2022 44th Annual International Conference of
the IEEE Engineering in Medicine & Biology Society (EMBC)
Scottish Event Campus, Glasgow, UK, July 11-15, 2022

2022 44th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC) | 978-1-7281-2782-8/22/$31.00 ©2022 IEEE | DOI: 10.1109/EMBC48229.2022.9871817

Drinking Gesture Detection Using Wrist-Worn IMU Sensors with Multi-Stage Temporal Convolutional Network in Free-Living Environments

Chunzhuo Wang[1,2], T. Sunil Kumar[1], Walter De Raedt[2], Guido Camps[3], Hans Hallez[4], and Bart Vanrumste[1]

Abstract— **Maintaining adequate hydration is important for health. Inadequate liquid intake can cause dehydration problems. Despite the increasing development of liquid intake monitoring, there are still open challenges in drinking detection under free-living conditions. This paper proposes an automatic liquid intake monitoring system comprised of wrist-worn Inertial Measurement Units (IMUs) to recognize drinking gesture in free-living environments. We build an end-to-end approach for drinking gesture detection by employing a novel multi-stage temporal convolutional network (MS-TCN). Two datasets are collected in this research, one contains 8.9 hours data from 13 participants in semi-controlled environments, the other one contains 45.2 hours data from 7 participants in free-living environments. The Leave-One-Subject-Out (LOSO) evaluation shows that this method achieves a segmental F1-score of 0.943 and 0.900 in the semi-controlled and free-living datasets, respectively. The results also indicate that our approach outperforms the convolutional neural network and long-short-term-memory network combined model (CNN-LSTM) on our datasets. The dataset used in this paper is available at https://github.com/Pituohai/drinking-gesture-dataset/.**

Clinical relevance— **This automatic liquid intake monitoring system can detect drinking gesture in daily life. It has the potential to be used to record the frequency of drinking water for at-risk elderly or patients in the hospital.**

I. INTRODUCTION

Water balance is essential for health and life, as the principal constituent of the human body is water [1], [2]. However, drinking water is frequently overlooked due to the fast pace of work in daily life. Inadequate water intake is one of the common causes of dehydration which is associated with multiple acute and chronic diseases [3], [4]. Older people have a higher risk of dehydration because of the diminution of the sense of thirst, the reduction of water proportions in the body, and the decrease of mobility [5].

To prevent dehydration, water intake monitoring is critical. Water intake records mainly rely on self-report logs clinically. This approach is time-consuming and is prone to making mistakes. An automatic liquid intake monitoring system consisting of sensors and machine learning techniques can address this issue. The systems can be broadly categorized into two groups: ambient-based and wearable-based systems.

The ambient-based systems utilize cameras fixed in environments (vision-based) or sensors embedded in containers (container-based) [6]–[8]. The wearable-based implementations detect water intake by acoustic sensors or IMUs [9]–[12]. Gomes and Sousa [11] developed an approach to detect hand to mouth (HtM) movement using a wrist-worn IMU and a random forest (RF) classifier. This approach detected HtM with an F1-score of 85% in free-living environments (Eating activities were excluded in their dataset). Senyurek et al. [13] proposed a convolutional neural network and long-short-term-memory network combined model (CNN-LSTM) for detecting drinking gesture using IMU data acquired from smartwatches. This approach achieved an F1-score of 87% on a publicly available dataset collected from 11 participants in the Leave-One-Subject-Out (LOSO) scheme. Their data were collected in group conversation scenarios.

The performance of methods using wrist-worn IMUs is promising in constrained or semi-constrained environments, e.g., laboratory conditions, limited duration, or limited type of daily activity. A more challenging problem arises when dealing with practical scenarios. In free-living environments, drinking is a sparse activity that distributes all over the day (the duration ratio is less than 1/100) [14]. The drinking gesture detection in a free-living environment, which is characterized by a longer period and more complex null class, is still an open question. The drinking gesture is defined as a movement from raising the container to the mouth until putting away the container and the null class embodies all the other non-drinking activities in free-living environments.

