Multi-sensor physical activity recognition in free-living

Katherine Ellis
UC San Diego, Electrical and Computer Engineering
9500 Gilman Drive
La Jolla, CA 92023 USA
kellis@ucsd.edu

Suneeta Godbole
UC San Diego, Family and Preventive Medicine
9500 Gilman Drive
La Jolla, CA 92023 USA
sgodbole@ucsd.edu

Jacqueline Kerr
UC San Diego, Family and Preventive Medicine
9500 Gilman Drive
La Jolla, CA 92023 USA
jkerr@ucsd.edu

Gert Lanckriet
UC San Diego, Electrical and Computer Engineering
9500 Gilman Drive
La Jolla, CA 92023 USA
gert@ece.ucsd.edu

UbiComp'14 Adjunct, September 13 - 17, 2014, Seattle, WA, USA

ACM 978-1-4503-3047-3/14/09...$15.00.
http://dx.doi.org/10.1145/2638728.2641673

Abstract
Physical activity monitoring in free-living populations has many applications for public health research, weight-loss interventions, context-aware recommendation systems and assistive technologies. We present a system for physical activity recognition that is learned from a free-living dataset of 40 women who wore multiple sensors for seven days. The multi-level classification system first learns low-level codebook representations for each sensor and uses a random forest classifier to produce minute-level probabilities for each activity class. Then a higher-level HMM layer learns patterns of transitions and durations of activities over time to smooth the minute-level predictions.

Author Keywords
Activity recognition; Linear dynamical system; Codebook; Accelerometer; GPS

ACM Classification Keywords
1.5.4 [Pattern Recognition Applications]: .

Introduction
Accurate and unobtrusive monitoring of physical activity in free-living populations (*i.e.* people performing their normal daily routines) is an area of research which extends to a variety of applications. Public health researchers are interested in how the type, frequency, intensity, and

associated behaviors of physical activity are related to diseases such as cancer, heart disease, and diabetes. Additionally, specific information about when and how people engage in physical activity can inform interventions. Real-time prediction of behaviors can enable "just-in-time" interventions that encourage people to be more active at certain times to maximize the effectiveness of the intervention. For example, a person might be more receptive to encouragement to exercise when they are watching TV at home rather than when in a work meeting. More generally, activity monitoring has applications for personalized and context-aware recommendation systems, targeted advertising, assistive technologies, automatic journaling or life-logging, personalized medicine and more.

The variety of sensors in mobile phones, including accelerometers, gyroscopes and GPS allow many opportunities for advancements in activity prediction. Stand-alone sensors, particularly accelerometers, have long been used to measure movement in physical activity research. In the future, as hardware improves and sensors can be made smaller and more portable, the range of sensors available will increase even more. With the advent of these sensors, effective frameworks for combining the diverse information provided by each sensor are needed.

Many previous studies in these areas have used datasets collected from prescribed activities that are performed in a laboratory or controlled setting [9] (*e.g.*, a researcher gives the participant a specific list of activities to perform and oversees the activities). Many studies show high accuracy in activity classification, but when activities are performed in daily life they may be performed with greater variety and introduce noisier data. Studies that compare performance between free-living data and controlled data indicate that performance measured on prescribed datasets may not translate to real-world performance [6, 7]. Because the goal of developing these systems is to implement them on real-world populations, it is essential to test their performance in a realistic situation. To this end, we have collected a large free-living dataset from participants going about their daily lives. The dataset was collected from a population of overweight and obese breast cancer survivors in a study at the University of California, San Diego. Participants wore tri-axial accelerometers on their hip and wrist, a GPS unit, and heart rate monitor for seven days. Ground truth information about their behaviors was obtained using a wearable camera that was later manually annotated by researchers.

In this paper, we present a classifier to predict basic postures and movements (sedentary, standing, walking/running, in a vehicle) from a hip accelerometer and GPS. We present a system that identifies the physical activities performed by a participant when we have no individual-specific training data nor prior knowledge about the participants habits. While previous work has shown that training a classifier on individual-specific data improves performance [1], the added burden of obtaining this individualized training data makes it prohibitive for many applications.

