Food Intake Detection from Inertial Sensors using LSTM Networks

Konstantinos Kyritsis, Christos Diou, and Anastasios Delopoulos

Multimedia Understanding Group, Information Processing Laboratory,
Aristotle University of Thessaloniki, Greece
{kokirits,diou}@mug.ee.auth.gr adelo@eng.auth.gr
https://mug.ee.auth.gr/

Abstract. Unobtrusive analysis of eating behavior based on Inertial Measurement Unit (IMU) sensors (e.g. accelerometer) is a topic that has attracted the interest of both the industry and the research community over the past years. This work presents a method for detecting food intake moments that occur during a meal session using the accelerometer and gyroscope signals of an off-the-shelf smartwatch. We propose a two step approach. First, we model the hand micro-movements that take place while eating using an array of binary Support Vector Machines (SVMs); then the detection of intake moments is achieved by processing the sequence of SVM score vectors by a Long Short Term Memory (LSTM) network. Evaluation is performed on a publicly available dataset with 10 subjects, where the proposed method outperforms similar approaches by achieving an F1 score of 0.892.

Keywords: Food intake; eating monitoring; wearable sensors; LSTM

1 Introduction

Recent reports[1] from the World Health Organization (WHO) point out the global epidemic status that obesity has reached by doubling the affected population worldwide since 1980. In particular, overweight and obesity are two of the most prevalent *preventable* causes of death, alongside smoking tobacco and sexually transmitted diseases, and are responsible for over 2.5 million deaths per annum since 2001 [11]. Thus, the ability to unobtrusively monitor eating behavior plays a key role in the study and treatment of obesity.

Several devices have been introduced specifically for measuring meal eating behavior, e.g. by weight scale [8] or based on sound [9]. In this paper, we are interested in detecting eating moments *during the course of a meal* using general purpose IMU sensors. This enables us to automatically measure in-meal eating behavior in terms of number of bites, bite frequency and bite frequency acceleration or deceleration, thus approximating the food intake curve of [8].

[1] http://who.int/mediacentre/factsheets/fs311/en/

Several approaches use multiple sensors to achieve high detection accuracy. In particular, the work of [1] involve the usage of multiple body-mounted accelerometers with the goal of detecting eating related gestures, whereas the authors of [6] combine a number of audio and motion sensors in order to detect bites and estimate intake weight. The main drawback of these methods, however, is the low usability compared to using a single, commercially available device.

Less obtrusive approaches exist, that employ the IMU sensors of a single smartwatch. Specifically, the authors of [12] propose the dissection of a feeding gesture into two sub-feeding movements, namely food-to-mouth and back-to-rest. Following the authors' proposed gesture recognition scheme, a clustering approach is used to detect the final eating moments, resulting in 0.757 F1 score on a laboratory controlled dataset. The work of [10] makes use of the sequential dependency between a small number of gestures leading to a bite of food. Moreover, the authors propose the usage of Hidden Markov Models (HMM) to capture the temporal evolution of eating. The results show the high performance of the proposed approach in manually segmented sequences in a large dataset. However, no results on non-segmented sequences are presented. A gyroscope-based approach is introduced in [2]. The authors make use of a characteristic wrist roll pattern that is exhibited during a meal to detect biting moments.

In our previous work [4], we showed how classification of hand movements into five meal-related gestures, followed by two discrete HMMs, can be used to characterize a food intake cycle. In this paper, we improve on this approach, by modeling hand micro-movements as an SVM score vector and by subsequently using an LSTM network to classify each sequence as an intake or non-intake cycle. Experimental results on our publicly available *Food Intake Cycle*[2] (FIC) dataset show the effectiveness of this method.

Following the introduction, Section 2 introduces the terminology and presents the steps of the method towards the detection of food intake cycles. Information about the dataset is presented in Section 3, whereas Section 4 presents the conducted experiments and their results. Finally, Section 5 concludes the paper.

2 Proposed approach

The work presented in this paper aims at identifying *food intake cycles* during a meal session. Each food intake cycle consists of a series of hand *micro-movements*. The relation between meal session, food intake cycle and micro-movement is depicted in Figure 1.

