Modeling Wrist Micromovements to Measure In-Meal Eating Behavior from Inertial Sensor Data

Konstantinos Kyritsis, Christos Diou, *Member, IEEE,*, and Anastasios Delopoulos, *Member, IEEE*

***Abstract*—Overweight and obesity are both associated with in-meal eating parameters such as eating speed. Recently, the plethora of available wearable devices in the market ignited the interest of both the scientific community and the industry towards unobtrusive solutions for eating behavior monitoring. In this paper we present an algorithm for automatically detecting the in-meal food intake cycles using the inertial signals (acceleration and orientation velocity) from an off-the-shelf smartwatch. We use 5 specific wrist micromovements to model the series of actions leading to and following an intake event (i.e. bite). Food intake detection is performed in two steps. In the first step we process windows of raw sensor streams and estimate their micromovement probability distributions by means of a Convolutional Neural Network (CNN). In the second step we use a Long-Short Term Memory (LSTM) network to capture the temporal evolution and classify sequences of windows as food intake cycles. Evaluation is performed using a challenging dataset of 21 meals from 12 subjects. In our experiments we compare the performance of our algorithm against three state-of-the-art approaches, where our approach achieves the highest F1 detection score (0.913 in the Leave-One-Subject-Out experiment). The dataset used in the experiments is available at https://mug.ee.auth.gr/intake-cycle-detection/.**

***Index Terms*—biomedical signal processing, wearable sensors**

I. INTRODUCTION

OVERWEIGHT and obesity is the consequence of *energy imbalance* or, in other words, the result of the positive difference between energy intake and energy expenditure over a prolonged period of time [1].

In clinical settings, eating habits are typically recorded in food diaries. Although food diaries can provide relevant information about one's eating behavior, they can be very inaccurate [2], [3] and cannot measure parameters of healthy eating behavior such as eating speed [4]. Objective measurement and monitoring of eating behavior is therefore important for understanding individual behavior, for achieving weight loss and for preventing obesity.

Driven by the need for tools to objectively measure eating parameters, researchers have pursued numerous technical approaches which can be found in recent literature [5], [6], including methods that detect bites, chewing, or swallowing.

Methods that approach food intake monitoring via the chewing mechanism are capable of detecting chewing episodes [7]–[11] and estimating the weight of the bite [12]. Such methods typically make use of acoustic sensors such as in-ear microphones [7]–[9], strain sensors [10], or a combination of sensors such as accelerometry and audio [11]. Approaches based on swallowing detection measure food intake by capturing the movements of the muscles located in the pharynx and esophagus. Multiple solutions have been proposed in the literature that successfully make use of piezoelectric [13], physiological [14] and acoustic [15], [16] sensors.

Besides wearable-based approaches, eating behavior monitoring methods have also been proposed that make use of more complex sensors such as cameras. Such methods are capable of estimating the volume of the food portion [17], [18], recognizing food types [19], [20] and estimating caloric intake [21]. The use of plate scales for measuring the total food intake, the eating rate and the food intake curve has also been proposed, e.g. in [22].

In addition to measurement effectiveness, an important and often overlooked aspect of eating behavior monitoring with wearable sensors is usability and comfort. Sensor obtrusiveness and usage complexity may lead to low compliance [23] with sensors being partially used, or not used at all.

In this work we present an algorithm for monitoring the in-meal eating behavior by detecting the food intake (i.e. bite) moments using the acceleration and orientation velocity signals from a single off-the-shelf smartwatch. To achieve this, we apply a Convolutional Neural Network (CNN) to windows of raw sensor streams to estimate the probability distribution of five wrist *micromovements* that correspond to the wrist motions that typically appear when taking a bite. Subsequently, we model the temporal evolution of the windows leading to a bite event by means of a Long-Short Term Memory (LSTM) network. We evaluate our algorithm in our challenging, publicly-available *Food Intake Cycle* (FIC) dataset that contains 21 complete meal sessions from 12 unique subjects, recorded under real life conditions. In addition, the performance of our proposed algorithm is compared against three state-of-the-art approaches [24]–[26].

The rest of the paper is organized as follows. Section II presents the related work regarding the detection of eating moments using wrist-mounted inertial sensors. Section III introduces the terminology and present the proposed food intake cycle detection algorithm. Section IV describes the dataset, the conducted experiments, the evaluation scheme and the limitations of our approach. Finally, the paper concludes with Section V.

II. RELATED WORK

Detection of food intake using inertial signals from wrist-mounted sensors enables unobtrusive, low-cost solutions for

K. Kyritsis, C. Diou and A. Delopoulos are with the Multimedia Understanding Group, Department of Electrical and Computer Engineering, Aristotle University of Thessaloniki, Greece. E-mail: kokirits@mug.ee.auth.gr, diou@mug.ee.auth.gr, adelo@eng.auth.gr

food intake monitoring that have received a lot of attention recently. Such approaches focus on recognizing individual or combinations of hand gestures that are part of the process of delivering food from the plate to the mouth area.

One of the most well-known approaches is the one proposed by Dong *et al.* [24] (also used to produce the results reported in [27]). In their study, the authors solely use a single channel of the gyroscope sensor with the purpose of detecting the characteristic roll motion of the wrist that typically appears when performing a bite. The algorithm relies on four parameters that control: a) the positive angular velocity threshold that needs to be surpassed for the initial roll motion of the wrist, b) the corresponding negative angular velocity threshold, c) the minimum amount of time between the two roll movements that belong to one bite event and d) the minimum amount of time between consecutive bites. Using the method reported in [24], the authors of [27] achieve a sensitivity/positive predictive value (ppv) of 0.81/0.86 (corresponding to an F1 score of 0.83) in their laboratory setting and a sensitivity/ppv of 0.81/0.83 (corresponding to an F1 score of 0.82) in their realistic cafeteria setting.

The work of Ramos-Garcia *et al.* [28] presents a method that uses the information from wrist-mounted triaxial accelerometers and gyroscopes with the purpose of recognizing certain coarse in-meal gestures including: a) rest, b) utensiling, c) drink, d) bite and e) other. The authors propose a *gesture-to-gesture* Hidden Markov Model (HMM) approach for capturing the temporal dependencies between gestures. Experiments in their dataset of 25 meals from 25 unique subjects reveal the potential of their approach by achieving a high recognition accuracy for the bite gesture when using manually pre-segmented data. No results are reported for bite detection in a continuous fashion.

Zhang *et al.* [29] also propose an algorithm for detecting feeding gestures using the acceleration and orientation velocity signals from a typical smartwatch. Their approach mainly focuses on a scheme for segmenting candidate regions that are later classified as feeding gestures by means of a Random Forest classifier. Experimental results on their relatively small unscripted dataset of 5 meals belonging to 5 unique subjects, show the potential of their approach by achieving a mean classification precision of 0.87 and a mean classification F1 score of 0.30.

There are also efforts that make use of the characteristic hand motions that occur when certain cutlery is used. For example, the works of [30], [31] use accelerometer data from wrist mounted sensors to deal with the recognition of Asian-style eating events that mainly involve the use of the spoon and chopsticks in contrast to the western eating styles that mostly use the fork. More specifically, in [30] the authors use a CNN to achieve an accuracy of 0.879 towards the classification of three eating activities involving the usage of spoon, chopsticks and cup in their intra-subject experiments on a dataset of 7 meals. The work of [31] elaborates on the same approach by identifying a total of 29 Asian-style eating actions, such as "taking chopsticks" or "picking toppings & putting in mouth" using 3 types of utensils. Using a Bagging (i.e. bootstrap aggregating) classification scheme, the authors achieve a classification accuracy of 0.75 towards the recognition of the utensil involved (3 classes) and an accuracy of 0.28 regarding the recognition of eating actions (29 classes). More specifically, the F1 score of bite-related hand movements such as "picking side dish & putting in mouth" or "scoop soup & putting in mouth" is reported as 0.133 and 0.128 respectively.

