collected while shaking hands with someone are very interesting, whereas images collected during the sixth continuous minute of typing are almost never interesting.

One obvious shortcoming is the one-minute granularity. A purely naturalistic protocol will not capture sufficient samples of short duration activities like shaking or clapping hands.

For these activities, we use seminaturalistic training in which naturally occurring activity is purposefully captured and annotated. Currently, the database for this work is 24 hours of data from 11 sessions, which reflects a fair sample of the everyday life of a student.

Figure 1. The InSense life logging system, showing sensor placement.

Table 1. Four classification categories with labels.

Location	Speech	Posture	Activities
Office	No speech	Unknown	No activity
Home	User speaking	Lying	Eating
Outdoors	Other speaker	Sitting	Typing
Indoors	Distant voices	Standing	Shaking hands
Restaurant	Loud crowd	Walking	Clapping hands
Car	Laughter	Running	Driving
Street		Biking	Brushing teeth
Shop			Doing the dishes

Classification architecture

We rely on acceleration for the posture and activities categories, audio for speech classes, and audio and WiFi for location classes. As Figure 2 shows, an initial preclassification step uses four separate classifiers to determine probabilities $\boldsymbol{p}(i, j)$ for each state i within each category j, assuming equal priors for each state independent of the probabilities in other categories. Then, in a final classification step, a common-sense model combines the probabilities from the four category classifiers to output a final, joint probability estimate $\boldsymbol{p}^*(i, j)$ for all the states in all the categories by incorporating the conditional probability relationships $\boldsymbol{p}(i, j \mid k, l)$ that have been observed between all the states (i, k) in the different categories (j, l). For example, $\boldsymbol{p}^*(i, j) = 1/m\, \boldsymbol{p}(i, j) \sum \boldsymbol{p}(i, j \mid k, l)\, \boldsymbol{p}(k, l)$, where the summation is over all k,l and m is the number of categories. The wearer's current state is then taken to be the set of maximum probability states in each category.

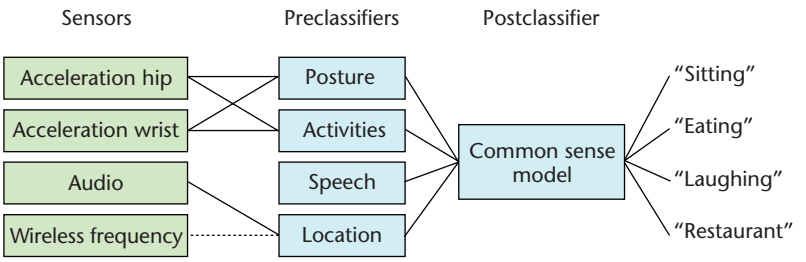

Figure 2. Classification architecture showing the successive stages of processing.

Table 2. Activities confusion matrix.

Classification	A	B	C	D	E	F	G	H	Accuracy (Percentage)
A = no activity	5,585	497	1,005	5	1	173	11	23	77
B = eating	84	490	95	0	0	0	0	1	73
C = typing	177	46	1,676	0	0	1	0	0	88
D = shaking hands	8	0	0	48	1	0	0	1	83
E = clapping hands	1	1	0	2	41	0	0	0	91
F = driving	41	1	4	0	0	198	0	0	81
G = brushing teeth	5	2	0	0	0	0	48	0	87
H = doing dishes	43	0	2	0	0	0	0	41	48
Class average									78.5
Overall average									78.5

For preclassification, we evaluated several classifiers and feature sets. In view of the selected architecture and considering that the system is to be run in real time on small, low-power platforms, we focused the preclassifiers not on reaching the highest-possible accuracy for each category, but rather on speed, storage requirements, and flexibility. Because of the limited program memory, we deemed approaches that rely on saved test data—such as *k*-nearest-neighbors or histogram-based Bayes classifiers—as unsuitable for this application.

The classification methods we studied were

- a naive Bayes classifier using a Gaussian model,
- a naive Bayes classifier using mixtures of Gaussians,
- C4.5 decision trees, and
- hidden Markov models (HMMs).

The highest overall accuracies were reached using the C4.5 algorithm. However, the decision trees' large size and the comparatively low class average accuracies made it apparent that the algorithm was overfitting the data. We tested HMMs mainly for activity classification. The results for ergodic two-state HMMs were good, with accuracy being 76.8 percent overall and 85.4 percent being the class average. However, comparable results were reached with a naive Bayes classifier using a mixture of three Gaussians, which is much faster to compute. For the categories posture and speech, we selected naive Bayes and a single Gaussian due to simplicity, speed, and the immunity to overfitting. Thus, for the preclassification step, all of the categories employed a naive Bayes classifier using Gaussian probability distributions.