In its ideal form, a food intake cycle starts by manipulating a utensil to pick up food from a plate, continues with an upwards movement of the hand operating the utensil towards the mouth, followed by inserting the food in the mouth and concluding with a downwards motion of the hand away from the mouth. However, in real meals we observe repetitions of certain hand movements, unrecognized hand movements, or no hand movement at all. In the same context, the term

[2] https://mug.ee.auth.gr/intake-cycle-detection/

micro-movement is used to describe a hand movement of limited duration that is related with the food intake cycle. A typical micro-movement example is the upwards movement of the hand operating the utensil from the plate towards the mouth. The micro-movements that we used in this study originate from the FIC dataset and are presented in Table 1.

Table 1. Table listing the selected micro-movements

Micro-movement	**Description**
Pick food	Hand manipulates a utensil to pick food from a plate
Upwards	Hand moves upwards, towards the mouth area
Downwards	Hand moves downwards, away from the mouth area
Mouth	Hand inserts food in mouth
No movement	Hand exhibits no movement
Other movement	Every other hand movement

The proposed method uses the acceleration and gyroscope signals of a smartwatch with the purpose of detecting food intake moments within a meal session. An array of binary (one-versus-one) SVMs is used to represent the initial signals as micro-movement score vectors; whereas an LSTM network is used to classify sequences of micro-movement score vectors as intake or non-intake cycles. An overview of the proposed system architecture is presented in Figure 2.

2.1 Data pre-processing

Initially, the synchronized 3D accelerometer $(a_x[n], a_y[n], a_z[n])$ and gyroscope $(g_x[n], g_y[n], g_z[n])$ sensor streams of a meal session are individually smoothed by a 5^{th} order median filter. Furthermore, since the accelerometer sensor captures both the acceleration caused by the hand's movement as well as the the acceleration due to the earth's gravitational field, the next step is to remove the gravity from the acceleration signal. To this end, we use the method proposed by [5]. More specifically, the gyroscope samples are used to estimate the rotation of the smartwatch with respect to a reference frame. We use the first sample as the reference frame (i.e. the position of the smartwatch when recording starts). Then, by assuming that the smartwatch is initially still, gravity can be removed by subtracting the first acceleration sample from the rotated sequence.

2.2 Feature extraction

Given the pre-processed accelerometer and gyroscope streams feature extraction is performed by extracting frames of length w_l and step w_s corresponding to 0.2 and 0.1 seconds respectively. Let $\boldsymbol{w}_{a_x}^i$ be the i-th extracted frame from $a_x[n]$ channel of the accelerometer signal. For each $\boldsymbol{w}_{a_x}^i$ a number of both time and frequency domain features are calculated, including i) the number of

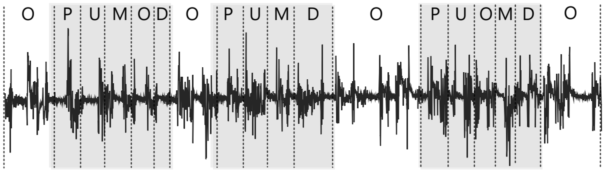

Fig. 1. Segmentation of a meal session (solid line) into intake cycles (shaded area) and micro-movements (dotted line).

zero crossings, ii) the mean, iii) the standard deviation, iv) the variance, v) the maximum value and minimum value, vi) the range of values, vii) the normalized energy and viii) the first $\frac{w_l}{2} + 1$ Discrete Fourier Transform coefficients. These features are also extracted for the rest of the accelerometer and gyroscope channels. Furthermore, the simple moving average is also calculated by $SMA_a^i = \frac{1}{w_l}\sum_{j=k}^{w_l+k} |w_{a_x}^i[j]| + |w_{a_y}^i[j]| + |w_{a_z}^i[j]|$ for the acceleration stream and in a similar manner for the gyroscope. The result of feature extraction is the representation of the $a_x[n]$, $a_y[n]$, $a_z[n]$, $g_x[n]$, $g_y[n]$ and $g_z[n]$ time series as a series of L-dimensional feature vectors $\boldsymbol{f}_i$.

