Automatic Annotation of Daily Activity from Smartphone-Based Multisensory Streams

Jihun Hamm[1], Benjamin Stone[2], Mikhail Belkin[1], and Simon Dennis[2]

[1] The Ohio State University, Dept. Computer Science and Engineering,
Columbus, OH 43210, USA

[2] The Ohio State University, Dept. Psychology, Columbus, OH 43210, USA

Abstract. We present a system for automatic annotation of daily experience from multisensory streams on smartphones. Using smartphones as platform facilitates collection of naturalistic daily activity, which is difficult to collect with multiple on-body sensors or array of sensors affixed to indoor locations. However, recognizing daily activities in unconstrained settings is more challenging than in controlled environments: 1) multiples heterogeneous sensors equipped in smartphones are noisier, asynchronous, vary in sampling rates and can have missing data; 2) unconstrained daily activities are continuous, can occur concurrently, and have fuzzy onset and offset boundaries; 3) ground-truth labels obtained from the user's self-report can be erroneous and accurate only in a coarse time scale. To handle these problems, we present in this paper a flexible framework for incorporating heterogeneous sensory modalities combined with state-of-the-art classifiers for sequence labeling. We evaluate the system with real-life data containing 11721 minutes of multisensory recordings, and demonstrate the accuracy and efficiency of the proposed system for practical lifelogging applications.

Keywords: mobile computing, lifelogging, activity recognition, automatic annotation.

1 Introduction

With an ever increasing number of smartphones today which are capable of recording motion, location, vision and audio, there are numerous applications that can make use of the multisensory data, such as context-aware services, health monitoring, augmented memory, and lifelogging. Lifelogging refers to a long-term process of automatically collecting sensory data using wearable devices, and storing the data into a personal multimedia form for browsing, annotating, and searching [14,20,39,7,10]. A lifelogging system will be particularly useful if it is capable of recognizing high-level experiences of users from low-level sensory streams without requiring the user to manually annotate the huge amount of data. In this paper, we present a lifelogging system for capturing a user's daily experience by collecting multisensory streams on a smartphone and automatically annotating the daily activity with high-level tags.

Automatic annotation is closely related to activity recognition, which has a rich literature across various fields including embedded and mobile systems, pervasive and ubiquitous computing, sensor networks, multimedia, intelligent systems, and machine

D. Uhler, K. Mehta, and J.L. Wong (Eds.): MobiCase 2012, LNICST 110, pp. 328–342, 2013.

learning. (We refer the reader to [22,23] for a survey.)[1] The majority of previous research on activity recognition used custom on-body sensors and embedded systems [3,25,32,48,8,17,1,26,10,24,36], or used external sensors such as surveillance cameras/microphones, object-attached sensors, and RFID tags [27,42,43,34,16,19].

Recently, some researchers have focused on using smartphones as a platform [4,7,12,44,49], which allows them to collect naturalistic data from a user who carries out everyday activity without interruption. The range of activities resulting from such a setting is much broader that atomic sets of actions performed by a user in a controlled setting such as postures/locomotion-types (sitting, lying, walking, running, standing up, sitting down, up/down the stairs, etc), and opens up new opportunities for studying everyday human behavior.

However, recognizing daily activities in unconstrained settings is more challenging than in controlled environments in two aspects – sensory data and activity labels. Firstly, various sensors equipped in generic smartphones are noisier and have to operate under a limited battery capacity, resulting in sparser samples than those from a surveillance system for example. More importantly, multisensory streams from heterogeneous sensors can be asynchronous, vary in sampling rates and have missing data occasionally. In this work, we propose a multisensory bag-of-word representation to handle these problems. The bag-of-words model, originally used for document analysis, has proven useful in other domains such as visual scene analysis [13] and activity recognition in particular [17,7,44]. We build on the previous work and present our multisensory bag-of-word framework that can be combined with various classification algorithms in the subsequent stage.

The second challenge with unconstrained daily activities is that they are continuous, can occur concurrently, and have fuzzy onset and offset boundaries. Furthermore, since the ground-truth labels for activities cannot be acquired on the fly but only after the collection by the user who reviews the logged images at the end of each day, the resultant ground-truth labels are prone to errors and are accurate only in a time scale of minutes rather than seconds or milliseconds. Various classifiers have been used for activity recognition including Fuzzy Logic, Neural Network, Naive Bayes, Bayesian Network, Nearest Neighbor, Decision tree, Support Vector Machines, boosting and bagging (refer to [23]). More recently, continuous recognition of activity was posed as a sequence labeling problem [37,41,46,9,45,31], using temporal models such as Hidden Markov Model [35]), Conditional Random Field (CRF)[21], and structured Large-Margin classifiers[2,40]. The continuous nature of daily activity makes the temporal models potentially more appropriate for handling noisy multisensory streams and labels from smartphones.

In this paper, we describe our system for acquiring naturalistic data from smartphones, and present the multisensory bag-of-words framework combined with state-of-the-art classifiers for automatic annotation of daily activity. We evaluate our approach using 42 days (corresponding to 11721 minutes) of recordings from a volunteer by comparing the performance of various classifiers that represent temporal vs non-temporal and generative vs discriminative approaches, and demonstrate the feasibility of automatic annotation of unconstrained daily activity.

[1] We omit the discussion of activity recognition approaches based on continuous videos.

The rest of the paper is organized as follows. In Section 2, we describe our lifelogging system and its components. In Section 3, we present our framework for handling multisensory streams. In Section 4, we introduce several generative and discriminative approaches for activity recognition. In Section 5, we describe the experiments and report evaluation results, and conclude the paper in Section 6.

2 System

Our system consists of a mobile app, a server infrastructure, and a user interface. We describe them in the following sections.

