Detecting Granular Eating Behaviors From Body-worn Audio and Motion Sensors

Mark Mirtchouk
Department of Computer Science
Stevens Institute of Technology
Hoboken, USA
mmirtcho@stevens.edu

Samantha Kleinberg
Department of Computer Science
Stevens Institute of Technology
Hoboken, USA
samantha.kleinberg@stevens.edu

***Abstract*—Wearable sensor technology has made it possible to gain insight into dietary activity, learning not only when people are eating, but identifying fine-grained behaviors such as chews per minute, and causes of food choices. This may enable interventions to maintain health and assist individuals with chronic diseases such as diabetes (e.g. by providing insulin dosing assistance). However, existing work on dietary monitoring has focused on identifying meal times, rather than fine grained behavior such as chewing. A key barrier is the difficulty of obtaining granular ground truth. In free-living environments it is difficult to obtain the high-quality video needed, and annotating large datasets is labor intensive and does not scale well. To address this, we introduce a new multi-stage initialization approach for Stochastic Variational Deep Kernel Learning (SVDKL) that enables learning from data with a mix of coarse labels (meal times) and granular ones (chews, intakes). Our approach outperforms the state of the art on both free-living and laboratory datasets, with 84% recall and 67% precision for detecting chews compared to prior results of 73% precision and 34% recall on the same data. Ultimately, our work may enable more types of human activity recognition from real-world environments at a lower cost.**

I. Introduction

Nutrition is essential for maintaining health and managing many chronic conditions such as diabetes, and much work in nutritional epidemiology aims to uncover how specific foods are linked to health. However unlike physical activity, which can be monitored in daily life with many consumer devices, it has remained difficult to obtain long-term large scale dietary data. With such data, though, it may be possible to learn how specific foods affect blood glucose in individuals with diabetes, and how meal features such as chewing speed relate to bodyweight. Doing this requires not only finding meal periods, but identifying foods consumed and other granular components of eating such as chewing.

Existing methods for automated dietary monitoring (ADM), such as those based on body-worn audio and motion sensors [1], [2], have mainly focused on identifying meal periods. While these works have high accuracy for identifying meal times, and have been used in both controlled and free-living (FL) environments [3], they provide insight into only one aspect of diet. In recent years, ADM has expanded to identifying foods consumed [4] and other aspects of nutrition such as fluid intake [5]. However a core challenge remains: as these works all use supervised learning, they require ground truth labels. As ADM moves outside the lab to identify activity in the environments of daily life and with large diverse populations, obtaining highly detailed labels is a significant barrier. Deep learning has excelled at other time series classification tasks [6], but dietary datasets tend to be small and imbalanced, with few eating events during an entire day of data. Further, individuals vary significantly in their behavior, making it important to leverage personalized data when available [7].

To address this we introduce a new approach that leverages data with highly granular (e.g. chews, intakes) and coarse (e.g. meal times) labels for classification of granular eating activities. FL data lacks detailed ground truth but provides insight into a wider variety of eating behavior, while lab data is generally smaller but with more trustworthy labels. We introduce init-SVDKL, an initialization procedure for Stochastic Variational Deep Kernel Learning (SVDKL) [8]. While SVDKL learns temporal dependencies (e.g. a food intake is likely to be followed by chewing), it cannot naturally make use of data with different label granularities or consider data relevance in its training. Our extension, init-SVDKL, makes use of individual training data and makes the model more likely to converge in fewer epochs. We evaluate this approach on multimodality (audio, motion) sensor data from free-living (coarse ground truth) and laboratory (granular ground truth) environments [7], [9]. On the lab data, our approach significantly increases recall of granular events (84% chew, 78% food intake, 88% drink intake) compared to prior work on this data (34%, 26% and 19% respectively) [9], with similar or higher precision. On FL data, we significantly increase precision.

II. Related Work

ADM has used a variety of body-worn and environmental sensors to detect eating behaviors. Body-worn microphones [1] and motion sensors (mounted on the head [10] or wrist) [2] have been used to detect eating periods or activities such as chews. While microphones can pick up on chewing, and motion sensors can be used to find intakes, multiple sensing modalities are needed to capture all activities [9], and to identify food type and amount consumed [11]. This information could be used to guide insulin dosing for individuals with diabetes by linking it to an artificial pancreas. Image-based

approaches such as with photos taken by users or by egocentric cameras can be used to identify meal periods and food type consumed [12], [13], but they cannot be used in real time to identify the fine-grained behaviors we aim to infer.