This paper aims to detect drinking gesture in free-living environments by using wrist-mounted IMU sensors. To this end, two datasets are collected: one contains 8.9 hours data from 13 participants in semi-controlled environments, the other one contains 45.2 hours data from 7 participants in the free-living environments. We propose an end to end approach by applying a novel multi-stage temporal convo-

*This work was a part of the project funded by the China Scholarship Council(CSC), China (Grant number: 202007650018).

[1]Chunzhuo Wang, T. Sunil Kumar, and Bart Vanrumste are with the e-Media Research Lab, and also with the ESAT-STADIUS Division, KU Leuven, 3000 Leuven, Belgium chunzhuo.wang@kuleuven.be; sunilkumar.telagamsetti@kuleuven.be; bart.vanrumste@kuleuven.be

[2]Walter De Raedt and Chunzhuo Wang are with the Life Science Department, IMEC, 3001 Heverlee, Belgium Walter.DeRaedt@imec.be

[3]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 guido.camps@wur.nl

[4]Hans Hallez is with the M-Group, DistriNet, Department of Computer Science, KU Leuven, 8200 Sint-Michiels, Belgium hans.hallez@kuleuven.be

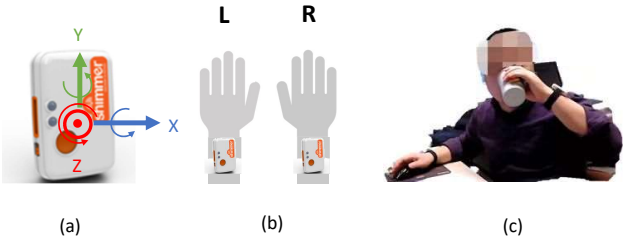

Fig. 1. Example of IMU sensor and experiment scene. Panel (a) indicates the coordinate frame of IMU, panel (b) presents the location of the IMU on hands, and panel (c) shows the scene of the experiment.

lutional network (MS-TCN) to detect the drinking gesture from IMU signals. Additionally, this method can segment the time interval of each drinking gesture.

II. Methods

A. Data Collection

The off-the-shelf IMU sensor Shimmer3[1] is used in this study. The Shimmer3 IMU contains a 3-axis accelerometer unit, a 3-axis gyroscope unit, and a 3-axis magnetometer unit. The signals from the accelerometer and gyroscope were used in this study, hence 6 Degrees-of-Freedom (6 DoF). Two IMUs were attached to the left and right wrists of the participant, as presented in Fig. 1. The sampling frequency is 64 Hz. During the experiment, the data were stored in the SD card embedded in the sensor. After the experiment, the data were transferred to a laptop via Shimmer docker and the software Consensys[2]. The ethics committee of KU Leuven has approved this research (Reference number: G-2021-4025-R4), and the written informed consent from each participant was collected. Two datasets were collected in the experiment:

1) DX-I: The first dataset was collected in semi-controlled environments from 13 participants. The total duration of this dataset is 8.9 hours. The average duration of each participant is 35.6±28.6 min. In each drinking session, Water/tea/cola was provided to participants. A camera was used to record the entire drinking session as the ground truth annotation. They were asked to drink while sitting on chairs and a sofa, and also drink while standing. Both hands could be used to take the cups and drink, no matter whether the participant is a left-hander or a right-hander. The data were collected in the real work environment or home environment. They can work on a laptop, talk, walk, eat chips, watch TV or smoke. To collect more drinking gestures in this session, they were required to drink with a higher frequency. This dataset contains 410 drinking gestures (left hand : right hand : two hands = 101:266:43).

2) DX-II: The second dataset was collected in free-living environments. The data were taken at the locations that are preferred by the participants, including work place and home. Seven participants took part in this experiment. Three out of the seven in DX-II also participated in DX-I. Each participant joined the experiment for 6.5±2.0 consecutive hours (from morning to afternoon or evening). The total duration of this dataset is 45.2 hours. The participants can drink water at their own pace. Each session contains daily life-related activities including, but not limited to, drinking water, eating snacks, eating lunch, working with laptops, watching smartphones and walking. A camera was placed on the desk in the office or at home to capture the drinking gesture. In total, there are 304 drinking gestures in DX-II dataset (left hand : right hand : two hands = 142:152:10).