2.1 Mobile Application and Server

We developed a Java app that acquires multisensory data on Android-based smartphones. The app acquires image, audio, GPS, accelerometer, and other information in regular but changeable time intervals, and stores them until it connects to the server. It runs in the background as a service so as not to disturb normal usage of the phone, although we used the smartphones only for data acquisition purposes in this paper.

Users carried the phone daily from morning till evening. The phone was carried inside a pouch attached to a neck strap to allow an unobstructed view for the camera. The app runs continuously on a standard Android phone for six hours before running out of battery, and it can last much longer with an extended battery. The collected data is sent automatically to a remote server, usually once a day in the evening, when the phone detects WiFi and is connected to a charger. The data is sent in batch mode via SFTP protocol for added security and remains inaccessible to other users in the system.

2.2 Sensory Data and Raw Features

Accelerometer. We use a tri-axial accelerometer with a maximum sampling of 16 Hz. The actual sampling rate obtained from the phone varies over time, so we took only contiguous samples whose actual rates are within a tolerable range. In the literature, simple time-domain features (mean, variance, zero-crossing rate, autocorrelation, etc) and especially frequency-domain features (FFT, spectral entropy, etc) are used for activity recognition [3,25,26,16]. We also perform 16 sample-long FFT on the accelerometer signals from each axis to get a sequence of 27-dimensional raw features.

GPS. The 2D GPS coordinates are obtained after a picture is taken from the camera, and therefore have a similar sampling rate as the images ($\sim$ 1 per minute). The GPS unit is turned off after acquiring the coordinates to preserve battery life and has to lock-on to satellite signals each time. The coordinates are often unavailable due to the failure to lock on inside a building, and they are treated as missing data.

Image. We use 24-bit color JPEG images of size 480 x 640, although much higher resolution is available in current phones. To handle a large amount of data per day, we took a sparse number of pictures ($\sim$ 1 per minute). Original images are stored in the server for lifelogging purpose, but they cannot not be used directly in analysis for privacy reasons.

Table 1. List of 30 tags in three categories describing daily activity

Category	Tags
Activity	other activity, walk, drive/inside a vehicle, eat/drink, talk/chat/discuss, chores (cook/clean/laundry/etc), exercise/play sports, listen to a lecture, give a lecture, shop in a store, tend to baby, use a computer, watch tv/movie, pick up/drop off, read/write on paper/board
Place	other place, my home, my office, classroom/meeting room, other's office, restaurant/cafe, store, public places, outdoor
People	other people, my family, friend(s), colleague(s), stranger(s), crowd

A simple feature extraction can be performed to remove identity-revealing information from the images. Kim et al [20] used CIELAB-space color map and orientation map to find salient regions, and then computed SIFT features [28] as visual descriptors. Doherty et al [11] used scalable color and edge histogram, which is a subset of MPEG-7 feature descriptors [5]. We opt for the latter approach, and use 64-bin HSV-space color histogram and 80-dimensional edge histogram, which leave us with 144-dimensional raw features per image.

Audio. Audio is recorded with a 11,025 Hz sampling rate in 16 bits PCM format. Various audio features have been used for activity recognition. Lester et al [25] used linear and log-scale FFT frequency coefficients, cepstral coefficients, spectral entropy, band-pass filter coefficients, correlations, integrals, means, and variances. Pärkkä et al [32] used a speech/music discriminator [33] to detect speech from the audio. Kim et al [20] used 26-dim MFCC features along with zero crossing rate, linear predictive coding, volume standard deviation, non-silence ratio, and spectral centroid and pitch, which are a subset of MPEG-7 audio descriptors [29]. Here we use MFCC with 20 filter banks and 13 cepstral coefficients with a 25 ms time window.

Although spectral features such as MFCC are popular in audio processing, they contain enough information to partially reconstruct the speech contents, and cannot be considered as privacy-protecting. One solution to the problem is to use summary features from further processing, such as spectral entropy or energy [47]. In [8], these features were computed on-the-fly on a dedicated device. Our solution to the privacy issue is to use a sparse, short-duration sampling of audio. We sample 250 ms audio fragments for every 5 seconds, effectively discarding 95% of the data before we compute features. The 250 ms fragment, which is shorter than the typical duration of a word, and the sparse sampling together make the overall speech unreconstructible. To get the final feature, we stack MFCC coefficients from non-overlapping windows, which results in 126-dimensional raw features for each audio fragment.

2.3 User Segmentation and Annotation

Through a visual user interface, a user reviews his or her visual log of daily recordings, segments each day into a few ($\sim$10) meaningful events, and annotates each event with

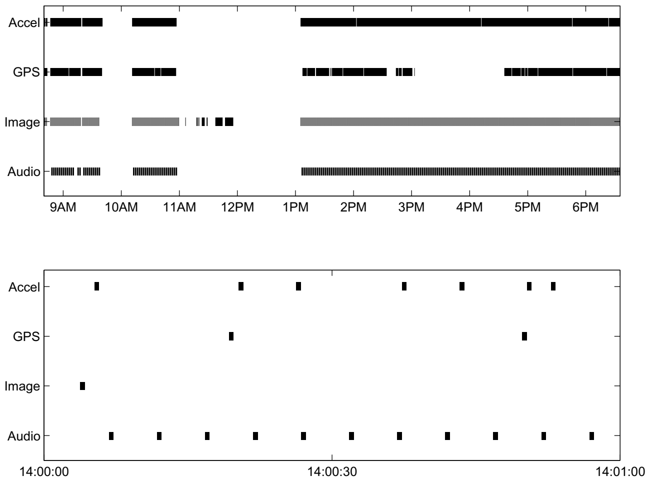

Fig. 1. The figure exemplifies several difficulties of handling heterogeneous multisensory data. (Top) Arrival times of multisensory streams for one day. The figure shows chunks of data are missing for all or some of the sensors. (Bottom) A one-minute portion of the streams around 2 PM, which shows asynchronous arrival times from different sensors. Also images arrive at one sample per minute, while accelerometer and audio samples arrive in higher and more irregular rates.

multiple tags. We use high-level tags that can cover a wide range of daily events, categorized into activity, place, and people (see Table 1). The tags are not mutually exclusive, e.g., a user can select [“eating” and “talking”] in a [“restaurant”] with [“friend” and “family”], or any combination of tags that best describe the event. In our experience, manual annotation of one day by segmentation followed by tagging can be performed in about ten minutes per day, which makes it practical to collect long-term annotated data from non-expert users.

