EatingTrak: Detecting fine-grained eating moments in the wild using a wrist-mounted IMU

EatingTrak: Detecting Fine-grained Eating Moments in the Wild Using a Wrist-mounted IMU

RUIDONG ZHANG, Cornell University, USA
JIHAI ZHANG, Shanghai Jiao Tong University, China
NITISH GADE, Cornell University, USA
PENG CAO, Massachusetts Institute of Technology, USA
SEYUN KIM, Carnegie Mellon University, USA
JUNCHI YAN, Shanghai Jiao Tong University, China
CHENG ZHANG, Cornell University, USA

In this paper, we present EatingTrak, an AI-powered sensing system using a wrist-mounted inertial measurement unit (IMU) to recognize eating moments in a near-free-living semi-wild setup. It significantly improves the SOTA in time resolution using similar hardware on identifying eating moments, from over five minutes to three seconds. Different from prior work which directly learns from raw IMU data, it proposes intelligent algorithms which can estimate the arm posture in 3D in the wild and then learns the detailed eating moments from the series of estimated arm postures. To evaluate the system, we collected eating activity data from 9 participants in semi-wild scenarios for over 113 hours. Results showed that it was able to recognize eating moments at three time-resolutions: 3 seconds and 15 minutes with F-1 scores of 73.7% and 83.8%, respectively. EatingTrak would introduce new opportunities in sensing detailed eating behavior information requiring high time resolution, such as eating frequency, snack-taking, on-site behavior intervention. We also discuss the opportunities and challenges in deploying EatingTrak on commodity devices at scale.

CCS Concepts: • **Human-centered computing → Ubiquitous and mobile computing systems and tools**.

Additional Key Words and Phrases: Eating detection, neural networks, wrist-mounted IMU, arm posture estimation

ACM Reference Format:
Ruidong Zhang, Jihai Zhang, Nitish Gade, Peng Cao, Seyun Kim, Junchi Yan, and Cheng Zhang. 2022. EatingTrak: Detecting Fine-grained Eating Moments in the Wild Using a Wrist-mounted IMU. *Proc. ACM Hum.-Comput. Interact.* 6, MHCI, Article 214 (September 2022), 22 pages. https://doi.org/10.1145/3546749

1 INTRODUCTION

Journaling eating activities is frequently recommended as the first step to combat unhealthy eating behaviors [16]. Traditional approaches for journaling eating activities usually require the user

Authors' addresses: Ruidong Zhang, rz379@cornell.edu, Cornell University, Ithaca, NY, USA, 14853; Jihai Zhang, yunfan243332345@sjtu.edu.cn, Shanghai Jiao Tong University, Shanghai, China; Nitish Gade, ng375@cornell.edu, Cornell University, Ithaca, NY, USA, 14853; Peng Cao, pengcao@mit.edu, Massachusetts Institute of Technology, Cambridge, MA, USA; Seyun Kim, seyunkim@andrew.cmu.edu, Carnegie Mellon University, Pittsburgh, PA, USA; Junchi Yan, yanjunchi@sjtu.edu.cn, Shanghai Jiao Tong University, Shanghai, Shanghai, China; Cheng Zhang, chengzhang@cornell.edu, Cornell University, Ithaca, NY, USA, 14853.

2573-0142/2022/9-ART214 $15.00
https://doi.org/10.1145/3546749

to manually log eating activities on paper or a smartphone app. It heavily relies on the user's self-motivation and determination, which may not be always reliable [16].

To free users from tedious self-reporting tasks, a variety of wearable technologies were invented to automatically identify eating moments in the wild, which is the first step towards a fully automated eating activity journaling system. However, the existing solutions have two limitations: 1) They either require using a new hardware, which may not immediately applicable [5, 9, 30, 60, 64]; 2) The ones using commodity wearables (e.g., smartwatch) can not recognize eating moments in high time resolution. For instance, Thomaz et al. [48] used a wrist-mounted IMU to recognize eating moments in the wild in time resolutions of 5 minutes, 30 minutes, 60 minutes with F1 scores of around 30%, 40%, and 76.1% respectively. Other work [17, 46] can only recognize eating moments in the unit of a meal. In other words, these systems can not recognize the detailed eating moment within a meal, such as the eating speed or frequency, which are critical indicators for healthy problems, such as overweight and obesity problems [34]. Furthermore, many eating activities are much shorter than a meal. For instance, eating snacks or fruits can happen in a few seconds to a few minutes. With the time resolution of 5 minutes [48] or a meal [17, 46], detecting short eating moments would be extremely challenging for the above eating sensing systems using a commodity smartwatch. Therefore, there is an imminent need for a sensing system on a commodity smartwatch that can recognize eating moments with a high time resolution. Based on this need, we developed the key research question of this paper:

- *Is it possible to significantly improve the time resolution of eating moment detection in the wild on a commodity smartwatch without requiring any additional hardware?*

To address this research question, we developed EatingTrak, an AI-powered eating moment detection method, which is able to identify eating moments as short as *three seconds* in a near-free-living scenario using a wrist-mounted IMU. EatingTrak first estimates the arm posture in 3D using the data of the wrist-mounted IMU. Then the estimated arm posture is fed into a customized deep neural network classifier to learn the latent temporal-spacial patterns of eating activities. To evaluate EatingTrak, we collected over 110 hours of eating activities with a near-free-living semi-wild setup from 9 participants where ground-truth were recorded using chest-mounted cameras and labeled manually. Applying EatingTrak on this dataset, we found that it was able to recognize eating moments with a semi-wild setup at time-resolutions of three seconds and 15 minute, with F-1 scores of 73.7% and 83.8% respectively. To the best of our knowledge, this is the *best time resolution* on recognizing eating moments with a comparable setup using a wrist-mounted IMU. To examine the effectiveness of EatingTrak's arm posture estimation algorithm, we conducted a study with 7 participants showing that EatingTrak works reliably especially in estimating arm posture for eating/drinking activities. To further explore the capabilities of EatingTrak, we conducted an additional snack-detection study and a food/utensil type recognition study in a controlled environment. Based on the encouraging results, we discuss the opportunities and challenges of deploying EatingTrak on commodity smartwatches at scale to improve the eating journaling experience for users.

The main contributions of this paper are:

- The first system that recognizes eating moments with a semi-wild setup by learning and estimating 3D arm postures.
- An evaluation on a wrist-mounted IMU dataset on 113.8 hours of eating activities with a semi-wild setup collected from 9 participants with manually labeled ground truth.
- Significantly improved the time-resolution of eating moment detection with comparable setup using a commodity smartwatch (IMU) from five minutes [48] or a meal [17, 46] to three seconds (F-1 73.7%).

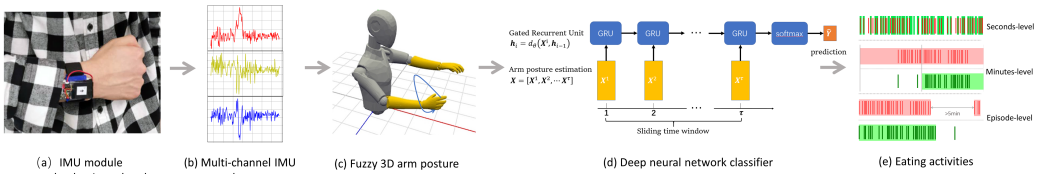


Fig. 1. Overview of EatingTrak. EatingTrak uses a single IMU sensor mounted on the dominant hand to estimate the 3D arm posture, and then uses a deep neural network classifier to detect eating activities at various time resolutions. The evaluation metrics with different time-resolution in (e) are specified in section 5.2.2. (a) IMU module and wristband. (b) Collected multi-channel IMU data. (c) Estimated 3D arm posture. (d) Deep neural network classifier. (e) Detected eating activities.

- An exploratory study on detecting snack-taking (chips) in a controlled environment.
- A discussion on how to apply EatingTrak on commodity smartwatches to improve food journaling experience.

2 RELATED WORK

Since the eating-activity recognition in EatingTrak relies on arm posture estimation, we discuss the related works in two sections: 1) previous works in estimating body posture or using body posture for activity recognition and 2) previous works that recognize eating activities using different sensing methods and form factors.

Ref	Applied sampling rate	Time resolution	F-1 score
Dong et al. [17]	15Hz	a meal	81.5%
Stankoski et al. [46]	25Hz	a meal *	82%
Sharma et al. [42]	15Hz	an EA	23.5%**
Thomaz et al. [48]	25Hz	5min TS	~25%***
		15min TS	~40%***
		60min TS	76.1%
Morshed et al. [33]	25Hz	a meal	87.3%
EatingTrak	8Hz	3 seconds	73.7%(69.9% in UI setting)
		15min TS	86.2%(83.8% in UI setting)

Table 1. Summary of related work using single wrist-mounted IMU sensor in free-living scenarios. EP = episode, TS = time segment, details specified in Section 5.2.2. Please note that the performance figures are obtained on different datasets with different algorithms. This is a summary of comtribution instead of comparison of performance.
*: Although the evaluation was conducted based on seconds-level window, the ground truth recording were based on participants' meal-level self reports.
**: EA: eating activity, a periods of continuous eating, average length of an EA is 14min. F-1 score derived from the confusion matrix presented in the paper.
***: Estimated from the figure presented in the paper.

2.1 Activity recognition using body postures

2.1.1 Body postures estimation. Most of the existing techniques that estimate human body skeleton/postures are from images captured by cameras. Traditional camera-based motion capturing solutions either require users to wear a lot of markers, or require users to be captured by cameras

from multiple views [45, 55, 57]. More recent projects successfully used deep convolutional networks to estimate 2D [12, 13, 21, 35, 50, 51, 53] and 3D [11, 28, 29, 47, 65] body postures from a single camera. However, these camera-based systems cannot be used to capture the user's body posture in the wild where cameras are not available.