All sensor types have been used in both lab and FL environments, but the types of ground truth available in each differ significantly. Labs can be outfitted with video cameras and semi-wild studies can set up mobile cameras [14], but this can change participant behavior [15] and could pose privacy concerns in FL. Further, it is labor intensive to annotate such video, making it challenging to collect large-scale datasets with fine-grained ground truth. As a result, most prior work focuses on identifying meal periods, which can be logged by participants, and fine-grained activities such as chewing have been studied mainly in lab. This means that FL accuracy is highly dependent on finding the beginning and end times of a meal (both of which may have noisy ground truth, due to logging delays or omissions) and this approach precludes identification of individual intakes and eating events such as chewing. Prior work [4] has shown that it is possible to label intakes with foods consumed in unconstrained settings with unconstrained food choices, but it relied on correctly identifying the intakes and the pattern of chews that follow them. Thus to fully realize the potential for automatically created meal logs and understanding of fine-grained eating behavior, we need approaches that do not require large sets of finely labeled training data.

III. Methods

We propose a novel training method that classifies micro events (e.g. chews) during macro events (e.g. meals) in time series data and allows learning from data labeled at both levels of granularity. We focus on learning from dietary data collected from lab (annotated with fine-grained ground truth), and free-living (annotated at a coarse level of granularity) environments. Our approach may be applied to many human activity recognition tasks where data is labeled at varying levels of granularity, reducing labeling costs. Below we briefly overview SVDKL before discussing the components of our initialization procedure and finally the full pipeline.

A. SVDKL

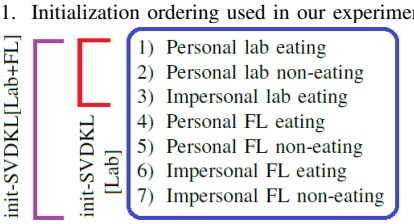
We build on SVDKL as it is robust to missingness, learns temporal dependencies, and can accurately perform multi-class classification [8]. SVDKL uses a combination of a deep neural network (DNN) and multiple Gaussian Processes (GPs), one per feature, to do multiclass classification. The DNN is used for feature reduction and the GPs are used to learn patterns within the data. SVDKL has been shown to improve the state of the art in image classification on digits such as MNIST and CIFAR10, but it has not been extended to time-series data, nor to problems where training data may vary in relevance. In the following sections we discuss how we extend SVDKL.

B. Notation

ADM and activity recognition often use a mix of data from other individuals (impersonal data) and from the individual whose activities are being classified (personal data). Personal data can significantly improve results, but we often face a choice between personal data with coarse grained annotations and impersonal data with finer grained labels. For the ADM task in this paper, macro labels are meal start and end times and micro labels are events that occur within a meal.

Data thus falls into four categories: [personal, impersonal] $\times$ [macro, micro]. The micro events are: discrete intake (I), continuous intake (Q), and chew (C). An I event has a single timepoint, such as when taking a bite of food. A Q event has a duration, and thus captures continuous events such as sipping soup. The only macro event here is "meal," which encompasses all eating and drinking (e.g. snacks, drinks).

Fig. 1. Initialization ordering used in our experiments.



C. Model initialization

Prior work has shown that pre-training can increase model accuracy compared to learning on an unordered set of data. We propose using multiple initialization stages, similar to hyperparameter tuning, guided by the structure of ADM. For the data used here we have three dimensions to consider: label granularity [micro, macro], personalization [personal, impersonal], and relevance [eating, non-eating]. In this dataset, all lab data has micro labels and all FL data has only macro labels. By relevance we consider whether data is during a meal (eating) or outside of meals (non-eating). Non-meal data is useful for training a classifier to distinguish eating from the other activities that may be confounded with it, but given the class imbalance found in real-world data (where most of a day is not spent eating), both types of data should not be given the same weight. Figure 1 shows our proposed ordering for initializing the model. We begin with micro-labelled data (lab). Within that category, we order data by relevance, beginning with personal lab eating, personal lab non-eating, and finally impersonal lab eating data. The rationale is that the model first learns micro meal events before learning to distinguish between meal and non-meal activity. Next, we use macro-labeled data (FL), beginning first with personal data (eating then non-eating) and finally impersonal data.