3 Multisensory Bag-of-Words

Processing multisensory data streams in real-life recordings from a smartphone poses several difficulties: 1) samples of different modalities arrive at different times with different rates, e.g., $\sim$1/60 Hz for images vs $\sim$12 Hz for audio fragments; 2) sampling rates of a single modality can change over time, e.g., when other processes on the phone are taking up the CPU resources; 3) chunks of data can be randomly missing for some duration, e.g., when the GPS fails to lock-in indoors, or the camera view is obstructed inside a pocket. Figure 1 shows an example of a day’s recording demonstrating the problems.

These problems can be efficiently handled by the bag-of-words representation. The bag-of-words model assumes that a datum(=document) is a bag of unordered symbols (=words), and that the information about the data is contained fully in the frequency of symbols. Our motivation for the sensory bag-of-words representation is explained by

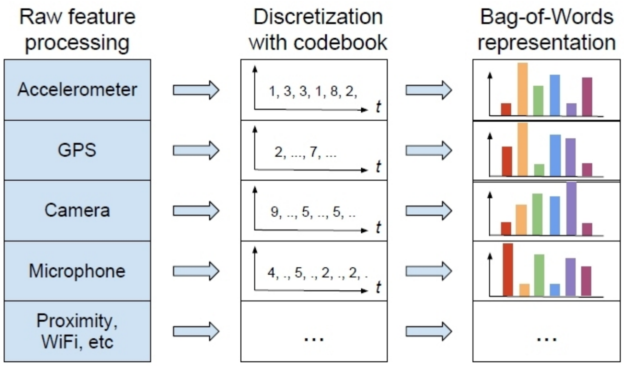

Fig. 2. Multisensory bag-of-words feature representation. Multisensory streams from a smartphone are processed into raw features, quantized into symbolic sequences, and represented as the intermediate bag-of-words features for a given time window.

the following example: during an event in which a person is “eating lunch at a restaurant”, it is unrealistic to assume the existence of a single low-level “eating feature” or a “restaurant feature” that can uniquely and reliably identify eating or restaurant. Rather, it is more reasonable to assume that “eating” is a unique *distribution* of multiple low-level features such as walking to the place, standing in line, sitting at a table, and moving upper body; similarly, “restaurant” is also a distribution of low-level features such as a certain location, ambient sounds, and images of people and tables. Therefore, the distribution(=frequency) of low-level features in the bag-of-words representation serves as an intermediate feature that links the raw sensory features with high-level concepts of events. There is also an ample evidence of the utility of the bag-of-words model for activity recognition. Huynh et al. [17] used the bag-of-words representation of accelerometery with a probabilistic topic model, where the frequency of sensory symbols in a 30 minute window was viewed as a document. The authors used the model to classify daily routines such as {dinner, commuting, lunch, office work}. Chennuru et al. [7] introduced an analogy between natural language and ‘activity language’, and used n-gram statistics of the symbol sequences as intermediate features. The approach was demonstrated with accelerometer data in a classification problem of three activities {walking, running, and cycling}. Wu et al. [44] further extended the approach to symbol-level fusion of GPS and accelerometer, and applied it to classification of 13 activities from a 5 day recording.

In this work, we further promote the multisensory bag-of-word representation as a general framework for transforming any number of heterogeneous sensory streams into homogeneous and normalized features at any desired times scale, that can be combined with any subsequent classification stage (see Figure 2 for illustration.)

To use the model, we first build sensor-specific codebooks (=vocabularies) of K words from the unlabeled collection of data in the training phase. We use K-means clustering to find K dominant modes of low-level features. In the testing phase, raw features are discretized to a symbolic sequence using the codebook of K words. We used

$K = 50$ for all modalities, but it is possible for different modalities to have different values of K. We create one bag-of-words feature for every minute – which we will call a frame – which is the sampling rate of the slowest modality (=image). Missing data can be handled by two ways – by imputing the missing portion from a modality with unbiased prior values (=mean histograms), or by assigning special 'missing' classes to the missing portion of data. Here we use the former approach.

4 Algorithms

We consider automatic annotation as the problem of predicting the presence or absence of each tag at each time frame. The presence or absence of a tag such as an activity or a location likely to be persistent in time rather than changing for every minute. This motivates the use of temporal models for capturing the temporal dependency of nearby states and treating the prediction problem as a sequence labeling problem. In this section, we review several state-of-the-art classifiers that we compare during evaluation.

4.1 Generative Models

Multinomial Naive Bayes. A Multinomial Naive Bayes (MNB)[30] classifier assumes a generative bag-of-words model. Here we adapt the model to our multisensory case. Assume that the data consist of T data frames $W_{1:T} = \{W_1, ..., W_T\}$ and the corresponding T labels $y_{1:T} = \{y_1, ..., y_T\}$. Each frame W_t consists of a collection of observed words from M sensory vocabularies. In MNB, the probability of observing a word v from m-th sensory modality given the state y is

$$p(w^m = v|y = s) = \frac{\exp \phi^m_{v,s}}{\sum_v \exp \phi^m_{v,s}}.$$

Under the 'naive' assumption, the probability of observing all words from M modalities in frame t is

$$p(W_t|y_t = s) = \prod_{m=1}^{M} \prod_{v=1}^{K^m} \left\{ \frac{\exp \phi^m_{v,s}}{\sum_v \exp \phi^m_{v,s}} \right\}^{N^m_v}, \quad (1)$$

where K^m is the size of vocabulary for modality m, and N^m_v is the count of word v from modality m in the frame. The model is trained by finding the maximum likelihood estimates of the parameters $\phi^m_{v,s}$ and the prior probability of a frame belonging to state s: $p(y_t = s) = \pi_s$. Using $p(W|y)$ and $p(y)$, the classification of a test frame W is performed by calculating the posterior probability of each state given the data dictated by the Bayes' rule

$$p(y = s|W) = \frac{\pi_s \cdot p(W|y = s)}{\sum_W \pi_s \cdot p(W|y = s)},$$

and by selecting the state with the highest probability $\arg\max_s p(y = s|W)$.