To capture body posture in the wild, several wearable-based systems have been developed. However, most of them require the user to wear multiple motion sensors on different parts of the body to capture body movement. For instance, Von Marcard et al. [52] and Huang et al. [22] use 6 body-worn IMUs to estimate full body posture through optimization and learning-based methods respectively. Wearing multiple sensors can be less practical and inconvenient for a user in daily activities. Researchers have also explored body posture tracking focusing on limb motion tracking [20], some of which only needs a smartwatch-like wrist-mounted IMU device [26, 43, 54]. We draw a lot of inspiration from ArmTrak [43] and ArmTroi [26], which demonstrated that they can estimate the arm posture using a commodity smartwatch. However, their experiments focused on a controlled lab setting, where the body direction is opportunistically estimated. It is unclear how these systems would perform in the wild, where the body direction can constantly varying. To address this issue, EatingTrak proposes a method which uses maximum likelihood estimation to estimate the body direction in the wild.

2.1.2 Activity recognition from 3D body postures. Because body posture contains rich information about the users' activities, they have been widely used for human activity recognition. For instance, Piyathilaka et al. [37] used a skeleton extracted from RGBD sensors to recognize 12 human activities. Cippitelli et al. [15] used the body skeleton extracted from RGBD sensors and evaluated activity recognition performances on activities related to active and assistive living-related actions. Nunez et al. [36] explores using deep-learning methods to perform human activity and gesture recognition based on full body skeletons.

To the best of our knowledge, we have not found any work that explores using the 3D arm postures to recognize eating activities in the wild. The research question of whether using body posture as the additional context info would improve the performance on recognizing eating activities has not been explored.

2.2 Eating recognition with wearables

2.2.1 Eating recognition using customized form factor. In order to detect eating activities, researchers have explored using sensors attached to different parts of the body to capture the motion or sound associated with eating activities. For instance, many projects place sensors around the head to detect eating episodes. Researchers have built earpieces with acoustic [3, 9], proximity [7, 8], and motion [7] sensors to recognize eating periods. Researchers have also tried to put microphones on the neck [10, 56] to capture eating gestures (e.g., chewing, drinking, swallowing). Using glasses equipped with multiple sensors is another choice. Glasses equipped with EMG [58–61], load cell [14], piezoelectric sensor [18]. Multi-modal sensors on glasses [5, 38] and necklaces[64] are also used. Some even demonstrates promising performance in detecting free-living eating activities[6]. Fusing sensors placed at various locations such as on upper and lower arms [2, 23], wrists, ear and head [30–32], wrist and neck [62] have also demonstrated promising results. Specifically, cameras in various form factors are widely used to analyse eating event from images [5, 30–32, 38, 49]. The recent papers using head-mounted device have already achieved promising performance on detecting eating moments in the wild. However, the above systems all require designing a new form factor or adding hardware components, which is known to be time consuming. In other words, they may not be immediately available to users in the near future.

2.2.2 Eating recognition using a wrist-mounted IMU. Smartwatches or wristbands are one of the most common commercial wearable products. All these commodity devices have built-in IMUs to detect the motion on the wrist. Researchers have explored using the IMU on a wrist-mounted device to detect eating activities. Some systems have achieved encouraging performance on recognizing eating activities in controlled or in-lab settings [24, 63], or during a meal in less structured environments [39, 40, 44]. Recent work conducted by Luktuke and Hoover et al [27] analyzed the eating intake gestures within a meal in a restaurant using a wrist-mounted IMU. However, it did not distinguish eating moments with other daily activities in the wild. Recognizing eating activities with a single wrist-mounted device in the wild has known to be challenging due to the diversity of eating and other activities [30, 42, 62]. As a result, the time resolution of detecting eating moments is relatively low. Thomaz et al. [48] can recognize eating activities with a time-resolution of 5 to 30 minutes with a F1 score around 30% to 45% respectively. Most recent work can detect eating activities in the unit of a meal (self-reported) [17, 41, 42, 46], some with impressive performance under large-scale dataset [33]. However, research has shown that only about 40% of time during a meal involves actual eating gestures [25]. Only detecting the start and end of a meal misses critical information on health, such as eating frequency, which is critical indicator for overweight or obesity [34]. Furthermore, in order to recognize eating activities with shorter period such as snacks, fruits, a much higher time resolution is needed. Based on the best of our knowledge, none of the prior work [17, 41, 42, 46, 48] has tried to detect the eating period as short as a few seconds.

To facilitate the comparison, we summarize the setting and performances of the related work using wrist-mounted IMU for eating detection in Table 1. It is clear that EatingTrak has significantly improved the time resolution of eating detection to 3 seconds with F1 score of 73.7% and 69.9% using user-dependent and user-independent models respectively. If we adopted the same metrics as as used in [48] with a time segment window size of 15 minutes, the F-1 score on recognizing eating activities using our system climbed to 83.8%, significantly outperforming $\sim$ 40% in [48] (15min time segment in the table. Admittedly, a direct comparison is not possible due to the adoption of different dataset, such gap is still encouraging in improving time resolution. The details of how we calculate this result is explained in Section 5.2.2).

3 METHOD

In this section, we first describe how EatingTrak operates in principle. We then present EatingTrak's processing pipeline.

3.1 Theory of Operation

EatingTrak was developed based on the observation that one key challenge of identifying eating moments from miscellaneous body movements in the wild using a wrist-mounted IMU (acceleration and angular velocity), compared to other systems using customized hardware [5, 9, 30, 60, 64], is the lack of enough contextual information of body posture. As a result, the eating gesture (raising the wrist towards the mouth) can be highly similar to other wrist movements in daily activities, if only examining the movement on the wrist [17, 46, 48]. For instance, picking up a cup from the desk to the mouth has similar movements with part of the waving gesture, except that the arm is higher than the shoulder in the latter case. Since lacking accurate contextual information of the wrist motion, in order to increase the accuracy of eating detection, the system requires more data or a larger window to make a confident prediction. Many of these confusions can be clarified if the system has more contextual information on the eating gestures, such as the arm posture in 3D including the movements and positions of wrists and the elbow, as the arm posture intuitively distinguishes eating behaviors from others. For instance, eating with utensils typically involves moving the arm towards the mouth, rotating the wrist, and moving the arm back down.

With contextual information of the relative position and orientation of the arm against the body, such gesture sequences have distinct features. Based on this observation, we propose the working hypothesis of EatingTrak:

- *The arm posture, if can be calculated, can significantly improve the performance of detecting eating moments in the wild.*

To verify this hypothesis, we implement EatingTrak, which contains two steps. First, it estimates the arm postures in 3D body space according to data collected by the IMU. Second, it trains a supervised customized deep neural network classifier on the arm posture estimation to capture the latent temporal-spacial patterns of eating activities. The structure of the system is presented in Figure 1.

3.2 Estimating 3D arm posture from a wrist-mounted IMU in the wild

An arm posture is defined as a combination of the elbow orientation oeb relative to the body and the wrist orientation owb relative to the body [4] (e as in elbow, w as in wrist, b as in body, E as in Earth). In order to estimate the 3D arm posture of the wearing arm using data from the IMU, we need 3 steps: 1) estimate the body direction d, 2) use the estimated body direction d and the wrist orientation relative to the Earth owE to calculate owb, and 3) use owb to estimate oeb. We propose the *weighted arm posture estimation* to implement step 3), and *body direction estimation* is proposed to implement step 1). Step 2) is mathematically determined. Our arm posture estimation technology is developed based on previous work of ArmTrak [43], which was only evaluated in controlled environments with the assumption that body direction does not change much. Unfortunately, it does not work well in the free-living condition, where the body direction constantly changes. Based on their work, we designed weighted arm posture estimation and body direction estimation to enable it to be applied in free-living scenarios.

3.2.1 Weighted arm posture estimation. Because the human body skeleton is fixed, given a certain wrist orientation owb_k, the arm posture estimation process can be viewed as finding the most likely elbow orientation $\hat{oeb}_k$, which can be calculated as

$$\hat{oeb}_k = \mathbb{E}_{p(oeb|owb_k)}(oeb) = \int oebp(oeb|owb_k)\mathrm{d}oeb \tag{1}$$

To achieve this, we first iterate all possible human arm postures based on the empirical movement ranges of human joints as described by Andriluka [4] and use the algorithm proposed by Akhter et al. [1] to remove invalid postures. Through this process, we can obtain a set which contains all the possible elbow orientations conditioned on owb_k, $O_k^{e,body} = \{o_1^{e,body}, o_2^{e,body}, \cdots, o_{n_k}^{e,body}\}$. Hence, the arm posture estimation process can be described as:

$$\hat{oeb}_k = \sum_{i=1}^{n_k} oeb_i p(oeb_i|owb_k) \tag{2}$$

From the equation above we can see that, to calculate the estimation, the key step is determining the probability distribution $p(oeb|owb_k)$. We use the empirical distribution to represent this probability distribution in eating activities. To achieve this, we conducted a pilot study to record every owb_k and every corresponding oeb_i in eating activities. One of the researchers wore one IMU sensor on the lower arm and one IMU sensor on the upper arm to record owb and oeb respectively while eating for about 6 hours in total. The IMU on the lower arm was worn in the same way as in the user study described in Section 4.1. The IMU on the upper arm was worn in a way such that all its axes were aligned with the IMU on the lower arm when standing upright with the arm resting

tightly at the side of the body. We then calculate the estimated conditional probability:

$$p(oeb_i|owb_k) = \frac{f(oeb_i)}{\sum_{j=1}^{n_k} f(oeb_j)} \tag{3}$$

where f denotes frequency. For the oeb_i that are not recorded, we estimate its conditional probability as the following:

$$p(oeb_i|owb_k) = p(\epsilon = oeb_i - oeb_j)\ ,\ \epsilon \sim \mathcal{N}(p(oeb_j|owb_k), \theta) \tag{4}$$

where oeb_j is oeb_i's neighboring elbow orientation recorded and θ is a manually set hyperparameter.

