Eating Speed Measurement Using Wrist-Worn IMU Sensors in Free-Living Environments

Chunzhuo Wang, T. Sunil Kumar, Walter De Raedt, Guido Camps, Hans Hallez, Bart Vanrumste

Abstract—Eating speed is an important indicator that has been widely scrutinized in nutritional studies. The relationship between eating speed and several intake-related problems such as obesity, diabetes, and oral health has received increased attention from researchers. However, existing studies mainly use self-reported questionnaires to obtain participants' eating speed, where they choose options from slow, medium, and fast. Such a non-quantitative method is highly subjective and coarse in individual level. In this study, we propose a novel approach to measure eating speed in free-living environments automatically and objectively using wrist-worn inertial measurement unit (IMU) sensors. Specifically, a temporal convolutional network combined with a multi-head attention module (TCN-MHA) is developed to detect bites (including eating and drinking gestures) from free-living IMU data. The predicted bite sequences are then clustered to eating episodes. Eating speed is calculated by using the time taken to finish the eating episode to divide the number of bites. To validate the proposed approach on eating speed measurement, a 7-fold cross validation is applied to the self-collected fine-annotated full-day-I (FD-I) dataset, and a hold-out experiment is conducted on the full-day-II (FD-II) dataset. The two datasets are collected from 61 participants in free-living environments with a total duration of 513 h, which are publicly available. Experimental results shows that the proposed approach achieves a mean absolute percentage error (MAPE) of 0.110 and 0.146 in the FD-I and FD-II datasets, respectively, showcasing the feasibility of automated eating speed measurement. To the best of our knowledge, this is the first study investigating automated eating speed measurement.

Index Terms—Eating speed, food intake monitoring, eating gesture detection, inertial sensor, free-living

1 Introduction

Eating speed is considered as an important factor associated with body mass index (BMI), obesity and diabetes, which has been widely investigated [1]–[3]. Participants with faster eating speeds are considered more likely to have higher BMI, a higher risk of obesity, diabetes, and cardiovascular disease. Furthermore, deviations in eating speed are also correlated with eating disorders [4]. Currently the most popular method for investigating eating speed is through self-reported questionnaires [2], [3]. In the work of Kudo *et al.* [3], participants were asked to choose their own eating speed from options such as slow, medium, and fast, by responding to questions like "How fast do you eat compared to others around same ages? (Faster, Normal, Slower)." The questionnaire based eating speed estimation is highly subjective, and there is no standard reference to define an appropriate objective eating speed. While self-reported eating speed may be sufficient at a group level, especially with a large population, it is an unreliable approach to assess an individual's eating speed, particularly in the context of precision healthcare [5]. There is a call for an automated and objective approach to measure eating speed.

Recently, automated food intake monitoring has drawn lots of attention, plenty of approaches haven been proposed to detect bites [6]–[9] during meal sessions, detect eating episodes in free-living scenarios [10], [11], and estimate calorie intake using various sensors (e.g., inertial, camera, microphone, proximity) and machine learning techniques. However, to date, there has been no research that focus on automated eating speed detection in free-living environments.

In this study, we use the term *bite* to refer to eating gestures and drinking gestures. The definitions of eating gestures and drinking gestures are consistent with the work in [12]. Specifically, they are defined as the action of raising the hand to the mouth with cutlery or a water container until the hand is moved away from the mouth. The definition of objective eating speed is the number of bites divided by the time taken to finish the eating episode (bites/min) [1], [13]. Based on this definition, a straightforward approach for eating speed estimation is to combine the bite detection (to count the number of bites) and eating episode localization (to obtain the time duration of the consumed eating episode). However, the reason that hinder automated eating speed estimation is two-fold: Firstly, existing bite detection approaches only focus on in-meal scenarios, it is challenging to detect bites in free-living environments. Secondly, current eating episode localization approaches cannot precisely segment the boundary of detected eating episode, whereas eating speed detection requires accurate

This project is funded in part by KU Leuven under grant C3/20/016, and in part by the China Scholarship Council (CSC) under grant 202007650018.
Chunzhuo Wang and Bart Vanrumste are with the e-Media Research Lab, and also with the ESAT-STADIUS Division, KU Leuven, 3000 Leuven, Belgium (e-mail:chunzhuo.wang@kuleuven.be; bart.vanrumste@kuleuven.be).
T. Sunil Kumar is with Vellore Institute of Technology, Chennai, India (e-mail: suneel457.ece@gmail.com).
Walter De Raedt and Chunzhuo Wang are with the Life Science Department, IMEC, 3001 Heverlee, Belgium (e-mail: walter.deraedt@gmail.com).
Guido Camps is with the Division of Human Nutrition and Health, Department of Agrotechnology and Food Sciences, Wageningen University and Research, 6700EA Wageningen, and also with the OnePlanet Research Center, 6708WE Wageningen, The Netherlands (e-mail: guido.camps@wur.nl).
Hans Hallez is with the M-Group, DistriNet, Department of Computer Science, KU Leuven, 8200 Sint-Michiels, Belgium (e-mail: hans.hallez@kuleuven.be).

boundaries detection for each eating episode.

In this research, we extend the bite detection from in-meal scenarios to free-living environments, and further cluster the detected bite sequences into eating episodes, to facilitate the automated eating speed estimation. The main contributions of this research can be summarized as follows:

- A complete framework for eating speed measurement in free-living environments is proposed: A sequence-to-sequence (seq2seq) temporal convolutional network combined with a multi-head attention (TCN-MHA) model is designed to process inertial measurement unit (IMU) data for detecting food intake gestures and segmenting the time interval of intake gestures in free-living environments. The obtained bite sequences are clustered into eating episodes to calculate eating speed. To our best knowledge, this is the first work to automatically estimate eating speed in free-living environments.
- An intensive comparison between our approach and existing works has been implemented. Five existing deep learning models are implemented for comparison. Additionally, the proposed eating speed measurement method is validated in two studies: 7-fold cross validation on the well-annotated FD-I dataset, and a hold-out validation on the coarsely annotated FD-II dataset.
- We make two datasets collected in this study publicly available[1]. Specifically, the FD-I dataset contains IMU data collected from 34 participants in free-living environments with fine annotation. This is the first full-day IMU dataset that contains eating and drinking gesture annotations not only during meal sessions, but also outside of meal sessions. The FD-II dataset serves as a hold-out dataset, containing IMU data from 27 participants in free-living environments.

2 Related work

Automated eating speed detection relies on two essential tasks: bite detection to count the number of bites and eating episode detection to predict the duration of each meal. In this section, we firstly introduce existing approaches for in-meal bite detection; then, we discuss eating episode detection approaches. Thirdly, we present a few studies measuring eating speed objectively (but not automated). Finally, deep learning models for time-series signal processing are introduced.

2.1 Bite Detection

Bite detection has been widely investigated using various sensors. Cameras have been used to detect bites [14], [15] and food types [7] by analyzing video signals. Acoustic sensors can be used for this task through processing chewing sounds [8]. Mertes *et al.* [16] developed a strain gauge-based smart plate to detect bites based on the weight change of food. In our recent work [12], a novel FMCW radar-based system was proposed for in-meal bite detection. This system was validated using our public Eat-Radar dataset, which includes 70 meals from 70 participants. The photoplethysmography (PPG) sensor [17] and electromyography (EMG) sensor [18] have also been explored for bite detection. Apart from these sensors, to date, the wrist-worn IMU sensor is a popular choice for bite detection because of its least burdensome and most acceptable. Dong *et al.* [19] developed a rule-based approach to detect bite using the rotation velocity of wrist. Shen *et al.* [20] further evaluated Dong's approach on Clemson dataset containing 488 eating episodes. Kyritsis *et al.* [6] proposed an end-to-end based approach using a convolutional neural network and long-short-term-memory network hybrid model (CNN-LSTM) model to detect bite automatically on FIC dataset. Rouast *et al.* [21] further developed single-stage ResNet based CNN-LSTM architecture for bite detection on the OREBA dataset, which contains 100 meals. Wei *et al.* [22] developed an energy-efficient approach, specifically, an optimized multicenter classifier (O-MCC) and an Android application, to detect intake gestures with low inference time.

The aforementioned bite detection approaches have shown promising performances. However, it should be noted that these approaches only focus on bite detection in-meal sessions (10-20 min). Although such in-meal detection is the key and fundamental element in food intake monitoring, a more challenging but meaningful scenario, bite detection in free-living environments ($\geq$ 6 h), has yet to be broadly investigated. There are several obstacles impeding the detection in free-living environments. Firstly, it is troublesome to obtain bite-level annotation information outside meal sessions. Secondly, in the scale of full-day duration, bites are extremely sparse, leading a highly imbalanced dataset compared to in-meal datasets, making bite detection more challenging.