Our system uses a multi-level classifier to capture both specific patterns of movement over the scale of a few seconds as well as longer term patterns on the scale of an entire day of behaviors. The low-level classifier learns a quantized representation of the accelerometer and GPS data and uses a random forest classifier to assign probabilities to each activity class. The high-level classifier uses a Hidden Markov Model (HMM) to model the

probabilities of transitioning between activities a produce a complete segmentation of activity predictions for a day.

Dataset

We collected a free-living dataset from a population of 40 overweight and obese breast cancer survivors. These participants were recruited from a group of women who were ineligible for a random control trial on weight loss and breast cancer risk at University of California, San Diego. Participants agreed to wear the sensors during waking hours for seven days. At the completion of data collection, participants were given an opportunity to view and delete any images that they did not want included in the study. All study procedures were approved by the research ethics board of the University of California, San Diego.

Sensors

Participants wore two GT3X+ accelerometers: one on their right hip and one on their non-dominant wrist. They also wore a Qstarz BT1000X GPS device on their hip and a heart rate monitor. The accelerometers sampled 3-axis acceleration at 30Hz. Due to storage restraints, the GPS was set to sample every 15 seconds. Additionally, participants wore a SenseCam — a small camera that is worn on a lanyard around the neck and automatically snaps images from the point of view of the wearer. The images taken by the SenseCam were used to manually annotate the dataset with ground truth annotation of the activities the participant was engaging in. The SenseCam takes an image every 10 to 15 seconds, when an onboard sensor is activated (*e.g.* by a change in movement, light, temperature or presence of another person). If the sensors are not triggered, a photo is taken every 50 seconds. More than 3000 wide-angle low-resolution images can be collected in 1 day. Participants were required to charge the device every night and received daily reminder texts to comply with the protocol. Participants were also instructed on how to use a privacy button on the device, which turns off image collection for up to 7 minutes. Participants were advised to remove the SenseCam in locations where cameras were not permitted (*e.g.* fitness facilities), and to use the privacy button for activities such as bathroom visits and banking. Participants were also encouraged to ask others for permission to record images during private or confidential meetings. In Figure 1 a few examples of the images are shown.

Annotation

SenseCam image data were downloaded and imported into the Clarity SenseCam browser. A standardized protocol was developed for annotating the images with activity labels. A group of researchers and undergraduate interns annotated the images according to the protocol. Interrater reliability of image annotation was established using an iterative cycle of annotation followed by discussion, with all disagreements resolved by group consensus. This yielded a set of annotated images from which additional annotators could be trained and certified. Approximately 10% of all subsequently annotated images were checked by a second annotator. Annotators also received additional training in protecting the privacy, confidentiality and security of the images. The full annotation protocol is available from the authors upon request.

Annotations were divided into two categories: posture labels and behavior labels. Table 1 lists the set of labels used in this dataset. Each image was assigned exactly one posture label. Sedentary posture (sitting or lying) was detected based on knee and leg positions visible in the image, hands resting on a table, or camera angles that were lower than other people who were standing.

Standing posture was detected based on height and distance to other furniture or standing people, and absence of knees or legs in the image. Subsequent images were used to judge the presence of movement. When objects in the image appeared in the position from one image to the next the label "standing still" was applied. If some movement was detected, but without significant forward progress, the image was labeled "standing moving". If progress toward a distant point was observed, the image was labeled "walking/running." It is very difficult to estimate speed from image sequences, so we did not attempt to differentiate between walking and running. An image was annotated as bicycling when handlebars were present.

After posture labels were assigned, behavior labels were assigned to each image. These labels included household activity, self care, conditioning exercise, sports, manual labor, leisure, administrative activity, riding in a car, riding in other vehicles, watching TV, other screen use, and eating. Images could be annotated with any number of behavior labels, including no label. Images where the camera lens was obstructed or the annotators could not determine the participant's activity were labeled as "uncodable." Subsequent images with identical labels were grouped into an activity bout, with start and end times provided by the timestamp of each image. If there was a gap of more than 4 minutes between identically annotated images, the sequence was broken into separate bouts.