[1]https://shimmersensing.com/product/shimmer3-imu-unit/
[2]https://shimmersensing.com/product/consensyspro-software/

B. Data Preprocessing

1) Downsampling and annotation: The sampling frequency of the data is 64Hz, which results in high redundancy for signal processing and requires a high computation cost; hence, the data were down-sampled to 16Hz. The data from each hand is annotated into two classes: Drinking (labelled as 1) and Null (labelled as 0). The movement from raising the left/right hand to the mouth with a container until putting away the container from the mouth is considered as a drinking gesture. The Null class contains all the other daily activities during the experiment. ELAN [15] was used to annotate the IMU signal. The total time spent on non-drinking activities was much longer than the time consumed on the drinking activity, which leads to the data being unbalanced, especially for dataset DX-II. The duration ratio of the annotated drinking gestures to null class is 1/27 and 1/196 in DX-I and DX-II, respectively. Fig. 2 shows the example of the annotated signal.

2) Hand mirroring: It is common for people to drink water using either of their hands to hold the container. That is the reason we employ IMUs on left and right wrists. We considered the participant's right hand as the reference and adjusted the orientation of the left-hand IMU coordinate frame to the right-hand reference [16].

$$\widetilde{\mathbf{H}}_r = \mathbf{H}_l \times \begin{bmatrix} -1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & 0 & 0 & -1 \end{bmatrix} \quad (1)$$

Where $\mathbf{H}_l$ is the original left-hand IMU signal and $\widetilde{\mathbf{H}}_r$ is the mirrored data by reversing the direction of $\mathbf{a}_x$, $\mathbf{g}_y$ and $\mathbf{g}_z$ from the left hand. Fig. 3 presents the example of hand mirroring.

C. MS-TCN Model Architecture

An MS-TCN [17] is established by stacking several single-stage TCNs (SS-TCN) [18] sequentially. Lea at al. [18] first developed the SS-TCN by utilizing dilated convolution and skip connection to recognize long-range temporal sequences in vision-based action segmentation. To date, it has been applied to the healthcare and disease diagnosis domain to process time-series signals [19], [20]. The experiment [21]

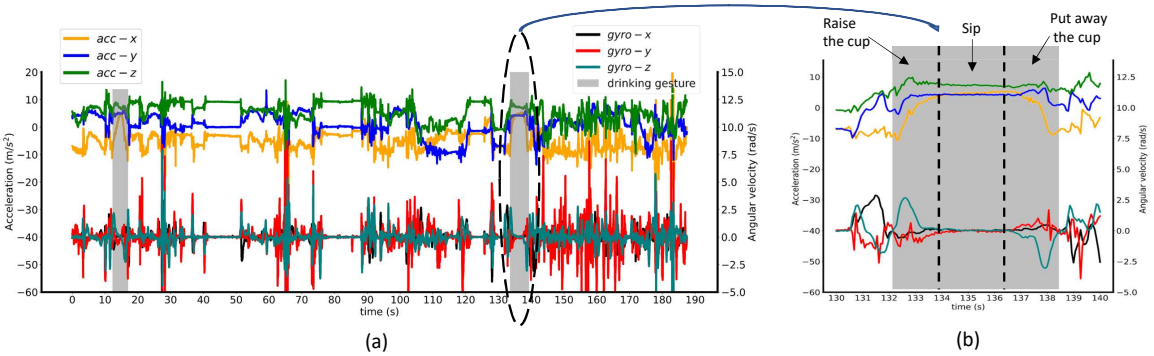

Fig. 2. Example of drinking using the right hand in free living environments. Subfigure (a) presents the 3 min segment which contains two drinking gestures (highlighted in grey box). Subfigure (b) shows the detail waveform of a drinking gesture from the right hand.