Another work of Zhang *et al.* [25] explores the performance effects of selecting the optimal feature set and classifier towards the detection of feeding gestures using two smartwatches (one in each hand). In their work, the authors characterize a wrist motion as a feeding gesture if it is either a "food-to-mouth" or "back-to-rest" motion. In detail, the authors propose a window-based feature extraction scheme, a classification scheme involving the aforementioned motions and a density-based clustering scheme (DBSCAN) for the final bite moment detection. Evaluation of their approach is performed on a dataset of 15 subjects performing scripted eating and non-eating activities in a lab environment. The authors report an average F1 score of 0.757 in their Leave-One-Subject-Out (LOSO) experiments using the most descriptive feature subset coupled with the AdaBoost classifier. Similarly, the recent work of Thomaz *et al.* [32] also models the process of eating as a bimanual (i.e. two hand) task. Using their dataset of 14 participants performing a total of 8 eating (e.g. use fork) and 7 non-eating (e.g. watch a movie trailer) activities in a scripted fashion, the authors show that there is significant gain in performance when taking into consideration the sensor data (acceleration and orientation velocity) from both hands in contrast to just a single hand. More specifically, the authors achieve an F1 score of 0.763 towards the recognition of eating related activities when using sensor data from both hands coupled with a L2 normalization scheme as preprocessing and the Random Decision Forest classifier. No results are reported however regarding the detection of each individual activity.

In our previous works [33] and [26] we used 5 micromovements to model the sequence of wrist motions leading to a food intake event. More specifically in the latest work [26], we made use of the acceleration and orientation velocity signals from a single off-the-shelf smartwatch and proposed a feature extraction scheme based on a sliding window. Subsequently, we modeled the extracted features as 10-dimensional score vectors using a multiclass Support Vector Machine, with 10 one-versus-one classifiers and the Radial Basis Function (RBF) kernel. These scores refer to the respective distance from the separating hyperplane produced by each of the one-versus-one SVMs. The sequences of score vectors are then modeled by a similar LSTM network. LOSO experiments in our 10 subject dataset achieved an F1 score of 0.892.

The current work significantly improves this approach by introducing CNN and LSTM models of sensor windows and their sequences. It also presents a comprehensive evaluation which shows the effectiveness of our method in 21 meals recorded by 12 subjects in real-world conditions. The evaluation includes experiments involving both the first stage (estimating the micromovement distribution of each window) and the complete food detection pipeline. More specifically, the main contributions of this paper are:

1) A method for in-meal bite detection involving a two-stage modeling of wrist movements and its cross-subject evaluation that indicates superior performance compared to other existing state-of-the-art methods.
2) A CNN architecture capable of extracting features from the raw inertial signals (acceleration and orientation velocity) and modeling signal windows as micromovement probability distribution vectors (see Section III-C). This eliminates the need for hand-crafted features.
3) A publicly-available challenging dataset recorded under unscripted real-life scenarios involving a wide spectrum of different food types and eating styles. Other than the inertial signals originating from the wrist-mounted sensors, the dataset includes annotations at micromovement and food intake level (see Section IV-A).

III. Food intake cycle detection algorithm

A. Definitions & method outline

In this paper we aim at the detection of *food intake cycles* that appear during the course of a meal. Each food intake cycle is modeled as a sequence of specific wrist *micromovements*.

The term micromovement corresponds to a simple and short duration movement of the wrist that operates the utensil during the course of a meal. A typical example of a micromovement is the upwards movement of wrist towards the mouth area. Table I presents the identified micromovements along with a short description. Subsequently, we define the food intake cycle as the period that starts by picking up food from the plate (*p*), continues with an upwards motion of the wrist towards the mouth area (*u*), progresses with the placement of the food into the mouth (*m*) and finishes with a downwards motion of the wrist away from the mouth area (*d*). The previous intake cycle definition refers to the ideal scenario; in practice however, the intake cycle can contain micromovements other than the previously mentioned ones (i.e., other movement - *o* and no movement - *n*) and/or repetitions of micromovements. We use the term *cycle* to emphasize the *quasiperiodic* nature of food intakes during a meal. A visual representation of the relation between micromovements and intake cycles as well as two indicative examples can be found in Figure 1.

The method presented in this paper processes fixed-size overlapping windows of the accelerometer and gyroscope streams with the purpose of detecting food intake cycles during the course of a meal. We propose a two-step approach. The first step deals with the estimation of the micromovement distribution that takes place during each window of the raw sensor streams by means of a CNN. Instead of propagating to the next step the multiclass classification output (hard decision) of the first step, we provide all micromovement scores as computed by the last layer (softmax) of the CNN. These outputs correspond to the probabilities of the input window belonging to each micromovement, and sum to one. The second steps deals with the classification of window sequences as food intake cycles or not, via an LSTM network. Figure 2 depicts the overall pipeline of our method. Note that while Figure 1 shows the start and stop moments for each micromovement, in our approach we use multiple fixed-size overlapping windows during each micromovement. This notion is further depicted in Figure 3.

TABLE I
Table listing the identified micromovements.

Micromovement	Description
Pick food	Wrist manipulates a utensil to pick food from a plate
Upwards	Wrist moves upwards, towards the mouth area
Downwards	Wrist moves downwards, away from the mouth area
Mouth	Wrist inserts food in mouth
No movement	Wrist exhibits no movement
Other movement	Every other wrist movement

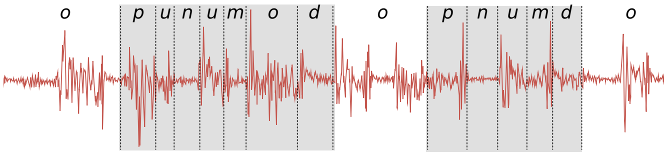


Fig. 1. Example of two intake cycles (shaded areas) and micromovements (dotted lines). Each micromovement is identified by the letter in the upper part of the figure, which matches the initial letter of the micromovement in Table I. The 1-D signal depicted in the figure is a single channel of the accelerometer stream. The leftmost food intake cycle example depicts: i) the repetition of the *u* micromovement inside the same intake cycle because the upward motion was interrupted by the *n* micromovement (e.g. pre-loading the utensil while still chewing the previous bite) and ii) the inclusion of wrist motions not related to eating (*o*) before the final downward (*d*) movement (e.g. participating in a conversation with the utensil mid-air).

B. Preprocessing

Let a_x^n, a_y^n, a_z^n and g_x^n, g_y^n, g_z^n, with $n = 1, \dots, N$ be the x, y and z streams of the 3D accelerometer and 3D gyroscope sensors captured during a meal. The length N of the streams is defined as $N = t \cdot f_s$, where t is the meal's duration in seconds and f_s the sampling frequency of the accelerometer and gyroscope sensors in Hz. A meal can be then represented by the $N \times 6$ data matrix $\mathbf{M}$ which is defined as $\mathbf{M} = [\mathbf{a}_x, \mathbf{a}_y, \mathbf{a}_z, \mathbf{g}_x, \mathbf{g}_y, \mathbf{g}_z]$, where $\mathbf{a}_z, \mathbf{a}_y, \mathbf{a}_z, \mathbf{g}_z, \mathbf{g}_y, \mathbf{g}_z$ are the N-length column vectors of the accelerometer and gyroscope sensor streams.

The first preprocessing step is to smooth any short and sudden fluctuations of each individual sensor stream. Experimentation with a 5^{th} order median filter provided satisfactory smoothing results. Furthermore, since the accelerometer sensor captures both the acceleration caused by the voluntary movement of the wrist as well as the acceleration caused due to the Earth's gravitational field, we convolve the $\boldsymbol{a}_x, \boldsymbol{a}_y$ and $\boldsymbol{a}_z$ columns of $\mathbf{M}$ with a high-pass FIR filter with a 512 tap delay line and a cutoff frequency of 1 Hz.