2.2 Eating Episodes Detection

Eating episodes detection is another popular research topic in food intake monitoring. Such a system mainly focus on the detection of eating episodes under free-living environments. A standard process pipeline involves cutting the full-day data into minute-level segments, and machine learning is used to predict if each segment belongs to eating episode or not. Sharma *et al.* [10] collected a data set contains 354 days of 6-axis IMU data from 351 subjects. The IMU sensor were mounted on the dominant wrist. Doulah *et al.* [11] developed the AIM-2 system, a pair of eyeglasses mounted with a camera, and a 3-axis accelerometer to detect eating episodes. The AIM-2 was validated on a dataset collected from 30 volunteers.

Unlike the above methods that directly predict eating episodes, another routine is first to detect basic elements of eating episode such as chewing, swallowing, and hand-to-mouth events, and then combine them together as an eating episode using various merging techniques. Bedri *et al.* [23] proposed the FitByte eyeglass-based system to detect eating/drinking events and eating episodes. They detected the episodes by merging any detected intakes that are within 5 minutes from each other. Zhang *et al.* [24] developed Necksense system to detect chewing sequences, then clustered the detected sequences into eating episodes using a density-based spatial clustering of applications with noise

1. We plan to add the link of this dataset in the revision stage.

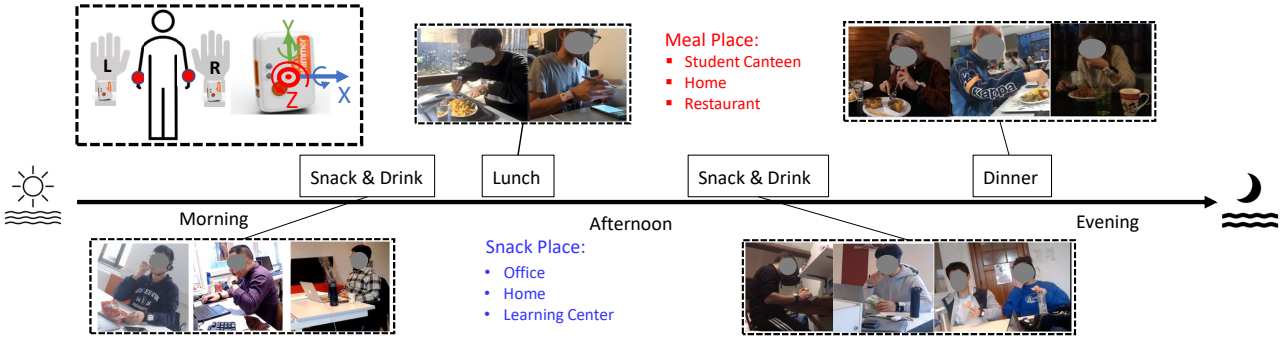

Fig. 1. Examples of wrist-worn IMU data collection. Two IMU sensors were mounted on both hands. Participants completed their daily activities without restriction. This figure only shows the food intake related scenes, their other daily activities, such as studying, walking, talking, running, cooking, were recorded by IMU sensor as well.

(DBSCAN) algorithm [25]. Kyritsis *et al.* [6] first detected bites on FreeFIC dataset, then applied a Gaussian filter on the bite sequence to generate meal regions. It should be noted that the model used to detect bites is trained on in-meal FIC dataset, there is no bite-level label in FreeFIC dataset, so the bite detection performance is unclear on FreeFIC dataset.

2.3 Eating Speed Estimation

Two studies have been found to objectively (but not automated) measure eating speed. Woodward *et al.* [5] provided a 550 g meal to each participant and used a stopwatch to record the meal time. Their experiment was conducted in lab environments. Alshurafa *et al.* [13] measured eating speed by using wearable fish-eye camera in free-living environments. A trained annotator manually counted the number of intake gestures, selected the boundaries of meal sessions by viewing the video. The eating speed was then calculated by using meal duration to divide the number of bites. The aforementioned approaches require manual processing, which are time-consuming and labor-intensive.

2.4 Temporal Sequence Models

The commonly used model in intake gesture detection is CNN with recurrent neural networks (CNN-RNNs). In this architecture, CNN is typically used to extract features from time-series data, and the feature sequences are then fed into RNNs to process temporal dependencies. Although RNNs can effectively process time-series data, they struggle to memorize long-term interdependencies due to the gradient vanishing problem. Two solutions have been proposed in the literature: the temporal convolutional network (TCN) and self-attention.

2.4.1 TCN

Lea *et al.* [26] proposed the TCN by utilizing dilated convolution and residual connection. Stacking a series of convolution layers with different dilation factors enables the model to incorporate short-term and long-term dependencies. Additionally, dilated convolutional layers increase the length of receptive fields without substantially increasing the number of parameters. Due to its superiority, studies have been conducted to further evolve the architecture (e.g. MS-TCN [27], MS-GCN [28]) and to exploit it into various scenarios.

2.4.2 Self-Attention

The self-attention module was proposed by Vaswani *et al.* [29] to compose the transformer architecture in natural language processing (NLP) domain, showcasing its superior ability. Motivated by its success, several approaches have been proposed by integrating the self-attention mechanism into CNN and RNN architectures to further improve the capability of time-series signal modeling [30], [31].

3 Methods

3.1 Sensors

As this experiment aims to record data in free-living environments, food intake events can happen with both hands, thus, two shimmer3 IMU wristbands[2] were mounted on both hands of participants. The battery duration of shimmer3 IMU is 24 h, which satisfies the requirement of this experiment. The sampling frequency was set to 64 Hz. The 3-axis accelerometer and 3-axis gyroscope units were activated to generate 6 channels IMU data. The data were stored in SD card of shimmer, and can be downloaded into laptop via Consensys software[3] for further data processing. A camera was used to record the experiment for annotation. The sensor deployment and data collection example are shown in Fig. 1.

3.2 Full-Day Data Collection

This research was approved by the ethical committee of KU Leuven with project number: G-2021-4025-R4. Informed consent was obtained from all participants. Two datasets were collected for this study, namely the well-annotated full-day-I (FD-I) dataset, and hold-out full-day-II (FD-II) dataset. There is no participants overlap among the two datasets.

2. https://shimmersensing.com/product/shimmer3-imu-unit/
3. https://shimmersensing.com/product/consensyspro-software/

3.2.1 FD-I Dataset

The FD-I dataset contains 34 days of IMU data from 34 participants (6 of them are from our previous study on drinking activity detection [32]). On the data collection day, our research assistants met the participant, instructed the participant to wear IMU wristbands. Participants were free to engage in their normal daily activities. The daily activity of each participant was recorded by a camera for annotation. Experiment locations contained participant's home (apartments, student residence), restaurants, library, university learning center, and campus rest areas. Our research assistants were responsible for the recording when participants change their locations to ensure that all eating and drinking gestures were captured. There were no limitations on the participants' activity during data collection. At least one meal was collected from each participant, and the minimum data collection duration was 6 h. Both eating alone and social eating scenarios were included in the dataset. The participants received restaurant voucher (20 euro) as experiment compensation after the data collection. A total of 251 h two-hand IMU data were collected, which contains 74 eating episodes, with 4,568 eating and 1,100 drinking gestures. The dataset contains four eating styles including forks & knives, chopsticks, spoon, and hands.

3.2.2 FD-II Dataset

The FD-II dataset contains 27 days of IMU data from 27 participants. The experiment protocol was the same as the FD-I dataset. However, in this dataset, only meal sessions were recorded by cameras (Some videos were collected by participants' own smartphones). All other eating/drinking gestures outside of meal sessions were not recorded. Therefore, the ground truth information only contains the bite information during meals and the meal boundaries. The FD-II dataset is considered as a hold-out dataset, which contains 48 meals with 2,723 eating gestures (including four eating styles) over a total duration of 262 h.

3.3 Datasets for Training-Only

The FD-I and FD-II datasets are highly unbalanced, with the target classes (eating and drinking) being the minority. To further include more target data, we included two datasets containing IMU data collected in meal sessions, specially, the self-collected meal-only (MO) dataset and the external OREBA dataset [33] as part of the training set.

3.3.1 Meal-Only Dataset

The MO dataset contains 48 meal sessions from 48 participants, with a total of 2,894 eating and 763 drinking activities. It should be noted that part of this dataset was collected together with our previous Eat-Radar project [12] and there is no participant overlap between MO and FD datasets.

3.3.2 Public OREBA Dataset

The OREBA dataset [33] contains 100 meal sessions data from 100 participants, with 4,496 eating and 406 drinking gestures. The data were collected from both hands using two IMU wristbands. It should be noted that the coordinate system (direction of x, y, and z axis) of the sensor used in OREBA is the same as ours, allowing us to integrate this data into the training set.

