Intake Gesture Detection With IMU Sensor in Free-Living Environments: The Effects of Measuring Two-Hand Intake and Down-Sampling

2023 IEEE 19th International Conference on Body Sensor Networks (BSN)

Wang, Chunzhuo; Kong, Jiaze; Cai, Yutong; Kumar, T.S.; De Raedt, Walter et al

https://doi.org/10.1109/BSN58485.2023.10331032

This publication is made publicly available in the institutional repository of Wageningen University and Research, under the terms of article 25fa of the Dutch Copyright Act, also known as the Amendment Taverne.

Article 25fa states that the author of a short scientific work funded either wholly or partially by Dutch public funds is entitled to make that work publicly available for no consideration following a reasonable period of time after the work was first published, provided that clear reference is made to the source of the first publication of the work.

This publication is distributed using the principles as determined in the Association of Universities in the Netherlands (VSNU) 'Article 25fa implementation' project. According to these principles research outputs of researchers employed by Dutch Universities that comply with the legal requirements of Article 25fa of the Dutch Copyright Act are distributed online and free of cost or other barriers in institutional repositories. Research outputs are distributed six months after their first online publication in the original published version and with proper attribution to the source of the original publication.

You are permitted to download and use the publication for personal purposes. All rights remain with the author(s) and / or copyright owner(s) of this work. Any use of the publication or parts of it other than authorised under article 25fa of the Dutch Copyright act is prohibited. Wageningen University & Research and the author(s) of this publication shall not be held responsible or liable for any damages resulting from your (re)use of this publication.

For questions regarding the public availability of this publication please contact openaccess.library@wur.nl

2023 IEEE 19th International Conference on Body Sensor Networks (BSN) | 979-8-3503-3841-6/23/$31.00 ©2023 IEEE | DOI: 10.1109/BSN58485.2023.10331032

Intake Gesture Detection With IMU Sensor in Free-Living Environments: The Effects of Measuring Two-Hand Intake and Down-Sampling

Chunzhuo Wang[1,3], Jiaze Kong[1], Yutong Cai[1], T. Sunil Kumar[2], Walter De Raedt[3], Guido Camps[4], Hans Hallez[5], Bart Vanrumste[1]

[1] The e-Media Research Lab, and the ESAT-STADIUS Division, KU Leuven, 3000 Leuven, Belgium

[2] Vellore Institute of Technology, Chennai, India

[3] The Life Science Department, IMEC, 3001 Heverlee, Belgium

[4] The Department of Agrotechnology and Food Sciences, Wageningen University and Research, and the OnePlanet Research Center, Wageningen, Netherlands

[5] The M-Group, DistriNet, Department of Computer Science, KU Leuven, 8200 Sint-Michiels, Belgium

chunzhuo.wang@kuleuven.be

***Abstract*—Food intake monitoring plays an important role in personal dietary systems. Numerous approaches have been proposed to automatically detect eating gestures using various sensors and machine learning. However, existing eating gesture detection approaches mainly focus on meal sessions. Such a task is still challenging in free-living environments due to longer monitoring duration and more non-feeding activities. This paper proposes a wearable Inertial Measurement Unit (IMU) based method to detect eating and drinking gestures in free-living environments. Two important factors that impede intake gesture detection in free-living environments are addressed: 1) how to handle IMU data from two hands, and 2) what is the impact of downsampling sensor data on performance. To integrate two-hand data, we propose a solution that combines hand mirroring and temporal concatenation techniques. The multi-stage temporal convolutional network (MS-TCN) is applied to effectively recognise intake gestures. A dataset contains 12 subjects with 67.5 h data is collected for validation. Moreover, IMU data with different sampling frequencies are processed to test performance. Validated by Leave-One-Subject-Out (LOSO) method, our approach (with 16 Hz sampling frequency) achieves a segmental F1-score of 0.826 and 0.893 for recognizing eating and drinking gestures, respectively. Results show that the proposed solution outperforms existing two-hand data combination approaches. Moreover, in our case, a higher sampling frequency does not always mean better performance.**

***Index Terms*—eating gesture detection, free-living environments, food intake monitoring, hand mirroring, down-sampling**

I. Introduction

Automatic food intake monitoring systems have drawn plenty of attention due to their potential applications in nutrition studies and precision healthcare. The current methods used in clinical settings, such as 24-hour recall and self-report questionnaires, are labor intensive, prone to error, and not feasible for long-term monitoring [1]. The automatic approach can detect eating/drinking gestures automatically and objectively. Numerous approaches have been investigated for this purpose [2]–[5]. One of the popular approach is using wearable Inertial Measurement Unit (IMU) sensor for eating gesture detection [6], [7]. While the feasibility of the IMU-based approach has been demonstrated in meal sessions, eating gesture detection in free-living environments remains an open question. In free-living environments, eating/drinking are sporadic and sparse activities in full-day data (the duration ratio is less than 1/20) [8]. Furthermore, the eating gesture can be more versatile as the participant can use different utensils throughout the day (use hand to eat snack, use fork&knife to eat meal), which adds complexity for eating gesture detection.