Hidden Markov Model. A Hidden Markov Model (HMM) assumes that 1) the hidden label state of a frame is conditionally independent of the state of other labels given the state of the preceding frame (a first-order Markov assumption), and 2) the observed feature at a frame is conditionally independent of other states or observations given the state of the frame. The joint probability of a HMM under these assumptions is

$$p(W_{1:T}, y_{1:T}) = p(y_1) \prod_{t=1}^{T} p(W_t|y_t) \cdot \prod_{t=2}^{T} p(y_t|y_{t-1}).$$

The model is specified by the three probabilities: initial probability of the state $p(y_1)$, probability of transition $p(y_t|y_{t-1})$, and emission probability $p(W_t|y_t)$ which models the probability of observing multisensory words in a frame given its state in Eq. 1. Note that if we assume that the state of a frame is independent of the state of any other frame, then HMM simply reduces to MNB.

The parameters of HMM are also learned from maximum likelihood estimation. When the hidden states of the training data is known, as in our case, the optimal parameters of the model can be computed in a closed form. In testing, the most likely sequence of hidden states $y_{1:t}$ of a new sequence $x_{1:t}$, is efficiently found by Viterbi decoding ([35]).

4.2 Discriminative Models 1

Logistic Regression. A Logistic Regression (LogReg) is a log-linear model for discriminative classification of vector data. Let the data consist of T real-valued feature vectors $x_{1:T} = \{x_1, ..., x_T\}$ and T labels $y_{1:T} = \{y_1, ..., y_T\}$, where each feature vector is a concatenated normalized histogram of multisensory words of the frame:

$$x_t = \left[\frac{N_1^1}{N^1}, ..., \frac{N_{K^1}^1}{N^1}, \; \cdots, \frac{N_1^M}{N^M}, ..., \frac{N_{K^M}^M}{N^M}\right]', \tag{2}$$

where K^M and N_v^m are defined the same as in Eq. 1, and N^M is the count of all words for modality m in the frame. Using this feature, the probability of the frame x_t belonging to state s is $p(y_t = s|x_t) = \exp(w_s' x_t)/\sum_s \exp(w_s' x_t)$, where w_s is the vector parameter for state s. For a two-class problem, only one vector w_1 is needed. Training involves learning the parameter w_s that maximizes the conditional log-likelihood of the T training data:

$$\log p(y_{1:T}|x_{1:T}) = \sum_t w_s' x_t - \sum_t \log \sum_s \exp(w_s' x_t).$$

Classification of a test data x is performed by evaluating the conditional likelihoods and selecting the most likely state: $\arg\max_s p(y = s|x)$.

Conditional Random Field. A Conditional Random Field (CRF)[21] is also a log-linear model for discriminative learning of arbitrary undirected graphical models, which was originally proposed to overcome the label-bias problem of HMMs for sequence labeling. In a (linear chain) CRF, the conditional probability of a state sequence $y_{1:T}$ given a data sequence $x_{1:T}$ is

$$p(y_{1:T}|x_{1:T}) = \frac{1}{Z(x_{1:T})} \exp(\sum_{j,t} w_j f_j(y_t, y_{t-1}, x_{1:T}))$$

where $Z(\cdot)$ is the partition function $Z(x_{1:T}) = \sum_{y_{1:T}} \exp(\sum_{j,t} w_j f_j(y_t, y_{t-1}, x_{1:T}))$. Note that $f_j(y_t, y_{t-1}, x_{1:T})$ generalizes the emission probability $\log p(x_t|y_t)$ and the transition probability $\log p(y_t|y_{t-1})$ from an HMM, in that it need not be a normalized probability, and that it can use data from all frames $x_{1:T}$ and not just the features of the same frame x_t. In the case where the features depend only on the state and the data of the same frame, that is, $f_j(y_t, y_{t-1}, x_{1:T}) = f_j(y_t, x_t)$, then CRF is similar to LogReg. We use two types of features: the multisensory histogram (Eq. 2) and the persistence of neighboring labels $I(y_{t-1} = y_t)$, where $I(\cdot)$ is an indication function.

For a linear chain CRF, training involves direct maximization of the conditional likelihood over the parameters w_j. In testing, the most likely sequence of the states $y_{1:t}$ of a new sequence $x_{1:t}$ is efficiently found by either an exact or approximate inference [21].

4.3 Discriminative Models 2

Support Vector Machine. A Support Vector Machine (SVM) is a discriminative, large-margin classifier which finds the separating hyperplane $\{w'x + b = 0\}$ for two classes of data. Assume we have the same T feature vectors $x_{1:T}$ and labels $y_{1:T}$ as in LogReg.