3.2.2 Body direction estimation from a wrist-mounted IMU data. As shown above, to estimate the arm posture, we have to know the orientation of wrist relative to the body owb, which can be calculated given a certain wrist orientation relative to the Earth owE and a body direction d. Since the IMU sensor can only provide owE, a body direction estimation is needed. The previous work [43] used opportunistic sensing to estimate the body direction. They obtain the body direction from the wrist orientation, when the arm falls down. However, in the wild, if the body direction changes frequently, and the arm does not always fall down, this opportunistic sensing approach may not work well. In order to address this issue, we use Maximum Likelihood Estimation to estimate the body direction.

Given a certain wrist orientation relative to the Earth, the corresponding wrist orientation relative to the body is a function of the body direction: $owb_k = g(d)$. Therefore,

$$p(owb = owb_k) = p(owb_k; d) \tag{5}$$

Assuming all the recorded wrist orientations relative to the body are $owb_1, owb_2, \cdots, owb_K$, it is obvious that,

$$\log \mathcal{L}(d) = \sum_{k=1}^{K} \log p(owb_k; d) = \sum_{k=1}^{K} \log \frac{f(owb_k)}{\sum_{j=1}^{K} f(owb_j)} \tag{6}$$

where f denotes frequency. According to Maximum Likelihood Estimation, it is easy to know that we can get the estimation of d when $\log \mathcal{L}(d)$ reaches its maximum. Since the estimation is performed per frame, a Kalman filter is applied subsequently to remove the noise in estimation.

We demonstrate the tracking performance of our arm tracking algorithm in Section 5.1. We then further demonstrate in our ablation study that our 3D arm posture estimation can significantly improve eating activity detection performance, especially in the free-living scenarios where the users are engaged in complicated activities apart from eating. Details and further discussions of the ablation study is specified in Section 5.6.

3.3 Deep neural network

Because an eating intake gesture contains a time-series of arm postures in 3D [19], we decided to use a deep neural network classifier to learn the temporal-spatial patterns of eating activities from the arm posture estimation.

3.3.1 Data preprocessing. Given the series of data on arm posture estimation, we use a sliding time window to segment the arm posture estimation data into multivariate time series with the same shape of $\tau \times M$. Each data segment $\mathbf{X}$ consists of multidimensional feature values in each frame of the sliding time window: $\mathbf{X} = [\mathbf{X}^1, \mathbf{X}^2, \cdots, \mathbf{X}^\tau]$, where τ is the length of the sliding time window. $\mathbf{X}^i = [x_1, x_2, \cdots, x_M]$, where M is the number of features and x_i is the value of the feature in the i^{th} dimension. The features we used are specified in Table 2. Then, we add labels to the data segments according to the ground truth and get the dataset $D = \{(\mathbf{X}_1, \mathrm{Y}_1), (\mathbf{X}_2, \mathrm{Y}_2), \cdots, (\mathbf{X}_N, \mathrm{Y}_N)\}$. $\mathrm{Y}_i = 1$ if there is a ground truth intake in the middle of the time window of $\mathbf{X}_i$, where k is the class

of the intake; $\mathbf{Y}_i = 0$ if there is no intake in the time window of $\mathbf{X}_i$. In this form we consider the eating gesture detection as a problem of time series classification.

Features from the reconstruction
Wrist's acceleration relative to the body (3D vector)
Wrist position relative to the body (3D coordinates)
Elbow position relative to the body (3D coordinates)
Palm's facing direction relative to the body (3D vector)
Palm's extreme facing direction relative to the body (3D vector) *
Body's facing direction relative to initial facing direction (angle)
Estimated 5 DoFs **

Table 2. Features from the Reconstruction.
* For the right hand, this is the facing direction of the palm if rotate the lower arm clockwise to the extreme extent without changing the position of the elbow and wrist.
** The 5 degrees of freedom of the arm that represent the arm posture are described by Andriluka [4].

3.3.2 Classifier implementation. Given the dataset D, we need to train a classifier to project the input $\mathbf{X}$ to the probability distribution over the labels. We use a Recurrent Neural Network d_θ as the information extractor to perform the following operation:

$$\mathbf{h}_i = d_\theta(\mathbf{X}^i, \mathbf{h}_{i-1}) \quad for\ i = 1, \cdots, \tau \tag{7}$$

$\mathbf{h}_\tau$ contains important information of the input $\mathbf{X}$. In our experiments, d_θ is parametrized by Gated Recurrent Units and initialized with zero vectors for $i = 0$. A fully connected layer with softmax as activation function is used to output the probability distribution over the classes:

$$\hat{\mathbf{Y}}_i(\mathbf{X}) = \frac{\exp(\mathbf{h}_\tau \mathbf{w}_i + b_i)}{\sum_{k=1}^{K} \exp(\mathbf{h}_\tau \mathbf{w}_k + b_k)} \tag{8}$$

where $\hat{\mathbf{Y}}_i$ denotes the probability of $\mathbf{X}$ belonging to class i out of K classes. w_i and b_i denote the weights and the bias term of the fully connected layer.

4 USER STUDY

In this section, we present the setup and procedures of the user study that we have conducted. We collected eating activity data in a near-free-living semi-wild setup without any constraints on the participants.

4.1 Procedures

We used a wristband with an IMU sensor (Adafruit BNO055[1]) mounted on an ESP32 Feather board[2]. The data was saved to a on-board micro SD card. Although the data collection device was not a commodity smartwatch, IMU sensors are easily available on commodity smartwatches. Therefore, this wristband is equivalent to a commodity smartwatch in the context of collecting IMU sensor data. Participants were asked to wear the wristband on their dominant hand. The wristband was worn such that the x-axis of IMU was pointing from the elbow towards the wrist and the z-axis was pointing outwards. A chest-mounted GoPro (pointing upwards towards the head) was used to collect ground truths.

[1] https://learn.adafruit.com/adafruit-bno055-absolute-orientation-sensor
[2] https://www.adafruit.com/product/3405

We recruited 9 participants (P1-P9, 6 male, 3 female) for the study. We collected data for 6 sessions in different days for each participant. In each session, after setting up the device, the participant was free to leave the lab and was asked to return the devices after at least 2 hours. The only instruction we gave was that they needed to have at least one meal during this time period and no other restrictions were imposed. In other words, the participants were free to eat any food at any place at any time or do any other activities.

4.2 Ground truth annotation

We define one intake gesture as the process in which the participant puts the food using their hands or utensil towards the mouth, and then relaxes the arm. The empirical average duration of one intake in our user study was about 3 seconds. The researchers manually labeled the video recorded by GoPro to identify the time point that the food/utensil is deepest into the mouth as the intake time. In this way, every intake gesture is represented by a single timestamp. The timestamps of the GoPro and IMU modules were synchronized.

4.3 Dataset statistics

We examined the composition of the dataset. In total, we collected 6831.3 minutes (113.8 hours) of data, with an average sampling rate around 30 Hz. The total recording duration of all participants ranges from 719 to 795 minutes (std=23.4min). In total, 3178 intakes were captured. On average, every participant had 58.9 intakes every session (min=7, max=277, std=52.0), showing great variance in how much intakes they took in different sessions.

We then examine how much eating/drinking took up during the entire recording. To do this, we consider every eating period to start 30 seconds before the first intake and end 30 seconds after the last. The time in between is considered to be in a eating/drinking event. Using this method, 13.1% of all data recorded were eating/drinking events. Across different participants, the time that he/she was eating/drinking ranges from 58.6 to 152.9 minutes (std=37.6min). In order to understand the distribution of eating/drinking events better under different time granularities, we apply similar calculations but with the metrics specified in Section 5.2.2. While limiting the duration of an eating/drinking event to 3 seconds, only 2.3% of all data were eating/drinking. Using the 5-minute eating episode formation method as described in Section 5.2.2, 13.6% of all recorded data were during an eating episode. Similarly, 16.8% of all recorded data were in an time segment with eating/drinking activities using a 15-minute time segment window as specified in Section 5.2.2.

We then tried to manually label the types of eating/drinking events of each intake based on the ground truth video captured by a chest-mouted GoPro. In total, we identified 8 different types of eating/drinking events: drinking with a bottle/cup/straw, eating with spoon/fork/hand(s)/finger/chopsticks, among them, eating with chopsticks/spoon/fork were most commonly seen. The distribution of these types is illustrated in Table 3. We also tried to identify the food types that the participants consumed. However, due to the vast diversity in food types, special camera angle (from chest point towards the face) and lighting conditions, we were not able to identify all food types. Among them, noodles, liquid (soups, milk, etc.), burger/sandwiches, salad, and rice were most commonly consumed. Similarly, due to the limitation of the view angle, we were unable to clearly determine all situations of eating/drinking events. Based on limited observations, dining table and desk are the most common places to eat, but other places such as couch and even eating while walking were also observed. These results show that although each session of our dataset was capped at about 2 hours, the data we collected are still very diverse.