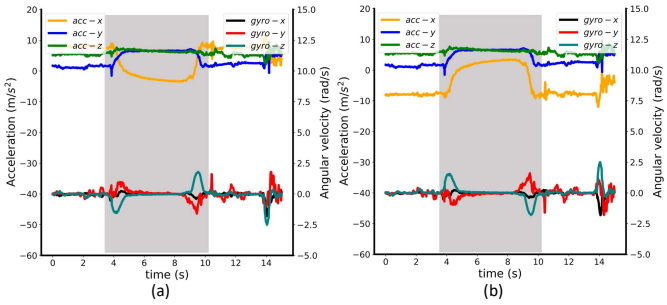

Fig. 3. Example of left hand signal transform to right hand signal. Panel (a) is the initial left hand signal, panel (b) is the mirrored signal. The grey section represents one drinking gesture.

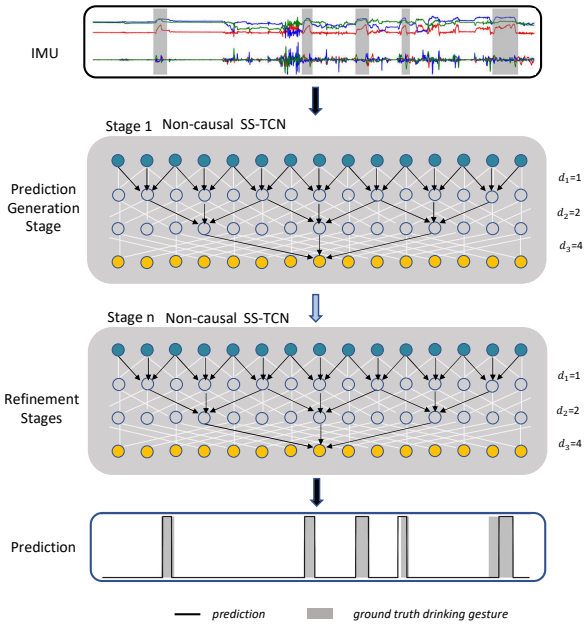

Fig. 4. The framework of drinking gesture detection. From top to bottom, the data from IMU are sent into the SS-TCN's input layer firstly. Secondly, a dilated non-causal MS-TCN model architecture is presented. The initial prediction sequence generated from the first stage is refined by the subsequent stages. The black line indicates the refined predictions, while the grey section represents the ground truth drinking gesture. In the figure, the dilation factor of each layer is shown by $d_l(l = 1, 2, 3...L)$.

demonstrates its outstanding ability on time-series signal processing.

Fig. 4 shows the structure of the model. The first stage is the prediction generation stage, which takes each sample points of IMU signals as input and generates the initial prediction as output. The subsequent stages are refinement stages which process the prediction generated by the previous stage and refine it. It is to be noted that the input channels in the first stage are different to those of the refinement stages. The input channel in the first stage is the dimension of the input IMU signal (6 dimensions), and the channel in the remaining stages is defined by the number of classes (2 classes).

In every stage, the SS-TCN is comprised of a series of L dilated residual layers (L is the number of layers). The dilation factor is doubled at each layer such that $d_l = 2^{l-1}$ $(1 \leq l \leq L)$. Each dilated residual layer consists of dilated convolutions with RELU activation, a residual connection that adds the input of the current layer and the convolution result together, as shown in Fig. 5. The SS-TCN in Fig. 4 is a non-causal type, which means the result depends not only on the data in the past, but also on the data in the future. After the last dilated residual layer, a softmax activation is applied to generate prediction according to the feature extracted from previous layers. The length of the receptive field for non-causal SS-TCN (kernal size is 3) is calculated as $r(L) = 2^{L+1} - 1$, where L represents the depth of the network (number of layers) in each stage.

The loss function of the MS-TCN model is a combination of the classification loss and the smoothing loss from each stage. Firstly, a cross entropy loss is applied as the classification loss of each stage:

$$\mathcal{L}_{cls} = \frac{1}{T} \sum_{t,c} -y_{t,c} log(\hat{y}_{t,c}) \tag{2}$$

where $y_{t,c}$ represents the ground truth label, $\hat{y}_{t,c}$ is the predicted output for class c at time t.