Existing research on full-day scenario focus more on eating episodes detection [6], [9], rather than fine-grained gesture-level detection. Few researchers explored intake gesture detection in free-living environments. One relevant study conducted by Kyritsis *et al*. [6] attempted to use predicted eating event sequence throughout the day to estimate meal session in FIC-free dataset, however, the accuracy of eating gesture detection in free-living environment was not discussed.

To leverage the full potential of intake gesture detection in free-living environments, two important factors are investigated. Firstly, existing approaches in meal sessions focus on intake gesture detection from dominant hand [6], [7], whereas, in full-day scenarios, it becomes necessary to detect intake gestures from both hands, rather than solely dominant hand. Secondly, as the monitoring duration in a full day is significantly longer than a typical meal session (300 min vs 20 min), an appropriate sampling frequency is essential to balance factors such as power consumption, data storage, computation cost, and data quality. For example, an IMU sensor with 128 Hz needs 700 MB space for 6 h monitoring.

In this research, we propose a method for detecting eating /drinking gesture in free-living environments. An IMU-based dataset consisting 12 full-day recordings collected from 12 subjects with a total duration of 67.5 h has been collected.

Authorized licensed use limited to: Wageningen UR. Downloaded on March 18,2024 at 11:55:33 UTC from IEEE Xplore. Restrictions apply.

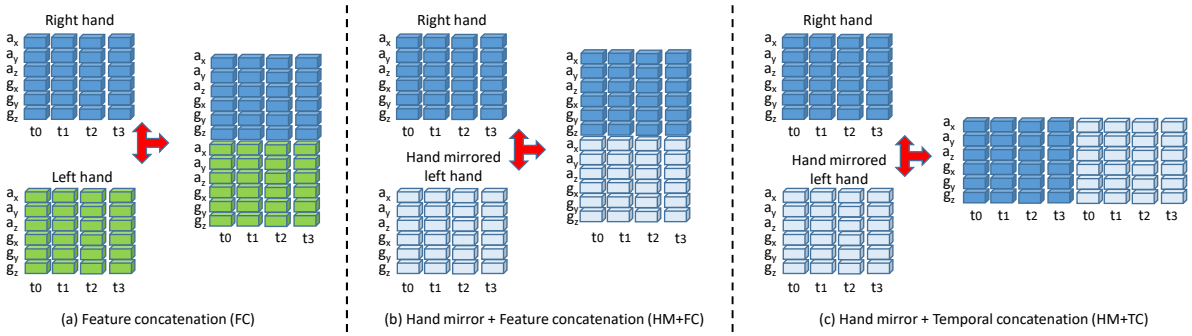

Fig. 1: Three different solutions for two-hand IMU data combination. Method in (a) uses feature concatenation directly. Method in (b) from [12] firstly hand mirrors left hand data, then applies feature concatenation. The proposed method (c) firstly uses hand mirror, then applies temporal concatenation.

In addition, the two-hand intake and sampling frequency selection problem are also addressed in the paper.

II. Methods

A. Data Collection and Annotation

The Shimmer3 IMU wristbands were used to collect data, with the sensors' sampling frequency set at 64 Hz. 12 participants were recruited in this study. The ethic committee of KU Leuven has approved this experiment (G-2021-4025). The signed inform consent is collected from all participants. Participants were students from KU Leuven University who live in Leuven City. On the experiment day, two IMU wristbands were mounted on participant's wrists. One research assistant accompanied the participant and used a camera to record eating and drinking activities. Participants were required to consume 2 meals (lunch and dinner) and at least one snack session. The meal venue can be participant's home, restaurant, learning center, or university cafeteria, based on their own preference. They did their daily activities freely outside meal sessions. In total, 67.5 h of data were collected.

The ELAN tool [10] is used to annotate data. The eating/drinking gesture is specifically defined as the action of raising either the left or right hand to the mouth with the fork/knife/spoon/chopsticks/water container until putting away the hand from the mouth. Importantly, the annotation also includes hand information (from left hand or right hand). The detailed data statistics is shown in Table I.