Event type	Drinking			Eating				
	bottle	cup	straw	spoon	fork	hand	fingers	chopsticks
Portion	3.8%	1.9%	1.6%	29.4%	17.4%	4.0%	0.5%	41.5%

Table 3. Composition of eating/drinking events in the dataset

5 EVALUATION

In this section, we first evaluated the arm posture estimation algorithm. We then evaluated our eating detection system with different setups including time resolution, user-dependent (UD) versus user-independent (UI) models using on the data we collected from the user study. We report the performance in three time resolutions: seconds-level, defined as eating moment detection with time resolution of under 1 minute; minutes-level, defined as eating moment detection with time resolution of 5-30 minutes of time segment; and episode-level, defined as eating moment detection with time resolution of 5-minute eating episode. Details of these metrics are specified in Section 5.2.2. In order to investigate whether the estimated arm posture helped in the recognition task compared to using IMU data only, we conducted an ablation study with only raw IMU data to demonstrate the effectiveness of our proposed arm posture estimation algorithm.

5.1 Evaluation of the arm posture estimation algorithm

In order to examine the effectiveness of EatingTrak's arm posture estimation algorithm, we conducted an extra scripted study with 7 participants (average age 22.6, std=1.6, 5 female, all right-handed) evaluating the performance of 3D arm posture reconstruction in the context of eating detection.

During the study, we asked each participant to wear a wrist-mounted IMU (same as used in the semi-wild study) and conduct activities including eating with spoon/fork/hand, drinking, using the cellphone, chatting, using laptop and walking, each activity lasting about 3 minutes. We used mediapipe with a laptop RGB camera as the ground truth for arm posture. To make sure that the body was properly captured by the camera, while walking, the participants were instructed to walk within a trapezoid area (distance from the camera was 1.5-3.5m, area 6.5m^2) marked on the ground. In all other activities the participants sat about 1.5m in front of the camera. Frames where the arm was not fully captured were discarded. The lengths of the participants' upper and lower arms were measured and used to normalize the ground truth and reconstructed arm postures. The study lasted 25.2 minutes for each participant. We demonstrated that our arm posture tracking algorithm achieved a Median Absolute Error (MAE) of 5.9 and 11.0 centimeters in 4 eating and 4 non-eating activities, respectively. Errors were calculated as the error in distance represented by 3D coordinates after mapping the pixels according to measured arm lengths. Error of the elbow tracking was slightly lower than that of wrist position tracking. Specifically, results in Figure 2(f) indicates that EatingTrak estimates the arm posture quite reliably while the user is eating/drinking. In other activities, the error increases. Meanwhile, we observed that the estimation results were quite stable while eating/drinking but jumping around dramatically in activities especially like walking. This is because our weighted arm posture estimation was optimized for eating/drinking activities.

5.2 Evaluation setup

5.2.1 Experiment configuration. We conducted UD and UI experiments on the collected semi-wild dataset. We conducted leave-one-session-out (UD) and leave-one-participant-out (UI) cross-validations.

To test the performance of the system under lower sampling rate to save power, we down-sampled the input data to 8 Hz. When training, the length of the sliding time window used in data segmentation was 25 frames (about 3 seconds, the average length of an intake gesture observed in the study) and the stride was 5 frames. Specifically, we skipped any windows that overlap within a ±1.5s range of an intake moment to avoid confusing the classifier. We duplicated the intake data samples until intake/non-intake data samples were comparable. When testing, we used a sliding time window of the same size with a stride of 1 frame to generate testing data. If the prediction of the classifier was positive for a testing time window, we considered there was an intake detected at the middle of the time window. In this way, we generated the *frame-level prediction*. We set the dimension of the Gated Recurrent Unit hidden state to 512. In the ablation study, because the number of features was smaller, we enlarged the hidden state dimension until the number of total parameters is the same to make the comparison fair. We used the Adam optimizer with learning rate 0.001.

5.2.2 Evaluation metrics. Classification accuracy may not well represent the performance of the system due to the highly imbalanced dataset (the users were not eating for most of the time). Instead, we used precision, recall and F-1 score as the evaluation metrics. It is worth noting that to calculate the overall precision, recall, and F-1 score, we added up the TP, FP and FN of all participants, and then use the overall TP, FP and FN to calculate the metrics instead of simply averaging the metrics of different participants. This is because different participants have different number of intakes. Simply averaging the metrics of all participants may not represent the characteristics of the whole dataset.

5.3 Seconds-level evaluation

Seconds-level evaluation refers to eating moment detection with time resolution of under 1 minute. With this metric, we evaluate how well our system can detect eating moments with a short duration. To do this, we use a n_{length}-frame sliding window with a stride of 1 frame to transfer the frame-level prediction into intake-level prediction. If the number of detected intakes within the window exceeds a certain threshold $n_{threshold}$, then we consider the whole window as a positive prediction. Since the stride is 1 frame, there can be overlaps between neighboring positive windows. We merge all positive windows that overlapped with each other, and generate a series of positive prediction periods. We count TP, TN, FP as illustraed in Figure 3. We then calculate precision = $\frac{TP}{TP+FP}$, recall = $\frac{TP}{TP+FN}$, and F1 = $\frac{2\times\text{precision}\times\text{recall}}{\text{precision+recall}}$.

We set $n_{length} = 25$ and $n_{threshold} = 4$. In the UD setting, EatingTrak achieved an average detection precision, recall, and F-1 score of 76.2%, 71.3% and 73.7%, respectively. In the UI setting, the detection precision, recall, and F-1 score were 67.6%, 70.7%, and 69.1%, respectively, as demonstrated in Figure 2(a). Detection performance across different participants are shown in Figure 2(b), ranging from 49.8% (UD) and 36.4% (UI) for P9 to 88.9% (UD) and 88.9% (UI) for P5, with SD=12.6% (UD) and 17.7% (UI). Based on our observation, in most sessions with low F1 score, the participant usually raised both hands near the mouth to eat (e.g., burger, sandwich). In this way, the hand movement is much less obvious and has less patterns. We detail this discussion and potential solutions in Section 6.2.

5.3.1 Impact of window length. In the evaluation above, the window length of three seconds was empirically decided based on our observation of a typical intake gesture. This input window length decides how much the model can "see" in each data point. To find out the impact of the length of window on the performance of the system, we conducted another experiment with different lengths of windows ranging from 1 second to 9 seconds. The results presented in Figure 2(c) show that performance increases as window length increases. This is potentially because intakes are

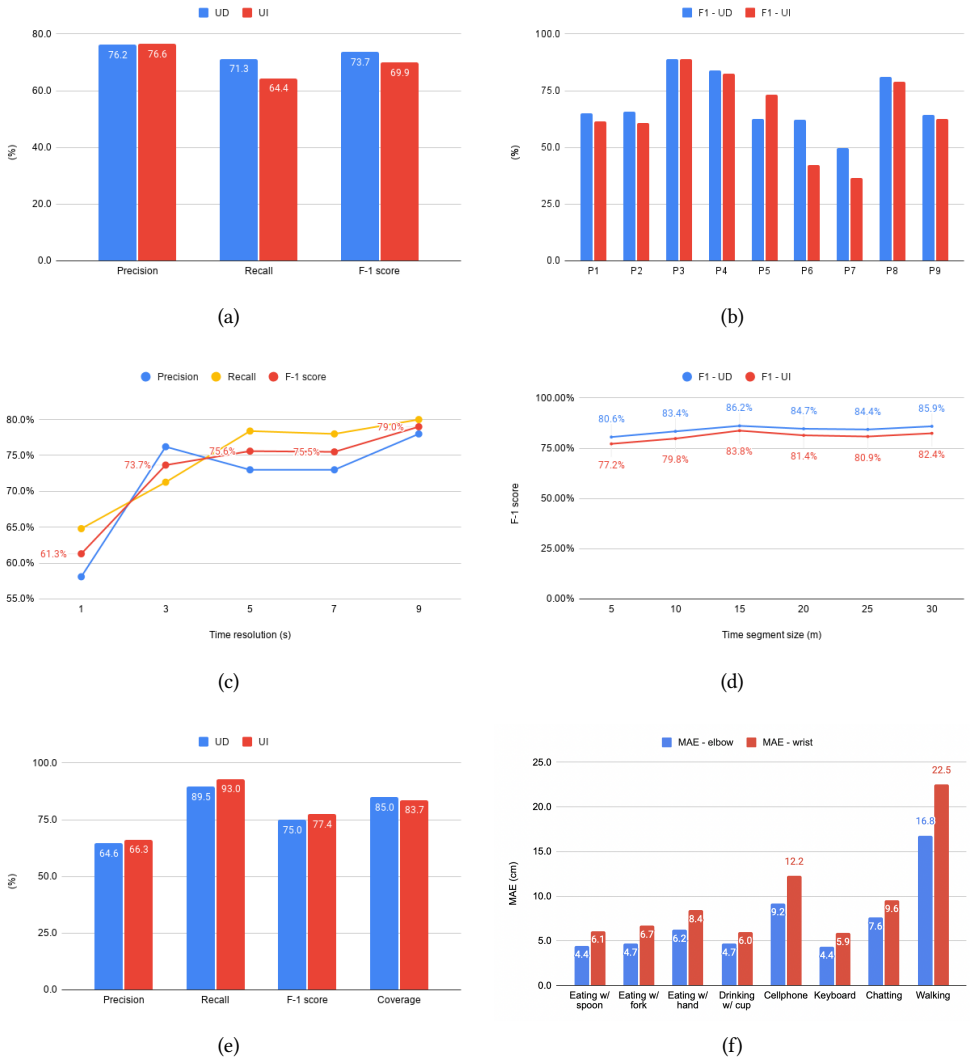


Fig. 2. Performance of eating/drinking detection with the semi-wild setup. UD: user-dependent, UI: user-independent. (a) Seconds-level. (b) Seconds-level performance across all participants. (c) Seconds-level performance with different window size. (d) Minutes-level performance with different segment window. (e) Episode-level. (f) Evaluation of arm posture estimation.

usually repeated. By "seeing" more of these repeated gestures, the model learns the pattern clearer. When the length of the window increases to 9 seconds, F1 score reaches 79.0%. This result also implies that the performance of EatingTrak can be higher with increased window length. However, longer window length also means the system requires more time before a decision can be made, which may not work well for certain applications, such as providing in-situ intervention. We plan to further explore this issue in the next step.

