A Deep Network for Automatic Video-based Food Bite Detection

Dimitrios Konstantinidis[1], Kosmas Dimitropoulos[1], Ioannis Ioakimidis[2], Billy Langlet[2] and Petros Daras[1]

[1] CERTH-ITI, 6th km Harilaou-Thermi, 57001 Thessaloniki, Greece
[2] Karolinska Institutet, Blickagången 16, 14183 Huddinge, Sweden

Abstract. Past research has now provided compelling evidence pointing towards correlations among individual eating styles and the development of (un)healthy eating patterns, obesity and other medical conditions. In this setting, an automatic, non-invasive food bite detection system can be a really useful tool in the hands of nutritionists, dietary experts and medical doctors in order to explore real-life eating behaviors and dietary habits. Unfortunately, the automatic detection of food bites can be challenging due to occlusions between hands and mouth, use of different kitchen utensils and personalized eating habits. On the other hand, although accurate, manual bite detection is time-consuming for the annotator, making it infeasible for large scale experimental deployments or real-life settings. To this regard, we propose a novel deep learning methodology that relies solely on human body and face motion data extracted from videos depicting people eating meals. The purpose is to develop a system that can accurately, robustly and automatically identify food bite instances, with the long-term goal to complement or even replace manual bite-annotation protocols currently in use. The experimental results on a large dataset reveal the superb classification performance of the proposed methodology on the task of bite detection and paves the way for additional research on automatic bite detection systems.

Keywords: Deep Learning, Bite Detection, Video Analysis, Motion Features.

1 Introduction

Food intake is the aggregate of a complex array of eating behaviors, such as bites, chews and inter-bite pauses [1]. In this work, we are primarily concerned with the identification of bite instances that occur when a person opens his mouth for food intake. Studies have shown that increased food intake rate is directly linked to obesity both in children and adults [2][3]. Thus, bite detection mechanisms that allow food bite quantification and meal analysis can be invaluable for nutritionists, dietary experts and food scientists in order to evaluate individuals and help them avoid health problems related to obesity [4].

Currently, the detection of bite instances is usually performed by human experts, who have to watch hours of videos in order to successfully annotate eating behaviors [1][5]. Although the annotation of human experts can be really accurate, the annotation procedure is time-consuming and prone to introduce errors due to the repetitive nature of the task. Thus, the need for the automation of the procedure has often been emphasized by experts in the field [1][6] in order to overcome these problems and speed up the bite detection procedure. In the past several methodologies have been proposed to achieve that, with their majority being based on weight, inertial, motion and visual sensors that facilitate the recognition of bite instances [7].

However, existing methodologies face various challenges that limit their use and potential for large scale evaluation deployments or real life settings. More specifically, the mediocre accuracy of existing weight scales can significantly affect bite detection results. Furthermore, a person can eat with both hands either simultaneously or interchangeably and therefore, wearable sensors should be placed on both hands or else they fail to detect all bite instances. Additionally, sensors without visual feedback are prone to errors as instances of wiping mouth with handkerchief or raising hand to scratch head can be erroneously recognized as bites. Moreover, sensors that monitor jaw movements can become obtrusive, while visual sensors, such as cameras can pose challenges to a bite detection method due to the variety in appearance of people and kitchen utensils used for food intake and the occlusions of body parts. Finally, although the combination of multiple sensors can lead to more sophisticated solutions, it can also reduce the usability of the proposed systems in everyday life scenarios [8].

To overcome the aforementioned limitations of current automatic bite detection methodologies, we propose a novel non-obtrusive deep learning based approach that is capable of achieving highly accurate bite detection results. To this end, we initially employ a deep network [9][10] to extract human motion features from video sequences. Subsequently, we propose a novel two-steam deep network that processes body and face motion data and combines the extracted information, thus taking advantage of both types of features simultaneously. We evaluate the proposed method on a large bite detection dataset and validate its bite detection performance. The main contributions of this work are summarized below:

- This is the first video-based deep learning approach that utilizes body and face features for the task of automatic bite detection.
- We propose a sophisticated deep network that extracts and combines spatiotemporal information from its inputs using convolutional neural networks (CNNs) and Long Short-Term Memory (LSTMs) units.
- We perform optimization of the hyper-parameters of the proposed deep network and explore data augmentation techniques to further improve its accuracy.

The remainder of the paper is organized as follows. Section 2 reviews work related to ours with respect to automatic bite detection. Section 3 presents our proposed methodology, while Section 4 presents the experimental results from the evaluation of our method. Finally, Section 5 concludes the work.