The optimization problem for a (linear, soft margin) SVM is to find a maximum margin solution

$$\min_{w,b,\xi} \ \frac{1}{2}w'w + C\sum_i \xi_i \ \text{ subject to}$$
$$y_i(w'x_i + b) \geq 1 - \xi_i, \ \ \xi_i \geq 0, \ \ \forall i,$$

where C is a parameter that allows one to manipulate the relative importance of increasing the margin versus classifying the training examples correctly, and ξ is the slack variable for allowing non-separability of classes. During training, the vector parameter w is found by the optimization, and C is determined by cross-validation. In testing, the state of the given frame x is predicted by the sign of the decision function $f(x) = w'x + b$. A nonlinear version of the SVM is obtained with a nonlinear mapping of samples $x \mapsto \phi(x)$ via a reproducing kernel.

Hidden Markov Support Vector Machine. A Hidden Markov Support Vector Machine (SVM-HMM) [2] is an example of a large-margin discriminative classifiers of structured data [40], with an HMM-like feature space.

Suppose the data is a collection of features sequences $X = \{\tilde{x}^1, ..., \tilde{x}^N\}$ and the corresponding collection of label sequences $Y = \{\tilde{y}^1, ..., \tilde{y}^N\}$, where a single feature sequence and a single label sequence are $\tilde{x}^i = x^i_{1:T} = \{x^i_1, ..., x^i_T\}$ and $\tilde{y}^i = y^i_{1:T} = \{y^i_1, ..., y^i_T\}$ as before.

The optimization problem for a (linear, soft margin) SVM-HMM is to find a maximum margin solution

$$\min_{w,\xi} \ \frac{1}{2}w'w + C\sum_i \xi_i \ \text{ subject to}$$

$$w'\varPhi(\tilde{x}^i, \tilde{y}^i) - w'\varPhi(\tilde{x}^i, \tilde{y}) \geq 1 - \xi_i, \quad \xi_i \geq 0, \ \ \forall i, \ \tilde{y} \neq \tilde{y}^i,$$

where $\varPhi(\tilde{x}^i, \tilde{y}^i)$ is a joint feature-output vector, and C is again a regularization parameter. In SVM-HMM, the joint feature-output vector captures the dependency between feature and label, and between labels of contiguous frames, such as those used in CRF: the multisensory histogram (Eq. 2) and $I(y_{t-1} = y_t)$.

The training and testing procedures are similar to SVM but they require evaluation of $\arg\max_{\tilde{y}} \ w'\varPhi(\tilde{x}, \tilde{y})$, which can be done efficiently with a Viterbi-like algorithm from HMM.

5 Experiments

5.1 Lifelogging Data

We collected lifelogging data from a volunteer, who carried the smartphone during daytime in weekdays and in weekends and provided segmentation and annotation of the data. We discarded segments with only private images or segments with less than 5 images, and finally collected 195 hours or 11721 minutes of continuous multisensory recordings of 42 days. The total number of user-defined segments in 42 days was 390, with an average of 9.3 segments per day. Out of the 30 tags (Table 1) that the user used, several tags occurred too infrequently to train and test classifiers reliably. We therefore collapsed those 11 tags that occurred less than 10 times into “other ...” tags, and used the remaining 19 tags in the experiments.

5.2 Comparison of Algorithms

We compare the classification accuracy of the six algorithms: MNB, LogReg, SVM, HMM, CRF, and SVM-HMM. For all algorithms, nearly the same features are used: multisensory word counts for generative models (MNB, HMM), and normalized multisensory histograms for discriminative models (LogRg, SVM, SVM-HMM). Although discriminative models can use more general features than generative models (e.g., histograms from overlapping time windows), we keep the features the same for comparison purposes. We used the packages LIBSVM [6] and the SVM^{hmm}[18], and in-house libraries for the other algorithms. Additional parameters for MNB and HMM are smoothing (=Dirichlet) hyper-parameters α and β for word emission and state prior parameters, which are set to 1. LogReg and CRF use additional regularization (=Gaussian) hyper-parameters, which we set to $\lambda = 10^{-12}$. SVM and HMM-SVM require selection of the coefficient C, which are chosen by five-fold cross-validation within the training data.

Performance is measured by leave-one-day-out classification accuracy for each tag, that is, a model is trained using 41 days and is used to predict the remaining one day.

Table 2. Per-frame classification accuracy of six algorithm for each tag. (mean and s.d. over 42 days.) Boldface means the best result.

		Non-temporal			Temporal		
	Tag	MNB	LogReg	SVM	HMM	CRF	SVM-HMM
Activity	other activity	0.774±0.177	0.804±0.168	**0.806±0.173**	0.784±0.204	0.769±0.230	0.801±0.225
	walk	0.941±0.058	0.968±0.037	0.969±0.040	0.943±0.060	0.950±0.061	**0.973±0.042**
	drive/inside a vehicle	0.904±0.091	0.966±0.055	0.973±0.049	0.883±0.151	0.972±0.045	**0.977±0.047**
	eat/drink	0.837±0.100	0.896±0.081	0.895±0.082	0.845±0.150	0.857±0.137	**0.921±0.094**
	talk/chat/discuss	0.726±0.106	0.797±0.112	0.798±0.113	0.771±0.139	0.824±0.155	**0.870±0.120**
	chores (cook/clean/laundry/etc)	0.846±0.150	0.981±0.039	0.984±0.039	0.772±0.222	0.947±0.097	**0.985±0.039**
	tend to baby	0.798±0.200	0.946±0.091	0.954±0.089	0.747±0.275	0.893±0.150	**0.955±0.089**
	use a computer	0.798±0.113	0.873±0.099	0.875±0.100	0.826±0.144	0.837±0.164	**0.905±0.108**
	read/write on paper/board	0.782±0.168	0.934±0.121	0.937±0.127	0.703±0.272	0.885±0.147	**0.940±0.129**
Place	other place	0.838±0.138	0.940±0.092	0.941±0.093	0.815±0.177	0.913±0.134	**0.942±0.099**
	my home	0.866±0.105	0.898±0.117	0.898±0.123	0.891±0.115	0.890±0.161	**0.935±0.139**
	my office	0.837±0.097	0.891±0.088	0.894±0.089	0.883±0.127	0.908±0.123	**0.941±0.079**
	classroom/meeting room	0.850±0.103	**0.930±0.081**	0.930±0.083	0.837±0.150	0.903±0.163	0.928±0.104
	other's office	0.834±0.143	0.937±0.104	0.939±0.099	0.777±0.229	0.902±0.137	**0.947±0.089**
	restaurant/cafe	0.929±0.062	0.954±0.053	0.958±0.049	0.976±0.032	0.948±0.093	**0.985±0.025**
	outdoor	0.945±0.053	0.971±0.033	0.972±0.034	0.946±0.061	0.951±0.064	**0.978±0.028**
People	other people	0.685±0.136	0.728±0.131	0.729±0.133	0.732±0.187	0.731±0.178	**0.763±0.175**
	my family	0.778±0.173	0.827±0.199	0.825±0.206	0.768±0.180	**0.840±0.193**	0.835±0.227
	colleague(s)	0.767±0.114	0.812±0.102	0.812±0.100	0.816±0.137	0.867±0.130	**0.894±0.100**
Average		0.828±0.070	0.898±0.072	0.899±0.073	0.827±0.077	0.884±0.063	**0.920±0.063**