Finally, each column of $\mathbf{M}$ is standardized, by subtracting the mean and dividing with its own standard deviation. The last step is important since having data in different scales in a learning scenario may lead to early saturation during the optimization procedure.

C. Learning the micromovements

Following the preprocessing step, a CNN is used to estimate the micromovement probability distribution that takes place

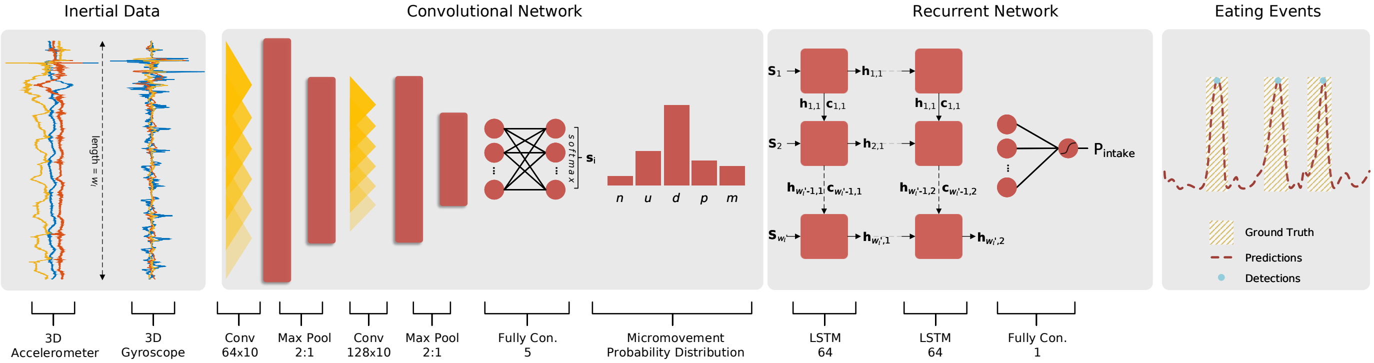


Fig. 2. Figure depicting the overall pipeline of the method. From left to right, the windowed accelerometer and gyroscope sensor streams of length w_l are transformed into the 5-dimensional micromovement probability distribution $\mathbf{s}_i$ by the convolutional neural network. Processing of additional sensor windows leads to the creation of the meal's probability distribution matrix $\mathbf{S}$. By processing parts of $\mathbf{S}$ with length w_l', the recurrent network outputs the probability $\mathbf{p}_{intake}$ that the given window sequence is a food intake cycle. The variables $h_{i,j}$ and $c_{i,j}$ are used to represent the i-th hidden output and cell state of the j^{th} LSTM layer respectively.

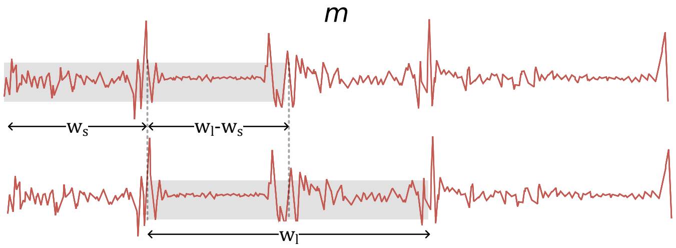


Fig. 3. Application of the sliding window approach in the leftmost m micromovement of Figure 1. The shaded areas represent two consecutive timesteps of the sliding window with length w_l, step size w_s. The quantity $w_l - w_s$ indicates the overlap between two windows.

during each window of the raw sensor streams. The proposed network architecture consists of two one-dimensional convolutional layers and a single fully connected layer with 5 units. Each of the convolutional layers is followed by a temporal max pooling operation with a decimation factor of 2. Furthermore, the fully connected layer takes as input the *flattened* output of the second max pooling operation. As the depth of the network increases, the number of filters in each convolutional layer is increased as well. Specifically, we use 64 and 128 filters in the convolutional layers. The size of the filters is retained across layers and equal to 10 tap delays, which corresponds to 0.1 seconds considering a sampling frequency of 100 Hz. All convolution operations produce an output of the *same* length as the provided input. In addition, each of the two one-dimensional convolutional layers uses the Rectified Linear Unit (ReLU) as activation, whereas the *softmax* is selected for the output layer.

We trained the network by extracting parts of the preprocessed matrix $\mathbf{M}$ (Section III-B) using the sliding window approach. The length w_l and step size w_s of the sliding window were set to 0.2 and 0.1 seconds (20 and 10 samples considering 100 Hz sampling frequency) respectively. Using this approach, the convolutional network can be trained using the extracted parts of $\mathbf{M}$, each with dimension 20×6, belonging to the: i) pick food (p), ii) upwards (u), iii) downwards (d), iv) mouth (m) and v) no movement (n) classes. The "other movement" (o) class was hard (and unnecessary) to model efficiently and was therefore excluded from the training procedure. This is due to its high intra-class variability, since it was used to classify any movement that is not p, u, d, m or n. Experiments (Section IV) indicate that the exclusion of the o class does not hurt the method's performance.

Finally, the convolutional network is trained by minimizing the categorical cross-entropy loss defined as:

$$\mathcal{L} = -\sum_{i \in B} \sum_{j \in J} y_{i,j} \log(p_{i,j}) \tag{1}$$

where $y_{i,j}$ is the optimization target, $p_{i,j}$ the predictions of the network, $J = \{p, u, d, m, n\}$ is the set of micromovements and B is the number of signal windows in the dataset. We used the *Adam* optimizer with a learning rate equal to 10^{-3} for 32 epochs with a mini-batch size of 32. Finally, we applied a 50% dropout chance to the weights of the fully connected layer to prevent overfitting during training [34].

Using the trained CNN, each $N \times 6$ data matrix $\mathbf{M}$ representing a meal session is effectively transformed into a $K \times 5$ matrix $\mathbf{S}$, with $K = \lfloor (N - w_l)/w_s \rfloor + 1$ being the number of overlapping signal windows. Essentially, the i-th row of $\mathbf{S}$ represents the micromovement probability distribution of the signal window of $\mathbf{M}$ indexed from $i \cdot w_s$ to $i \cdot w_s + w_l$.

D. Modeling the temporal evolution

To classify sequences of windows as food intake cycles we apply an LSTM network. LSTM networks [35] are gated Recurrent Neural Networks (RNN), that are specifically designed to overcome the long term dependency, vanishing gradient and exploding gradient problems. By using a combination of input, output and forget gates, LSTMs are able to retain information over a long period; making them efficient in modeling sequences of food intake cycles that deviate from the ideal scenario. For example, when there are wrist movements that do not lead to intakes in-between bites.

The proposed recurrent network takes as input overlapping sequences of micromovement distributions (rows of matrix $\mathbf{S}$, defined in Section III-C), with sequence length w_l' and a step w_s', and outputs a probability that the input sequence is a food intake cycle. The length w_l' is set to 35 samples while the step size w'_s is set to 1 sample, which correspond to 3.6 and 0.1 seconds for $w_l = 20$, $w_s = 10$ (Section III-C) and sensor sampling frequency of 100 Hz (see Section III-B). A window length of 3.6 seconds approximates the median food intake cycle duration in our dataset (which is equal to 3.55 seconds). The recurrent network consists of two consecutive LSTM layers each with 64 hidden cells. Both LSTM layers use the *hard sigmoid* function, defined as $\sigma(x) = \max(0, \min(1, x \cdot 0.2 + 0.5))$, as the non-linearity for the recurrent step and the *hyperbolic tangent* (tanh) for the activation of the gates. The output of the network is obtained by a fully connected layer with a single neuron using the sigmoid function. Similar to the CNN, we applied a 50% dropout chance to the weights of the fully connected layer to prevent overfitting during training [34].