The selected acceleration features were the means and variances of the X, Y, and Z axes of both accelerometers over a window of 4.4 seconds. Speech classification is based on the features' formant frequency, spectral entropy, energy maximum, and number of autocorrelation peaks, which we compute at 62.5 hertz (Hz). Again, the means and variances are taken over a 4.8-second window. We implemented the system in C++ and based it on the MIThril 2003 software architecture.[14]

The preclassification results are refined by a final classification step that combines the preclassification activity probabilities $\boldsymbol{p}(i, j)$ in each category with the interactivity conditional probabilities $\boldsymbol{p}(i, j \mid k, l)$ to establish a final classification for each of the activity categories. These commonsense relationships—for example, that driving implies that you are in a car—are captured by computing the pairwise conditional probabilities between all of the states in the activity, location, posture, and speech categories. In cases where there was too little data to establish the conditional probability, it was taken to be $\boldsymbol{p} = 0.05$.

Classification results

Tables 2 through 4 show the classification accuracies for the activity, speech, and posture categories. We omitted the results for the location category because this is simply taken to be the nearest WiFi access point.

Interesting moments

Obviously, not all 24 hours of a person's day are equally interesting. About a third of our time

we're sleeping, the vast part of our day is often spent at an office desk and long periods of time can be spent driving, sitting on a bus, reading a book, or watching TV. These activities can of course be interesting and should be part of a diary. However, memorable things usually happen when these recurring patterns are interrupted.

In this study we found that the user's notion of interesting moments could be captured by a rule-based system based on the user's context and activity. These rules include the following:

- A context such as typing, driving, or lying down is uninteresting.
- A context such as speech, a restaurant, or eating is moderately interesting.
- A context such as laughter, shaking hands, and clapping hands is extremely interesting.
- Long stretches of uninteresting context—like a 15-minute bike ride—need only be captured once, because numerous images won't increase the amount of information.
- Changes in context indicate possibly interesting interruptions or new activities.

Different users assign different weights and parameters for the rules. However, we can figure out these weights and parameters from the user's annotations of what they consider interesting.

Interest prediction algorithm

We implemented an algorithm that calculates the current level of interest based on the context classification. If that level exceeds a certain interest threshold, the system detects a moment of interest. It will capture an image (or video) and store it together with the current context information.

The algorithm combines three measures:

- accumulated static interest, based on the interest assigned to each user state;
- interest bonus for state transitions; and
- time since the last moment of interest.

The static interest is the sum of interest points that correspond to the current classification of location, speech, posture, and activities. The interest map in Table 5 (next page) shows the mapping between labels and interest points for the first author.

By default the interest threshold is set to 5. This means that as soon as a very interesting activity is detected—for example, shaking hands—the system takes a picture. To detect context transitions, the system stores classifications over the last 1 minute and computes the mode state. The mode

Table 3. Speech confusion matrix.

Classification	A	B	C	D	E	F	Accuracy (Percentage)
A = no speech	785	4	21	4	8	3	95
B = user speaking	7	104	65	0	9	2	56
C = other speaking	26	6	493	10	21	0	89
D = distant voices	76	0	41	6	2	0	5
E = loud crowd	16	1	6	1	46	2	64
F = laughter	3	4	6	0	3	37	70
Class average							63.0
Overall accuracy							80.9

Table 4. Posture confusion matrix.

Classification	A	B	C	D	E	F	G	Accuracy (Percentage)
A = unknown	53	1	5	2	0	0	0	87
B = lying	1	89	2	0	0	0	0	97
C = sitting	22	3	6,241	174	2	0	27	96
D = standing	8	0	304	924	43	1	100	67
E = walking	0	0	6	16	182	0	6	87
F = running	0	0	0	0	1	22	0	96
G = biking	0	0	6	17	2	0	547	96
Class average								89.3
Overall accuracy								91.5

Table 5. Assignment of interest points.