TABLE 1
Full Day Datasets Statistics

Parameter	FD-I	FD-II
# Participants	34	27
# Days	34	27
# Eating episodes	74	52
# Eating gestures	4,568	2,723
# Drinking gestures	1,100	-
Mean day duration (h)	7.40±2.13	9.71±3.79
Duration ratio of other : eating : drinking	142.52 : 2.51 : 1	116.43 : 1 : -

3.4 Annotation

3.4.1 Bite Annotation

Video recordings were viewed to annotate bites via ELAN [34]. The data were labeled into 3 classes, eating gesture, drinking gesture, and others. Three trained annotators labeled the datasets, with each annotator assigned to a specific portion of the dataset. The first author rechecked the annotation and made corrections when necessary. All annotators followed the same annotation instructions.

3.4.2 Eating Episodes Annotation

After annotating all the eating/drinking gesture from all-day data, the first eating gesture in an eating episode signifies the beginning boundary of the eating episode, and the last eating gesture in an eating episode is considered the ending boundary of the episode.

In free-living environments, it is normal for people to eat snacks outside meal sessions. However, snack eating exhibits high variability compared to meal eating. Some snack eating has frequent bites in very short duration, while other snack eating occurs over a longer period with a very low frequency (only one bite in several minutes). This variability makes the detection of snack session difficult; therefore, we focus on eating episodes that last at least 3 min. Snacking sessions with duration less than 3 min were neglected in this step.

3.4.3 Ground Truth Eating Speed

After annotating both eating gestures and eating episodes, we derived the number of bites in each eating episode and the duration of each respective eating episode. Consequently, the ground truth eating speed is obtained by computing the ratio of the number of bites in the eating episode and the time taken to finish the episode. The unit of eating speed is bites per minute (bite/min).

3.5 Data Preprocessing

Existing approaches in free-living environments typically process IMU data from the dominant hand, assuming that people only use one hand for daily food intake. However, in real-life, both hands can be used for eating food or drinking water, as illustrated in Fig. 2. Considering the IMU waveform for eating with left and right hand differs, we applied the two-hand combination method that combines hand mirroring and temporal concatenation to process two-hand IMU data, which was validated in our previous study [35]. The hand mirroring method has been applied in multiple IMU-based bite detection studies when the participant

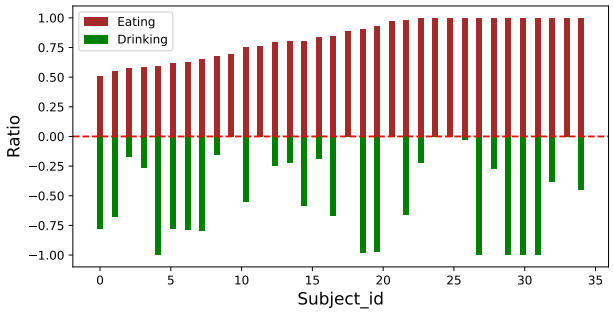

Fig. 2. The quantity ratio of eating and drinking gestures using only dominant hand for each participant on FD-I dataset. A ratio with 0.5 means half number of eating gestures are from the dominant hand, the others are from the non-dominant hand, a ratio with -1 means all drinking gestures are from the dominant hand.

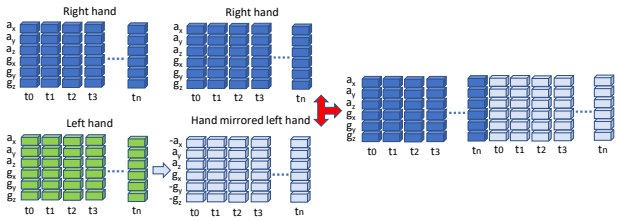

Fig. 3. Visualization of the two hand combination process. t_n represents the end time point of the eating episode.

is left hand dominant, which involves flipping the direction of a_x (in accelerometer), g_y and g_z (in gyroscope). The hand-mirrored left hand IMU data were than concatenated after right hand data. Hence, the preprocessed data have 6 channels, and the data length of each recording is doubled, as shown in Fig. 3. Meanwhile, to reduce the computation cost, the data were downsampled to 16 Hz.

3.6 Deep Learning Model

To explore the potential of utilizing both TCN and attention, the proposed TCN-MHA architecture is comprised of three parts: the TCN module using dilated convolution layer to process multi-scale temporal patterns, the multi-head attention (MHA) module to further focus on representative temporal features to improve the performance, and the fully connected network (FCN) to generate predictions, as shown in Fig. 4.

3.6.1 TCN Module

The TCN distinguishes classical CNN by using dilated convolutions [26]. A series of dilated convolution layers are stacked together to compose the TCN module. Each layer involves conv1d with a dilation factor $d_l = 2^{l-1}$ $(1 \leq l \leq L)$, where l, L are the number of the current layer and total layers, respectively. A residual connection is also applied in each layer to combine the input features of current layer and the processed features. The stacked dilated layers increased the receptive field without substantially increasing the number of parameters, allowing the TCN to effectively capture long-term temporal dependencies. The structure is shown in Fig. 4 (a). Given the input sequences $X \in \mathbb{R}^{T \times C_{in}}$, a 1×1 convolution layer is first applied to adjust the dimension to $T \times C_m$, where C_{in} is the input dimension, and C_m equals to the number of kernels in each dilated convolution layer (the number of kernels of each conv layer is the same). After L dilated conv layers, the output of the TCN module is $X_1 \in \mathbb{R}^{T \times C_m}$.

3.6.2 MHA Module

The MHA module is an essential component of the Transformer [29], showcasing considerable efficacy in capturing relationships within time sequence signals. Before being fed into the attention module, the positional encoding is added to the input feature map. The attention mechanism is considered as a mapping among the query ($Q = X_1 W_Q$), key ($K = X_1 W_K$), and value ($V = X_1 W_V$), where $X_1 \in \mathbb{R}^{T \times d_{in}}$ signifies the input of the module ($d_{in} = C_m$), W_Q,W_K and $W_V \in \mathbb{R}^{d_{in} \times d_{model}}$, with d_{in} representing the dimension of the output sequence from the TCN model, and d_{model} denoting the dimension of the attention module. The formula of the attention calculation is presented as follows:

$$Att(Q, K, V) = \text{Softmax}\left(\frac{QK^\top}{\sqrt{d_{model}}}\right)V \tag{1}$$

The aforementioned computation involves a singular attention graph; however, multi-head attention employs multiple attention graphs to further learn attention maps across diverse aspects. The parameter h denotes the number of heads, signifying the quantity of attention graphs in use.

$$\begin{aligned} MHA(Q, K, V) &= \text{Concat}(SA_1, \ldots, SA_h)W^O \\ \text{where } SA_i &= Att(QW_i^Q, KW_i^K, VW_i^V) \end{aligned} \tag{2}$$

where $W^O \in \mathbb{R}^{h \cdot d_h \times d_{model}}$, W_i^Q, W_i^K and $W_i^V \in \mathbb{R}^{d_{model} \times d_h}$, and d_h denotes the dimension of each individual head in the attention mechanism, here we keep $d_h = d_{model}/h$.

3.6.3 FCN Module

The output of the MHA module is fed into the final FCN classifier block containing 2 linear layers, yielding $Y \in \mathbb{R}^{T \times C_{out}}$, where C_{out} is the number of classes.

3.6.4 Loss Function

The model's loss function comprises both classification and smoothing loss. Specifically, the classification loss is implemented through a cross-entropy loss, while the smoothing loss involves a truncated mean squared error (MSE) calculated over sample-wise log-probabilities. Details on the integrated loss function can be referenced in [27].

3.6.5 Implementation Details

The TCN-MHA model is implemented within the PyTorch framework. Following each dilated layer, a dropout rate of 30% is applied. Based on experiments, the TCN module is constructed with a total of 9 layers, with each layer comprising 64 kernels. For the MHA module, 8 attention heads are employed, with each head characterized by a dimension of 16, resulting in a model dimension of 128 (d_{model}=128). The first layer of the FCN has 64 neurons with ReLU activation, the last layer contains 3 neurons with Softmax activation. For model training, an Adam optimizer