D. init-SVDKL

We now augment SVDKL with multiple initialization stages based on our ranking of the relevance of the data. These partitions are task specific, however, we expect that for other tasks sequencing the data by label type [micro, macro], and then within that by personalization [personal, impersonal] and then relevance [eating, non-eating]. To train on these multiple disjoint datasets, we use the output of training on one data set

as the initialization before training on the next dataset. Thus we begin by training an SVDKL on the most relevant data, then we use the parameters learned as initialization and train an SVDKL on the next most relevant data, until the least, and then finally on the training data.

E. Multi-pass approach

Our approach can misclassify certain events based on the nulls and other events that are outside macro events. A solution to this is to use a multi-pass approach: first classify micro events, combine the micro events to create macro events, and then classify micro events during only these macro events.

IV. Experiments

We test init-SVDKL on dietary datasets collected in the lab and in FL environments and evaluate it across multiple initializations and data orderings. We compare init-SVDKL to baselines used in previous work on these datasets [7], [9] and the deep learning baselines of long short-term memory (LSTM) and SVDKL. We further investigate the effect of initialization that uses only lab data and using both lab and FL data. We show that even though the FL data is noisy and coarsely labelled, our approach uses it to improve accuracy on detecting micro events in lab data, outperforming prior work and initialization using only lab data.

A. Datasets

We use multimodality datasets from laboratory and free living environments that have been described in prior work.

Lab [9]: 6 subjects (2 female, 4 male) aged 18-35 participated in two 6-hour data collection sessions. In total, 59 hours of data was recorded with 5.4 hours of eating across 30 meals.

FL [7]: 5 out of 6 lab participants (aged 18-35; 2 female, 4 male) participated in two data collection sessions ($\approx$12 hours each day) during their normal lives. In total 110.5 hours of data was recorded with 8.4 hours of eating across 30 meals. Data was also collected from 6 new participants aged 20-73 over either 2 days (5 participants) or 5 days (1 participants) during daily life. In total, 144.2 hours of data was recorded with 14.6 hours of eating across 51 meals. Thus we have 11 participants and 81 FL meals.

B. Sensors and processing

1) Sensors: We use the following multimodality sensors:

Audio: A custom earbud with 2 microphones (1 in-ear and 1 external) recorded audio data at 44.1 kHz. Two microphones were used for noise cancellation as described in [9]. Data was then down-sampled to 16 kHz.

Motion: Using an Android smartwatch (LG G Watch) on each wrist, 9-axis IMU data was recorded at a rate of 15Hz.

2) Feature extraction: Based on previous work on this data [7], [9], we extract the same features. We segment the raw motion data into 5s windows with a 100ms step size and raw audio data into 200ms frames with a 20ms step size and then extract audio (energy, spectral flux, zero-cross rate, 11 MFCC coefficients, centroid, spread, skewness, and kurtosis) and motion (mean, covariance, derivative, coefficients of 4th order polynomial fit to acceleration values, zero crossing rate of high-pass filtered acceleration components, and the standard deviation of the zero-crossing intervals) features.

3) Raw data: Deep learning methods usually perform best with raw data. Thus for SVDKL, init-SVDKL, and a version of LSTM, we segment both the raw motion data and audio data into non-overlapping 70ms windows, to ensure a maximum of one event per window.

4) Ground truth: Lab data was previously annotated with granular eating activities from video recordings. We use the chew, intake, and drink annotations as the other categories (e.g. swallow) are less prevalent. During training and testing, we create macro events by combining micro events with a gap of less than 1 minute, ensuring that all meals have either one second of drink or at least one chew and intake. FL data was annotated using participant logs of meal start and end times.