TABLE I: Intake data statistics

Parameter	Values
# Participants	12
# Eating gesture	2434 (L: 814, R: 1620)
# Drinking gesture	443 (L: 206, R: 237)
Duration ratio of other : eating : drinking	88.54:3.14:1
Total duration	4050 min

B. Two-hand Combination

In free-living environments, it is common to use both hands for eating and drinking activities. In our previous research [11], we proposed hand mirroring (HM) and temporal concatenation (TC) combined method (HM+TC) as a potential solution for two-hand combination. However, it is important to further evaluate its effectiveness and compare it with other commonly used solutions to determine the optimal one. Apart from our approach, another direct solution is feature concatenation (FC). Besides, the hand mirroring and feature concatenation method (HM+FC) is used in [12]. Different solutions are shown in Fig. 1. The hand mirroring, temporal concatenation, and feature concatenation techniques are explained below.

1) Hand Mirroring (HM): To achieve data uniformity, the hand mirroring method [6] is employed to transform left hand's IMU data to right hand. This has been used in several IMU-based eating gesture detection when the participant is left hand dominant. The hand mirroring involves flipping the direction of $\mathbf{a}_x$ (in accelerometer), $\mathbf{g}_y$ and $\mathbf{g}_z$ (in gyroscope).

2) Temporal Concatenation (TC): Temporal concatenation is to combine the data from two hands in the time dimension. This method remains the same number of features, but the data length is doubled.

3) Feature Concatenation (FC): Feature concatenation is to merge the two hands' data in the feature dimension. The data length remains same, the number of features is doubled.

C. Down Sampling

One of the objectives of this research is to examine the effect of different sampling frequencies on the performance of the proposed eating gesture recognition approach. To achieve this, the raw data are first filtered by an IIR lowpass filter and then downsampled to 32 Hz, 16 Hz, 8 Hz, and 4 Hz.

D. Deep Learning Models

The Multi-Stage Temporal Convolutional Network (MS-TCN) model used from [11] is adopted in this research for eating and drinking gesture detection due to its superior outcomes compared to CNN-LSTM and CNN-BiLSTM. The MS-TCN model is obtained by stacking multiple single-stage temporal convolutional network (TCN) [13]. Utilizing dilated convolution enhances the ability of TCN to effectively capture

Authorized licensed use limited to: Wageningen UR. Downloaded on March 18,2024 at 11:55:33 UTC from IEEE Xplore. Restrictions apply.