Location	Interest Points	Posture	Interest Points
Office	0	Unknown	0
Home	0	Lying	0
Outdoors	1	Sitting	0
Indoors	1	Standing	1
Restaurant	1	Walking	1
Car	0	Running	3
Street	1	Biking	0
Shop	1		
Speech	**Interest Points**	**Activities**	**Interest Points**
No speech	0	No activity	0
User speaking	2	Eating	2
Other speaker	2	Typing	0
Distant voices	1	Shaking hands	5
Loud crowd	2	Clapping hands	5
Laughter	5	Driving	0
		Brushing teeth	0
		Doing the dishes	0

state for each category corresponds to the label, which was classified most frequently during the minute. Each time there is a change in mode state in any context category, a transition bonus of 0.5 points is added.

Finally, to make sure pictures are taken every once in a while even when the interest level is below its threshold, the time since the last picture is taken into account. Every second, 1/120 of a point is added to the interest score, the equivalent to one point every 2 minutes or five points, and thus the system takes a picture every 10 minutes.

Each time the interest threshold is exceeded and an image and sound clip records, the system resets to zero the two counters for transition bonuses and time elapsed since the last picture. Additionally, we impose a hold-off period of 5 seconds after an image is taken so that there won't be large numbers of images taken during periods such as continuous laughter.

The most obvious result of this algorithm is the fact that pictures are taken at a low frequency when the user isn't engaged in anything interesting over a long period of time and a higher frequency during interesting activities. The numeric values were chosen such that in a typical recording, the average frequency of images taken is approximately one every 2 minutes. This varies, as mentioned, from one picture every 10 minutes for a user working on his computer in the office to several pictures per minute during a discussion in a restaurant over lunch.

Experimental results

A three-hour session was recorded with running classifiers to assess the generalizability of the interest algorithm across different people. This is important, because if we hope to share media between people based on how interesting it is, then the notion of what's interesting must be similar between different people. The subject (Mark Blum) started off with working at his desk. Then he met some friends at a restaurant for lunch. After lunch he took his bike to the supermarket for some shopping and brought the food home. On the bike ride back to the lab he stopped briefly at a shop. At the lab work continued for close to an hour. Then he lay down for a few minutes for a nap. At the end he was involved in a short discussion.

The result was two sets of images. Set A contains the interesting pictures that were initiated by the described algorithm. Set B includes pictures that were taken once every minute. The system took a total of 114 pictures for set A, and 178 for set B.

Two examples

The first example covers a time period of 15 minutes, in which the wearer was shopping for food at the local supermarket, bicycled home, and then unpacked his groceries. In the supermarket the wearer paid for the groceries and packed his bags, activities that were accompanied by conversation with the checkout person, then spent 4 minutes of biking to the porch of his house, went upstairs, and unpacked the groceries.

Figure 3. Images captured from set A, using the interest algorithm. These images reflect the subject's trip to the supermarket, carrying the groceries home, and unpacking the groceries.

