inSensed: Wearable Collection of Multimedia based on Interest

M. Blum, A. Pentland, G. Troester
ETH, Zurich Switzerland
MIT, Cambridge MA

Abstract

We present a wearable data collection system that allows users to collect their experiences into a continually growing and adapting multimedia diary. The system, called iSensed, 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. `Interest' is estimated by a user-trained rule relating the sequence of user activities/situations to their concept of `interest'. This allows the multimedia information to be structured in a manner similar to the user's episodic memory, e.g., unusual and potentially important events are recorded chronologically, allowing rapid browsing and automatic structuring of the multimedia diary. The system is based on commercially available mobile devices, either a PDA or a camera phones, together with two small wireless accelerometers worn on the wrist and belt.

I. Introduction

In 1945 Vannevar Bush proposed the MEMEX (short for memory extender) as a device for storing first-person information that is 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" has spawned a variety of modern projects such as the Rememberance Agent [2], the Familiar [3,4], myLifeBits [5], Memories for Life [6], and What Was I Thinking [7].

Each of these more recent projects focus on organizing, categorizing and searching a massive store of relatively unedited personal data. 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 record everything, then filtering the information to find items relevant and interesting to the user.

In contrast we have shifted the problem from offline analysis of collected data to online evaluation of a user's current situation. We are evaluating the context of the user in real time, and then using 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". Earlier examples of this approach are our `The Familiar' and first `iSensed' systems [3,4, 15], which structure multimedia on the fly, the `eyeBlog' system [8] which records video each time eye contact is established.

There are several reasons to make the change from record-and-analyze to annotate-on-the-fly. First, real-time annotation of multimedia allows real-time sharing between users: e.g., "here, take look at this, its interesting!" Second, online annotation means that data need not be moved off of the body to be used, an important privacy consideration especially when systems such as these are to be used when travelling 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, an interest prediction algorithm is used to assess the current situation. If a moment of interest is detected, a picture is taken and a short audio clip is stored.

II. Hardware Platform

The hardware platform used is based on low-cost sensors and leverages off commodity hardware. It consists of a PDA (Sharp Zaurus SL6000L), two wireless accelerometers and the matching receiver [9]. This provides the following sensing layout:

- Triaxial accelerometer on the left side of the hip (~90Hz, 10bit)
- Triaxial accelerometer worn on the wrist of the dominant hand (~90Hz, 10bit)
- Audio recorded from the wearer's chest (8kHz, 16bit)
- Images taken from the wearer's chest (1 per minute, 480x480 pixels)
- WiFi access point sniffing with the PDA (every 100 seconds)

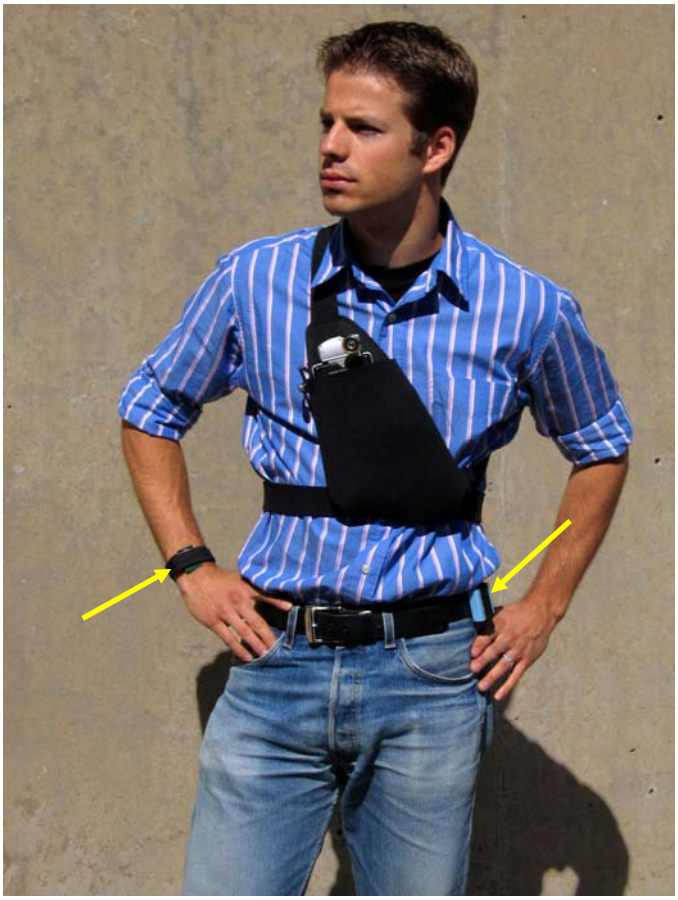

Figure 1: Sensor placement

We believe that this minimal set of sensors is sufficient to classify many interesting dimensions of context. This assumption is supported by previous work in wearable computing [10,11].