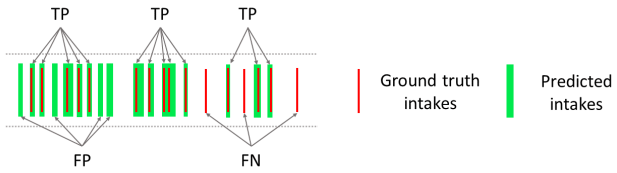


Fig. 3. Metric for seconds-level evaluation. Ground truth intakes (red lines) covered by predicted intakes (green rectangles) are considered true positives (TP). Ground truth intakes not covered by any predicted intakes are recorded as false negatives (FN). Predicted intakes that do not cover any ground truth intakes are recorded as false positives (FP).

5.3.2 Estimation of delay in just-in-time intervention. We are also interested in the delay in the seconds-level evaluation, since this is critical for applications such as just-in-time intervention would be useful. With our ground truth annotation method, every intake was represented by the moment that the food was deepest into the mouth. In our seconds-level evaluation, an intake is only counted as true positive if the predicted window covers the labeled moment. This means that if the system is successful in detecting an intake, it has to be able to detect it before the food is deepest into the mouth. In practice, we used a sliding window approach which looks 1.5s into the future, indicating a systematic delay of 1.5s. With this in mind, we then analyzed the runtime of our processing pipeline. We tested our pipeline on a Workstation with AMD Threadripper 3960X CPU and RTX 2080Ti GPU. Our system takes about 15ms to predict each frame. With all things considered, our system has a theoretical delay of about 1.5s if running in real-time at a predicting time-resolution of 3s. However, we would like to acknowledge that this delay was computed theoretically and might not reflect real life performance.

5.4 Minutes-level evaluation

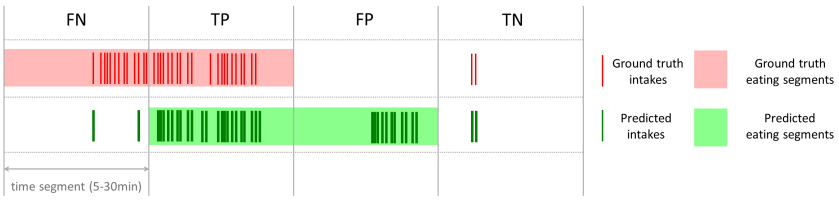


Fig. 4. Metric for minutes-level evaluation. Each time segment is 5-30min long. For a time segment, if the number of ground truth intakes exceeds certain amount, then it is considered a ground truth eating segment. Accordingly, if the number of predicted intakes exceeds a certain amount, then it is considered a predicted eating segment. TP, FP and FN are then calculated on a time segment basis.

Minutes-level evaluation refers to eating moment detection with a time resolution of 5-30 minutes, using a metric similar to the one used in Thomaz et al. [48]. This metric is designed to evaluate how our system can detect continuous eating activities that last for over 5 minutes such as a meal. We first divide the whole session into *time segments* [48] of equal length (5-30min). For each time segment, we examine the number of ground truth intakes and number of positive predicted frames. To focus on continuous eating periods and remove discrete random intakes, if a time segment of t_w minutes contains more than $\frac{t_w+20}{5}$ ground truth intakes (from 5 intakes within 5 minutes to 10 intakes within 30 minutes), then it is considered a *ground truth eating segment*. If more than

$(2t_w + 40)$ frames of all frames within a time segment of t_w minutes are predicted as positive, then this segment is considered as a *predicted eating segment* (predictions are made on a frame-by-frame basis, therefore, the number of positive frames are significantly larger than the number of ground truth intakes because multiple neighboring frames near an intake can all be predicted as positive). We then calculate TP, FP and FN accordingly, as illustrated in Figure 4. We then report the precision, recall and F1 score.

We set the time segment length from 5 minutes to 30 minutes, which is similar to [48]. We did not extend the segment length to 60 minutes because one session in our dataset lasted about 2 hours. Results showed that EatingTrak achieved an F-1 score of 77.2% (UI) and 80.6% (UD) at 5-minute segment length and 82.4% (UI) and 85.9% (UD) at 30 minutes. Performance slightly increased and then flattened with increasing segment lengths, as illustrated in Figure 2(d). Although a rigorous direct comparison is not possible due to the differences in the datasets, our results significantly outperformed the results from Thomaz et al. [48]. Due to lack of the specific sensor data (sensor orientation) required, we could not directly apply our algorithm on their dataset. We admit this comparison may not be completely fair, as many variables are different in the two studies, such as participants, the size of the dataset, and the machine learning pipeline. Our outstanding results can be introduced by a compounding effect of all the factors above. Thus, we intent to present all details of the experiment so that the researchers can estimate the improvements introduced by our proposed system.

5.5 Episode-level evaluation

In food journaling applications, it is also important to record eating/drinking episodes outside of a meal. To evaluate how well our system can detect eating episodes from both within and outside of a meal, we adopt an episode-level metric similar to the one used in FitByte [5]. We assume that two consecutive intake episodes should be at least 5 minutes apart. Therefore, we merge neighboring intakes that are within 5 minutes and form intake episodes using the method illustrated in Figure 5. We then drop the episodes that are less than 5 minutes long. We calculate precision, recall and F-1 score on an episode level. Similar to FitByte, we also calculate *coverage* which is defined as the percentage of intake episodes that are covered by predicted intake episodes.

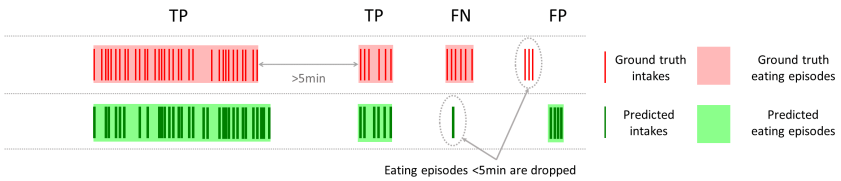


Fig. 5. Metric for episode-level evaluation. Ground truth intakes that are less than 5min apart are merged to form an eating episode, so are the predicted intakes. TP, FP and FN are then calculated on an episode basis.

We found EatingTrak achieved an overall precision, recall and F-1 score of 64.6%, 89.5%, and 75.0%, respectively, in UD settings. In UI settings, precision, recall and F-1 score were 58.9%, 93.0%, and 72.1%, respectively. The *coverage* in UD and UI settings were 85.0% and 83.7%, respectively. Compared with the UI results (precision 84.6%, recall 78.6%, F-1 score 81.5% and coverage 89%) in FitByte [5], our results were lower by around 5%-10%. However, EatingTrak has achieved arguably comparable performance on detecting eating episodes with the semi-wild setup using a single wrist-mounted IMU (equivalent to a smartwatch), compared to FitByte, which uses a customized glassframe with multiple sensors.

5.6 Ablation study

It is important to understand how much the estimated arm postures can improve the performance compared to previous systems [17, 48, 62] using the data from IMU only. To investigate this matter, we conducted an ablation study comparing the eating detection performance using estimated arm postures compared to only using the raw IMU data as shown in previous work[17, 48, 62]. We used the seconds-level evaluation as specified in Section 3 with a time resolution of 3 seconds. We use a UD approach for training and testing. The input data of the deep neural network pipeline with and without estimation are specified in Table 2 and Table 4, respectively. It is worth noting that the feature dimension in the ablation study is smaller. This is because our arm posture estimation introduces extra useful features. Apart from the input data, all other configurations were identical.

Features from the reconstruction
Wrist's acceleration relative to the Earth (3D vector)
Wrist's gravity (3D coordinates)
Wrist's orientation relative to the Earth (quaternion)
Wrist's orientation relative to the Earth (Euler angles)

Table 4. Features from the reconstruction

Results show that the average F-1 score was 60.2% using only raw IMU data (without arm posture estimation), much lower than 73.7% when arm posture estimation was applied. We conducted a one-way ANOVA test, showing that performance with arm posture estimation across all 54 sessions was significantly better than without ($F(1, 106) = 6.69$, $p = 0.011$). This result confirms positively about our *working hypothesis* that using the estimated arm posture can significantly improve the performance on detecting eating moments with the semi-wild setup.

5.7 Snack study

To examine EatingTrak's capability in detecting short eating period out of a meal, we conducted an additional preliminary study with 12 participants (P10-P21, 9 males and 3 females) in a controlled environment focusing on snack detection. The goal of this study was to demonstrate and evaluate a specific use case of EatingTrak - detecting snacks during various activities. Therefore, we limited the types of food and designed the procedure of activities. An Apple Watch Series 4 [3] was used as the data collect device via an sensor logging app (Sensor Log[4]). The same chest-mounted GoPro was used to capture ground-truth.