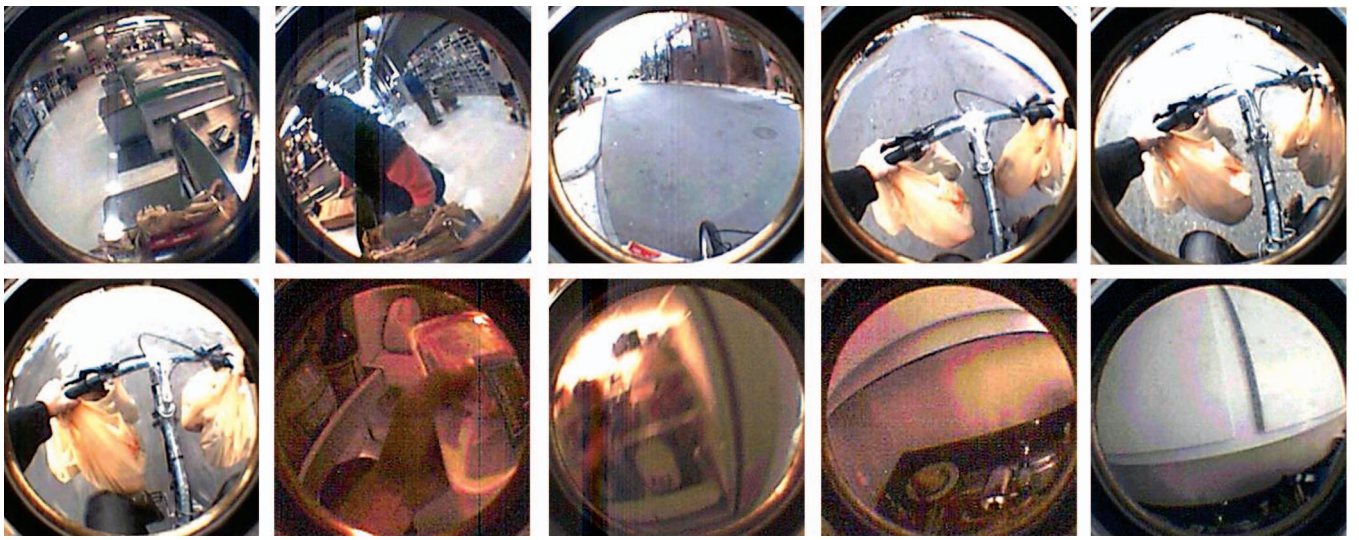

Figure 4. Images captured from set B, using regular sampling instead of the interest algorithm. These images are less interesting shots of the subject's trip to the supermarket, the bike ride home, and unpacking the groceries.

In set A (see Figure 3), which used the interest algorithm, two images were captured at the checkout counter, triggered by conversation and frequent small hand motions. A third image was captured on the exit from the supermarket, as the wearer began walking and transitioning from inside to outside.

No images of the bike trip home were captured, because this exit transition image reset all of the interest counters. The following 4 minutes of biking (2 points), plus the transition from the street via walking and biking (0.5 points), and the static interest of the street and outdoors (1 point) weren't enough to pass the media collection interest threshold.

Upon arriving home, the transition from biking to walking indoors triggered an image, and finally the transition from walking to standing with small hand motions triggered an image of unpacking the groceries.

In contrast, set B (regular sampling; see Figure 4) captured two much less interesting shots of the supermarket, missed the exit from the supermarket, captured four similar images of the bike ride home, missed the arrival home, and captured four less interesting images of the process of unpacking groceries.

The second example set involves a much longer example, beginning with typing that's interspersed with a short discussion with an office mate. Following this is a short nap, followed by more typing and a discussion that began with one person, and then was later joined by a second person. This discussion had several instances of laughter, which triggered media collection.

This time set A (see Figure 5, next page)—again using the interest algorithm—begins with one image of the computer, then one image of a short side conversation with the office mate, and then another typing image. It then collected one image during the nap, with the camera staring up at a ceiling light. The nine remaining images in this set are of the discussion. The large number of images is primarily due to the frequent occurrences of laughter, which always triggers media collection.

In set B (see Figure 6), with a regular collection of audio and images, there are six images of the laptop computer and typing (versus two in set A), with one image of the conversation with the office mate breaking up this sequence. The nap gets six images (versus one in set A), and the conversation at the end gets only three images (versus nine in set A).

Human judgment vs. interest algorithm

Beyond simply making comparisons of particular sequences, we also wanted to determine if the images of set A would be judged more interesting by a wide range of observers. To accomplish this, we conducted an experiment in which we asked people to make judgments about how interesting the two sets of images were.

To make the two sets comparable in size, every third picture in set B was dropped. The two sets of pictures were printed and displayed at the lab with voting slips that could be placed in an urn. The concept of the experiment was briefly explained, and people were asked which set they found more interesting and why. We also contacted and interviewed people via email, describing things similarly.

Figure 5. Another set of images captured from set A, using the interest algorithm. This time the images captured reflect a discussion, a nap, and laughter.

In this experiment, the algorithm clearly did a better job in distinguishing interesting moments. From a total of 28 votes received, 26 were for set A and only two for set B. About two-thirds of the people mentioned the ratio of laptop pictures that appeared in sets A and B, about half mentioned the surplus of images with people in set A and some found that set B had too many repetitive pictures—for example, biking. We'll discuss and explain some of these results in the following paragraphs.

Set A contains 15 laptop pictures versus 47 in set B. It should be noted that the ratio of laptop pictures was only 3 to 17 before the lunch but 12 to 30 after lunch, because the subject was typing near two people who were involved in a discussion. This case suggests the need for a measure to determine if the recognized speech actually involves the user.

The lunch scene was clearly better documented in set A (30 images) than in set B (11 images). What's particularly nice is that at the end of the lunch the subject shook hands with three people, and in two cases an image was taken.

It's also interesting to consider the number of bike ride pictures (see Table 6). In three of the four cases, set A needed fewer images to document the ride. However, in one case the algorithm clearly failed, as we previously noted.

Table 6. Number of biking images.