For this study, we combined these labels into a set of four mutually exclusive activities that cover the basic postures and motion states: sedentary, standing, walking/running, and riding in a vehicle.

Figure 1: Examples of SenseCam images and annotations

	minutes
Posture Labels	
Sedentary	79,571
Standing Still	7,762
Standing Moving	8,353
Walking/Running	6,832
Bicycling	112
Behavior Labels	
Household Activity	9,689
Self Care	813
Conditioning Exercise	800
Sports	82
Manual Labor	207
Leisure	2,040
Administrative Activity	6,136
Car	12,286
Other Vehicle	1,352
Television	25,325
Other Screen	25,875
Eating	5,825

Table 1: Annotations applied to the dataset and number of minutes collected for each annotation, grouped by posture and behavior labels. Posture labels are mutually exclusive; behavior labels can occur simultaneously.

Data Representation

For this exploratory study we used only the data from the hip accelerometer and GPS. These types of sensors have been successfuly used for activity recognition in previous studies. Future work will investigate methods to use the data from the wrist accelerometer and heart rate monitor. We used a half-overlapping sliding window to break the sensor streams into 1-minute windows. If the full window fell within a valid bout of annotated activity, we labeled the window with the corresponding activity label. If the window spanned multiple activities or contained time for which the true activity could not be determined, we left the window unlabeled.

We represent each window of sensor data using a quantized codebook representation, learning a separate codebook for each sensor. Codebooks can be learned from unlabeled data, which is very easy to obtain for wearable sensors. These codebooks are described in detail below.

GPS

For each window of GPS data we extracted the following features:

(1) The average and standard deviation of speed in the window. (2) The number of satellites and the signal to noise ratio. This gives a general idea of the quality of satellite reception, which can provide information about whether the participant was indoor or outdoor and subsequently more likely to be engaging in certain behaviors. (3) The distance traveled (both total distance over the window and net distance). Distance features can give an idea about the path traveled, whether it was direct path or winding. We quantized these features into 64 codewords using k-means, and represented each data window by the closest codeword.

Accelerometer

Linear Dynamical Systems

We first model the raw acceleration signal using linear dynamical systems (LDS). The LDS describes the acceleration signal values over time as the output of a latent dynamical process.

Specifically, a sequence $y_{1:\tau}$ of τ accelerometer samples is the output of an LDS:

$$x_t = Ax_{t-1} + v_t, \quad (1)$$
$$y_t = Cx_t + w_t + \bar{y}, \quad (2)$$

where the random variable $y_t \in \mathbb{R}^m$ encodes the acceleration at time t, and a lower dimensional hidden variable $x_t \in \mathbb{R}^n$ encodes the dynamics of the observations over time. The state transition matrix $A \in \mathbb{R}^{n \times n}$ encodes the evolution of the hidden state x_t over time, $v_t \sim \mathcal{N}(0, Q)$ is the driving noise process, the observation matrix $C \in \mathbb{R}^{m \times n}$ encodes the basis functions for representing the observations y_t, $\bar{y}$ is the mean of the observation vectors, and $w_t \sim \mathcal{N}(0, R)$ is the observation noise. The initial condition is distributed as $x_1 \sim \mathcal{N}(\mu, S)$.

LDSs have successfully been applied to music information retrieval [3], video annotation [2], surgical gesture recognition [10] and activity recognition from video data [8].

We use an LDS to represent each 5-second sample of accelerometer data. In this way the LDS captures a short pattern of motion such as the acceleration and deceleration that occurs during each step of walking. We learn a codebook by estimating the parameters of a mixture of LDSs using an EM algorithm. We learn 128 codewords from the pooled data in the training set. The parameters are directly estimated from the pooled acceleration data of all participants in the training dataset, using an approximate and efficient algorithm based on principal component analysis [4]. Each 5-second sample in the dataset is then represented by the most representative codeword, according to the conditional likelihood of the LDS given the data. This codebook representation is similar to a bag of words representation that is commonly used in natural language processing and computer vision, and LDSs as codewords has been used for automatic music annotation [5].