To train the network, we label as positive examples the subsequences of the micromovement probability density matrix $\mathbf{S}$ (Section III-C) that began with the "pick food", ended with the "downwards" and containing a "mouth" micromovement instance. Micromovement sequences appearing in-between two positive examples were considered as negative examples. The set of negative examples is further augmented by allowing the negative examples to partially overlap with the positive ones. The percentage of the overlap for each example in the augmented set is experimentally selected as a random number between 15% and 35% of the duration of the corresponding positive example. In addition to including transitions between positive and negative sequences, this augmentation step balances the number of training examples of the two classes, thus allowing for a more smooth optimization during training.

The recurrent network is trained by minimizing the binary cross-entropy loss defined as:

$$\hat{\mathcal{L}} = -\sum_{i \in B'} \Big(\hat{y}_i \log(\hat{p}_i) - (1 - \hat{y}_i) \log(1 - \hat{p}_i) \Big) \qquad (2)$$

where $\hat{y}_i \in \{0, 1\}$ indicates the optimization target and $\hat{p}_i$ the prediction of the i-th sequence in the dataset size B', using the RMSprop [36] optimizer with a learning rate equal to 10^{-3} for 6 epochs with a mini-batch size of 32.

Using the trained LSTM model, each $K \times 5$ micromovement distribution matrix $\mathbf{S}$ is transformed into the $K' \times 1$ predictions vector $\boldsymbol{p}$, with $K' = \lfloor (K - w'_l)/w'_s \rfloor + 1$ being the number of overlapping sequences. Essentially, the i-th element of $\boldsymbol{p}$ corresponds to the normalized probability that the input sequence of micromovement probability distributions of $\mathbf{S}$ indexed from $i \cdot w'_s$ to $i \cdot w'_s + w'_l$, with $i = 1, \ldots, K'$, is a food intake cycle.

E. Eating moment detection

Eating moment detection on a meal session is first performed by transforming the raw data matrix $\mathbf{M}$ into the micromovement probability distribution matrix $\mathbf{S}$ using the trained CNN network. Then, by providing the probability distribution matrix $\mathbf{S}$ as input the trained LSTM network we obtain the food intake predictions vector $\boldsymbol{p}$. Next, we perform thresholding on the series $\boldsymbol{p}$ by replacing with zeros the elements that are below a probability threshold p_t. By experimenting with a small subset of the FIC dataset (four out of the twenty-one meals selected at random) we selected p_t to be equal to 0.89 as this value yielded the highest F1 detection score for that small subset. The final food intake cycles are detected by performing a local maxima search in the thresholded series $\boldsymbol{p}$ with a minimum distance between successive peaks set at 2 seconds. The food intake moments correspond to the timestamps of the detected local maxima.

IV. Experiments & Evaluation

A. Dataset

In our experiments we use our publicly available *Food Intake Cycle* (FIC) dataset. The FIC dataset contains the triaxial acceleration and orientation velocity signals from 21 meal sessions provided by 12 unique subjects. All meals were recorded in the restaurant of Aristotle University of Thessaloniki using a commercial smartwatch, the Microsoft Band 2™ for ten out of the twenty-one meals and the Sony Smartwatch 2™ for the remaining meals.

Each participant was free to select the food of their preference, typically consisting of a starter soup, a salad, a main course and a desert. This led to recordings from a diverse set of main dishes, including meat, soup, pasta, rice and others. This is important because movements in the "other movement" category vary, depending on the type of meal (e.g., cutting meat with a knife, or holding a fork versus holding a spoon). Prior to the recording, the participant was asked to wear the smartwatch to the wrist that he typically uses in his everyday life to manipulate the fork and/or the spoon. A GoPro™ Hero 5 camera was already set at the table of the participant using a small, 23 cm in height, tripod facing the participant, including both the food tray and upper body part in it's field of view. The purpose of video recording was to obtain ground truth data by manually annotating the IMU sequences based on the video stream. Participants were also asked to perform a clapping hand movement both at the start and end of the meal, for synchronization purposes (as this movement is distinctive in the IMU signal). No other instructions were given to the participants. Figure 4 depicts four typical examples of subjects included in FIC. It should be noted that the FIC dataset does not contain instances related with liquid consumption or eating without the fork, knife and spoon (e.g. eating directly with hands).

The start and end moments of each food intake cycle as well as of each micromovement are annotated throughout the dataset. Out of the total number of meals, twelve meals were annotated by two raters with a intra-rater agreement of 99%, the rest of the meals were annotated by a single rater. The FIC dataset is publicly available on the Multimedia Understanding Group (MUG) website[1]. Technical information about the duration of the intake cycles, micromovements and meals can be found in Tables II and III.

B. Experiments & evaluation methodology

[1] The FIC dataset is available for download at https://mug.ee.auth.gr/intake-cycle-detection/

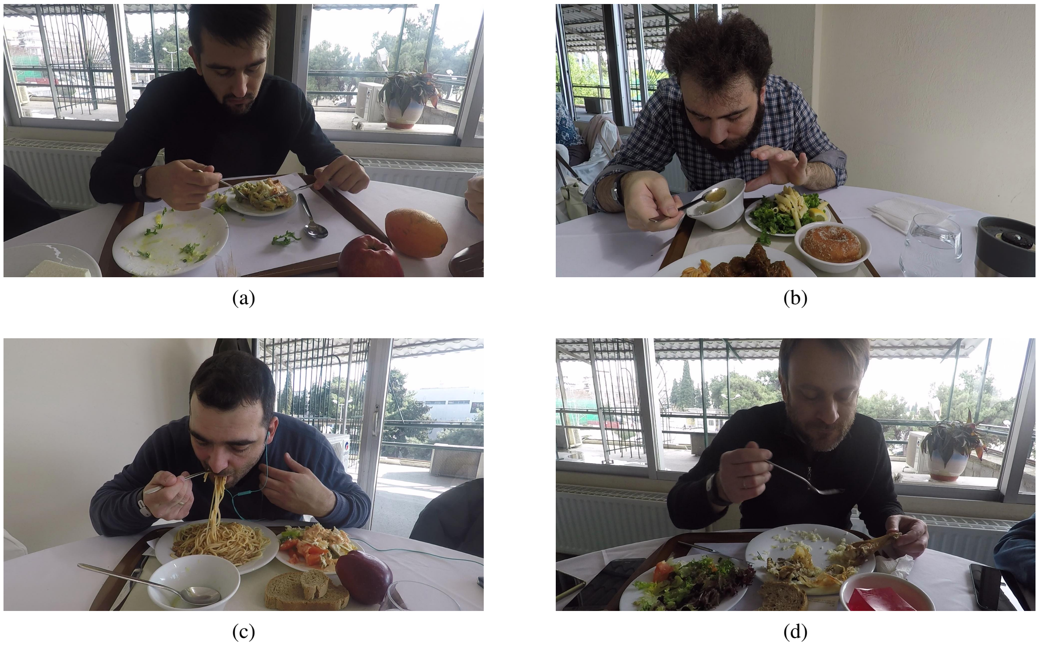

(a)
(b)
(c)
(d)

Fig. 4. Typical examples of subjects in the FIC dataset performing: a) pick food, b) upwards, c) mouth and d) downwards micromovements. The stills originate from the camera used to provide video annotations. The various eating styles and food types contained in FIC are also depicted.

TABLE II
FIC DATASET MEAL AND FOOD INTAKE CYCLE DATASET STATISTICS.

Type	#	Mean (s)	Std (s)	Median (s)	Total (s)
Meal sessions	21	703.56	186.18	717.88	$14,774.80$
Food Intake Cycles	$1,332$	4.52	3.22	3.55	$6,023.07$

TABLE III
FIC DATASET MICROMOVEMENT STATISTICS.