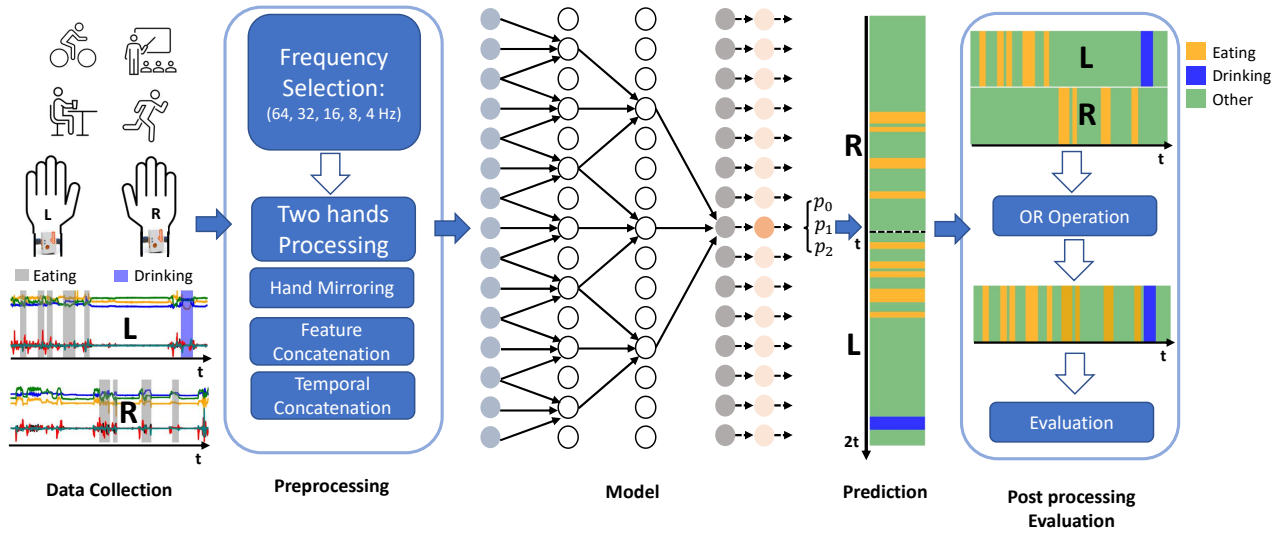

Fig. 2: The experiment pipeline from IMU sensing to prediction. In the preprocessing, the IMU data from two hands are downsampled and combined using different solutions. For the model, dilated convolution layers are utilized to increase the receptive field. A softmax activation is used after the last dilated layer to give predictions based on the feature extracted from preceding layers. It should be noted that the output of MS-TCN is point-wise; we only explicitly indicate the 3 class prediction (p_0, p_1, p_2) for one point. The post-processing step is only applied when temporal concatenation is used.

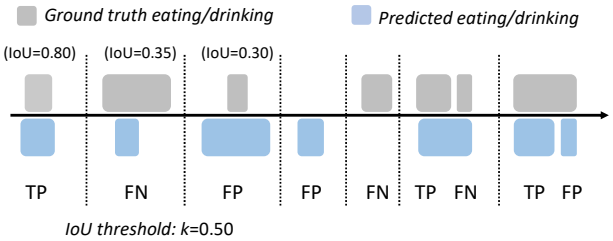

Fig. 3: Segmental evaluation examples for TPs, FPs, and FNs

long-term temporal dependencies by extending the effective receptive field. Fig. 2 illustrates the pipeline of the proposed approach. We maintain the identical hyperparameters of the architecture employed in [11], except the output dimension (from binary classification to 3-class classification). The MS-TCN contains 2 stages, and each stage comprises 11 layers, where each layer is composed of of 64 Conv1D with a kernel size of 3. The first stage is used to give initial prediction, while the second stage is used to refine the initial prediction. The network's output entails point-wise 3-class predictions. During the training phase, the input data's temporal length is set to 60 s. We use an Adam optimizer with a learning rate of 5e-4 for training. The batch size is 64, and the epoch is set as 100.

All training tasks were finished on an Intel 9-core Xeon Gold 6140 CPUs@2.3 GHz (Skylake) with 5 GB RAM per core, and one piece of NVIDIA P100-SXM2-16GB GPU. These computing resources were provided by the Vlaams Supercomputer Centrum (VSC) [1].

[1]See https://www.vscentrum.be/

E. Post Processing

For experiment that applies temporal concatenation, post processing is necessary before evaluation. The data length of output is doubled compared to the original length in HM+TC solution. Therefore the prediction needs to be divided into two sequences (left and right). The OR operation is applied between left and right prediction afterwards, as participants can hold two hands to hold the cup to drink.

F. Evaluation Scheme

The output of deep learning model is point-wise prediction since the model is seq2seq architecture. Considering our objective is eating/drinking gesture detection, a direct point-wise evaluation can not reflect to gesture-level performance. In this research, we use the segment-wise evaluation method from [11], as shown in Fig. 3. The evaluation process begins by computing the intersection over union (IoU) between each ground truth and predicted intake gesture. Based on a pre-selected IoU threshold, denoted as k, the segmental True Positives (TP), False Negatives (FN) and False Positives (FP) are determined after the comparison between IoU score and k (In our case, k=0.5). Afterwards, we calculated the segment-wise F1-score to evaluate the performance using determined TP, FN, and FP. In case of the 3-class evaluation, when the target evaluation class is eating gesture, the drinking gesture is included in the other class, and vice versa.

III. Results and Discussion

The Leave-One-Subject-Out (LOSO) method was used for validation. In addition, two common used model: CNN-LSTM and CNN-BiLSTM models were used for comparison.

Authorized licensed use limited to: Wageningen UR. Downloaded on March 18,2024 at 11:55:33 UTC from IEEE Xplore. Restrictions apply.