2.3 Modeling the micro-movements

From the list of micro-movements of Table 1, we observed that class O exhibits high inner class variance, since it is used to represent every hand movement other than P, U, D, M and N. As a result, all extracted features belonging in the O class are excluded from the learning procedure. The micro-movement learning process is achieved by employing an array of one-versus-one SVM classifiers with the Radial Basis Function (RBF) kernel. Given the features belonging in the five classes of interest, a total of ten one-versus-one classifiers are trained. In addition, since some micro-movements are inherently longer in duration than others (e.g. P and N) all classes are weighted according to their prior probabilities. Finally, prior to training, all features are linearly scaled in $[0, 1]$. Given the trained SVM models, each feature $\boldsymbol{f}_i$ extracted as in Section 2.2 is converted into a 10-dimensional vector $\boldsymbol{s}_i$ composed of the pair-wise prediction scores of the 10 one-versus-one SVM classifiers.

2.4 Learning the food intake sequences

We designed an LSTM network with the purpose of classifying sequences of $\boldsymbol{s}_i$ as intake or non-intake cycles. The LSTM network is an extension of the Recurrent Neural Network (RNN) specifically designed to solve the long term dependency and vanishing gradient problems, thus giving it the ability to effectively model large intra-dependent sequences such as micro-movement sequences. In contrast with Markov models where the current state depends solely on the previous state in time, LSTM networks use a combination of input, output and forget gates to retain information over a long period; thus, model more efficiently intake

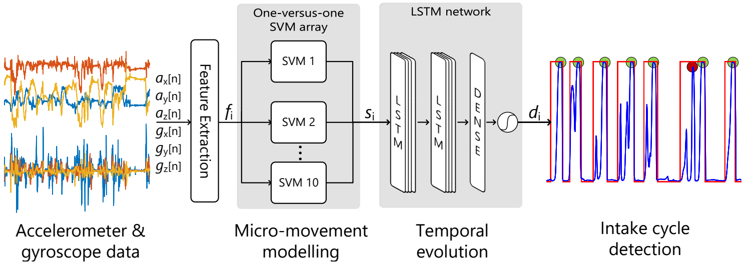

Fig. 2. Overview of the proposed system.

sequences that differ greatly from the *ideal* intake sequence due to the insertion of non intake-related micro-movements between intake-related micro-movements.

The proposed network's architecture consists of two consecutive LSTM layers with 128 hidden cells each, followed by a fully connected output layer with a single neuron. For the activation function of the recurrent steps we used the hard sigmoid defined as $\sigma(x) = \max(0, \min(1, x\,0.2 + 0.5))$, while for the output layer we used the sigmoid function. In a compact notation, the network can be written as $L(128) - L(128) - D(1)$, where $L(k_1)$ represents an LSTM layer with k_1 hidden cells and $D(k_2)$ a fully connected layer with k_2 neurons. The reason for using two LSTM layers stems from the work of Karpathy *et al.* [3], where the authors have shown that using a depth of at least two recurrent layers is beneficial when learning sequences.

Both intake and non-intake sequences are introduced to the network during training. Given the true label corresponding to each $\boldsymbol{s}_i$, a sequence of $\boldsymbol{s}_i, i = 1, 2, \ldots, n_j$ is considered an intake cycle if it starts with P (the first P in a sequence of P labels), ends with D (the last D in a sequence of D labels) and contains at least an M micro-movement. On the other hand, the remaining sequences that appear between consecutive intake cycles, are considered as non-intake cycles. We then represent each intake and non-intake sequence by their appropriate $n_j \times 10$ SVM score matrix. Since the input sequence of each LSTM layer is required to have a constant length, each sequence was pre-padded with zeros to a size $n' \times 10$, where $n' = \max\{n_j : j = 1, 2, 3 \ldots\}$. Thus, the input is long enough to contain every intake or non-intake sequence in the corpus. We used binary cross-entropy loss with the RMSprop optimizer (with 10^{-3} learning rate) that has demonstrated high effectiveness in a recurrent network topology [7]. Finally, the network is trained using an batch size M equal to 32 for 5 epochs.