This process is repeated for 42 times with a different held-out day. The presence or absence of 19 tags at each frame of the test day is predicted by a binary classifier for each tag instead of a multiclass classifier for all tags, since an arbitrary combination of multiple activities, places, and people can occur concurrently at each time frame.

5.3 Results

The per-frame classification rates of six classifiers are summarized in Table 2. The table shows the following trends:

- Accuracy-wise, MNB $\sim$ HMM $<$ CRF $\sim$ LogReg $\sim$ SVM $<$ SVM-HMM.
- Certain tags, such as "walk" and "drive" in activity, and "restaurant/cafe" and "outdoor" in places, are recognized very accurately (> 0.95) for all algorithms.
- Recognition of people is understandably more difficult than activity and place, since we have not used face-related visual features.

It is interesting to see that the two non-temporal models (MNB, LogReg) are not necessarily worse than their temporal counterparts (HMM, CRF). However, it is seen that discriminative models outperform generative models, which is a common observation across different problem domains, and that SVM-HMM outperforms HMM and CRF by a large margin, similar to the results from smart-home data [45].

Lastly, Figure 3 shows predicted tags from SVM-HMM during the course of a day. The predictions are overall satisfactory except for the middle section of the day.

6 Discussion

In this paper, we present a system for collecting and automatically annotating daily experience from multisensory streams on a smartphone. We use a flexible multisensory

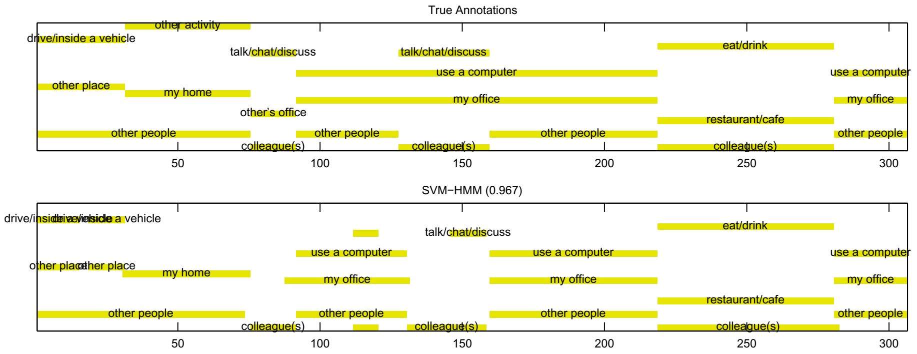

Fig. 3. True vs automatic annotations from data during one sample day. The x-axis is the time in minutes and y-axis is the 19 tags. The yellow dots indicate the positive prediction for each tag at each frame. To reduce visual clutter, we only show tags that last longer than 10 minutes.

bag-of-words representation for incorporating multiple heterogeneous streams, along with state-of-the-art classifiers to predict the tags with real-life data acquired with our system. In the evaluation of algorithms, we compared the performance of various classifiers, and achieved accurate predictions (>0.9) in the majority of tags (15 out of 19) with the SVM-HMM, which outperformed other algorithms by a large margin in almost all tags. The proposed system and algorithms can be immediately used for personal lifelogging applications. For example, using tags as keywords, frames of interest can be retrieved automatically from the vast amount of personal logs. For another example, personal statistics such as the amount of "walking", "driving", or "using a computer" can be calculated from the the predicted labels and provide the user with life-style related feedbacks.

There are several directions to improve the accuracy of automatic annotations with the proposed system. This includes extraction of better features from raw data, and adding more information such as time, Wi-Fi signals, assisted-GPS, ambient light, or proximity. The proposed multisensory framework can elegantly accommodate these inhomogeneous sensory data without modifying the system. The classifiers may also benefit from incorporating a longer-range label dependency or dependency among tags with Factorized HMM [15] or Dynamic CRF [38], as demonstrated in [46]. However, increased model complexity does not always translate to a better performance, and often results in inefficient model learning and inference, in contrast to the models in this paper for which only global optimization is required. While these improvements remain to be explored further, the current work represents a state-of-the-art system in collecting and annotating naturalistic daily experience with multisensory streams from smartphones.

Acknowledgements. The work was supported by the NSF grant IIS-1117707.