Secondly, a truncated mean squared error (MSE) over sample-wise log-probabilities [17] is used as the smoothing

loss:

$$\mathcal{L}_{T-MSE} = \frac{1}{TC}\sum_{t,c}\widetilde{\triangle}^2{}_{t,c} \tag{3}$$

$$\widetilde{\triangle}_{t,c} = \begin{cases} \triangle_{t,c} : \triangle_{t,c} \leq \tau \\ \tau : otherwise \end{cases} \tag{4}$$

$$\triangle_{t,c} = |log(\hat{y}_{t,c}) - log(\hat{y}_{t-1,c})| \tag{5}$$

where T is the temporal length of data, C is the number of classes, and $\hat{y}_{t,c}$ is the probability of class c at time t.

The complete model loss $\mathcal{L}$ is obtained to combine the two types of loss :

$$\mathcal{L}_n = \mathcal{L}_{cls} + \lambda\mathcal{L}_{T-MSE} \tag{6}$$

$$\mathcal{L} = \sum_n \mathcal{L}_n \tag{7}$$

where $\mathcal{L}_n$ is the loss at stage n, λ is a parameter to determine the weights of the two losses. We select $\tau = 4$ and $\lambda = 0.15$ for the loss function according to [17].

We adapted MS-TCN2 [17] to implement the 2-stage non-causal MS-TCN. There are 128 filters in each layer, and the kernel size is 3. The 30% dropout is applied after each layer. According to experiments, the depth in each stage is 9 layers, so the receptive field is 1023 sample points. An Adam optimizer with a learning rate of 0.0005 is applied to train the model. It should be noted that there is a time delay for prediction according to its non-causal architecture. The time delay is obtained by taking half of the receptive field divided by the sampling frequency (0.5×1023/16=32s).

D. Post-Processing

The outputs from the MS-TCN model are predictions on each single hand. If the participant uses two hands to drink simultaneously, the corresponding segments from both hands will be labelled as 1. To eliminate the repetitive counting, an OR operator is used to produce the final prediction.

E. Evaluation Scheme

The segmental F1-score is applied to assess the performance [17], [18]. In order to calculate the segmental F1-score, the intersection over union (IoU) of each predicted drinking gesture is first calculated. The IoU is defined as $\frac{A \cap B}{A \cup B}$, the overlap ratio between the time intervals of the ground truth segment (A) and the predicted segment (B). Fig. 6 depicts the definition of True Positive (TP), False Negative (FN) and False Positive (FP) in the segment-wise evaluation scheme. If a segment's IoU is higher than a selected threshold k, it is considered as a TP (Fig. 6 TP_1), otherwise it is an FN segment (Fig. 6 FN_1) or FP segment (Fig. 6 FP_1). The decision tree is as follows:

$$Segment = \begin{cases} TP, & IoU \geq k \\ FP, & IoU < k, length_{gt} < length_p \\ FN, & IoU < k, length_{gt} > length_p \end{cases} \tag{8}$$

Where $length_{gt}$ and $length_p$ are the temporal lengths of the ground truth drinking gesture and predicted drinking gesture, respectively. Three thresholds k are selected as 0.1, 0.25 and 0.5 according to [17], [18]. Furthermore, if more than one predicted segments exist within the interval of a single ground truth drinking gesture, only one is counted as a TP, while all others are FP (Fig. 6 TP_2 and FP_2). Conversely, if a predicted segment spans multiple ground truth drinking gestures, only one counts as TP, all others are considered as FN (Fig. 6 TP_3 and FN_2). The segmental F1-score is calculated as $\frac{2TP}{2TP+FP+FN}$. The advantages of the segment-wise evaluation scheme are twofold. Firstly, it penalizes over-segmentation errors (Fig. 6 FP_1) and under-segmentation errors (Fig. 6 FN_1); secondly, it allows minor temporal shifts between ground truth and prediction, which may be caused by annotation variability.