2.5 Food intake cycle detection

Given the trained LSTM network and a sequence of $\boldsymbol{s}_i$ that represents a meal session, intake cycle detection is performed by extracting 3 second frames from the sequence of $\boldsymbol{s}_i$ with a step of 0.2 seconds. The extracted frames are then pre-padded with zeros to the target size $n' \times 10$ and given as input to the LSTM network. The network output $d[m]$ (i.e. the output of the sigmoid function)

Table 2. Details of the exhibited micro-movements in the food intake cycle dataset

Label	**Instances**	**Total duration (sec)**	**Mean (± std) duration (sec)**
P	727	1613.43	2.21 (± 1.71)
U	700	678.28	0.96 (± 0.58)
D	694	518.37	0.74 (± 0.57)
M	695	311.96	0.44 (± 0.17)
N	161	965.67	5.99 (± 5.71)
O	742	3837.92	5.17 (± 7.42)

represents the normalized probability that an input frame is an intake cycle. Subsequently, by replacing with zeros the elements of the $d[m]$ series that are lower than a threshold T_d, the filtered series $d'[m]$ is created. Finally, food intake cycles are detected by performing a local maximum search in $d'[m]$, with the minimum distance between two successive peaks set to 3 seconds. In particular, the timestamp corresponding to each local maximum (i.e. intake cycle) is the timestamp of the middle of the frame that produced the local maximum.

3 Dataset

In this study we used our publicly available FIC dataset. The FIC dataset consists of recordings from 10 subjects performing one meal session each, with an average duration of 13.2 minutes, in the restaurant of Aristotle University of Thessaloniki. The accelerometer and gyroscope streams originate from the Microsoft Band 2 smartwatch and are provided at a sample rate of approximately $62\,Hz$. The ground truth is provided at a micro-movement level based on analysis of video sequences captured during each subject's meal session. No specific instructions were given to the participating subjects other than clapping their hands once in the beginning and once in the end of the session for video/smartwatch synchronization purposes. Thus, the participants were able to engage in activities such as talking to other individuals in their proximity, during the recording. Table 2 provides additional information regarding the appearances of micro-movements in the dataset. Additionally, the average food intake cycle duration (from P to D) and the average distance between two consecutive food intakes were 5.39 (±3.86) and 11.22 (±8.79) seconds, respectively.

4 Experiments & results

Given the true start and end moments of the i-th food intake cycle, t_s^i and t_e^i respectively, as well as t_d^j the moment of the j-th detected intake cycle in the same meal session, performance metrics were calculated by the following evaluation scheme. If for a given true intake cycle i, t_d^j is outside $[t_s^i, t_e^i]$ for any detected intake cycle j, then it counted as a false negative. Otherwise it counted as a true positive. However, every other occurrence of detected intake cycle in

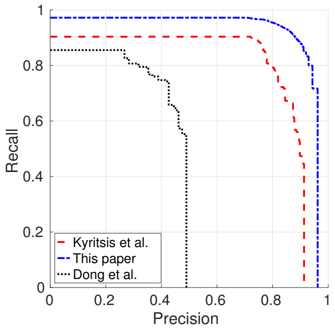

Fig. 3. Precision recall curves for the proposed approach (blue dash-dot line), the approach by [4] (red dash line) and by [2] (black dotted line).

the same $[t_s^i, t_e^i]$ interval counted as a false positive. Finally, if a detected intake cycle didn't belong in $[t_s^i, t_e^i]$ for any i, then it also counted as a false positive.

We used Leave One Subject Out (LOSO) cross validation for both training steps of the pipeline. As a result, for the evaluation of a single subject in the corpus, we trained ten SVM arrays, and one LSTM network. Since the LSTM is trained in a stochastic fashion, we repeated the LSTM training process for ten times, resulting in a total of 100 SVM arrays and 100 LSTM networks for the entire corpus. Experimentation with a small subset of the corpus led us to the selection of the C and γ parameters of the SVM to be equal to 100 and 0.1 respectively. Similarly, the threshold parameter T_d was set to 0.89 by picking the value that achieved the highest F1 score.