During the study, each participant was asked to perform a series of activities in the order of their choice, including walking up/down the stairs in the building, using smartphones for 1 to 2 minutes, typing on a laptop, cleaning up the table, and talking to the researcher. We placed a bag of chips on the table and told them that they were welcomed to eat snacks if they wanted to. In other words, the participants decided if, when, how (e.g., hold the bag in their hands or place the bag on the table, or take the snack with them while walking), and how much to eat the chips. Among all the participants, only one did not eat any chips. In total, we collected 218.7 minutes (3.6 hours) data with 283 snack intakes. Each session lasted 18.2 minutes (12.9-30.0min, std=5.1min).. The average sampling rate of the sensor was 49.9Hz.

We merged consecutive frame-level predictions and used the seconds-level metric. Results showed that EatingTrak achieved an average detection precision, recall, and F-1 score of 72.1%,

[3] https://www.apple.com/watch/

[4] https://apps.apple.com/us/app/sensorlog/id388014573

93.7% and 81.5%, respectively. F-1 score across different participants ranges from 55.6% to 97.0% (SD=14.7%). We conducted a similar ablation study, confirming that F1 score with arm posture (81.5%) was significantly better than without (63.0%, ANOVA: F(1, 20) = 6.96, p = 0.016). Specifically, the high recall shows that the system is sensitive to short eating activities as eating chips, indicating the feasibility of detecting short period eating activities using EatingTrak. This study provides a glimpse into a practical use case of our system. However, this result is preliminary given the size and setting of the study, further study is needed to draw a conclusive conclusion.

5.8 Exploring the feasibility of food/utensil recognition in a preliminary experiment

Eating moment detection is a critical first step towards automatic food journaling systems. One step further, food and utensil types also contain very important information. Although some other work has obtained decent performance for this task. They usually require multi-modal sensors worn on several locations [32]. In order to explore the feasibility of recognizing food/utensil types with EatingTrak, we conducted an additional and preliminary experiment in a controlled lab setting to collect eating behavior data using different utensils. 12 Participants (P10-P21, same from the snack study) were asked to eat 7 types of food under 4 utensil types as specified in Table 5. The same Apple Watch and data logging app in the snack study were used. Each participant completed 6 identical sessions arranged on 2 different days. In each session, they were asked to have 10 intakes for each food type. The participant could choose the order in which they ate the different types of food. A front-facing camera (GoPro) was used to capture the ground truth. In this experiment, 709.3 minutes (11.8h) of data were collected with average sampling rate of 49.9Hz.

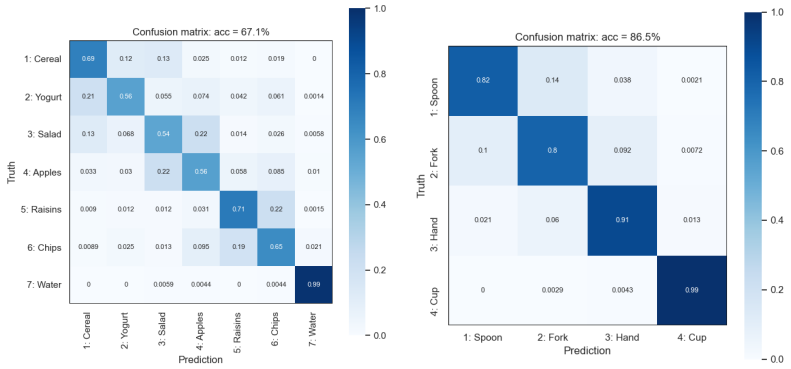
Utensil	Spoon	Fork	Hand	Cup
Food	Cereal	Salad	Raisins	Water
	Yogurt	Apples	Chips	

Table 5. Food and utensil types for the classification experiment

We first used the same seconds-level eating moment detection metric as used in the semi-wild experiment. Results show that F-1 score was 96.8% and 96.0% and consistent across different participants (SD=1.77% and 2.53%) using UD and UI setups, respectively. We then calculate the classification accuracy of 7 food types and 4 utensil types, defined as the percent of intakes correctly classified among those successfully detected. Results show that EatingTrak obtained 67.1% and 86.5% in classifying 7 food and 4 utensil types in the UD setting. In UI setting, the classification accuracy was 56.8% and 78.9%.

The confusion matrices of food and utensil type classification in UD settings are shown in Figure 6. In food type classification, the most common confusions were between salad and apples, raisins and chips, and cereal and yogurt, which were all food with the same utensil. Classification accuracy rose significantly when food types with the same utensil were combined in utensil classification. This confirms our assumption that there exists different patterns in using different utensils. It also shows that distinguishing food with the same utensil is more difficult than with different utensils. Furthermore, the discrepancy of classification performance on distinguishing food and utensil types between UI and UD setting were not surprising, because different users have different ways of using utensils, for example, how they hold the utensil and the speed of each of their intake.

Although the results are encouraging, we believe it is still extremely challenging to detect the food or utensil types in the wild. Because the user's eating behaviors (eating speed, frequency, hand gesture) vary significantly in the free-living condition. What we present in this study is preliminary. A much thorough study and experiment needs to be done to draw further conclusion.



(a) Food classification (b) utensil classification

Fig. 6. Confusion matrix for food and utensil type classification in a controlled scenario. Experiment conducted under the UD setting.

6 DISCUSSION

In this section, we discuss the challenges and opportunities of deploying EatingTrak on commodity wearable devices in the real-world settings.

6.1 Applications

To our knowlwdge, EatingTrak has provided the best performance (three seconds) in terms of the time-resolution on detecting eating moments using a wrist-mounted IMU (e.g., smartwatch, wristbands). The high time resolution on detecting eating moments can potentially improve eating activity journaling experience and enable new applications. For instance, it can potentially allow in-situ eating behavior intervention. EatingTrak, if integrated into a smartwatch-based eating journaling app, can potentially remind the user shortly after the eating activity is detected. Alternately, EatingTrak can be used to ask the users to provide detailed information of an eating episode shortly after the meal is over. Moreover, the system can be adjusted to be more specific (with less false alarms) or more sensitive (with less missed alarms) to cater to user's preferences and needs. EatingTrak can also be used to analyse users' eating styles and help users develop a healthier eating habit. Our high time-resolution detection makes eating frequency measurement possible, which is found to be positively associated with overweight/obesity [34].

6.2 Balancing precision and recall

Precision and recall represent how many FP (false alarm) and FN (missed alarm) errors the system may encounter. Depending on the applications, the two type of errors carry different consequences in real-world applications. For instance, in just-in-time interventions, if the user does not wish to be disturbed too frequently, reducing FP (higher precision) is more important. In food journaling, if EatingTrak is used to activate other sensors like cameras, then a higher sensitivity (less FN, higher recall) is preferred. While generating seconds-level predictions, a threshold $n_{threshold}$ is used to determine when a window is considered a positive or negative prediction as specified in Section 5.2.2. By adjusting $n_{threshold}$, EatingTrak can be adjusted between higher precision and higher recall, as demonstrated in Figure 7(b).

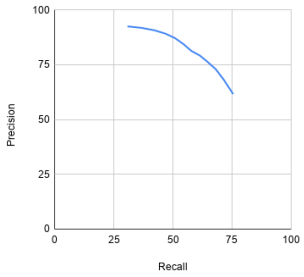


Fig. 7. Relationship between precision and recall. Experiments conducted on the free-living dataset under the UI setting. Adjusting the threshold $n_{threshold}$ can make EatingTrak alter between precise (higher precision) and sensitive (higher recall).

6.3 Challenges of eating/drinking detection with wrist-mounted IMU

EatingTrak has demonstrated promising performance in eating/drinking detection with a single wrist-mounted IMU, both in eating/drinking moment detection in semi-wild environment and in food/utensil/snack detection in a more constrained environment. However, we would like to acknowledge that there still remains significant challenges in eating/drinking detection with wrist-mounted IMUs. One of the key challenge is that not all eating/drinking activities involve significant hand movement. For instance, as pointed out in Section 5.2.2, EatingTrak did not work very well when participants ate food like burgers/sandwiches with both hands. We have observed that when the participants raised both their hands, they usually rested the elbow on the table and held the hands close to the mouth. In such a scenario, participants tend to lean the head forward to reach the food more than moving the food towards the mouth with the hands. In this way, there was little arm movement while eating. Therefore, it is difficult for the model to track eating moments if the eating intake gesture was mostly completed by head movements instead of arm movements. Drinking using a straw is another example of such a case. To address this issue, one natural next step we plan to explore is to fuse the movements on the head (e.g., earphone, glass) together with 3D arm postures to further improve the performance.

6.4 Challenges of deploying EatingTrak on commodity wearables

EatingTrak only requires a wrist-mounted IMU sensor as the hardware which is available on all commodity wristbands or smartwatches. Thus, it has a great potential to make an immediate impact on eating tracking using commodity wearables. However, there are challenges that need to be addressed before large scale deployments. First, our system needs IMU to continuously sample at 8 Hz. It is possible to only wake up the sensor when motion is detected using ultra-low-power accelerometer[5] to save battery. Second, the machine learning algorithm is currently deployed at cloud server. We plan to optimize The ML algorithms so that it can be deployed on phones in the future.