III. Data Collection and Annotation

Four concurrent categories – location, speech, posture and activities – were chosen to represent many diverse aspects of a user's context. The labels within each category are mutually exclusive and represent situations in everyday life.

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

Table 1: The four classification categories with labels

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. After the recording session an offline annotation tool is used, which presents at a time an image, the corresponding sound clip and a list of labels to choose from. This naturalistic approach reflects the statistics of the every day life of a user and apart from the annotated data also statistics on conditional probabilities between the subject's activities. That is, this `experience sampling' approach allows us to 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 the `interestingness' of the image and audio collected. These ratings allow us to learn an `interest operator' relating the user's context and activity to the `interestingness' of the collected images and sound. 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 6th 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 certain activities like shaking or clapping hands. For these short activities, semi-naturalistic training is necessary. Currently the database for this work is 24 hours of data from 11 sessions, which reflects fair sample of the everyday life of a student.

IV. Classification Architecture

We rely on acceleration for the categories posture and activities and on audio for the other two. In a pre-classification step four separate classifiers make a decision in their category. Then, in a post-classification step, a `common sense' model combines the information from all four categories to output a final classification.

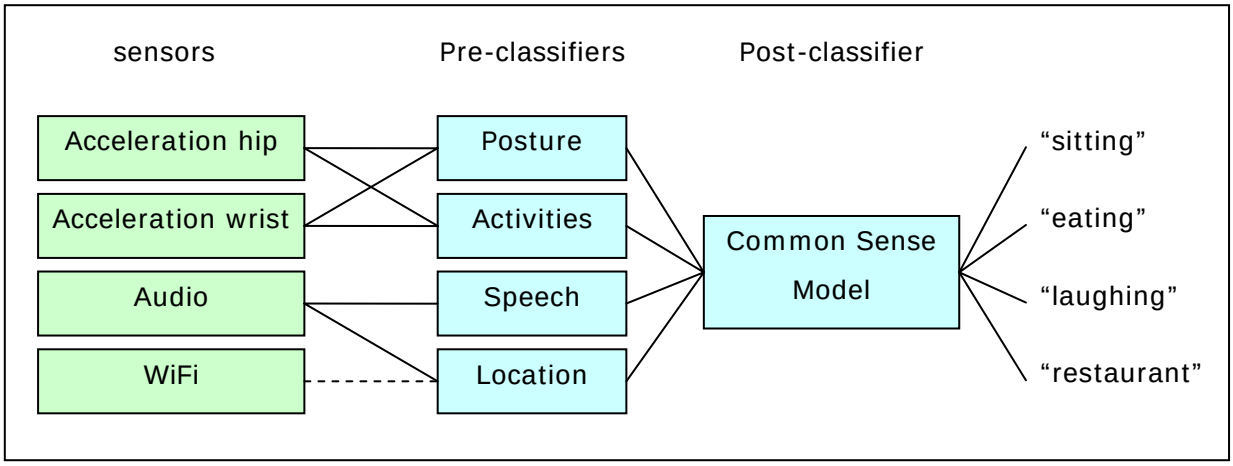

Figure 2: Classification architecture

For pre-classification several classifiers and feature sets were evaluated. A detailed discussion is beyond the scope of this article. We refer to the first author's thesis [12].
The selected acceleration features are the means and variances of X, Y and Z axis 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 62.5Hz (see [13] for details). Again the means and variances are taken over a 4.8-second window. The classifier reported here is a naive Bayes using Gaussian probability distributions.

The pre-classification results are further improved by taking into account the dependencies between the four categories. These `common sense' relationships, e.g. that driving implicates that you are in a car, are captured by computing the pairwise conditional probabilities between activities, locations, postures, and speech categories. The implementation was done in C and C++ and is based on the MIThril 2003 software architecture developed by our group [14].