Micromovement	#	Mean (s)	Std (s)	Median (s)	Total (s)
p	$1,376$	1.65	1.40	1.16	$2,275$
u	$1,369$	0.93	0.51	0.81	$1,274$
d	$1,343$	0.63	0.45	0.53	848
m	$1,344$	0.47	0.24	0.43	632
n	328	6.03	5.75	4.13	$1,978$
o	$1,517$	5.65	7.30	3.27	$8,576$

1) Food intake detection experiment - FI: The first experiment we conducted deals with the cross-subject effectiveness of our approach, i.e. how well the method performs for meals that belong to unseen subjects. To achieve this, we used all available data from FIC to train both the proposed convolutional and LSTM networks in a LOSO fashion by iteratively leaving-out meals belonging to a single subject. At each iteration, evaluation is performed on the meals of the left-out subject.

2) Micromovement experiment - MM: We also performed an experiment evaluating the effectiveness of estimating the micromovement in each sensor window in order to gain further insight regarding how the quality of this estimate may effect the overall food intake detection performance. To this end we used the complete dataset to extract a total of 155655 windows of length w_l using the sliding window approach described in III-C. Out of the total number of samples, 19720 samples (12.66%) belong in the n class, 12716 (8.16%) in u, 8487 (5.45%) in d, 22755 (14.61%) in p, 6327 (4.06%) in m and 85650 (55.02%) in o. We then trained the proposed convolutional network in a LOSO fashion initially using the samples from the five classes of interest (i.e. excluding the samples from the o class) and then using the samples from all six micromovement classes.

Regarding the FI experiment, we measure the performance of the method by calculating the true positive (TP), false positive (FP) and false negative (FN) metrics. In more detail, given the timestamps of the D detected moments t_i^d for $i = 1, 2, \ldots, D$ in a meal session and the G ground truth food intake intervals described by their start and end moments $[t_j^s, t_j^e]$, $j = 1, 2, \ldots, G$, we perform metric calculation in the following fashion. If a $t_i^d \in [t_j^s, t_j^e]$ exists for some i and j,

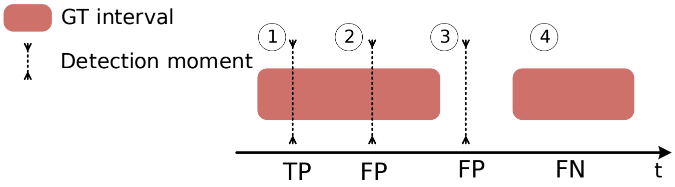


Fig. 5. Four representative examples that arise when calculating TPs, FPs and FNs using our proposed evaluation scheme.

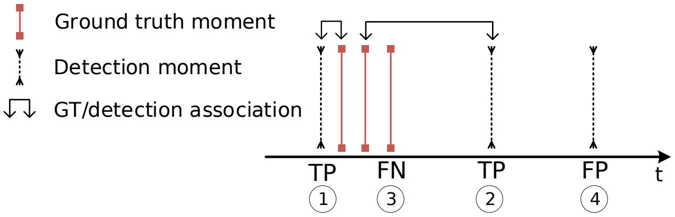


Fig. 6. Four representative examples that arise when calculating TPs, FPs and FNs using the evaluation scheme of Dong *et al.* [24].

then the i-th detection is being associated to j^{th} ground truth interval and it counts as a TP; unless a detection moment $t^d_{i'}$ with $i' < i$ has already been associated with the same ground truth interval. In the latter case it counts as a FP. Detection moments not associated with any ground truth interval (i.e. $t^d_i \notin [t^s_j, t^e_j]$ for all j) also count as FPs. Finally, ground truth intervals not associated with a detection t^d_i for any i, count as FNs. Figure 5 illustrates the proposed metric calculation scheme.

We also extend the evaluation of experiment FI by calculating the TP, FP and FN metrics using the evaluation scheme presented in the work of Dong *et al.* [24]. One fundamental difference between the two schemes is the definition of ground truth. In [24] the authors use as ground truth the *moment* of bite while in our approach we use the *interval* starting with p, ending with d and containing an m micromovement. In our experiments the moment of bite is approximated by the center of the m micromovement interval. Another difference is that the scheme proposed in [24] is significantly less strict when considering the temporal localization of the bite. For each detected bite, the authors consider the interval between the previous and the following detected bites. The first ground truth bite in this interval that is not associated with a detected bite is considered as TP. As the authors of [24] indicate, the reason is that in some cases it is possible for detections to occur prior to the actual placing of food into the mouth. If there are no ground truth bites in this interval then the detected bite is classified as FP. After classifying all detected bites any ground truth bites that are not associated with detected ones count as FNs. Figure 6 illustrates the evaluation scheme of [24].

For the MM experiment, evaluation is performed by comparing the hard decision (most probable class) of the micromovement classifier against the true label of each window.

C. Results & discussion

Early experimentation with a small subset of the FIC dataset, showed that CNN architectures that are shallower than the proposed are not as effective in terms of average micromovement recognition accuracy. Similarly, single-layer LSTM models proved to be suboptimal as well. The latter is further supported by the work of Karpathy *et al.* [37] where the authors have shown that using a depth of at least two recurrent layers is beneficial when learning sequences.

For comparison purposes we evaluate the cross-subject performance of our method against three state-of-the-art approaches (experiment FI, Section IV-B1). The first is the method proposed by Dong *et al.* [24], the second is the work by Zhang *et al.* [25] and the third is a previous work of our group [26]. An overview of these algorithms has been provided in Section II. We implemented and tuned both [24] and [25] according to the instructions provided by the authors. Specifically for [25], the best results were obtained by using a window size of 1 seconds with 0.7 seconds overlap (denoted “slow fine-grain” approach), the complete feature set, an SVM classifier with a linear kernel (regularization parameter $C = 10$) and finally $\textit{eps} = 2$ and $\textit{minimum samples} = 3$ for DBSCAN clustering. The method of [24] was optimized at each LOSO iteration in order to yield the best possible performance.

Regarding the MM experiment (Section IV-B2) we compare the micromovement probability estimation performance achieved by the proposed CNN against our previous work [26] that follows a window-based feature extraction scheme and a classification scheme that involves a multiclass SVM with 10 one-versus-one classifiers.

Since each repetition of the LSTM stochastic optimization process may lead to different minima in the loss surface and therefore influence the estimated food intake detection performance we repeated the LSTM training process 10 times. More specifically, regarding the FI experiment a total of 10 models were produced at *each* LOSO iteration leading to a total of $10 \cdot 12 = 120$ models, where 12 is the number of subjects. The metrics corresponding to the same meal are then averaged and the average metrics are then summed across different meals to produce the aggregated confusion matrix and calculate the final Precision $\left(\frac{\text{TP}}{\text{TP+FP}}\right)$, Recall $\left(\frac{\text{TP}}{\text{TP+FN}}\right)$ and F1 $\left(2 \cdot \frac{\text{Precision}\cdot\text{Recall}}{\text{Precision+Recall}}\right)$ metrics. We selected these retrieval metrics specifically (i.e. precision, recall and F1, their harmonic mean) because both the proposed evaluation scheme as well as the evaluation scheme proposed in [24] are unable to calculate true negatives (TNs). The decimals in the TP, FP, and FN metrics presented in this section are the result of averaging across experiment iterations for the same meal.

It should also be noted that despite the food intake cycles span on average 40% of the meal duration in our dataset, the prior probability of the food intake cycle detection problem is significantly lower. This is because only one detection must occur for each intake cycle, with additional detections counting as FPs. Specifically, out of the approximately 147748 total windows of length w'_l only 1332 (0.9%) correspond to the positive class. This imbalance between the positive and the negative class further increases the value of the F1 metric.

Tables IV and V present the results of the FI experiment that involves the proposed, [26], [25] and [24] methods, using the two evaluation schemes (ours and the one proposed in [24]).