References

1. Altun, K., Barshan, B.: Human Activity Recognition Using Inertial/Magnetic Sensor Units. In: Salah, A.A., Gevers, T., Sebe, N., Vinciarelli, A. (eds.) HBU 2010. LNCS, vol. 6219, pp. 38–51. Springer, Heidelberg (2010)
2. Altun, Y., Tsochantaridis, I., Hofmann, T.: Hidden markov support vector machines. In: Proceedings of the Twentieth International Conference on Machine Learning, pp. 3–10. AAAI Press (2003)
3. Bao, L., Intille, S.S.: Activity Recognition from User-Annotated Acceleration Data. In: Ferscha, A., Mattern, F. (eds.) PERVASIVE 2004. LNCS, vol. 3001, pp. 1–17. Springer, Heidelberg (2004)
4. Bieber, G., Voskamp, J., Urban, B.: Activity Recognition for Everyday Life on Mobile Phones. In: Stephanidis, C. (ed.) UAHCI 2009, Part II. LNCS, vol. 5615, pp. 289–296. Springer, Heidelberg (2009)
5. Blighe, M., le Borgne, H., O'Connor, N., Smeaton, A.F., Jones., G.: Exploiting context information to aid landmark detection in sensecam images. In: ECHISE 2006 - 2nd International Workshop on Exploiting Context Histories in Smart Environments - Infrastructures and Design, Ubicomp (2006)
6. Chang, C.-C., Lin, C.-J.: LIBSVM: A library for support vector machines. ACM Transactions on Intelligent Systems and Technology 2, 27:1–27:27 (2011)
7. Chennuru, S.K., Chen, P.-W., Zhu, J., Zhang, Y.: Mobile lifelogger - recording, indexing, and understanding a mobile user's life. In: Proceedings of the Second International Conference on Mobile Computing, Applications, and Services, Santa Clara, CA, USA (2010)
8. Choudhury, T., Borriello, G., Consolvo, S., Haehnel, D., Harrison, B., Hemingway, B., Hightower, J., Klasnja, P., Koscher, K., LaMarca, A., Landay, J.A., LeGrand, L., Lester, J., Rahimi, A., Rea, A., Wyatt, D.: The mobile sensing platform: An embedded activity recognition system. IEEE Pervasive Computing 7, 32–41 (2008)
9. Chung, P.C., Liu, C.-D.: A daily behavior enabled hidden markov model for human behavior understanding. Pattern Recogn. 41(5), 1589–1597 (2008)
10. Doherty, A.R., Caprani, N., Conaire, C., Kalnikaite, V., Gurrin, C., Smeaton, A.F., O'Connor, N.E.: Passively recognising human activities through lifelogging. Comput. Hum. Behav. 27(5), 1948–1958 (2011)
11. Doherty, A.R., Ó Conaire, C., Blighe, M., Smeaton, A.F., O'Connor, N.E.: Combining image descriptors to effectively retrieve events from visual lifelogs. In: Proceedings of the 1st ACM International Conference on Multimedia Information Retrieval, pp. 10–17. ACM, New York (2008)
12. Farrahi, K., Gatica-Perez, D.: Discovering routines from large-scale human locations using probabilistic topic models. ACM Trans. Intell. Syst. Technol. 2(1), 3:1–3:27 (2011)
13. Fei-Fei, L., Perona, P.: A bayesian hierarchical model for learning natural scene categories. In: 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, vol. 2, pp. 524–531 (2005)
14. Gemmell, J., Bell, G., Leuder, R., Drucker, S., Wong, C.: MyLifeBits: Fulfilling the Memex Vision. In: Proceedings of Multimedia 2002 (2002)
15. Ghahramani, Z., Jordan, M.I.: Factorial hidden markov models. Mach. Learn. 29(2-3), 245–273 (1997)
16. Gu, T., Wang, L., Wu, Z., Tao, X., Lu, J.: A pattern mining approach to sensor-based human activity recognition. IEEE Trans. on Knowl. and Data Eng. 23, 1359–1372 (2011)
17. Huynh, T., Fritz, M., Schiele, B.: Discovery of activity patterns using topic models. In: Proceedings of the 10th International Conference on Ubiquitous Computing, pp. 10–19. ACM, New York (2008)