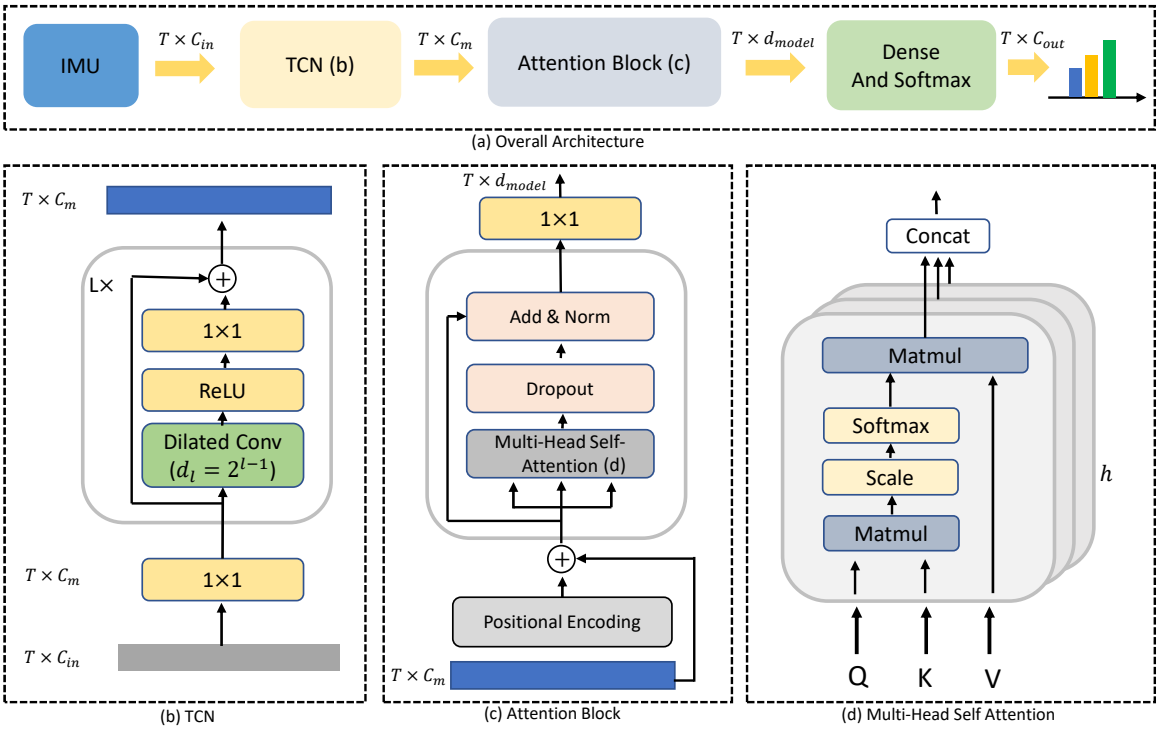

Fig. 4. The architecture of the proposed TCN-MHA model. Figure (a) shows the overall architecture of the model. The processed IMU data are first fed into the TCN module, followed by further processing of the TCN module's output through the Attention block. Finally, the FCN block is applied to process the output of the attention block to generate predictions. Figure (b) represents the architecture of TCN module. The architecture of attention block is illustrated in Figure (c). Figure (d) explains the mechanism of the multi-head attention module. d_l signifies the dilated factor, where l denotes the layer order, L represents the total number of TCN layers, h represents the number of attention heads.

is utilized with a learning rate set to 0.0005. It is important to note that the model exhibits a temporal lag in its predictions due to its non-causal nature. This temporal delay can be quantified by dividing half of the receptive field by the sampling frequency (0.5×1023/16=32s). The window length of input data is set to 60 s accordingly.

All experiments for training, validation, and testing were carried out on a computational node equipped with an Intel 18-core Xeon Gold 6140 CPUs@2.3 GHz (Skylake) with 5 GB RAM per core, and two NVIDIA P100-SXM2-16 GB GPUs provided by Vlaams Supercomputer Centrum (VSC) [4].

3.7 Bite Detection

The argmax is applied on the probability sequence generated by the TCN-MHA model to yield point-wise predictions. In order to mitigate the impact of noise in the predictions, adjacent bites intervals with identical values that are within 0.5 s are consolidated. After that, bites with duration less than 1 s are excluded. Subsequently, we obtain a set of detected bites $\mathcal{B} = \{b_1, b_2, ...b_N\}$, where each b_i corresponds to the interval $[t_i^l, t_i^r]$ $(1 \leq i \leq N)$ representing the left and right temporal boundaries of the i-th bite, N denotes the total number of detected bites. It should be noted that the value of the interval distinguishes the type of bite (Eating: 1, Drinking: 2).

4. See https://www.vscentrum.be/

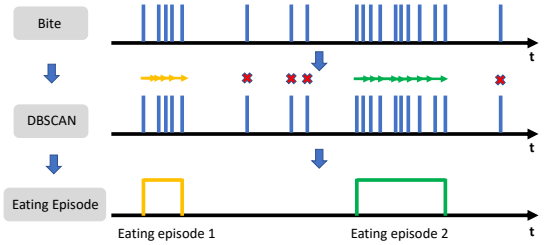

Fig. 5. The DBSCAN clustering example for eating episode detection.

3.8 Eating Episode Detection

In bite detection step, to achieve data uniformity, the data from the left hand is hand mirrored and then temporally concatenated after right hand data. Prior to meal session detection, the output data is divided into to two subsequences (left hand and right hand), then an OR operation is applied to integrate bites from both hands. As eating episodes mainly involves eating gestures, all detected drinking gestures are removed in this step.

The predicted bite sequence is then clustered by 1D density-based spatial clustering of applications with noise algorithm (1D-DBSCAN) [25] to compose eating episodes. The DBSCAN identifies clusters based on the density of points. The function *sklearn.cluster.DBSCAN* is employed with an epsilon parameter set to 3 min and a minimum samples parameter set to 5. In our case, the distance between two bites is the temporal proximity between the bites. Subsequently, sparse bites are filtered as noise, while bites

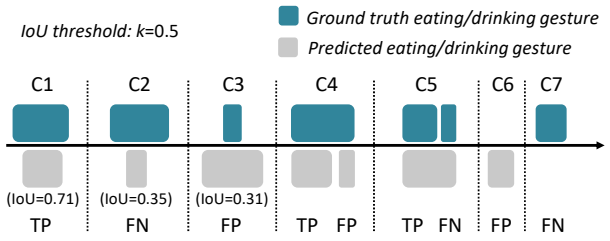

Fig. 6. Examples of segment-wise evaluation. When the evaluated class is drinking gesture, eating gestures are categorized as other, and vice versa.

sequence with high density are clustered together to compose eating episodes, as shown in Fig. 5. After clustering, we follow the same operations in [6], [36] to merge very close eating episodes and remove very short episodes. Specifically, if the distance between two eating episodes is less than 3 min, they are merged into one episode. Additionally, if the duration of the episode is less than 3 min after merging, the eating episode is removed. Finally, we obtain a set of detected eating episode $\mathcal{M} = \{m_1, m_2, ...m_Z\}$, where each m_j corresponds to the interval $[t_j^l, t_j^r]$ $(1 \leq j \leq Z)$ that corresponds the left and right boundaries of j-th eating episode, Z denotes the total number of detected eating episodes.

3.9 Eating Speed Estimation

After yielding eating episodes from the previous step, eating speeds are obtained by utilizing the previously detected bite set $\mathcal{B}$ and the eating episode set $\mathcal{M}$. The eating episode detection algorithm predicts the start point t_j^l and end point t_j^r of the j-th eating episode. If the detected bite b_i falls within the interval of the eating episode, this bite is considered to belong to that eating episode. At the end, the number of bites divided by the duration of the eating episode gives the eating speed.

4 Evaluation and Experiment

4.1 Evaluation Criteria

4.1.1 Evaluation on Bite Detection

The output of the proposed seq2seq model is point-wise multi-class prediction. As bite-related datasets are normally unbalanced, we choose to use the index Cohen Kappa [37] to represent the performance of point-wise classification. Although such results can indicate the performance of the model, it should be noted that the purpose of bite detection is to count the number of bites, whereas point-wise results are unable to reveal such information. To address this issue, we use a segment-wise evaluation method to evaluate the bite detection, which has been applied in previous study [12]. Fig. 6 shows examples of this evaluation. The evaluation method involves two steps. Firstly, the intersection over union (IoU) between each predicted bite and ground truth bite is calculated, as shown in Fig. 6 C1-C3. Secondly, the calculated IoU is compared to a selected threshold k to determine segment-wise true positive (TP), false negative (FN) and false positive (FP). Subsequently, the segmental F1-score is calculated for each class (eating and drinking).

The segment-wise evaluation scheme allows for short temporal shifts between ground truth and prediction, which maybe caused by annotation variability. Meanwhile, it furnishes straightforward information including the number of detected bites. Furthermore, by adjusting the threshold k, we can evaluate not only the detection performance, but also the segmentation performance. In this study, two thresholds are selected as 0.1 and 0.5.

4.1.2 Evaluation on Eating Episode Detection

The evaluation of eating episode detection mainly focuses on two aspects, specifically, the detection performance (how many eating episodes are detected), and the segmentation performance (how well the boundaries of each eating episode are determined). Therefore, the aforementioned segment-wise evaluation method is also used for eating episode detection ($k = 0.5$). Additionally, for each predicted eating episode, we utilize the IoU score to evaluate the segmentation performance.