C. Evaluation and baselines

We compare our approach against random forest (used extensively in prior work), LSTM (as a deep learning baseline) and SVDKL without our initialization procedure. For all methods we use leave one session out (LOSO) evaluation, training on all but the target session and testing on the held out one. To train on solely macro-labelled data (FL), we apply a classifier trained on micro data (Lab) to label micro-events in FL, then combine the lab data and micro-labeled FL data. While the labels may be noisy, they still help the classifiers avoid overfitting to the more homogeneous lab data.

1) Random Forest (RF): We use the same settings as prior work, with an RF classifier for each micro event with 100 trees in lab data, and added a meal classifier for FL data [7].

2) LSTM: As LSTM can perform better with raw data, we test LSTM (features), which uses the extracted features and LSTM (raw), which uses only the raw data. We use a stateful LSTM, with a batch size of 1024, categorical crossentropy loss, a softmax activation function, and train for 300 epochs.

3) SVDKL and init-SVDKL: Both SVDKL and init-SVDKL use raw data with a batch size of 1024 for 300 epochs. We use a Deep Kernel Learning feature extractor with fully connected layers and the following architecture: $d \rightarrow 1000 \rightarrow 1000 \rightarrow 500 \rightarrow 50 \rightarrow 4$, where d is the size of the original raw data for each time window. We train one GP per DKL feature and combine them using a softmax layer. A Cholesky variational distribution is used with a multitask variational strategy through grid interpolation. The covariance matrix is calculated with a scale kernel on top of an RBF kernel with a smoothed box prior. We use stochastic gradient descent as the optimizer, and Variational Evidence Lower Bound as the loss function. An init-SVDKL is an SVDKL that uses section III-D to initialize all parameters before training.

We evaluate the precision and recall of micro events using the same tolerances as for previous work on this dataset: 250ms for chew, 500ms for intake, and 1000ms for drink [9] and evaluate meals by comparing the overlap between ground truth and inferred eating times at a level of deci-seconds [7].

TABLE I
PRECISION (PREC) AND RECALL FOR LAB DATASET. FOR EACH TASK, THE BEST RESULT IS BOLDED. ^{MPA} IS DESCRIBED IN SECTION III-E

	Chew		Intake		Drink		Meal		
	Prec.	Recall	Prec.	Recall	Prec.	Recall	Prec.	Recall	F1
RF	**73**	34	44	26	47	19	85	**92**	**89**
LSTM (features)	68	26	39	27	**51**	27	88	87	87
LSTM (raw)	70	42	38	33	50	35	87	86	86
SVDKL	45	70	38	56	23	66	76	77	76
init-SVDKL[Lab]	52	71	45	60	35	**88**	**91**	88	**89**
init-SVDKL[Lab+FL]	69	83	55	69	44	**88**	90	89	**89**
init-SVDKL[Lab+FL] ^{MPA}	67	**84**	**64**	**78**	45	**88**	**91**	85	88

TABLE II
COMPARISON OF METHODS PRECISION, RECALL, AND F1 SCORE FOR MEALS ON FL. FOR EACH TASK, THE BEST RESULT IS BOLDED.

	Meal		
	Precision	Recall	F1
RF	28	85	42
LSTM (features)	45	84	59
LSTM (raw)	52	83	64
SVDKL	49	82	62
init-SVDKL[Lab]	53	84	65
init-SVDKL[Lab+FL]	**63**	**88**	**74**

V. RESULTS

Table I shows detailed results on the lab data, comparing the baselines and initialization approaches. First, while RF has high precision and recall at detecting meals, it has the lowest recall for individual events. This is because identifying meal periods depends mainly on finding events at the start and end of a meal. For chews and intakes, our approach has an increase of $\geq 50\%$ over RF. One difference is that our multi-pass approach (detecting eating first, then micro events within meals) significantly increases recall of intakes, by better identifying the start of a meal. Lastly, our approach leads to a large increase in recall of drink intakes. While RF only detects 19%, and LSTM finds 35%, we detect 88% of drinking events. Thus, the methods that achieve the best meal-level results do not necessarily identify granular events most accurately, and our approach significantly improves on the accuracy of SVDKL, allowing detection of granular eating activities.

Table II shows FL results. While we have only meal-level ground truth, our approach increases both precision and recall on this data. This is mainly due to detecting shorter meals that LSTM and RF tended to miss or classify as a longer duration, though init-SVDKL did still miss the shortest meal (10.6 seconds of eating chocolate).