V. Classification Results

The following tables show the classification accuracies. The results for the category location are omitted because this is simply taken to be the nearest WiFi access point.

classified as -->	a	b	c	d	e	f	g	accuracy
a = unknown	53	1	5	2	0	0	0	87%
b = lying	1	89	2	0	0	0	0	97%
c = sitting	22	3	6241	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 2: Posture confusion matrix

classified as -->	a	b	c	d	e	f	g	h	accuracy
a = no activity	5585	497	1005	5	1	173	11	23	77%
b = eating	84	490	95	0	0	0	0	1	73%
c = typing	177	46	1676	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 accuracy:	78.5%

Table 3: Activities confusion matrix

classified as -->	a	b	c	d	e	f	accuracy
a = no speech	785	4	21	4	8	3	95%
b = user speaking	7	104	65	0	9	2	56%
c = other speaker	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: Speech confusion matrix

VI. What are interesting moments?

Obviously, not all 24 hours of a person's day are equally interesting. About a third of our time we are sleeping, the vast part of daytime is often spent at an office desk and long periods of time can be spent driving, sitting in a bus, reading a book or watching TV. These activities can of course be interesting and should make part of a diary. However, memorable things usually very often happen when these reoccurring 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 are:

- There is uninteresting context such as typing, driving, or lying down.
- There is moderately interesting context such as speech, restaurant or eating.
- There is explicitly interesting context such as laughter, shaking hands and clapping hands.
- Long stretches of uninteresting context like a 15 minute bike ride need only be captured once, because numerous images will not 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 these weights and parameters can be learned from the user's annotations of `interestingness'.

VII. Interest prediction algorithm

An algorithm was implemented 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 and store it together with the current context information.

The algorithm combines three measures:

1. The accumulated static interest, based on an interest map
2. Interest bonus for state transitions
3. 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 below shows the mapping between labels and interest points for the first author.

	Interest points		Interest points
Location		**Posture**	
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		
		Activities	
Speech		no activity	0
no speech	0	eating	2
user speaking	2	typing	0
other speaker	2	shaking hands	5
distant voices	1	clapping hands	5
loud crowd	2	driving	0
laughter	5	brushing teeth	0
		doing the dishes	0

Table 5: Assignment of interest points

By default the interest threshold is set to 5. This means, that as soon as e.g. shaking hands is detected, a picture is taken.

Then, to detect context transitions, the classifications over the last one minute are stored and a super-state is computed. The super-state for each category corresponds to the label which was classified most during that minute. Each time there is a change in super-state in any context category, a transition bonus of 0.5 points is added.

Finally, in order 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. This is equivalent to one point every 2 minutes or 5 points, and thus a picture, every 10 minutes.

Each time a moment of interest is detected, the two counters for transition bonuses and time elapsed since last picture are reset to zero. In addition a hold-off period of 5 seconds will make sure pictures are not taken in masses for instance in the case of several seconds of laughter.

The most obvious result of this algorithm is the fact that pictures are taken at a low frequency when the user is not engaged in anything interesting over a long period of time and a higher frequency during interesting activities. The numeric values were chosen as such, that in a typical recording, the average frequency of images taken is approximately one every 1-2 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.

VIII. Experiment

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 is interesting must be similar between the different people. The subject (the first author) 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 as usual once every minute. A total of 114 pictures were taken for set A, and 178 pictures were taken for set B. To make the two sets comparable, 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. The same was done by means of email.

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 2 for set B. About two thirds of the people mentioned the ratio of laptop pictures, about half mentioned the surplus of images

with people in set A and some found that set B had too many repetitive pictures e.g. biking. In the following some of the results are discussed and explained.

There are 15 laptop pictures in set A versus 47 in set B. It should be noted that the ration of laptop pictures was only 3 to 17 before the lunch but 12 to 30 after lunch, mainly because my office mate was in a discussion with a colleague. This case suggests 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 is 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. One image of set B shows a short discussion with an office mate. The same discussion was documented with two pictures in set A. Six images in set B document the nap at the lab. In set A this was possible with only one image. In the encounter at the end, there was a lot of laughter which resulted in the algorithm taking 9 pictures of that 4 minute conversation instead of only 3 in set B.

Interesting is also the number of bike ride pictures:

	Set A	Set B
lunch → supermarket	1	2
supermarket → home	0	3 to 4
home → shop	2	2
shop → lab	2	4

Table 6: Number of biking images

In three of four cases set A needed less or equal pictures to document the ride. However in one case the algorithm clearly failed. A picture was taken on the way out of the supermarket, thus resetting the transition counters. The four minutes of biking (2 points) plus the transitions shop to street & walking to biking (0.5 points each) and the static interest of street/outdoors (1 point) were not enough to pass the threshold. It also needs to be said, that 7 pictures in the supermarket were initiated by the misclassification of clapping hands. Such false positives do of course affect the results directly.

Overall, the results are very pleasing and suggest that this approach, simple as it is, can increase the “amount of interest” in recorded pictures, and this notion of `interesting’ is at least someone generalizable across people, potentially allowing automatic sharing of media based on its `interestingness’. The interest operator can be customized to a specific user’s prefer-

ences by assigning different values to interest points and by adjusting the interest threshold. Something that remains to be studied is how this approach can scale downwards to taking only a handful of pictures per day. Will the most interesting moments still be captured? For that goal it will be important to incorporate behavioral patterns on a higher level, as discussed above.