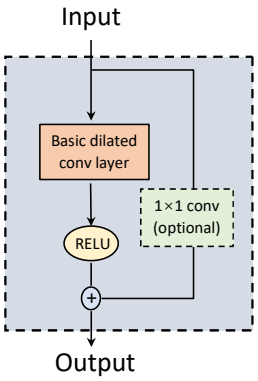

Fig. 5. The architectural elements of a dilated residual layer. In addition to the residual connection, an 1×1 convolution is employed if the current layer is the first layer from first stage (Input layer). In that case, the dimension of the input is different to that of the output after the RELU activation, the 1×1 convolution can adjust it to the same dimension to enable the residual connection.

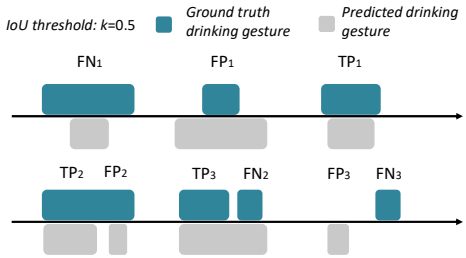

Fig. 6. Examples of the segment-wise evaluation. The threshold k is 0.5, if the calculated IoU is under 0.5, then comparing the length of the ground truth and prediction, FN_1 is under-segmentation, FP_1 is over-segmentation. The IoU of the third case is high than 0.5, so it is TP_1. If one ground truth segment spans two predicted segments, we only count one, so there is 1 TP_2 and 1 FP_2. Similarly, there is 1 TP_3 and 1 FN_2 for the fifth case. If there is a predicted segment without a ground truth segment or a ground truth segment without a predicted segment, then there is 1 FP_3 or 1 FN_3.

III. Results and Discussion

We performed two sets of experiments in this study. We first applied the MS-TCN model on DX-I and DX-II separately using LOSO cross-validation to evaluate the performance in semi-controlled environments and free-living environments. Then the CNN-LSTM approach from [13] was applied to our datasets as the benchmark. Table I

TABLE I

SEGMENT-WISE PERFORMANCE WITH DIFFERENT MODELS ON TWO COLLECTED DATASETS.

Dataset	Model	$k = 0.1$				$k = 0.25$				$k = 0.5$			
		TP	FP	FN	F1-score	TP	FP	FN	F1-score	TP	FP	FN	F1-score
DX-I	CNN-LSTM [13]	402	74	8	0.907	395	75	14	0.899	380	77	27	0.880
	MS-TCN	401	28	9	0.956	400	29	9	0.955	391	32	15	0.943
DX-II	CNN-LSTM [13]	292	84	12	0.859	280	85	23	0.838	258	85	45	0.799
	MS-TCN	294	34	10	0.930	286	37	15	0.917	277	38	25	0.900

presents the performance of segment-wise evaluation with three thresholds (k=0.1, 0.25, and 0.5). The proposed method outperforms the CNN-LSTM approach on two datasets. When k=0.5, the F1-score for CNN-LSTM is 0.880 in semi-controlled environments, lower than the F1-score for MS-TCN (0.943). The performance reduction for CNN-LSTM in free-living environments is significant (0.880→0.799), whereas the MS-TCN model obtains a much higher F1-score of 0.900. The number of FP segments is larger than that of FN segments on both datasets. By investigating the corresponding video and wrong predictions, we found that the model tends to recognize some eating gestures (i.e., eating food with a hand) as drinking gestures. The CNN-LSTM model suffers more on this.

IV. CONCLUSION

In this paper, we explored the MS-TCN model to detect drinking gesture in free-living environments via two IMU wristbands. The model was evaluated on two datasets. The first dataset collected in semi-controlled conditions and the second dataset collected in free-living environments were used to evaluate the model's performance. Experimental results show that the MS-TCN model has a good capability to detect drinking gesture in long-term free-living environments. The limitation of our approach is its inability to assess the volume of consumed water. In the future, we plan to collect more data in free-living environments to validate our model and apply our approach on public available dataset to further evaluate the performance.

ACKNOWLEDGMENT

The authors would like to thank the participants who participated in the experiments for their efforts and time. The authors would also like to thank Yiyuan Zhang for manuscript revision.

REFERENCES

[1] E. Jéquier and F. Constant, "Water as an essential nutrient: The physiological basis of hydration," Eur. J. Clin. Nutr., vol. 64, no. 2, pp. 115–123, Feb. 2010.