VI. CONCLUSION

We improve upon the state of the art in detecting eating behavior, using a new initialization procedure for SVDKL that better uses data of varying relevance, personalization, and label granularity. On laboratory data with ground truth we have significantly higher recall for detecting chewing (84% versus 34%), intakes (78% versus 26%), and drinking (88% versus 19%) with similar or higher precision and similar meal-level F1 scores. On FL data, we significantly increase precision over prior work, with strictly higher recall. Future work is needed to tune parameters to ensure even the briefest snacks are detected, and to determine how best to construct macro events from micro ones.

ACKNOWLEDGEMENTS

This work was supported in part by the NLM of the NIH under Award Number R01LM013308.

REFERENCES

[1] O. Amft, M. Stäger, P. Lukowicz, and G. Tröster, "Analysis of chewing sounds for dietary monitoring," in *UbiComp*, 2005.

[2] Y. Dong, A. Hoover, J. Scisco, and E. Muth, "A new method for measuring meal intake in humans via automated wrist motion tracking," *Appl Psychophysiol Biofeedback*, vol. 37, no. 3, pp. 205–215, 2012.

[3] A. Bedri, R. Li, M. Haynes, R. P. Kosaraju, I. Grover, T. Prioleau, M. Y. Beh, M. Goel, T. Starner, and G. Abowd, "Earbit: Using wearable sensors to detect eating episodes in unconstrained environments," *IMWUT*, vol. 1, no. 3, pp. 37:1–37:20, 2017.

[4] M. Mirtchouk, D. L. McGuire, A. L. Deierlein, and S. Kleinberg, "Automated estimation of food type from body-worn audio and motion sensors in free-living environments," in *Machine Learning for Healthcare*, 2019.

[5] T. Hamatani, M. Elhamshary, A. Uchiyama, and T. Higashino, "Fluidmeter: Gauging the human daily fluid intake using smartwatches," *IMWUT*, vol. 2, no. 3, 2018.

[6] H. I. Fawaz, G. Forestier, J. Weber, L. Idoumghar, and P.-A. Muller, "Deep learning for time series classification: a review," *Data Mining and Knowledge Discovery*, vol. 33, pp. 917–963, 2019.

[7] M. Mirtchouk, D. Lustig, A. Smith, I. Ching, M. Zheng, and S. Kleinberg, "Recognizing eating from body-worn sensors: Combining free-living and laboratory data," *IMWUT*, vol. 1, no. 3, 2017.

[8] A. G. Wilson, Z. Hu, R. R. Salakhutdinov, and E. P. Xing, "Stochastic variational deep kernel learning," in *NeurIPS*, 2016.

[9] C. Merck, C. Maher, M. Mirtchouk, M. Zheng, Y. Huang, and S. Kleinberg, "Multimodality sensing for eating recognition," in *Pervasive Health*, 2016.

[10] S. A. Rahman, C. Merck, Y. Huang, and S. Kleinberg, "Unintrusive Eating Recognition using Google Glass," in *Pervasive Health*, 2015.

[11] M. Mirtchouk, C. Merck, and S. Kleinberg, "Automated estimation of food type and amount consumed from body-worn audio and motion sensors," in *UbiComp*, 2016.

[12] J. Noronha, E. Hysen, H. Zhang, and K. Z. Gajos, "Platemate: Crowdsourcing nutritional analysis from food photographs," in *UIST*, 2011.

[13] G. Schiboni, F. Wasner, and O. Amft, "A privacy-preserving wearable camera setup for dietary event spotting in free-living," in *IEEE International Conference on Pervasive Computing and Communications Workshops*, 2018.

[14] K. Kyritsis, C. Diou, and A. Delopoulos, "Modeling wrist micromovements to measure in-meal eating behavior from inertial sensor data." *IEEE journal of biomedical and health informatics*, 2019.

[15] R. Alharbi, T. Stump, N. Vafaie, A. Pfammatter, B. Spring, and N. Alshurafa, "I can't be myself: Effects of wearable cameras on the capture of authentic behavior in the wild," *IMWUT*, vol. 2, no. 3, 2018.