TABLE IV
RESULTS OF THE FI EXPERIMENT USING THE PROPOSED EVALUATION SCHEME. THE $\star$ SYMBOL IN THE METHOD PROPOSED BY DONG *et al.* IS USED TO SIGNIFY THAT PARAMETER TUNING WAS PERFORMED BY OPTIMIZATION BASED ON OUR PROPOSED EVALUATION SCHEME.

Method	**TP**	**FP**	**FN**	**Prec**	**Rec**	**F1**
Proposed	1,241.8	144.5	90.2	0.895	0.932	**0.913**
Kyritsis *et al.* [26]	1,221.5	228.4	110.5	0.842	0.917	0.878
Zhang *et al.* [25]	944	431	388	0.686	0.708	0.697
Dong *et al.* [24]	707	794	625	0.471	0.530	0.499
Dong *et al.* [24] $\star$	772	746	560	0.508	0.579	0.541

TABLE V
RESULTS OF THE FI EXPERIMENT USING THE EVALUATION SCHEME PRESENTED IN [24]. THE $\star$ SYMBOL IN THE METHOD PROPOSED BY DONG *et al.* IS USED TO SIGNIFY THAT PARAMETER TUNING WAS PERFORMED BY OPTIMIZATION BASED ON OUR PROPOSED EVALUATION SCHEME.

Method	**TP**	**FP**	**FN**	**Prec**	**Rec**	**F1**
Proposed	1,263.4	122.9	68.6	0.911	0.948	**0.929**
Kyritsis *et al.* [26]	1,267.6	182.3	64.4	0.874	0.951	0.911
Zhang *et al.* [25]	1,102	233	230	0.825	0.827	0.826
Dong *et al.* [24]	1,190	311	142	0.792	0.893	0.840
Dong *et al.* [24] $\star$	1,214	304	118	0.799	0.911	0.851

Evaluation results indicate that the proposed method achieves the highest F1 score for both evaluation methods (Tables IV and V). An overall increase to F1 can also be observed across all experiments of Table IV when switching to the evaluation scheme of [24] scheme; 0.929 from 0.913 for the proposed, 0.911 from 0.878 for [26], 0.826 from 0.697 for [25] and 0.840 from 0.499 for [24]. Furthermore, a significant increase in the F1 score of method [24] can be observed when performing parameter tuning by optimizing using our proposed evaluation scheme, 0.541 up from 0.499 in Table IV and 0.851 up from 0.840 in Table V.

Row-wise comparison of Tables IV and V points out the differences in the obtained results produced by the two different evaluation schemes. Selection of the appropriate evaluation scheme depends on the desired application. On the one hand, if the goal is to count the number of bites in a meal session or to estimate the duration of a meal, then the less strict evaluation scheme presented in [24] suffices. On the other hand, if the goal is to investigate the structure of a meal then bites need to be detected in a more detailed fashion. As an example of the latter, the works of [38], [39], and [22] study the detailed meal structure including the eating rate and its temporal evolution. Correlation of the intervals between bites with eating behavior indicators is left as future research.

Table VI presents the *normalized* confusion matrices (including percentages instead of absolute numbers) as well as the average accuracy for each of the micromovement experiments. The results initially point out that both micromovement classifiers when trained using samples from the 5 classes of interest outperform in terms of accuracy their counterparts trained using all samples, including the ones from the *o* class. In addition, the proposed convolutional network achieves higher average accuracy when compared with the SVM that takes handcrafted features as input (0.791 in contrast to 0.737).

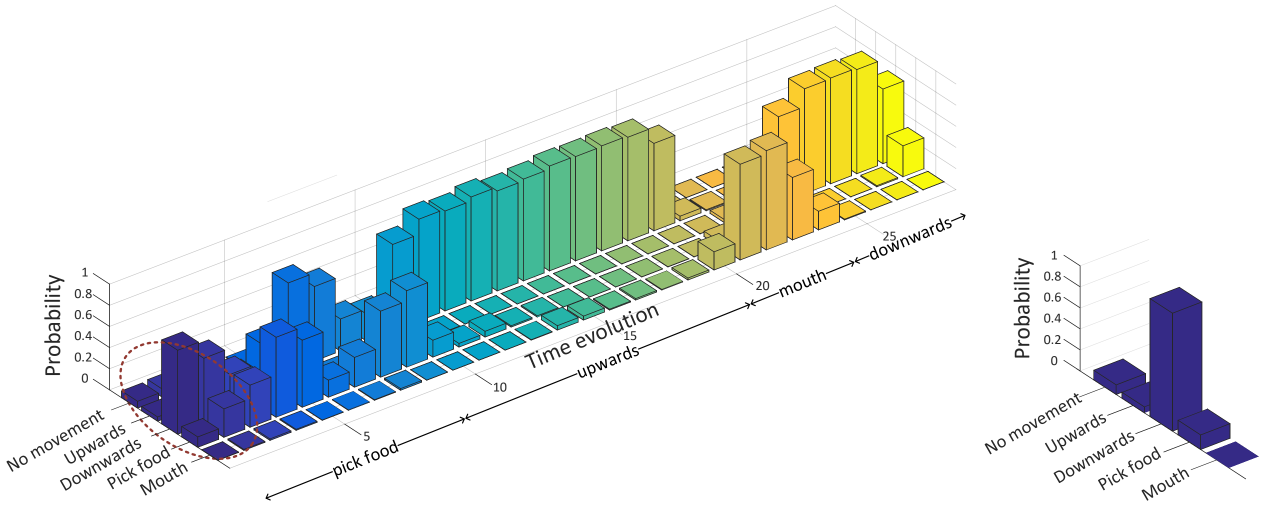
In our proposed approach instead of classifying raw signal windows as bite (i.e. *m*) events, we model them as micromovement probability distributions (by means of the CNN) and use their sequences as input to the sequence model (LSTM). This allows us to make use of the temporal evolution of the micromovement probability distributions leading to and after the bite event. Also, the input to the sequence model is the micromovement distribution and not a single micromovement (i.e. the hard decision). Thus, even if the wrong micromovement has the highest probability, there is usually significant probability mass to the correct one as well. Results from the FI experiment (Table IV) regarding the proposed approach indicate that despite the average micromovement recognition accuracy being mediocre (0.792, Table VI-a), sequence modeling by means of an LSTM network is robust. Figure 7 provides a graphical example of how an intake event of the left-out subject is modeled using sequences of micromovement distributions.

Experiment FI is closely related to MM in the sense that they follow the same micromovement learning setup. Comparing Tables IV and VI (a and c) regarding the proposed and the method of [26], it can be observed that while keeping the same temporal modeling and intake moment detection mechanisms, an increase to the micromovement recognition performance can lead to an overall increase of the food intake detection performance, even when the average micromovement recognition accuracy is mediocre.

D. Limitations

Experimental results in Section IV-C indicate that our method can effectively detect food intake cycles during the course of a meal. However, our method assumes that the start and end moments of the meal are known. In a practical scenario this would require that the user starts recording moments before eating and stops recording moments after the meal ends. Future research efforts will be directed towards the automatic detection of the meal start and end points.

In addition, our current work deals with the modeling of eating behavior by analyzing meal sessions that are centered around using the fork, the knife and the spoon. Thus, liquid intake instances (e.g. drinking water from a glass), eating with



(a)

(b)

Fig. 7. a) Example of modeling a food intake cycle as a sequence of micromovement probability distributions. The micromovement ground truth intervals are also marked below the figure. b) Zoomed-in view of the area marked by the dotted line in a) (a single micromovement probability distribution at $t = 0$).

TABLE VI

MM experiment results. Top row: Normalized confusion matrices for the proposed approach using 5 (a) and 6 (b) classes. Bottom row: Normalized confusion matrices for the multiclass SVM from [26] using 5 (c) and 6 (d) classes.