6.5 Further improvements on 3D arm posture estimation

The 3D arm posture estimation can potentially be improved by adopting the following approaches. First, the reconstruction depends on the facing direction estimation module. The current facing direction estimation algorithm does not take into account the correlation of facing direction between frames and the constraints of human movement. Second, the current reconstruction algorithm

[5]https://www.bosch-sensortec.com/news/ultra-low-power-accelerometer-bma400.html

assumes that the torso is mostly upright. However, information about the inclination of the torso could improve the accuracy of the reconstruction. Third, the algorithm to estimate the arm posture could benefit from using more temporal information. Currently our temporal model is only used to detect eating/drinking moments based on estimated arm posture, while the arm posture estimation did not utilize much temporal information (except a Kalman filter). Time series data processing model (e.g., Hidden Markov Model) can be adopted to further optimize the classification results. If collecting large-scale 3D arm posture ground truth data in free-living scenarios is possible, it is also promising to utilize deep-learning approaches such as time-series regression to estimate 3D arm postures.

6.6 Limitations and future work

Just like any other research project, EatingTrak also has limitations. One critical limitation of EatingTrak is that it requires users to wear the device on their dominant hand. This means that EatingTrak will not work for people who are not willing to wear a smart wristband or smartwatch on their dominant hand. However, we are interested in investigating whether users are willing to change the wrist for wearing the smartwatch if the device can track their eating activities. We think the people with strong motivation of journaling eating activities may benefit more from our technology. Second, each session in in-the-wild study only lasted for two hours, where the participant was expected to eat. This is limited by the battery life of the GoPro, which is critical for a fine-grained ground truth annotation. However, it is unclear that how the system would perform if the data collection session is significantly longer. For instance, if the user wear the device to conduct activities involving heavy arm movements (e.g., basketball, tennis), it is unclear how would the system perform. We leave this issue for future exploration.

7 CONCLUSION

In this paper, we present EatingTrak, a AI-powered sensing system, which can detect eating moments using a wrist-mounted IMU. It can detect eating moments as short as three seconds, by deep learning the estimated 3D arm postures from the IMU data. EatingTrak was able to identify eating activities from 113 hours of data collected in a near-free-living semi-wild setup, with a F1 score of 73.7% and 83.8% when the time resolution is 3 seconds and 15 minutes, respectively. It has significantly improve the time-resolution of identifying eating moments in the wild to three seconds. Because it does not require any additional hardware on wrist-mounted commodity wearables, it has the potential to be immediately deployed on millions of commodity smartwatches in the near future.

8 ACKNOWLEDGEMENTS

The authors thank all reviewers for their valuable opinions and suggestions. The authors would also like to thank all participants for participating in the study. This research is partially funded by Cornell China Center.

REFERENCES

[1] Ijaz Akhter and Michael J Black. 2015. Pose-conditioned joint angle limits for 3D human pose reconstruction. In Proceedings of the IEEE conference on computer vision and pattern recognition. 1446–1455.
[2] Oliver Amft, Holger Junker, and Gerhard Troster. 2005. Detection of eating and drinking arm gestures using inertial body-worn sensors. In Ninth IEEE International Symposium on Wearable Computers (ISWC'05). IEEE, 160–163.
[3] Oliver Amft, Mathias Stäger, Paul Lukowicz, and Gerhard Tröster. 2005. Analysis of chewing sounds for dietary monitoring. In International Conference on Ubiquitous Computing. Springer, 56–72.
[4] Mykhaylo Andriluka, Stefan Roth, and Bernt Schiele. 2009. Pictorial structures revisited: People detection and articulated pose estimation. In 2009 IEEE conference on computer vision and pattern recognition. IEEE, 1014–1021.

[5] Abdelkareem Bedri, Diana Li, Rushil Khurana, Kunal Bhuwalka, and Mayank Goel. 2020. Fitbyte: Automatic diet monitoring in unconstrained situations using multimodal sensing on eyeglasses. In Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems. 1–12.

[6] Abdelkareem Bedri, Richard Li, Malcolm Haynes, Raj Prateek Kosaraju, Ishaan Grover, Temiloluwa Prioleau, Min Yan Beh, Mayank Goel, Thad Starner, and Gregory Abowd. 2017. EarBit: using wearable sensors to detect eating episodes in unconstrained environments. Proceedings of the ACM on interactive, mobile, wearable and ubiquitous technologies 1, 3 (2017), 37.

[7] Abdelkareem Bedri, Apoorva Verlekar, Edison Thomaz, Valerie Avva, and Thad Starner. 2015. Detecting mastication: A wearable approach. In Proceedings of the 2015 ACM on International Conference on Multimodal Interaction. 247–250.

[8] Abdelkareem Bedri, Apoorva Verlekar, Edison Thomaz, Valerie Avva, and Thad Starner. 2015. A wearable system for detecting eating activities with proximity sensors in the outer ear. In Proceedings of the 2015 ACM International Symposium on Wearable Computers. ACM, 91–92.

[9] Shengjie Bi, Tao Wang, Nicole Tobias, Josephine Nordrum, Shang Wang, George Halvorsen, Sougata Sen, Ronald Peterson, Kofi Odame, Kelly Caine, et al. 2018. Auracle: Detecting eating episodes with an ear-mounted sensor. Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies 2, 3 (2018), 1–27.

[10] Yin Bi, Mingsong Lv, Chen Song, Wenyao Xu, Nan Guan, and Wang Yi. 2015. Autodietary: A wearable acoustic sensor system for food intake recognition in daily life. IEEE Sensors Journal 16, 3 (2015), 806–816.

[11] Federica Bogo, Angjoo Kanazawa, Christoph Lassner, Peter Gehler, Javier Romero, and Michael J Black. 2016. Keep it SMPL: Automatic estimation of 3D human pose and shape from a single image. In European Conference on Computer Vision. Springer, 561–578.

[12] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. 2017. Realtime multi-person 2d pose estimation using part affinity fields. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 7291–7299.

[13] Xianjie Chen and Alan L Yuille. 2014. Articulated pose estimation by a graphical model with image dependent pairwise relations. In Advances in neural information processing systems. 1736–1744.

[14] Jungman Chung, Jungmin Chung, Wonjun Oh, Yongkyu Yoo, Won Gu Lee, and Hyunwoo Bang. 2017. A glasses-type wearable device for monitoring the patterns of food intake and facial activity. Scientific reports 7, 1 (2017), 1–8.

[15] Enea Cippitelli, Samuele Gasparrini, Ennio Gambi, and Susanna Spinsante. 2016. A human activity recognition system using skeleton data from RGBD sensors. Computational intelligence and neuroscience 2016 (2016).

[16] Felicia Cordeiro, Daniel A Epstein, Edison Thomaz, Elizabeth Bales, Arvind K Jagannathan, Gregory D Abowd, and James Fogarty. 2015. Barriers and negative nudges: Exploring challenges in food journaling. In Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems. 1159–1162.

[17] Yujie Dong, Jenna Scisco, Mike Wilson, Eric Muth, and Adam Hoover. 2013. Detecting periods of eating during free-living by tracking wrist motion. IEEE journal of biomedical and health informatics 18, 4 (2013), 1253–1260.

[18] Muhammad Farooq and Edward Sazonov. 2016. Segmentation and characterization of chewing bouts by monitoring temporalis muscle using smart glasses with piezoelectric sensor. IEEE journal of biomedical and health informatics 21, 6 (2016), 1495–1503.

[19] Hassan Ismail Fawaz, Germain Forestier, Jonathan Weber, Lhassane Idoumghar, and Pierre-Alain Muller. 2019. Deep learning for time series classification: a review. Data Mining and Knowledge Discovery 33, 4 (2019), 917–963.

[20] Alessandro Filippeschi, Norbert Schmitz, Markus Miezal, Gabriele Bleser, Emanuele Ruffaldi, and Didier Stricker. 2017. Survey of motion tracking methods based on inertial sensors: A focus on upper limb human motion. Sensors 17, 6 (2017), 1257.

[21] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. 2017. Mask r-cnn. In Proceedings of the IEEE international conference on computer vision. 2961–2969.

[22] Yinghao Huang, Manuel Kaufmann, Emre Aksan, Michael J Black, Otmar Hilliges, and Gerard Pons-Moll. 2018. Deep inertial poser: learning to reconstruct human pose from sparse inertial measurements in real time. ACM Transactions on Graphics (TOG) 37, 6 (2018), 1–15.

[23] Holger Junker, Oliver Amft, Paul Lukowicz, and Gerhard Tröster. 2008. Gesture spotting with body-worn inertial sensors to detect user activities. Pattern Recognition 41, 6 (2008), 2010–2024.

[24] Konstantinos Kyritsis, Christos Diou, and Anastasios Delopoulos. 2018. End-to-end Learning for Measuring in-meal Eating Behavior from a Smartwatch. In 2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC). IEEE, 5511–5514.

[25] Konstantinos Kyritsis, Christos Diou, and Anastasios Delopoulos. 2019. Modeling wrist micromovements to measure in-meal eating behavior from inertial sensor data. IEEE journal of biomedical and health informatics 23, 6 (2019), 2325–2334.

[26] Yang Liu, Zhenjiang Li, Zhidan Liu, and Kaishun Wu. 2019. Real-time Arm Skeleton Tracking and Gesture Inference Tolerant to Missing Wearable Sensors. In Proceedings of the 17th Annual International Conference on Mobile Systems, Applications, and Services. 287–299.