4.1.3 Evaluation on Eating Speed Estimation

The mean absolute percentage error (MAPE) is used to evaluate the deviation between the estimated speed and the ground truth speed.

$$\mathrm{MAPE} = \frac{1}{\mathrm{z}} \sum_{\mathrm{i}=1}^{\mathrm{z}} \left| \frac{\hat{\mathrm{s}}_\mathrm{i} - \mathrm{s}_\mathrm{i}}{\mathrm{s}_\mathrm{i}} \right| \times 100\% \tag{3}$$

where z is the total number of truly detected meals (TP), $\hat{s}_i$ and s_i represent the estimated eating speed and ground truth eating speed, respectively.

For statistical quantitative analysis, the Pearson correlation coefficient (PCC) is also calculated to assess the correlation of the predicted eating speed with the ground truth objectively.

$$\mathrm{PCC} = \frac{\sum_{\mathrm{i}=1}^{\mathrm{z}} (\mathrm{s}_\mathrm{i} - \bar{\mathrm{s}})(\hat{\mathrm{s}}_\mathrm{i} - \bar{\hat{\mathrm{s}}})}{\sqrt{\sum_{\mathrm{i}=1}^{\mathrm{z}} (\mathrm{s}_\mathrm{i} - \bar{\mathrm{s}})^2 \cdot \sum_{\mathrm{i}=1}^{\mathrm{z}} (\hat{\mathrm{s}}_\mathrm{i} - \bar{\hat{\mathrm{s}}})^2}} \tag{4}$$

where $\bar{s}$ and $\bar{\hat{s}}$ represent the mean value of the ground truth speed and estimated speed, respectively. It should be noted that the MAPE and PCC are only calculated among eating speeds of successfully detected eating episodes.

4.2 Models for Benchmarking

To evaluate the efficacy of the proposed method, existing models for bite detection were chosen as comparative benchmarks. Specifically, the CNN-LSTM [6], ResNet-LSTM [21], and MS-TCN [32] were chosen. Additionally, the bidirectional type of LSTM layer (BiLSTM) was used to replace the LSTM layer to compose the CNN-BiLSTM and ResNet-BiLSTM models as extra models. All of these models can learn temporal context from time-series input data and have demonstrated their capability in human activity recognition. It should be noted that the output of these models are point-wise probabilities.

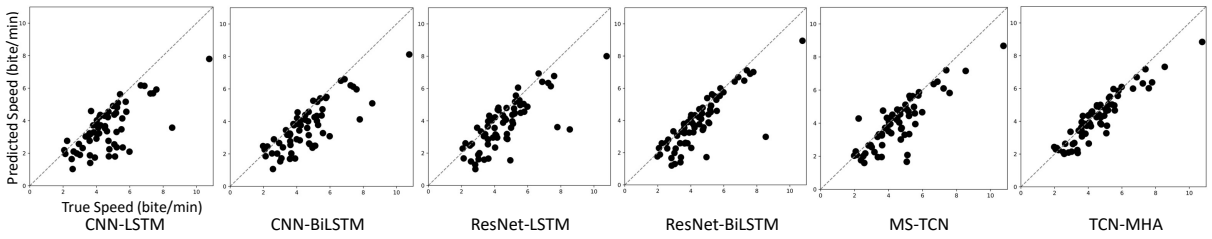

Fig. 7. The scatter plot for eating speed on FD-I dataset.

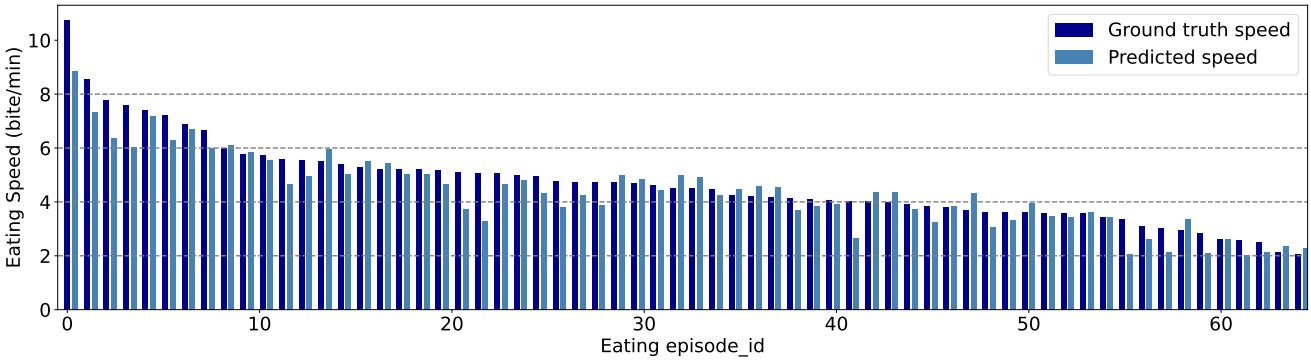

Fig. 8. The bar plot for eating speed on FD-I dataset.

5 Experimental Results

5.1 Experiments on FD-I Dataset

5.1.1 Bite Detection

The bite detection in free-living settings was tested on FD-I dataset. The 7-fold cross validation was used on FD-I dataset. The split of this dataset was on participant-level to avoid information leakage between the train set and test set. As mentioned in Section III-C, the FD-I dataset is highly unbalanced, hence, the MO and OREBA datasets were included as part of the training set. Results are shown in Table 2. For point-wise results, TCN-MHA obtained the highest kappa score (0.735). For segment-wise results, the TCN-MHA also achieved the highest F1-score for eating with the value of 0.849 and 0.781, for $k = 0.1$ and $k = 0.5$, respectively. However, MS-TCN yielded a higher F1-score for drinking compared to TCN-MHA (0.906→0.902, for $k = 0.1$).

5.1.2 Eating Episode Detection

The predicted bite sequences from previous step were clustered into eating episodes using the DBSCAN-based algorithm. The results of eating episodes detection are shown in Table 3. Among the 74 ground truth eating episodes, the TCN-MHA successfully detected 64 sessions with a mean IoU of 0.899.

5.1.3 Eating Speed

For eating speed estimation, the MAPE and PCC are shown in Table 3. The TCN-MHA model had the least MAPE of 0.110; the highest PCC value of 0.925. The scatter plots are drawn to graphically demonstrate the correlation between predicted and ground truth eating speed in the FD-I dataset (Fig. 7). Most eating episodes had an eating speed falling into the range of 2-6 bite/min. Additionally, the result for each individual episode on TCN-MHA model has been shown in Fig. 8.

5.2 Experiments on FD-II Dataset

To further validate the proposed method on eating speed measurement, the FD-II was used as the hold-out dataset. We utilized two in-meal datasets (OREBA, MO) and the FD-I dataset to train our model, then used the entire FD-II as the test set. It should be noted that we were only able to measure the eating speed during meal sessions, as we lack

TABLE 2
Bite Detection Performance on FD-I Dataset

Data	Model	Point-wise Kappa	Segmental Eating F1-score $k = 0.1$	Segmental Eating F1-score $k = 0.5$	Segmental Drinking F1-score $k = 0.1$	Segmental Drinking F1-score $k = 0.5$
FD-I	CNN-LSTM	0.557	0.738	0.583	0.811	0.619
	CNN-BiLSTM	0.666	0.788	0.694	0.859	0.762
	ResNet-LSTM	0.630	0.753	0.626	0.791	0.682
	ResNet-BiLSTM	0.704	0.790	0.701	0.849	0.779
	MS-TCN	0.702	0.824	0.761	**0.906**	0.853
	TCN-MHA	**0.735**	**0.849**	**0.781**	0.902	**0.858**

TABLE 3
Eating Episode Detection and Eating Speed Performance on FD-I Dataset.

Data	Model	Eating Episodes Detection TP	FP	FN	F1	IoU	Eating Speed MAPE	PCC
FD-I	CNN-LSTM	60	0	14	0.896	**0.900**	0.238	0.696
	CNN-BiLSTM	63	1	11	0.913	0.895	0.202	0.789
	ResNet-LSTM	65	6	9	0.897	0.863	0.197	0.782
	ResNet-BiLSTM	64	6	10	0.889	0.864	0.155	0.829
	MS-TCN	62	0	12	0.912	0.881	0.151	0.827
	TCN-MHA	64	0	10	**0.928**	0.899	**0.110**	**0.925**

TABLE 4
In-meal Bite Detection, Eating Episode Detection and Eating Speed Performance on FD-II Dataset

Data	Model	Eating Gesture F1-score[a]		Eating Episodes Detection					Eating Speed	
		$k = 0.1$	$k = 0.5$	TP	FP	FN	F1	IoU	MAPE	PCC
FD-II (Hold-out)	CNN-LSTM	0.731	0.555	43	53	9	0.581	0.746	0.231	0.683
	CNN-BiLSTM	0.783	0.622	47	35	5	0.701	0.786	0.173	0.860
	ResNet-LSTM	0.752	0.584	48	42	4	0.676	0.809	0.201	0.780
	ResNet-BiLSTM	0.798	0.650	49	26	4	0.766	0.827	**0.142**	0.910
	MS-TCN	0.814	0.636	47	22	5	0.777	0.831	0.155	0.886
	TCN-MHA	**0.820**	**0.651**	48	7	4	**0.897**	**0.841**	0.146	**0.924**

[a] The eating segmental F1-scores in FD-II dataset only show the results of bite detection in meal sessions. Eating gestures from the outside of meals are not labelled, hence is unable to be evaluated.