Combined representation

Finally, each minute of accelerometer data is represented as a histogram over the the LDS-codewords representating each 5-s subsample. This is concatenated with the GPS features described in the previous section to obtain a combined representation for each minute of data.

Low-level classifier

The low-level classifier operates on the minute level, producing a prediction score for each data window. We learn a random forest classifier over the combination of GPS and accelerometer features described in the previous sections. Preliminary experiments showed that the random forest classifier produced higher accuracy than other classifiers such as SVMs and logistic regression. A random forest classifier combines the output of many randomized decision trees. Random forests have been successfully applied to activity recognition problems [6]. Each decision tree is learned from a random subset of training examples and a random subset of features. The output of each decision tree in the forest is combined using majority voting to obtain a prediction. We learned a random forest consisting of 50 trees with 10,000 training examples and 25 features per tree.

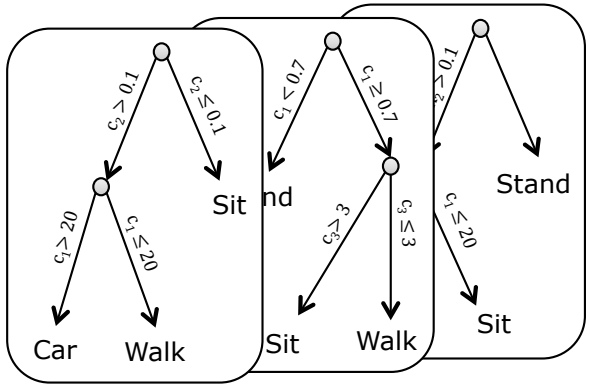

Figure 2: The low-level layer uses a random forest to predict probabilities for each minute of activity.

High-level classifier

The second level classifier is a Hidden Markov model (HMM) that models activity bouts over minutes. Figure 3 shows a graphical representation of an HMM. Each hidden state $u_t, t = 1, 2, \ldots, T$ belongs to one of M discrete states, corresponding to the activities we would like to predict. Each observed state $v_t, t = 1, 2, \ldots, T$ also belongs to one of M discrete states, corresponding to the activity predicted by the low-level classifier. Each state corresponds to one data window, which in this case is one minute. The $M \times M$ transition matrix B represents the probabilities of transitioning between each hidden state, *i.e.*, $B_{mn} = Pr(u_{t+1} = n | u_t = m)$. The $M \times M$ observation matrix D represents the probabilities of each observation given each hidden state, *i.e.*, $D_{km} = Pr(v_t = k | u_t = m)$. The initial state u_0 is distributed according to a probability distribution pi.

We learn the parameters B, D and pi using maximum likelihood estimate according to the training data. To classify a test sequence of predictions from the low-level classifier, we use the viterbi algorithm to generate the most likely sequence of activity states. The HMM layer improves performance of the low-level classifier by explicitly modeling the probability of transitioning between activities. For example, it is very unlikely to transition directly from "sedentary" to "vehicle" without a small bout of walking in between. The HMM also models the duration of an activity bout via the self-transition probability — the probability that the activity state will remain in state m for τ timesteps follows a geometric distribution with parameter $1 - B_{mm}$. Applying the high-level classifier smooths abrupt transitions between low-level activity predictions and produces a segmentation of activity bouts that is aligns with realistic daily activity patterns.

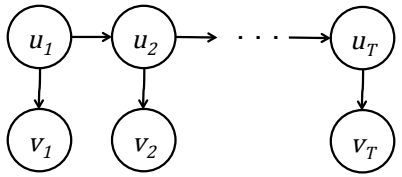

Figure 3: The high-level layer uses an HMM to segment the minute-level probabilities into bouts of activities.

Results

Table 2 shows the results of our activity classification system using leave-one-subject-out cross (LOSO) validation. LOSO validation simulates the real-world scenario in which we would like to train a classifier from a large population of training data, and apply it to a previously unseen participant. The overall accuracy was 85.6%.