TABLE II: F1-score for different sampling frequency (k=0.5)

Class	Model	Frequency				
		64Hz	32Hz	16Hz	8Hz	4Hz
Eating	MS-TCN	0.811	0.817	0.826	0.801	0.745
	CNN-BiLSTM	0.769	0.776	0.784	0.783	0.743
	CNN-LSTM	0.687	0.709	0.723	0.727	0.721
Drinking	MS-TCN	0.876	0.862	0.893	0.860	0.807
	CNN-BiLSTM	0.833	0.868	0.866	0.854	0.792
	CNN-LSTM	0.720	0.743	0.792	0.769	0.786

TABLE III: F1-score for various two hand solutions (k=0.5)

Class	Model	Two hands solution		
		FC	HM+FC	HM+TC
Eating	MS-TCN	0.776	0.762	0.826
	CNN-BiLSTM	0.774	0.771	0.784
	CNN-LSTM	0.689	0.696	0.723
Drinking	MS-TCN	0.801	0.818	0.893
	CNN-BiLSTM	0.766	0.804	0.866
	CNN-LSTM	0.687	0.695	0.792

1) Downsampling Experiments: Five sampling frequencies (64,32,16,8,4 Hz) were selected for comparison. The HM+FC solution was applied for combining the data from two hands. The same temporal lengths of input data were maintained for each model to make sure that input data contains same time-series movement information. Table II shows the performance. The data with 16 Hz obtains the highest performance in MS-TCN and CNN-BiLSTM model in eating, and 8 Hz in CNN-LSTM. Significantly reduced performance was observed when the frequency dropped below 8 Hz.

2) Two-hand Combination Experiments: Three different two-hand combination solutions (see Fig. 1) were compared. The sample frequency was 16 Hz in this experiment. Table III shows the results. The procedure that hand mirroring left hand data and then temporal concatenation obtained the highest performance. Feature Concatenation methods obtained the lowest performance.

The results from Table III indicate that the proposed method for two-hand IMU combination achieves the highest performance. Meanwhile, compared to FC and HM+FC, the proposed solution can also be used to predict data on single IMU case, providing more flexibility. Wearing two IMU wristbands in daily life can be a limiting factor, as people used to wear smartwatch/fitness tracker on one hand only. However, two-hand approach can be used in hospital or healthcare center. Furthermore, the prerequisite of the hand mirroring and temporal concatenation is that the used hand information (left/right) should be clear before processing. Surprisingly, data with the highest sampling frequency (64 Hz) did not result in the best performance across all models. One potential reason is that high-frequency data contains redundant information or more noise. The choice of the sampling frequency is a trade-off between the model's performance and computation efficiency.

IV. Conclusion

In this paper, we investigated fine-grained eating/drinking gesture detection in free-living environments using two IMU wristbands. The two-hand combination and sampling frequency factors were addressed. Various two-hand solutions and sampling frequencies were tested on a dataset collected from 12 subjects in free-living environments. The method that combines hand mirroring and temporal concatenation was validated as a feasible solution for the two-hand problem. Sampling frequency with 16 Hz is considered as an optimal trading-off between model performance and computation cost in our case. For future work, larger data will be collected to further validate the intake gesture detection in free-living environments. The detected eating gesture in full-day period will also be explored to localize meal sessions.

Acknowledgment

The authors express their gratitude to the participants who dedicated their efforts and time to participate in the experiments. This project is funded by the China Scholarship Council (CSC), China (Grant number: 202007650018).

References

[1] D. A. Schoeller, "How accurate is self-reported dietary energy intake?" *Nutrition Rev.*, vol. 48, no. 10, pp. 373–379, Oct. 1990.

[2] S. He *et al.*, "A comprehensive review of the use of sensors for food intake detection," *Sensors and Actuators A: Physical*, vol. 315. Elsevier B.V., p. 112318, Nov. 01, 2020

[3] P. A. Neves *et al.*, "Thought on Food: A systematic review of current approaches and challenges for food intake detection," *Sensors*, vol. 22, no. 17, pp. 1–21, 2022.

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

[5] C. Wang, T. S. Kumar, W. De Raedt, G. Camps, H. Hallez, and B. Vanrumste, "Eat-Radar: Continuous fine-grained eating gesture detection using FMCW radar and 3D temporal convolutional network," pp. 1–14, 2022, [Online]. Available: http://arxiv.org/abs/2211.04253.

[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] 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.

[8] 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.

[9] 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

[10] 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.

[11] 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.

[12] H. Heydarian, P. V. Rouast, M. T. P. Adam, T. Burrows, C. E. Collins, and M. E. Rollo, "Deep learning for intake gesture detection from wrist-worn inertial sensors: The effects of data preprocessing, sensor modalities, and sensor positions," *IEEE Access*, vol. 8, pp. 164936–164949, 2020

[13] 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.

Authorized licensed use limited to: Wageningen UR. Downloaded on March 18,2024 at 11:55:33 UTC from IEEE Xplore. Restrictions apply.