(A)

(B)

This short discussion with an office mate was documented with two pictures in set A instead of only one in set B.

This picture was taken by the interest algorithm just after shaking hands.

The lunch scene was documented with three times as many pictures in set A than in set B.

In most cases the interest algorithm needed fewer pictures to capture a bike ride.

IX. Summary

Most prior work in the field of automated diaries has conceptualized the problem of categorizing and searching user data as an offline process. The approach presented here uses information on the user's context to evaluate the current situation in real-time, collecting images and sound clips that are likely to be `interesting' and annotating them with the user's context and activity state.

For this purpose a wearable computing and sensing platform was developed. A large amount of naturalistic user data was collected using interval-contingent experience sampling. Classifiers were trained on this data to recognize several situations in everyday life, the `common sense' conditional probabilities between them, and the `interestingness' of the collected images and sound clips.

An algorithm was developed that predicts the current level of `interestingness' based on the user's state and recent history. If a moment of high possible interest is detected, an image is taken and audio is recorded. This process was tested in a three-hour recording session. The images taken by the algorithm were compared against the same number of images taken at regular intervals. An overwhelming majority of people voted for the images and sound clips selected by interest prediction algorithm. This supports the idea that the idea of `interestingness' is common among at least acquaintances, potentially allowing for automatic sharing of `interesting' media clips. For more detailed information about the algorithms and performance, we refer to the first author's thesis [12].

References

[1] Vannevar Bush's MEMEX (1945) http://www.theatlantic.com/doc/194507/bush

[2] B. Rhodes, T. Starner (1996) "Remembrance Agent: A continuously running automated iformation retrieval system" The First International Conference on The Practical Application Of Intelligent Agents and Multi Agent Technology *(PAAM '96), pp. 487-495.*

[3] B. Clarkson, A. Pentland (1999) "Unsupervised Clustering of Ambulatory Audio and Video" IEEE ICASSP 1999, March 15-19, Phoenix, AZ, vol. 6, pp. 3037-3040

[4] B. Clarkson, K. Mase, A. Pentland (2001) "The familiar: a living diary and companion" ACM CHI '01, Seattle, Washington pp. 271 - 272

[5] J. Gemmell, G. Bell, R. Lueder, S. Drucker, C. Wong, (2002) "MyLifeBits: Fulfilling the Memex Vision", ACM Multimedia '02, December 1-6, Juan-les-Pins, France, pp. 235-238

[6] A. Fitzgibbon, E. Reiter, (2003) " 'Memories for life': Managing information over a human lifetime", UK Computing Research Committee Grand Challenge proposal

[7] S. Vemuri, W. Bender, (2004) "Next-generation personal memory aids", BT Technology Journal, Vol 22, No 4, October.
Project: http://web.media.mit.edu/~vemuri/wwit/wwit-overview.html

[8] C. Dickie, R. Vertegaal, D. Chen, D. Fono, D. Cheng, C. Sohn, (2004) "Augmenting and Sharing Memory with eyeBlog", Proceedings of 1st ACM Workshop on Continuous Archival and Retrieval of Personal Experiences. NYC: ACM Press

[9] E. M. Tapia, N. Marmasse, S. S. Intille, K. Larson, (2004) "MITes: Wireless portable sensors for studying behavior", Proceedings of Extended Abstracts, Ubiquitous Computing,

[10] L. Bao, S. S. Intille, (2004) "Activity recognition from user-annotated acceleration data", Pervasive Computing: Proc. of the 2nd Int'l Conference. pp. 1–17

[11] D. Wyatt, M. Philipose, T. Choudhury, (2005) "Unsupervised Activity Recognition using Automatically Mined Common Sense", Proceedings of the Twentieth National Conference on Artificial Intelligence (AAAI 2005)

[12] M. Blum, (2005) "Real-time Context Recognition", M.Sc. thesis in Information Technology and Electrical Engineering, ETH Zurich

[13] S. Basu, (2002) "Conversational Scene Analysis", PhD thesis MIT, Dept. of Electrical Engineering and Computer science.

[14] R. DeVaul, M. Sung, J. Gips, A. Pentland, (2003) "MIThril 2003: Applications and Architecture", Proceedings IEEE ISWC, October 2003.

[15] B. Clarkson, (2003) `Life Patterns', PhD thesis, MIT, Program in Media Arts and Sciences.