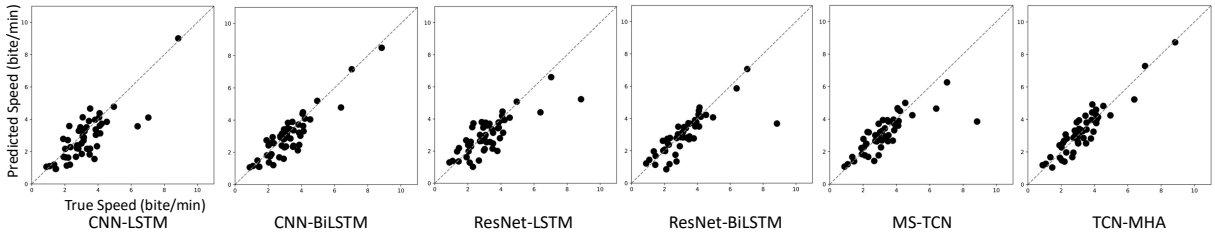

Fig. 9. The scatter plot for eating speed on FD-II dataset.

labels for out-of-meal sessions. Hence, we did not evaluate the predicted snack sessions (eating episodes with duration less than 7 min). The performance of in-meal bite detection, meal detection, and speed measurement are shown in Table 4. For in-meal bite detection, the TCN-MHA model yielded the highest segmental F1-score with 0.820 when $k = 0.1$. The TCN-MHA obtained the best performance in meal detection, which successfully detected 48 meal sessions (7 FPs, 4 FNs), with an F1-score of 0.897 and a mean IoU of 0.841. For eating speed measurement, the TCN-MHA had the MAPE of 0.146 and the PCC of 0.924. The ResNet-BiLSTM had the highest number of TP meals, however the number of FPs was also higher than that of the TCN-MHA. Scatter plots to show the deviation between predicted and true eating speed (only for meal sessions) in the FD-II dataset are shown in Fig. 9.

6 Discussion

In this study, we first tested several models on free-living datasets (FD-I and FD-II) for bite detection. Results from Table 2 show that bite detection in free-living environments is feasible. Meanwhile, Table 3 also indicates a tendency for models to overlook certain eating episodes, as evidenced by a higher number of FNs compared to FPs. Upon comparing the output and the annotation video, we found that FNs are mainly from snack sessions, implying that the clustering of snack sessions is more challenging than detecting meals. One reason is that the eating patterns of snacking can vary widely in terms of frequency and duration, making it difficult to cluster all snacking gestures into specific episodes. Other eating episode detection studies also suffer from this [10], [38].

When comparing our work to prior studies on food intake monitoring in free-living environments (as shown in Table 5), the dataset exhibits a comparable size to others, except for the dataset used by Sharma *et al*. [10], [44], which is substantially larger than the rest. However, it's worth noting that existing free-living datasets mainly focus on eating episode detection, which only requires labeling the starting and ending times of eating episodes. In contrast, our data for eating speed measurement requires bite-level annotation, which demands additional efforts for data collection and annotation. The detection granularity of this study is shown in Fig. 10. Meanwhile, compared to eyeglass-based and necklace based approaches focusing on mouth-throat movements for food intake monitoring, one limitation of wristband-based approach is that we need to wear IMUs on both hands. Wearing the IMU only on the dominant hand may lead to the omission of some eating or drinking gestures in free-living environments, as illustrated in Fig. 2. However, the advantage of the wrist-worn IMU sensor lies in its wide integration into smartwatch products, making it more readily accepted by users compared to other solutions. Table 5 can also show the prevalence of wrist-worn IMU.

In food intake monitoring, several wrist-worn IMU based bite detection datasets have been published, however, they are only used separately to benchmark performance of different models. In this study, we examined the feasibility of combining different datasets for training. When performing bite detection on the FD-I dataset, our own MO dataset and the external public OREBA dataset were included as part of training set. This is the first attempt to integrate different datasets for food intake monitoring, resulting in a 0.7% increase of F1-score in bite detection, 3.4% increase of PCC in eating speed detection. However, it is important to note that the prerequisite of this integration is that the coordinates of the IMU sensors used in both datasets are the same. The coordinates of IMU in OREBA datasets were adjusted to make sure the orientation of x, y, z axis of the two datasets were consistent (ours is depicted

TABLE 5
Existing studies on Food Intake Monitoring in Free-living Environment

Work	Position[a]	Sensor[b]	# Participants	# Days	# Hours	Eating Episode detection	Bite annotation & detection[c]	Eating speed
Dong *et al*. (2014) [39]	P1	S1	43	43	449	✓	-	-
Fontana *et al*. (2014) [40]	P1, P2, P3	S2, S3, S4	12	12	-	✓	-	-
Thomaz *et al*. (2015) [41]	P1	S1	8	37	-	✓	-	-
Mirtchouk *et al*. (2017) [38]	P1, P2, P4	S1, S5	11	25	257	✓	-	-
Bedri *et al*. (2017) [36]	P2, P5	S1, S5, S6	10	-	45	✓	-	-
Sen *et al*. (2018) [42]	P1	S1, S7	9	-	52	✓	-	-
Schiboni *et al*. (2018) [43]	P1	S1	7	35	345	-	✓	-
Sharma *et al*. (2020) [10], [44]	P1	S1	351	351	4,068	✓	-	-
Doulah *et al*. (2020) [11]	P4	S2, S7, S8	30	60	-	✓	-	-
Zhang *et al*. (2020) [24]	P5	S1, S6, S9	20	-	271	✓	-	-
Bedri *et al*. (2020) [23]	P4	S1, S6, S7	23	8	91	✓	✓	-
Kyrtisis *et al*. (2020) [6]	P1	S1	12	12	113	✓	-	-
Ours	**P1**	**S1**	**61**	**61**	**513**	✓	✓	✓

[a] P1: Wrist, P2: Ear, P3: Chest, P4: Head, P5: Neck.
[b] S1: IMU, S2: Accelerometer, S3: Piezo, S4: RF, S5: Microphones, S6: Proximity, S7: Camera, S8: Flex, S9: Light.
[c] It should be noted that some approaches use in-meal datasets to train the model to detect bite, then use the trained model to process free-living datasets, but these free-living datasets do not contain bite-level label, so they are considered no bite detection & evaluation in free-living scenarios.

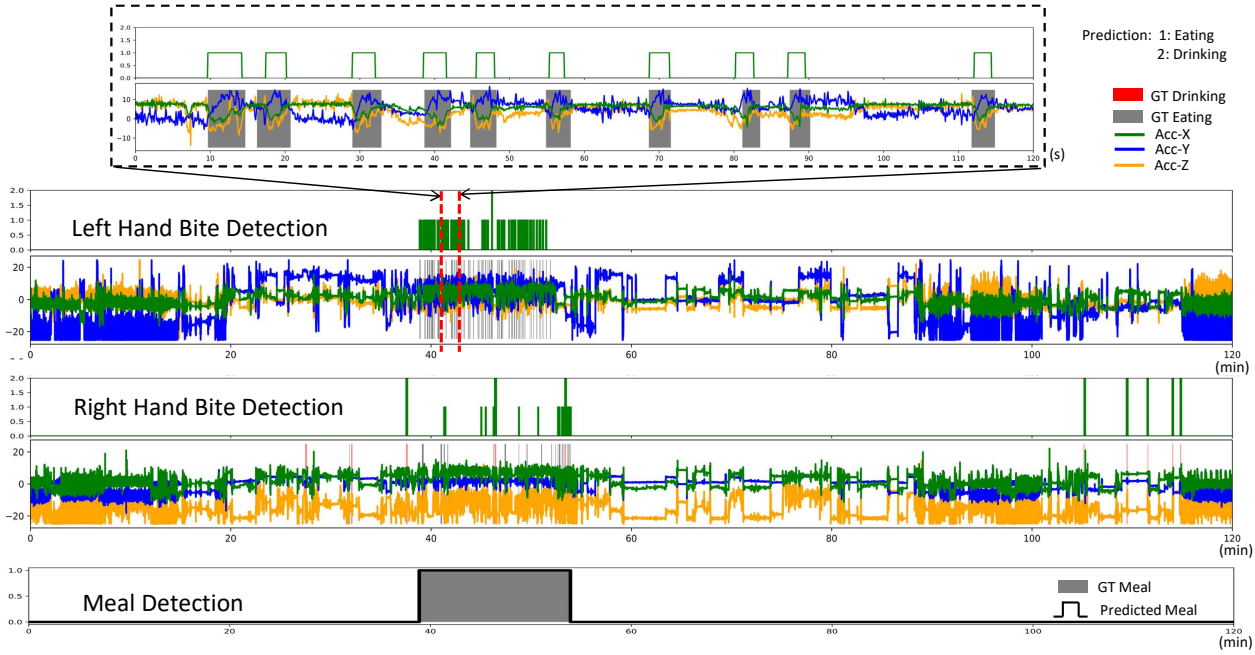

Fig. 10. The 2 h IMU data segment from both hands and the corresponding bite and meal detection examples. Additionally, a 2-min in-meal segment is also selected to show the detailed bite detection.

in Fig. 1). Meanwhile, the two-hand combination method also facilitates the training of multiple datasets, as it allows one-hand datasets and two-hands datasets. Specifically, to handle data from both hands in free-living environments, we applied the hand mirroring + temporal concatenation method to combine IMU data from two hands. This solution also remains flexibility to process data from single hand IMU.