We used precision and recall for evaluation. The approaches of [2] and [4] were also implemented and evaluated against the same dataset. Parameter selection for those approaches was performed according to the authors' suggestions. Figure 3 depicts the precision-recall curves for all approaches, while Table 3 provides numerical results for the top F1 score. The decimals in the TP and FN columns arise from the averaging over the ten LSTM training repetitions.

5 Conclusions

We presented a method for detecting food intake cycles during a meal, using an off-the-shelf smartwatch. Results on a 10-subject publicly available corpus indicate that the combination of multiple micro-movement SVMs and an LSTM network for score sequence classification is highly effective and outperforms similar approaches found in the literature.

Acknowledgments. The work leading to these results has received funding from the European Community's Health, demographic change and well-being Programme under Grant Agreement No. 727688 (`http://bigoprogram.eu`),

Table 3. Evaluation results.

Method	**TP**	**FP**	**FN**	**Prec**	**Rec**	**F1**
Proposed approach	623.7	89	60.3	0.875	0.911	0.892
Approach by [4]	603	193	81	0.757	0.881	0.814
Approach by [2]	508	683	176	0.426	0.742	0.541

01/12/2016 - 30/11/2020. We gratefully acknowledge the support of NVIDIA Corporation with the donation of the Tesla K40 GPU used for this research.

References

1. Amft, O., Junker, H., Troster, G.: Detection of eating and drinking arm gestures using inertial body-worn sensors. In: Ninth IEEE International Symposium on Wearable Computers. pp. 160–163 (2005)
2. Dong, Y., Hoover, A., Scisco, J., Muth, E.: A new method for measuring meal intake in humans via automated wrist motion tracking. Applied psychophysiology and biofeedback 37(3), 205–215 (2012)
3. Karpathy, A., Johnson, J., Li, F.: Visualizing and understanding recurrent networks. CoRR abs/1506.02078 (2015), http://arxiv.org/abs/1506.02078
4. Kyritsis, K., Tatli, C.L., Diou, C., Delopoulos, A.: Automated analysis of in meal eating behavior using a commercial wristband imu sensor. In: 2017 39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC) (2017)
5. Madgwick, S.O.H., Harrison, A.J.L., Vaidyanathan, R.: Estimation of imu and marg orientation using a gradient descent algorithm. In: 2011 IEEE International Conference on Rehabilitation Robotics. pp. 1–7 (2011)
6. Mirtchouk, M., Merck, C., Kleinberg, S.: Automated estimation of food type and amount consumed from body-worn audio and motion sensors. In: Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing. pp. 451–462 (2016)
7. Ordez, F.J., Roggen, D.: Deep convolutional and lstm recurrent neural networks for multimodal wearable activity recognition. Sensors 16(1) (2016)
8. Papapanagiotou, V., Diou, C., Langlet, B., Ioakimidis, I., Delopoulos, A.: A parametric probabilistic context-free grammar for food intake analysis based on continuous meal weight measurements. In: 2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC) (2015)
9. Papapanagiotou, V., Diou, C., Zhou, L., van den Boer, J., Mars, M., Delopoulos, A.: A novel chewing detection system based on ppg, audio, and accelerometry. IEEE Journal of Biomedical and Health Informatics 21(3), 607–618 (May 2017)
10. Ramos-Garcia, R.I., et al.: Improving the recognition of eating gestures using intergesture sequential dependencies. IEEE Journal of Biomedical and Health Informatics 19(3), 825–831 (2015)
11. World Health Organization: Global health risks: mortality and burden of disease attributable to selected major risks. World Health Organization (2009)
12. Zhang, S., et al.: Food watch: Detecting and characterizing eating episodes through feeding gestures. In: Proceedings of the 11th EAI International Conference on Body Area Networks. pp. 91–96 (2016)