18. Joachims, T.: SVM-HMM : Sequence tagging with structural support vector machines (2008)
19. Kim, E., Helal, S.: Modeling Human Activity Semantics for Improved Recognition Performance. In: Hsu, C.-H., Yang, L.T., Ma, J., Zhu, C. (eds.) UIC 2011. LNCS, vol. 6905, pp. 514–528. Springer, Heidelberg (2011)
20. Kim, I.-J., Ahn, S.C., Ko, H., Kim, H.G.: Automatic lifelog media annotation based on heterogeneous sensor fusion. In: Proceedings of IEEE International Conference on Multi Sensor Fusion and Integration for Intelligent Systems, Seoul, Korea, August 20-22 (2008)
21. Lafferty, J.D., McCallum, A., Pereira, F.C.N.: Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In: Proceedings of the Eighteenth International Conference on Machine Learning, pp. 282–289. Morgan Kaufmann Publishers Inc., San Francisco (2001)
22. Lane, N.D., Miluzzo, E., Lu, H., Peebles, D., Choudhury, T., Campbell, A.T.: A survey of mobile phone sensing. Comm. Mag. 48, 140–150 (2010)
23. Lara, O.D., Labrador, M.A.: A survey on human activity recognition using wearable sensors. Submitted to IEEE Communications Surveys and Tutorials (2012)
24. Lara, O.D., Perez, A.J., Labrador, M.A., Posada, J.D.: Centinela: A human activity recognition system based on acceleration and vital sign data. In: Pervasive and Mobile Computing (2011)
25. Lester, J., Choudhury, T., Borriello, G.: A Practical Approach to Recognizing Physical Activities. In: Fishkin, K.P., Schiele, B., Nixon, P., Quigley, A. (eds.) PERVASIVE 2006. LNCS, vol. 3968, pp. 1–16. Springer, Heidelberg (2006)
26. Li, M., Rozgić, V., Thatte, G., Lee, S., Emken, A., Annavaram, M., Mitra, U., Spruijt-Metz, D., Narayanan, S.: Multimodal physical activity recognition by fusing temporal and cepstral information. IEEE Transactions on Neural Systems and Rehabilitation Engineering 18(4), 369–380 (2010)
27. Logan, B., Healey, J., Philipose, M., Tapia, E.M., Intille, S.: A Long-Term Evaluation of Sensing Modalities for Activity Recognition. In: Krumm, J., Abowd, G.D., Seneviratne, A., Strang, T. (eds.) UbiComp 2007. LNCS, vol. 4717, pp. 483–500. Springer, Heidelberg (2007)
28. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. Int. J. Comput. Vision 60(2), 91–110 (2004)
29. Manjunath, B.S., Salembier, P., Sikora, T. (eds.): Introduction to MPEG-7: Multimedia Content Description Language. Wiley (April 2002)
30. McCallum, A., Nigam, K.: A comparison of event models for naive bayes text classification, pp. 41–48. AAAI Press (1998)
31. Nazerfard, E., Das, B., Holder, L.B., Cook, D.J.: Conditional random fields for activity recognition in smart environments. In: Proceedings of the 1st ACM International Health Informatics Symposium, pp. 282–286. ACM, New York (2010)
32. Pärkkä, J., Ermes, M., Korpipää, P., Mäntyjärvi, J., Peltola, J., Korhonen, I.: Activity classification using realistic data from wearable sensors. IEEE Transactions on Information Technology in Biomedicine 10(1), 119–128 (2006)
33. Penttilä, J., Peltola, J., Seppänen, T.: A speech/music discriminator-based audio browser with a degree of certanity measure. In: Proc. Infotech Oulu Int. Workshop Information Retrieval, pp. 125–131 (2001)
34. Pijl, M., van de Par, S., Shan, C.: An event-based approach to multi-modal activity modeling and recognition. In: Eigth Annual IEEE International Conference on Pervasive Computing and Communications, PerCom 2010, Mannheim, Germany, March 29 - April 2, pp. 98–106 (2010)
35. Rabiner, L.R.: A tutorial on hidden Markov models and selected applications in speech recognition. Proceedings of the IEEE 77(2), 257–286 (1989)
36. Riboni, D., Bettini, C.: Cosar: hybrid reasoning for context-aware activity recognition. Personal and Ubiquitous Computing 15, 271–289 (2011)

37. Sminchisescu, C., Kanaujia, A., Li, Z., Metaxas, D.: Conditional random fields for contextual human motion recognition. In: Proceedings of the Tenth IEEE International Conference on Computer Vision, vol. 2, pp. 1808–1815. IEEE Computer Society, Washington, DC (2005)
38. Sutton, C., McCallum, A., Rohanimanesh, K.: Dynamic conditional random fields: Factorized probabilistic models for labeling and segmenting sequence data. J. Mach. Learn. Res. 8, 693–723 (2007)
39. Takata, K., Ma, J., Apduhan, B.O., Huang, R., Jin, Q.: Modeling and analyzing individual's daily activities using lifelog. In: Proceedings of the 2008 International Conference on Embedded Software and Systems, pp. 503–510. IEEE Computer Society, Washington, DC (2008)
40. Tsochantaridis, I., Joachims, T., Hofmann, T., Altun, Y.: Large margin methods for structured and interdependent output variables. J. Mach. Learn. Res. 6, 1453–1484 (2005)
41. Vail, D.L., Veloso, M.M., Lafferty, J.D.: Conditional random fields for activity recognition. In: Proceedings of the 6th International Joint Conference on Autonomous Agents and Multiagent Systems, pp. 235:1–235:8. ACM, New York (2007)
42. van Kasteren, T., Noulas, A., Englebienne, G., Kröse, B.: Accurate activity recognition in a home setting. In: Proceedings of the 10th International Conference on Ubiquitous Computing, pp. 1–9. ACM, New York (2008)
43. Wang, H., Huang, M., Zhu, X.: A generative probabilistic model for multi-label classification. In: Proceedings of the 2008 Eighth IEEE International Conference on Data Mining, pp. 628–637. IEEE Computer Society, Washington, DC (2008)
44. Wu, P., Peng, H.-K., Zhu, J., Zhang, Y.: SensCare: Semi-automatic Activity Summarization System for Elderly Care. In: Zhang, J.Y., Wilkiewicz, J., Nahapetian, A. (eds.) MobiCASE 2011. LNICST, vol. 95, pp. 1–19. Springer, Heidelberg (2012)
45. Wu, T.Y., Hsu, J.Y.J., Chiang, Y.T.: Continuous recognition of daily activities from multiple heterogeneous sensors. In: AAAI Spring Symposium: Human Behavior Modeling, pp. 80–85. AAAI (2009)
46. Wu, T.-Y., Lian, C.-C., Hsu, J.-Y.: Joint Recognition of Multiple Concurrent Activities using Factorial Conditional Random Fields. In: AAAI Workshop on Plan, Activity, and Intent Recognition, Technical Report WS-07-09. The AAAI Press, Menlo Park (2007)
47. Wyatt, D., Choudhury, T., Kautz, H.: Capturing spontaneous conversation and social dynamics: A privacy sensitive data collection effort. In: Proc. of ICASSP (2007)
48. Zappi, P., Stiefmeier, T., Farella, E., Roggen, D., Benini, L., Tröster, G.: Activity recognition from on-body sensors by classifier fusion: Sensor scalability and robustness. In: 3rd Int. Conf. on Intelligent Sensors, Sensor Networks, and Information Processing, pp. 281–286 (2007)
49. Zhu, Y., Arase, Y., Xie, X., Yang, Q.: Bayesian nonparametric modeling of user activities. In: Proceedings of the 2011 International Workshop on Trajectory Data Mining and Analysis, pp. 1–4. ACM, New York (2011)