To assess the model's complexity, the number of parameters for each model and their floating point operations (FLOPs) are indicated in Table 6. The number of parameters in ResNet-(Bi)LSTM-based models were significantly larger than CNN-LSTM (0.134 M) and TCN-MHA models (0.203 M). Additionally, a test was carried out to assess the latency for processing 1 min data using the TCN-MHA model. Utilizing a laptop equipped with an Intel Core i7 10750 CPU @2.6 GB, 6 cores (no GPU configuration), the TCN-MHA required 37.61 ms to generate predictions.

The proposed wrist-worn IMU-based approach has the potential to replace the questionnaire-based eating speed surveys in nutrition studies, which offers more accurate quantitative results, thereby advancing the analysis of the correlation between eating speed and obesity-related problems. Additionally, such an eating behavior-related digital biomarker can be applied to individuals interested in documenting their daily dietary habits for long-term dietary

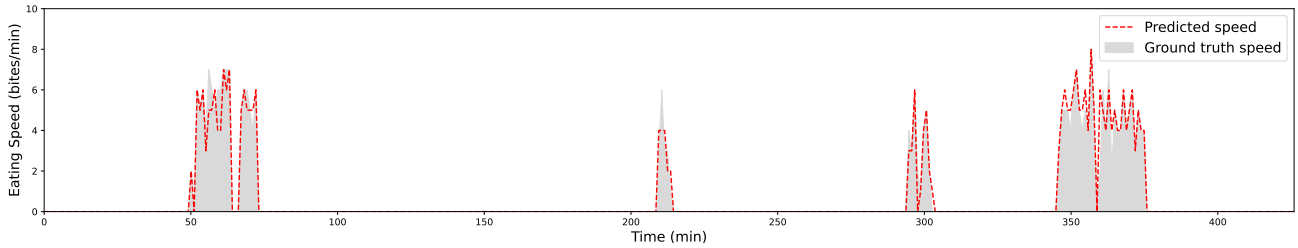

Fig. 11. The minute-level eating speed distribution from one participant in FD-I dataset.

TABLE 6
Model Complexity

Model	#Paras (M)	# FLOPs (G)
CNN-LSTM	0.134	0.129
CNN-BiLSTM	0.241	0.233
ResNet-LSTM	3.078	2.955
ResNet-BiLSTM	3.415	3.280
MS-TCN	0.298	0.284
TCN-MHA	0.203	0.194

TABLE 7
Minute-level Eating Speed

Model	FD-I		FD-II	
	MAPE	PCC	MAPE	PCC
CNN-LSTM	0.256	0.702	0.304	0.650
CNN-BiLSTM	0.221	0.781	0.272	0.758
ResNet-LSTM	0.246	0.747	0.326	0.710
ResNet-BiLSTM	0.219	0.793	0.273	0.797
MS-TCN	0.184	0.805	0.224	0.816
TCN-MHA	**0.181**	**0.840**	**0.212**	**0.834**

health assessment.

This study mainly focus on eating speed measurement for each eating episode, resulting in an averaged speed per episode, which is a well-accepted definition. Additionally, we also explored the feasibility of measuring minute-level eating speed. To achieve this, the number of detected bites in each minute is considered as minute-level eating speed. Fig. 11 shows the minute-level eating speed distribution through one day. Results are shown in Table 7. The TCN-MHA had the MAPE of 0.181 and the PCC of 0.840 on FD-I, and had the MAPE of 0.212 and the PCC of 0.834 on FD-II. The meal-level eating speed represents a holistic view of an eating session, whereas minute-level speed provides insights into eating patterns at a more granular level and is suitable for studying immediate speed changes during a meal. On the other hand, the performance of minute-level speed detection (Table 7) implies that the minute-level eating speed detection is more challenging compared to meal-level speed detection.

This research utilized IMU wristband to measure eating speed in free-living environment. The results were promising, however, it should be noted that the proposed method were based on off-line processing. To maximize its application scenarios, it is worthwhile to exploit the feasibility of implementing this method to smartwatch-smartphone setup for daily eating speed monitoring. Another limitation is that obtaining fine-annotated ground truth data is troublesome. In our approach, research assistants had to follow participants activities to record videos. Existing wearable-based cameras [13] can be used for recording to minimize the effort in future study. Furthermore, to quantify the actual food intake, the wearable IMU system has the potential to be combined with the smart plate [16] and the smart snack box [45] to estimate the calorie intake in real life.

7 Conclusion

In this work, we presented a comprehensive framework for automated measurement of eating speed in free-living environments. To the best of our knowledge, this is the first of its kind. This framework has the potential to extend the application scope of the automated food intake monitoring field. The framework mainly relies on two essential parts: bite detection and eating episode detection in free-living environments. The success of bite detection paves the way for eating episode detection and segmentation, resulting in a good capability for eating speed measurement. The hold-out experiments further underscore its robustness in meal-level eating speed detection.

Acknowledgment

The authors thank the participants involved in the experiments for their dedicated contributions of time and effort.

References

[1] E. Kolay *et al.*, "Self-reported eating speed is associated with indicators of obesity in adults: A systematic review and meta-analysis," *Healthc.*, vol. 9, no. 11, pp. 1–18, 2021.

[2] S. Sasaki, A. Katagiri, T. Tsuji, T. Shimoda, and K. Amano, "Self-reported rate of eating correlates with body mass index in 18-y-old Japanese women," *Int. J. Obes.*, vol. 27, no. 11, pp. 1405–1410, 2003.

[3] A. Kudo *et al.*, "Fast eating is a strong risk factor for new-onset diabetes among the Japanese general population," *Sci. Rep.*, vol. 9, no. 1, pp. 1–8, Dec. 2019.

[4] P. Fagerberg *et al.*, "Fast eating is associated with increased bmi among high-school students," *Nutrients*, vol. 13, no. 3, pp. 1–19, 2021.

[5] E. Woodward, J. Haszard, A. Worsfold, and B. Venn, "Comparison of self-reported speed of eating with an objective measure of eating rate," *Nutrients*, vol. 12, no. 3, pp. 18–24, 2020.

[6] K. Kyritsis, C. Diou, and A. Delopoulos, "A data driven end-to-end approach for in-the-wild monitoring of eating behavior using smartwatches," *IEEE J. Biomed. Heal. Informatics*, vol. 25, no. 1, pp. 22–34, 2020.

[7] J. Qiu, F. P. W. Lo, S. Jiang, Y. Y. Tsai, Y. Sun, and B. Lo, "Counting bites and recognizing consumed food from videos for passive dietary monitoring," *IEEE J. Biomed. Heal. Informatics*, vol. 25, no. 5, pp. 1471–1482, 2021.

[8] K. S. Lee, "Joint audio-ultrasound food recognition for noisy environments," *IEEE J. Biomed. Heal. Informatics*, vol. 24, no. 5, pp. 1477–1489, 2020.

[9] M. Tufano, M. Lasschuijt, A. Chauhan, E. J. M. Feskens, and G. Camps, "Capturing eating behavior from video analysis: A systematic review," *Nutrients*, vol. 14, no. 22. pp. 1–14, 2022.

[10] S. Sharma and A. Hoover, "Top-Down detection of eating episodes by analyzing large windows of wrist motion using a convolutional neural network," *Bioengineering*, vol. 9, no. 2, pp. 20–23, 2022.

[11] A. Doulah, T. Ghosh, D. Hossain, M. H. Imtiaz, and E. Sazonov, "'Automatic ingestion monitor version 2' - A novel wearable device for automatic food intake detection and passive capture of food images," *IEEE J. Biomed. Heal. Informatics*, vol. 25, no. 2, pp. 568–576, 2021.