		Predicted				
		n	u	d	p	m
Actual	n	0.892	0.013	0.018	0.069	0.005
	u	0.042	0.749	0.008	0.167	0.031
	d	0.068	0.013	0.649	0.193	0.075
	p	0.052	0.057	0.043	0.826	0.019
	m	0.028	0.112	0.089	0.141	0.627

(a) Average accuracy= 0.791

		Predicted					
		n	u	d	p	m	o
Actual	n	0.348	0.003	0.001	0.004	0.000	0.640
	u	0.002	0.655	0.001	0.058	0.022	0.260
	d	0.002	0.003	0.312	0.006	0.047	0.628
	p	0.003	0.040	0.003	0.257	0.005	0.689
	m	0.002	0.097	0.068	0.021	0.486	0.323
	o	0.053	0.017	0.012	0.039	0.005	0.871

(b) Average accuracy= 0.651

		Predicted				
		n	u	d	p	m
Actual	n	0.922	0.008	0.006	0.060	0.001
	u	0.060	0.754	0.002	0.173	0.009
	d	0.098	0.011	0.619	0.210	0.061
	p	0.090	0.086	0.047	0.750	0.024
	m	0.061	0.160	0.127	0.407	0.242

(c) Average accuracy= 0.737

		Predicted					
		n	u	d	p	m	o
Actual	n	0.824	0.002	0.000	0.001	0.000	0.170
	u	0.038	0.660	0.000	0.061	0.000	0.237
	d	0.036	0.006	0.317	0.007	0.009	0.627
	p	0.041	0.050	0.004	0.184	0.000	0.718
	m	0.019	0.138	0.067	0.066	0.036	0.671
	o	0.213	0.029	0.020	0.048	0.001	0.686

(d) Average accuracy= 0.581

different utensils (e.g. chopsticks) or eating using bare hands are not taken under consideration.

While the recording conditions coincide with the typical way that someone eats in the cafeteria of Aristotle University of Thessaloniki, the FIC dataset does not cover the whole spectrum of eating and non eating-related gestures that can be performed during meals. Nevertheless, there are multiple instances where for example the participant engaged in conversation with individuals in the same table (with or without a "loaded" utensil mid-air) or used his personal smartphone.

The relatively high memory and computational requirements of our method may impose limitations on current wearable hardware. Out of the four approaches that we presented in our experimental section, only the approach presented in [24] can be integrated in a wearable device as the current state of wearable hardware does not allow for demanding real-time processing. The average processing times, in seconds, for a meal in our dataset are: 0.334 for [24], 3.243 for the proposed method, 13.842 for [25], and 72.735 for our previous method presented in [26]. In our calculations we took under consideration the processing time required, where applicable, for preprocessing, feature extraction, model inference and postprocessing. All experiments were conducted using an Intel(R) Xeon(R) CPU E5-2650 clocked at 2.30GHz and coupled with an NVIDIA Tesla K40 GPU. It is worth noting that although these methods cannot be implemented on current wearable hardware, offline processing can be performed by transmitting the raw IMU signal externally.

V. Conclusions

We have presented an algorithm for measuring the in-meal eating behavior by performing temporal localization of the intake moments (i.e. bites) using the inertial signals from off-the-shelf smartwatches. In our approach we use five specific wrist micromovements to model the sequence of actions leading to and after an intake event. Past the preprocessing step, our algorithm initially uses a CNN to recognize and model the micromovements. Next, the temporal evolution of the micromovements is modeled by a recurrent network with two LSTM layers.

We evaluate the performance of our algorithm on our challenging, publicly-available dataset of 21 meals from 12 subjects. In addition, we compare the performance of our approach against three state-of-the-art methods using two evaluation schemes. Results from our LOSO food intake detection experiments indicate that our approach yields higher F1 score (0.913) when compared with other state-of-the-art approaches. Finally, we experimentally show the relation between the overall intake detection and the micromovement detection performance.

These results are promising, and indicate that the proposed method is sufficiently robust to be used as a measurement tool for in-meal eating event detection, both for scientific research as well as for dietary monitoring applications. Future research involves the development of improvements to overcome the limitations outlined in Section IV-D, as well as the extensive evaluation of the proposed method in larger datasets.

Acknowledgments

The work leading to these results has received funding from the EU Commission under Grant Agreement No. 727688 (http://bigoprogram.eu, H2020), and Grant Agreement No. 690494 (http://www.i-prognosis.eu, H2020).

We gratefully acknowledge the support of NVIDIA Corporation with the donation of the Tesla K40 GPU used for this research.

References

[1] World Health Organization, *Obesity: preventing and managing the global epidemic*. World Health Organization, 2000, no. 894.

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

[3] W. J. Korotitsch and R. O. Nelson-Gray, "An overview of self-monitoring research in assessment and treatment." *Psychological Assessment*, vol. 11, no. 4, p. 415, 1999.

[4] K. Maruyama, S. Sato, T. Ohira, K. Maeda, H. Noda, Y. Kubota, S. Nishimura, A. Kitamura, M. Kiyama, T. Okada *et al.*, "The joint impact on being overweight of self reported behaviours of eating quickly and eating until full: cross sectional survey," *Bmj*, vol. 337, p. a2002, 2008.

[5] T. Vu, F. Lin, N. Alshurafa, and W. Xu, "Wearable food intake monitoring technologies: A comprehensive review," *Computers*, vol. 6, no. 1, 2017. [Online]. Available: http://www.mdpi.com/2073-431X/6/1/4

[6] R. Steele, "An overview of the state of the art of automated capture of dietary intake information," *Critical Reviews in Food Science and Nutrition*, vol. 55, no. 13, pp. 1929–1938, 2015, pMID: 24950017. [Online]. Available: https://doi.org/10.1080/10408398.2013.765828

[7] V. Papapanagiotou, C. Diou, and A. Delopoulos, "Chewing detection from an in-ear microphone using convolutional neural networks," in *2017 39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, July 2017, pp. 1258–1261.

[8] O. Amft, "A wearable earpad sensor for chewing monitoring," in *2010 IEEE Sensors*, Nov 2010, pp. 222–227.

[9] S. Päßler and W. J. Fischer, "Acoustical method for objective food intake monitoring using a wearable sensor system," in *2011 5th International Conference on Pervasive Computing Technologies for Healthcare (PervasiveHealth) and Workshops*, May 2011, pp. 266–269.

[10] E. S. Sazonov and J. M. Fontana, "A sensor system for automatic detection of food intake through non-invasive monitoring of chewing," *IEEE Sensors Journal*, vol. 12, no. 5, pp. 1340–1348, May 2012.

[11] 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 Journal of Biomedical and Health Informatics*, vol. 21, no. 3, pp. 607–618, May 2017.

[12] O. Amft*, M. Kusserow, and G. TrÖster, "Bite weight prediction from acoustic recognition of chewing," *IEEE Transactions on Biomedical Engineering*, vol. 56, no. 6, pp. 1663–1672, June 2009.

[13] H. Kalantarian, N. Alshurafa, and M. Sarrafzadeh, "A wearable nutrition monitoring system," in *2014 11th International Conference on Wearable and Implantable Body Sensor Networks*, June 2014, pp. 75–80.

[14] M. Farooq, J. M. Fontana, and E. Sazonov, "A novel approach for food intake detection using electroglottography," *Physiological measurement*, vol. 35, no. 5, p. 739, 2014.

[15] E. Sazonov, S. Schuckers, P. Lopez-Meyer, O. Makeyev, N. Sazonova, E. L. Melanson, and M. Neuman, "Non-invasive monitoring of chewing and swallowing for objective quantification of ingestive behavior," *Physiological measurement*, vol. 29, no. 5, p. 525, 2008.