[2] EFSA Panel on Dietetic Products, Nutrition, and Allergies (NDA), "Scientific Opinion on Dietary Reference Values for water," EFSA J., vol. 8, no. 3, pp. 1–48, Mar. 2010.

[3] A. M. El-Sharkawy, O. Sahota, and D. N. Lobo, "Acute and chronic effects of hydration status on health," Nutr. Rev., vol. 73, no. 2, pp. 97–109, Sep. 2015.

[4] B. M. Popkin, K. E. D'Anci, and I. H. Rosenberg, "Water, hydration, and health," Nutr. Rev., vol. 68, no. 8, pp. 439–458, Aug. 2010.

[5] R. Cohen, G. Fernie, and A. R. Fekr, "Fluid Intake Monitoring Systems for the Elderly : A Review of the Literature," Nutrients, vol. 13, no. 6, pp. 1–28, Jun. 2021.

[6] J. S. Tham, Y. C. Chang, and M. F. A. Fauzi, "Automatic identification of drinking activities at home using depth data from RGB-D camera," 2014 Int. Conf. Control. Autom. Inf. Sci. (ICCAIS), 2014, pp. 153–158.

[7] K. C. Liu, C. Y. Hsieh, H. Y. Huang, L. T. Chiu, S. J. P. Hsu, and C. T. Chan, "Drinking Event Detection and Episode Identification Using 3D-Printed Smart Cup," IEEE Sens. J., vol. 20, no. 22, pp. 13743–13751, Nov. 2020.

[8] C. Zimmermann, J. Zeilfelder, T. Bloecher, M. Diehl, S. Essig, and W. Stork, "Evaluation of a smart drink monitoring device," 2017 IEEE Sensors Appl. Symp., 2017, pp. 1–5.

[9] A. Wellnitz, J. P. Wolff, C. Haubelt, and T. Kirste, "Fluid intake recognition using inertial sensors," PervasiveHealth Pervasive Comput. Technol. Healthc., 2019, no. 4, pp. 1-7.

[10] H. Y. Huang, C. Y. Hsieh, K. C. Liu, S. J. P. Hsu, and C. T. Chan, "Fluid intake monitoring system using a wearable inertial sensor for fluid intake management," Sensors (Switzerland), vol. 20, no. 22, pp. 1–17, Nov. 2020.

[11] D. Gomes and I. Sousa, "Real-time drink trigger detection in free-living conditions using inertial sensors," Sensors (Switzerland), vol. 19, no. 9, May. 2019.

[12] K. S. Chun, N. Streeper, A. B. Sanders, D. E. Conroy, R. Adaimi, and E. Thomaz, "Towards a generalizable method for detecting fluid intake with wrist-mounted sensors and adaptive segmentation," Int. Conf. Intell. User Interfaces, Proc. IUI, 2019, pp. 80–85.

[13] V. Y. Senyurek, M. H. Imtiaz, and N. Hassan, "Detection of Drinking via Inertial Sensor †," in Proceedings of the 6th International Electronic Conference on Sensors and Applications, 2019, doi: 10.3390/ecsa-6-06590.

[14] 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.

[15] 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.

[16] 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, Jan. 2021.

[17] 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.

[18] 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.

[19] B. Filtjens, P. Ginis, A. Nieuwboer, P. Slaets, and B. Vanrumste, "Automated freezing of gait assessment with marker-based motion capture and deep learning approaches expert-level detection," 2021, [Online]. Available: http://arxiv.org/abs/2103.15449.

[20] D. Jarrett, J. Yoon, and M. Van Der Schaar, "Dynamic Prediction in Clinical Survival Analysis Using Temporal Convolutional Networks," IEEE J. Biomed. Heal. Informatics, vol. 24, no. 2, pp. 424–436, Feb. 2020.

[21] S. Bai, J. Z. Kolter, and V. Koltun, "An Empirical Evaluation of Generic Convolutional and Recurrent Networks for Sequence Modeling," 2018, [Online]. Available: http://arxiv.org/abs/1803.01271.