[12] C. Wang, T. S. Kumar, W. De Raedt, G. Camps, H. Hallez, and B. Vanrumste, "Eat-Radar: Continuous Fine-Grained Intake Gesture Detection Using FMCW Radar and 3D Temporal Convolutional Network with Attention," *IEEE J. Biomed. Heal. Informatics*, doi: 10.1109/JBHI.2023.3339703.

[13] N. Alshurafa, S. Zhang, C. Romano, H. Zhang, A. F. Pfammatter, and A. W. Lin, "Association of number of bites and eating speed with energy intake: Wearable technology results under free-living conditions," *Appetite*, vol. 167, no. September 2020, p. 105653, 2021.

[14] P. V. Rouast and M. T. P. Adam, "Learning deep representations for video-based intake gesture detection," *IEEE J. Biomed. Heal. Informatics*, vol. 24, no. 6, pp. 1727–1737, 2020.

[15] C. Wang, T. S. Kumar, G. Markvoort, H. Hallez, and B. Vanrumste, "Eating activity monitoring in home environments using smartphone-based video recordings," in *2022 International Conference on Digital Image Computing: Techniques and Applications (DICTA)*, 2022, pp. 1–5.

[16] G. Mertes, L. Ding, W. Chen, H. Hallez, J. Jia, and B. Vanrumste, "Measuring and localizing individual bites using a sensor augmented plate during unrestricted eating for the aging population," *IEEE J. Biomed. Heal. Informatics*, vol. 24, no. 5, pp. 1509–1518, 2020.

[17] V. Papapanagiotou, C. Diou, L. Zhou, J. Van Den Boer, M. Mars, and A. Delopoulos, "A novel chewing detection system based on PPG, audio, and accelerometry," *IEEE J. Biomed. Heal. Informatics*, vol. 21, no. 3, pp. 607–618, 2017.

[18] R. Zhang, S. Bernhart, and O. Amft, "Diet eyeglasses: Recognising food chewing using EMG and smart eyeglasses," in *BSN 2016 - 13th Annual Body Sensor Networks Conference*, 2016, pp. 7–12.

[19] 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.

[20] Y. Shen, J. Salley, E. Muth, and A. Hoover, "Assessing the accuracy of a wrist motion tracking method for counting bites across demographic and food variables," *IEEE J. Biomed. Heal. Informatics*, vol. 21, no. 3, pp. 599–606, 2017.

[21] P. V. Rouast and M. T. P. Adam, "Single-stage intake gesture detection using CTC loss and extended prefix beam search," *IEEE J. Biomed. Heal. Informatics*, vol. 25, no. 7, pp. 2733–2743, 2021.

[22] B. Wei, S. Zhang, X. Diao, Q. Xu, Y. Gao, and N. Alshurafa, "An end-to-end energy-efficient approach for intake detection with low inference time using wrist-worn sensor," *IEEE J. Biomed. Heal. Informatics*, vol. 27, no. 8, pp. 3878–3888, 2023.

[23] A. Bedri, D. Li, R. Khurana, K. Bhuwalka, and M. Goel, "FitByte: Automatic diet monitoring in unconstrained situations using multimodal sensing on eyeglasses," in the *CHI Conference on Human Factors in Computing Systems*, 2020, pp. 1–12.

[24] S. Zhang *et al.*, "NeckSense: A multi-sensor necklace for detecting eating activities in free-living conditions," *Proc. ACM Interactive, Mobile, Wearable Ubiquitous Technol.*, vol. 4, no. 2, 2020, pp. 1–26.

[25] M. Ester, H. P. Kriegel, J. Sander, and X. Xu, "A density-based algorithm for discovering clusters in large spatial databases with noise," in *Proc. 2nd Int. Conf. Knowl. Discovery Data Mining*, Portland, OR, USA, Aug. 1996, pp. 226–231.

[26] C. Lea, M. D. Flynn, R. Vidal, A. Reiter, and G. D. Hager, "Temporal convolutional networks for action segmentation and detection," In *Proc. 30th IEEE Conf. Comput. Vis. Pattern Recognition (CVPR)*, 2017, pp. 1003–1012.

[27] Y. A. Farha and J. Gall, "MS-TCN: Multi-stage temporal convolutional network for action segmentation," In *Proc. 32th IEEE Conf. Comput. Vis. Pattern Recognition (CVPR)*, 2019, pp. 3570–3579.

[28] B. Filtjens, B. Vanrumste, and P. Slaets, "Skeleton-based action segmentation with convolutional neural networks," *IEEE Trans. Emerg. Top. Comput.*, vol. PP, pp. 1–11, 2022.

[29] A. Vaswani *et al.*, "Attention is all you need," in *Proc. Neural Inf. Process. Syst. (NIPS)*, Long Beach, CA, USA, vol. 30, 2017, pp. 1–11.

[30] Y. Luo, J. Li, K. He, and W. Cheuk, "A hierarchical attention-based method for sleep staging using movement and cardiopulmonary signals," *IEEE J. Biomed. Heal. Informatics*, vol. 27, no. 3, pp. 1354–1363, 2022.

[31] S. P. Singh, M. K. Sharma, A. Lay-Ekuakille, D. Gangwar, and S. Gupta, "Deep convLSTM with self-attention for human activity decoding using wearable sensors," *IEEE Sens. J.*, vol. 21, no. 6, pp. 8575–8582, 2021.

[32] C. Wang, T. S. Kumar, W. De Raedt, G. Camps, H. Hallez, and B. Vanrumste, "Drinking gesture detection using wrist-worn IMU sensors with multi-stage temporal convolutional network in free-living environments," *Proc. Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. (EMBC)*, 2022, pp. 1778–1782.

[33] P. V. Rouast, H. Heydarian, M. T. P. Adam, and M. E. Rollo, "OREBA: A dataset for objectively recognizing eating behavior and associated intake," *IEEE Access*, vol. 8, pp. 181955–181963, 2020.

[34] H. Sloetjes and P. Wittenburg, "Annotation by category - ELAN and ISO DCR," In *Proc. 6th Int. Conf. Lang. Resour. Eval. Lr.*, 2008, pp. 816–820.

[35] C. Wang *et al.*, "Intake Gesture Detection With IMU Sensor in Free-Living Environments: The Effects of Measuring Two-Hand Intake and Down-Sampling," in *2023 IEEE 19th International Conference on Body Sensor Networks (BSN)*, 2023, pp. 1–4.

[36] A. Bedri *et al.*, "EarBit: Using wearable sensors to detect eating episodes in unconstrained environments," *Proc. ACM Interactive, Mobile, Wearable Ubiquitous Technol.*, vol. 1, no. 3, 2017, pp. 1–20.

[37] J. Cohen, "A coefficient of agreement for nominal scales," *Educ. Psychol. Meas.*, vol. 20, no. 1, pp. 37–46, 1960.

[38] 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," *Proc. ACM Interactive, Mobile, Wearable Ubiquitous Technol.*, vol. 1, no. 3, 2017, pp. 1–20.

[39] Y. Dong, J. Scisco, M. Wilson, E. Muth, and A. Hoover, "Detecting periods of eating during free-living by tracking wrist motion," *IEEE J. Biomed. Heal. Informatics*, vol. 18, no. 4, pp. 1253–1260, 2014.

[40] J. M. Fontana, M. Farooq, and E. Sazonov, "Automatic ingestion monitor: A novel wearable device for monitoring of ingestive behavior," *IEEE Trans. Biomed. Eng.*, vol. 61, no. 6, pp. 1772–1779, 2014.

[41] E. Thomaz, I. Essa, and G. D. Abowd, "A practical approach for recognizing eating moments with wrist-mounted inertial sensing," *UbiComp 2015 - Proc. 2015 ACM Int. Jt. Conf. Pervasive Ubiquitous Comput.*, 2015, pp. 1029–1040.

[42] S. Sen, V. Subbaraju, A. Misra, R. Balan, and Y. Lee, "Annapurna: Building a real-world smartwatch-based automated food journal," *19th IEEE Int. Symp. a World Wireless, Mob. Multimed. Networks, WoWMoM 2018*, 2018, pp. 1–6.

[43] G. Schiboni and O. Amft, "Sparse natural gesture spotting in free living to monitor drinking with wrist-worn inertial sensors," *Proc. Int. Symp. Wearable Comput. ISWC*, 2018, pp. 140–147.

[44] S. Sharma, P. Jasper, E. Muth, and A. Hoover, "The impact of walking and resting on wrist motion for automated detection of meals," *ACM Trans. Comput. Healthc.*, vol. 1, no. 4, 2020, pp. 1-19.

[45] F. J. de Gooijer, A. van Kraaij, J. Fabius, S. Hermsen, E. J. M. Feskens, and G. Camps, "Assessing snacking and drinking behavior in real-life settings: Validation of the SnackBox technology," *Food Qual. Prefer.*, vol. 112, no. May, p. 105002, 2023.