Biking Location	Set A	Set B
From lunch to the supermarket	1	2
From the supermarket to home	0	4
From home to the shop	2	2
From the shop to the lab	2	4

Conclusion

Overall, the results are pleasing and suggest that this approach, simple as it is, can increase the amount of interest in recorded pictures. It also shows that we can begin to somewhat quantify and generalize the notion of what's interesting to people, potentially allowing automatic sharing of media based on its interestingness. We can also customize the interest operator to a specific user's preferences by assigning different values to interest points and by adjusting the interest threshold. For more detailed information about the algorithms and performance in this article, please refer to Blum.[15]

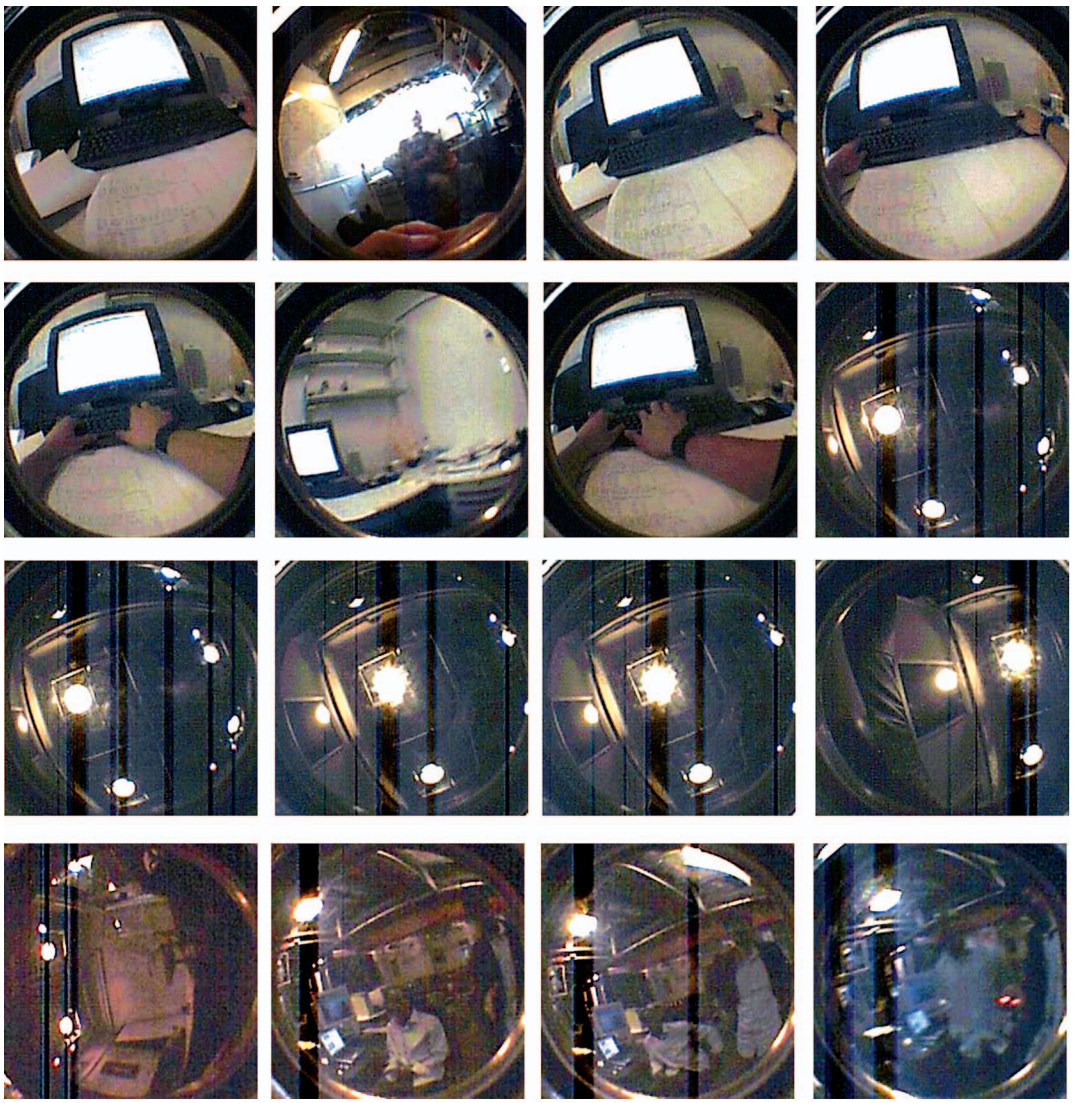

Figure 6. More images captured from set B, using regular sampling instead of the interest algorithm. This version captures more images of the nap with fewer images focusing on the subject's interactions with others.