[16] E. S. Sazonov*, O. Makeyev, S. Schuckers, P. Lopez-Meyer, E. L. Melanson, and M. R. Neuman, "Automatic detection of swallowing events by acoustical means for applications of monitoring of ingestive behavior," *IEEE Transactions on Biomedical Engineering*, vol. 57, no. 3, pp. 626–633, March 2010.

[17] M. Puri, Z. Zhu, Q. Yu, A. Divakaran, and H. Sawhney, "Recognition and volume estimation of food intake using a mobile device," in *2009 Workshop on Applications of Computer Vision (WACV)*, Dec 2009, pp. 1–8.

[18] F. Zhu, M. Bosch, I. Woo, S. Kim, C. J. Boushey, D. S. Ebert, and E. J. Delp, "The use of mobile devices in aiding dietary assessment and evaluation," *IEEE Journal of Selected Topics in Signal Processing*, vol. 4, no. 4, pp. 756–766, Aug 2010.

[19] M. M. Anthimopoulos, L. Gianola, L. Scarnato, P. Diem, and S. G. Mougiakakou, "A food recognition system for diabetic patients based on an optimized bag-of-features model," *IEEE Journal of Biomedical and Health Informatics*, vol. 18, no. 4, pp. 1261–1271, July 2014.

[20] F. Zhu, M. Bosch, N. Khanna, C. J. Boushey, and E. J. Delp, "Multilevel segmentation for food classification in dietary assessment," in *2011 7th International Symposium on Image and Signal Processing and Analysis (ISPA)*, Sept 2011, pp. 337–342.

[21] F. Kong and J. Tan, "Dietcam: Automatic dietary assessment with mobile camera phones," *Pervasive and Mobile Computing*, vol. 8, no. 1, pp. 147 – 163, 2012. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1574119211001131

[22] V. Papapanagiotou, C. Diou, I. Ioakimidis, P. Sodersten, and A. Delopoulos, "Automatic analysis of food intake and meal microstructure based on continuous weight measurements," *IEEE Journal of Biomedical and Health Informatics*, pp. 1–1, 2018.

[23] J. van den Boer, A. van der Lee, L. Zhou, V. Papapanagiotou, C. Diou, A. Delopoulos, and M. Mars, "The user-informed development of the SPLENDID eating detection sensor," *JMIR mHealth and uHealth*, 2018 (preprint).

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

[25] S. Zhang, R. Alharbi, W. Stogin, M. Pourhomayun, B. Spring, and N. Alshurafa, "Food watch: Detecting and characterizing eating episodes through feeding gestures," in *Proceedings of the 11th EAI International Conference on Body Area Networks*, ser. BodyNets '16. ICST, Brussels, Belgium, Belgium: ICST (Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering), 2016, pp. 91–96. [Online]. Available: http://dl.acm.org/citation.cfm?id=3068615.3068638

[26] K. Kyritsis, C. Diou, and A. Delopoulos, "Food intake detection from inertial sensors using LSTM networks," in *International Conference on Image Analysis and Processing*. Springer, 2017, pp. 411–418.

[27] 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 Journal of Biomedical and Health Informatics*, vol. 21, no. 3, pp. 599–606, May 2017.

[28] R. I. Ramos-Garcia, E. R. Muth, J. N. Gowdy, and A. W. Hoover, "Improving the recognition of eating gestures using intergesture sequential dependencies," *IEEE Journal of Biomedical and Health Informatics*, vol. 19, no. 3, pp. 825–831, May 2015.

[29] S. Zhang, W. Stogin, and N. Alshurafa, "I sense overeating: Motif-based machine learning framework to detect overeating using

wrist-worn sensing," *Information Fusion*, vol. 41, pp. 37 – 47, 2018. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1566253517304785

[30] J. Cho and A. Choi, "Asian-style food intake pattern estimation based on convolutional neural network," in *2018 IEEE International Conference on Consumer Electronics (ICCE)*, Jan 2018, pp. 1–2.

[31] H. J. Kim, M. Kim, S. J. Lee, and Y. S. Choi, "An analysis of eating activities for automatic food type recognition," in *Proceedings of The 2012 Asia Pacific Signal and Information Processing Association Annual Summit and Conference*, Dec 2012, pp. 1–5.

[32] E. Thomaz, A. Bedri, T. Prioleau, I. Essa, and G. D. Abowd, "Exploring symmetric and asymmetric bimanual eating detection with inertial sensors on the wrist," in *Proceedings of the 1st Workshop on Digital Biomarkers*, ser. DigitalBiomarkers '17. New York, NY, USA: ACM, 2017, pp. 21–26. [Online]. Available: http://doi.acm.org/10.1145/3089341.3089345

[33] K. Kyritsis, C. L. Tatli, C. Diou, and A. Delopoulos, "Automated analysis of in meal eating behavior using a commercial wristband IMU sensor," in *Engineering in Medicine and Biology Society (EMBC), 2017 39th Annual International Conference of the IEEE*. IEEE, 2017, pp. 2843–2846.

[34] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *J. Mach. Learn. Res.*, vol. 15, no. 1, pp. 1929–1958, Jan. 2014. [Online]. Available: http://dl.acm.org/citation.cfm?id=2627435.2670313

[35] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997. [Online]. Available: https://doi.org/10.1162/neco.1997.9.8.1735

[36] T. Tieleman and G. Hinton, "Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude," *COURSERA: Neural networks for machine learning*, vol. 4, no. 2, pp. 26–31, 2012.

[37] A. Karpathy, J. Johnson, and F. Li, "Visualizing and understanding recurrent networks," *CoRR*, vol. abs/1506.02078, 2015. [Online]. Available: http://arxiv.org/abs/1506.02078

[38] M. Zandian, I. Ioakimidis, C. Bergh, U. Brodin, and P. Södersten, "Decelerated and linear eaters: Effect of eating rate on food intake and satiety," *Physiology & Behavior*, vol. 96, no. 2, pp. 270 – 275, 2009. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0031938408003193

[39] M. Zandian, I. Ioakimidis, C. Bergh, and P. Södersten, "Linear eaters turned decelerated: Reduction of a risk for disordered eating?" *Physiology & behavior*, vol. 96, no. 4-5, pp. 518–521, 2009.

Konstantinos Kyritsis is a PhD student in the Multimedia Understanding Group (MUG) of the Information Processing Laboratory (IPL), Department of Electrical and Computer Engineering, Aristotle University of Thessaloniki (AUTh). He recieved his diploma from the Department of Electrical and Computer Engineering of the University of Thessaly (UTh) in 2014. His research interests focus on the areas of digital signal processing, wearable sensors and machine learning.

Christos Diou received his diploma and PhD from the Electrical and Computer Engineering Department of the Aristotle University of Thessaloniki, in 2004 and 2010 respectively. Since 2011 he has been working as a Research Associate at the Information Processing Laboratory of the Aristotle University of Thessaloniki in the areas of machine learning and signal processing, for applications such as multimedia analysis and retrieval, analysis of human eating and physical activity behavior, analysis of electrical energy consumption behavior of small-scale consumers as well as machine learning for computer security applications.

Dr. Anastasios Delopoulos was born in Athens, Greece, in 1964. He graduated from the Department of Electrical Engineering of the National Technical University of Athens (NTUA) in 1987, received the M.Sc. from the University of Virginia in 1990 and the Ph.D. degree from NTUA in 1993. From 1995 till 2001 he was a senior researcher in the Institute of Communication and Computer Systems of NTUA. Since 2001 he is with the Electrical and Computer Engineering Department of the Aristotle University of Thessaloniki where he serves as an associate professor. His research interests lie in the areas of machine learning, signal and multimedia processing and computer vision. On the applied domain he works in the areas of multimedia retrieval, biomedical engineering and behavioral informatics. He is the (co)author of more than 80 journal and conference scientific papers. He has participated in 21 European and National R&D projects related to application of his research to entertainment, culture, education and health sectors. Dr. Delopoulos is a member of the Technical Chamber of Greece and the IEEE.