Table 3 shows the confusion matrix for our classification system. The highest rate of misclassification was walking

that was misclassified as standing (32%). This may be an understandable error, as we grouped the "standing moving" category with standing rather than walking, some instances of walking may look closer to standing moving. In free-living situations, very short bouts of walking happen frequently, and standing perfectly still is rare, which might lead to higher error rates between these two classes than is seen in prescribed studies.

		Predicted label		
True Label	Sed	Stand	Walk	Vehicle
Sed	0.960	0.038	0.002	0.000
Stand	0.276	0.672	0.048	0.005
Walk	0.040	0.322	0.633	0.004
Vehicle	0.058	0.045	0.004	0.894

Table 3: Confusion matrix using leave-one-subject-out cross-validation. Values are reported as percentages of the number of true examples for each activity.

Conclusion

We have presented an activity recognition system that classifies free-living accelerometer and GPS data into four motion states. Future work will focus on predicting more detailed behaviors such as household activities, conditioning exercises and administrative activities. Toward this aim, we will incorporate the use of an additional wrist accelerometer, which may be essential for predicting activities mainly characterized by arm movements (*e.g.*, lifting weights). Incorporating location prediction into the model may help with predicting more specific behaviors as well, as certain behaviors are more likely to occur in recurring locations (*e.g.*, lifting weights at the gym).

Predicting these specific activities is a difficult task in free-living data because they tend to be very rare (for example, only 112 minutes of bicycling data, from one participant, was collected in this dataset). On average, the women in this dataset spent 74% of their day in a sedentary behavior.

References

[1] Bao, L., and Intille, S. S. Activity recognition from user-annotated acceleration data. In *Pervasive computing*. Springer, 2004, 1–17.

[2] Chan, A. B., and Vasconcelos, N. Modeling, clustering, and segmenting video with mixtures of dynamic textures. *Pattern Analysis and Machine Intelligence, IEEE Transactions on 30*, 5 (2008), 909–926.

[3] Coviello, E., Chan, A. B., and Lanckriet, G. Time series models for semantic music annotation. *Audio, Speech, and Language Processing, IEEE Transactions on 19*, 5 (2011), 1343–1359.

[4] Doretto, G., Chiuso, A., Wu, Y. N., and Soatto, S. Dynamic textures. *Intl. J. Computer Vision 51*, 2 (2003), 91–109.

[5] Ellis, K., Coviello, E., Chan, A., and Lanckriet, G. A bag of systems representation for music auto-tagging.

[6] Ellis, K., Godbole, S., Chen, J., Marshall, S., Lanckriet, G., and Kerr, J. Physical activity recognition in free-living from body-worn sensors. In *Proceedings of the 4th International SenseCam & Pervasive Imaging Conference*, ACM (2013), 88–89.

[7] Ermes, M., Parkka, J., Mantyjarvi, J., and Korhonen, I. Detection of daily activities and sports with wearable sensors in controlled and uncontrolled conditions. *Information Technology in Biomedicine, IEEE Transactions on 12*, 1 (2008), 20–26.

[8] Kellokumpu, V., Zhao, G., and Pietikäinen, M.

		Sitting	Standing	Walking/Running	Vehicle	**Average**
Low-level	P	0.796	0.754	0.819	0.982	0.838
	R	0.963	0.558	0.656	0.832	0.752
	F	0.871	0.641	0.729	0.901	0.786
High-level	P	0.860	0.749	0.784	0.992	0.846
	R	0.960	0.672	0.633	0.894	0.790
	F	0.907	0.708	0.701	0.940	0.814

Table 2: F-scores for each activity class after the low-level RF classifier and the high-level HMM classifier.

Human activity recognition using a dynamic texture based method. In *BMVC* (2008), 1–10.

[9] Staudenmayer, J., Pober, D., Crouter, S., Bassett, D., and Freedson, P. An artificial neural network to estimate physical activity energy expenditure and identify physical activity type from an accelerometer. *Journal of Applied Physiology 107*, 4 (2009), 1300–1307.

[10] Zappella, L., Béjar, B., Hager, G., and Vidal, R. Surgical gesture classification from video and kinematic data. *Medical image analysis 17*, 7 (2013), 732–745.