Something that remains to be studied is how we can scale this approach down to taking only a handful of pictures per day. Will the most interesting moments still be captured? We speculate that to accomplish such a dramatic summarization of the interesting points in a day will probably involve incorporating higher-level behavioral patterns. **MM**

References

1. V. Bush, "As We May Think," *The Atlantic Monthly*, July 1945; http://www.theatlantic.com/doc/194507/bush.
2. B. Rhodes and T. Starner, "Remembrance Agent: A Continuously Running Automated Information Retrieval System," *Proc. 1st Int'l Conf. Practical Application of Intelligent Agents and Multi-Agent Technology*, 1996, pp. 487-495.
3. B. Clarkson and A. Pentland, "Unsupervised Clustering of Ambulatory Audio and Video," *Proc. IEEE Int'l Conf. Acoustics, Speech, and Signal Processing,* IEEE CS Press, vol. 6, 1999, pp. 3037-3040.
4. B. Clarkson, K. Mase, and A. Pentland, "The Familiar: A Living Diary and Companion," *Proc. ACM Conf. Computer–Human Interaction*, ACM Press, pp. 271-272.
5. J. Gemmell et al., "MyLifeBits: Fulfilling the Memex Vision," *Proc. ACM Multimedia*, ACM Press, 2002, pp. 235-238.
6. A. Fitzgibbon and E. Reiter, "'Memories for Life': Managing Information over a Human Lifetime," UK Computing Research Committee Grand Challenge proposal, 2003.
7. S. Vemuri and W. Bender, "Next-Generation Personal Memory Aids," *BT Technology J.*, vol. 22, no. 4, 2004; http://web.media.mit.edu/~vemuri/wwit/wwit-overview.html.

8. B. Clarkson, "Life Patterns," doctoral dissertation, Program in Media Arts and Sciences, Massachusetts Inst. of Technology, 2003.
9. C. Dickie et al., "Augmenting and Sharing Memory with eyeBlog," *Proc. 1st ACM Workshop on Continuous Archival and Retrieval of Personal Experiences*, ACM Press, 2004.
10. J. Gemmell et al., "Passive Capture and Ensuing Issues for a Personal Lifetime Store," *Proc. Continuous Archival and Retrieval of Personal Experiences*, ACM Press, 2004.
11. E.M. Tapia et al., "MITes: Wireless Portable Sensors for Studying Behavior," *Proc. Extended Abstracts, Ubiquitous Computing*, 2004.
12. L. Bao and S.S. Intille, "Activity Recognition from User-Annotated Acceleration Data," *Pervasive Computing: Proc. 2nd Int'l Conf.*, 2004, pp. 1-17.
13. D. Wyatt, M. Philipose, and T. Choudhury, "Unsupervised Activity Recognition Using Automatically Mined Common Sense," *Proc. 20th Nat'l Conf. Artificial Intelligence*, 2005.
14. R. DeVaul et al., "MIThril 2003: Applications and Architecture," *Proc. IEEE Int'l Semantic Web Conf.*, IEEE CS Press, 2003, p. 4.
15. M. Blum, "Real-Time Context Recognition," master's thesis, Dept. of Information Technology and Electrical Eng., ETH Zurich, 2005.

Mark Blum has an MS in electrical engineering and information technology from the Swiss Federal Institute of Technology in Zürich (ETH). His primary interests are context recognition with wearable computing.

Alex (Sandy) Pentland is the Massachusetts Institute of Technology's (MIT's) Toshiba Professor of Media Arts and Sciences, and director of Human Dynamics Research. He's a pioneer in wearable computers, health systems, smart environments, and technology for developing countries. He's a cofounder of the wearable computing research community, the autonomous mental development research community, the Center for Future Health, and was the academic head of the MIT Media Laboratory.

Gerhard Tröster is a professor and the head of the Electronics Laboratory at ETH. His field of research includes wearable computing, reconfigurable systems, smart textiles, and electronic packaging. Tröster received his MS from the Technical University of Karlsruhe, Germany, and his PhD from the Technical University of Darmstadt, Germany, both in electrical engineering.

Readers may contact Sandy Pentland at pentland@media.mit.edu.