[27] Yadnyesh Y Luktuke and Adam Hoover. 2020. Segmentation and recognition of eating gestures from wrist motion using deep learning. In 2020 IEEE International Conference on Big Data (Big Data). IEEE, 1368–1373.
[28] Dushyant Mehta, Oleksandr Sotnychenko, Franziska Mueller, Weipeng Xu, Srinath Sridhar, Gerard Pons-Moll, and Christian Theobalt. 2018. Single-shot multi-person 3d pose estimation from monocular rgb. In 2018 International Conference on 3D Vision (3DV). IEEE, 120–130.
[29] Dushyant Mehta, Srinath Sridhar, Oleksandr Sotnychenko, Helge Rhodin, Mohammad Shafiei, Hans-Peter Seidel, Weipeng Xu, Dan Casas, and Christian Theobalt. 2017. Vnect: Real-time 3d human pose estimation with a single rgb camera. ACM Transactions on Graphics (TOG) 36, 4 (2017), 1–14.
[30] Mark Mirtchouk, Drew Lustig, Alexandra Smith, Ivan Ching, Min Zheng, and Samantha Kleinberg. 2017. Recognizing eating from body-Worn sensors: combining free-living and laboratory data. Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies 1, 3 (2017), 85.
[31] Mark Mirtchouk, Dana L McGuire, Andrea L Deierlein, and Samantha Kleinberg. 2019. Automated estimation of food type from body-worn audio and motion sensors in free-living environments. In Machine Learning for Healthcare Conference. 641–662.
[32] Mark Mirtchouk, Christopher Merck, and Samantha Kleinberg. 2016. Automated estimation of food type and amount consumed from body-worn audio and motion sensors. In Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing. 451–462.
[33] Mehrab Bin Morshed, Samruddhi Shreeram Kulkarni, Richard Li, Koustuv Saha, Leah Galante Roper, Lama Nachman, Hong Lu, Lucia Mirabella, Sanjeev Srivastava, Munmun De Choudhury, et al. 2020. A Real-Time Eating Detection System for Capturing Eating Moments and Triggering Ecological Momentary Assessments to Obtain Further Context: System Development and Validation Study. JMIR mHealth and uHealth 8, 12 (2020), e20625.
[34] Kentaro Murakami and M Barbara E Livingstone. 2015. Eating frequency is positively associated with overweight and central obesity in US adults. The Journal of nutrition 145, 12 (2015), 2715–2724.
[35] Alejandro Newell, Kaiyu Yang, and Jia Deng. 2016. Stacked hourglass networks for human pose estimation. In European conference on computer vision. Springer, 483–499.
[36] Juan C Nunez, Raul Cabido, Juan J Pantrigo, Antonio S Montemayor, and Jose F Velez. 2018. Convolutional neural networks and long short-term memory for skeleton-based human activity and hand gesture recognition. Pattern Recognition 76 (2018), 80–94.
[37] Lasitha Piyathilaka and Sarath Kodagoda. 2013. Gaussian mixture based HMM for human daily activity recognition using 3D skeleton features. In 2013 IEEE 8th conference on industrial electronics and applications (ICIEA). IEEE, 567–572.
[38] Shah Atiqur Rahman, Christopher Merck, Yuxiao Huang, and Samantha Kleinberg. 2015. Unintrusive eating recognition using Google Glass. In 2015 9th International Conference on Pervasive Computing Technologies for Healthcare (PervasiveHealth). IEEE, 108–111.
[39] Raul I Ramos-Garcia, Eric R Muth, John N Gowdy, and Adam W Hoover. 2014. Improving the recognition of eating gestures using intergesture sequential dependencies. IEEE journal of biomedical and health informatics 19, 3 (2014), 825–831.
[40] Jenna L Scisco, Eric R Muth, and Adam W Hoover. 2014. Examining the utility of a bite-count–based measure of eating activity in free-living human beings. Journal of the Academy of Nutrition and Dietetics 114, 3 (2014), 464–469.
[41] Surya Sharma, Phillip Jasper, Eric Muth, and Adam Hoover. 2016. Automatic Detection of Periods of Eating Using Wrist Motion Tracking. In 2016 IEEE First International Conference on Connected Health: Applications, Systems and Engineering Technologies (CHASE). IEEE, 362–363.
[42] Surya Sharma, Phillip Jasper, Eric Muth, and Adam Hoover. 2020. The impact of walking and resting on wrist motion for automated detection of meals. ACM Transactions on Computing for Healthcare 1, 4 (2020), 1–19.
[43] Sheng Shen, He Wang, and Romit Roy Choudhury. 2016. I am a smartwatch and i can track my user's arm. In Proceedings of the 14th annual international conference on Mobile systems, applications, and services. ACM, 85–96.
[44] Yiru Shen, Eric Muth, and Adam Hoover. 2018. The Impact of Quantity of Training Data on Recognition of Eating Gestures. arXiv preprint arXiv:1812.04513 (2018).
[45] Takaaki Shiratori, Hyun Soo Park, Leonid Sigal, Yaser Sheikh, and Jessica K Hodgins. 2011. Motion capture from body-mounted cameras. In ACM SIGGRAPH 2011 papers. 1–10.
[46] Simon Stankoski, Marko Jordan, Hristijan Gjoreski, and Mitja Luštrek. 2021. Smartwatch-Based Eating Detection: Data Selection for Machine Learning from Imbalanced Data with Imperfect Labels. Sensors 21, 5 (2021), 1902.
[47] Bugra Tekin, Pablo Márquez-Neila, Mathieu Salzmann, and Pascal Fua. 2016. Fusing 2d uncertainty and 3d cues for monocular body pose estimation. arXiv preprint arXiv:1611.05708 2, 3 (2016).
[48] Edison Thomaz, Irfan Essa, and Gregory D Abowd. 2015. A practical approach for recognizing eating moments with wrist-mounted inertial sensing. In Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing. ACM, 1029–1040.

[49] Edison Thomaz, Aman Parnami, Irfan Essa, and Gregory D Abowd. 2013. Feasibility of identifying eating moments from first-person images leveraging human computation. In Proceedings of the 4th International SenseCam & Pervasive Imaging Conference. 26–33.
[50] Jonathan J Tompson, Arjun Jain, Yann LeCun, and Christoph Bregler. 2014. Joint training of a convolutional network and a graphical model for human pose estimation. In Advances in neural information processing systems. 1799–1807.
[51] Alexander Toshev and Christian Szegedy. 2014. Deeppose: Human pose estimation via deep neural networks. In Proceedings of the IEEE conference on computer vision and pattern recognition. 1653–1660.
[52] Timo von Marcard, Bodo Rosenhahn, Michael J Black, and Gerard Pons-Moll. 2017. Sparse inertial poser: Automatic 3d human pose estimation from sparse imus. In Computer Graphics Forum, Vol. 36. Wiley Online Library, 349–360.
[53] Shih-En Wei, Varun Ramakrishna, Takeo Kanade, and Yaser Sheikh. 2016. Convolutional pose machines. In Proceedings of the IEEE conference on Computer Vision and Pattern Recognition. 4724–4732.
[54] Wenchuan Wei, Keiko Kurita, Jilong Kuang, and Alex Gao. 2021. Real-time 3D arm motion tracking using the 6-axis IMU sensor of a smartwatch. In 2021 IEEE 17th International Conference on Wearable and Implantable Body Sensor Networks (BSN). IEEE, 1–4.
[55] Liguang Xie, Mithilesh Kumar, Yong Cao, Denis Gracanin, and Francis Quek. 2008. Data-driven motion estimation with low-cost sensors. (2008).
[56] Koji Yatani and Khai N Truong. 2012. BodyScope: a wearable acoustic sensor for activity recognition. In Proceedings of the 2012 ACM Conference on Ubiquitous Computing. 341–350.
[57] Yasuyoshi Yokokohji, Yuki Kitaoka, and Tsuneo Yoshikawa. 2005. Motion capture from demonstrator's viewpoint and its application to robot teaching. Journal of Robotic Systems 22, 2 (2005), 87–97.
[58] Rui Zhang and Oliver Amft. 2017. Monitoring chewing and eating in free-living using smart eyeglasses. IEEE journal of biomedical and health informatics 22, 1 (2017), 23–32.
[59] Rui Zhang and Oliver Amft. 2018. Free-living eating event spotting using EMG-monitoring eyeglasses. In 2018 IEEE EMBS International Conference on Biomedical & Health Informatics (BHI). IEEE, 128–132.
[60] Rui Zhang and Oliver Amft. 2020. Retrieval and timing performance of chewing-based eating event detection in wearable sensors. Sensors 20, 2 (2020), 557.
[61] Rui Zhang, Severin Bernhart, and Oliver Amft. 2016. Diet eyeglasses: Recognising food chewing using EMG and smart eyeglasses. In 2016 IEEE 13th International Conference on Wearable and Implantable Body Sensor Networks (BSN). IEEE, 7–12.
[62] Shibo Zhang, Rawan Alharbi, Matthew Nicholson, and Nabil Alshurafa. 2017. When generalized eating detection machine learning models fail in the field. In Proceedings of the 2017 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2017 ACM International Symposium on Wearable Computers. 613–622.
[63] Shibo Zhang, William Stogin, and Nabil Alshurafa. 2018. I sense overeating: Motif-based machine learning framework to detect overeating using wrist-worn sensing. Information Fusion 41 (2018), 37–47.
[64] Shibo Zhang, Yuqi Zhao, Dzung Tri Nguyen, Runsheng Xu, Sougata Sen, Josiah Hester, and Nabil Alshurafa. 2020. Necksense: A multi-sensor necklace for detecting eating activities in free-living conditions. Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies 4, 2 (2020), 1–26.
[65] Xiaowei Zhou, Menglong Zhu, Spyridon Leonardos, Konstantinos G Derpanis, and Kostas Daniilidis. 2016. Sparseness meets deepness: 3D human pose estimation from monocular video. In Proceedings of the IEEE conference on computer vision and pattern recognition. 4966–4975.

Received February 2022; revised May 2022; accepted June 2022