2 Related Work

So far, there is limited work in the literature about automatic bite detection as obsolete sensor technology put great challenges to automatic bite detection systems. However, recent research works [2][3] revealed the strong correlation between food intake and obesity, thus intensifying the efforts towards the development of methods that study eating behaviors in order to prevent health related problems. Moreover, the development of modern sensors and the advancements in the classification techniques has sparked interest towards automatic bite detection methodologies that overcome the tediousness of having experts manually annotate bite instances.

In [11], the authors used a piezoelectric strain gauge sensor to detect the movement of the lower jaw that can characterize the eating behavior. On the other hand, in [5] and [12], the authors used high-precision food weight scales that can model the reduction of food from a subject’s plate, while in [13], the authors employ smart-glasses to detect bite instances. More recent studies employ inertial sensors, such as accelerometer and gyroscope, located in wearable devices, in order to automatically extract bite instances [8][14]. Finally, numerous studies develop methodologies that rely on the combination of audio and motion sensors [15][16] or the combination of multiple motion and gesture sensors [17][18] in order to robustly monitor eating behavior.

As far as classification techniques are concerned, early methods employ spectral segmentation and Random Forest classification [19] and Hidden Markov Models that are able to capture the temporal dependencies between hand gestures and bite instances [20]. Most recent bite detection systems take advantage of the success of deep learning on several classification tasks in order to propose more accurate and robust solutions. More specifically, the authors in [8] and [14], employ CNNs and recurrent neural networks in order to capture temporal dependencies between the outputs of inertial sensors and bite instances.

Our proposed methodology attempts to overcome the challenges of current state-of-the-art methods by proposing a video-based deep learning approach that extracts, processes and combines body and face motion data from video sequences. To our knowledge, our work comprises the first attempt to process videos and, more importantly, combine or fuse information extracted from body and face motion data to achieve accurate and robust automatic bite detection results.

3 Proposed Methodology

In this section, we analyze the proposed automatic bite detection methodology. Initially, we employ OpenPose [9][10] that is able to process videos and extract body and face features from each video frame, along with the confidence of the algorithm on its predictions. Afterwards, we propose a deep network that takes as input only the most relevant features for the task of bite detection out of those computed in the first step.

The proposed two-stream deep network is presented in Fig. 1 and gets as input 2 types of features: a) upper body and b) face features. More specifically, we employ the nose and hand features’ x- and y-coordinates and the distances between the selected numbered features, as shown in Fig. 2. For the face features, we employ only the x- and y-coordinates of the mouth features that we believe are the most relevant for the modelling of bite instances. We perform averaging between similar neighboring mouth features and end up with 3 features that describe the middle points of the upper and lower lips and the point where the lips converge (i.e., corner of mouth). Due to the fact that the videos of our dataset depict people eating from a side view, we select only the mouth corner with the highest confidence (A or C in Fig. 2). Furthermore, we also compute the distances between the 3 mouth features that are shown either as the triangle ABD or BCD in Fig. 2, based on the selected mouth corner. These mouth features are adequate in describing the basic movements of mouth during eating.

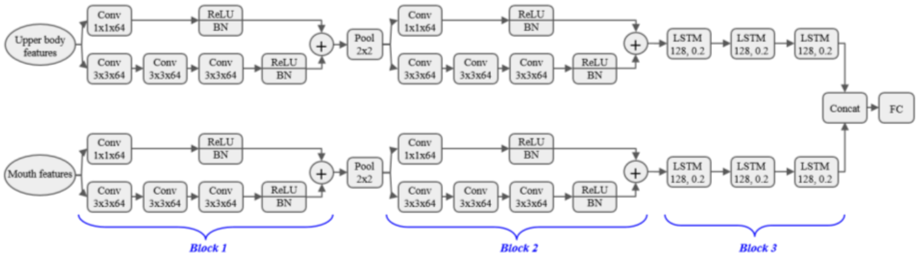

Fig. 1. Proposed deep network for automatic bite detection.

The proposed deep network processes the input features and extracts new discriminative ones that can better model the underlying spatiotemporal information. More specifically, the first two blocks of the proposed deep network employ stacked convolutional layers to compute spatial features (i.e., interactions between neighboring features both in time and space). The max-pooling layer between blocks 1 and 2 is responsible for down-sampling the feature space and improve the robustness of the network. Furthermore, the upper parts of blocks 1 and 2 are shortcut connections that tackle the vanishing gradient problem and improve the convergence of the network. The third block of the network employs a series of LSTM units that extract temporal features from the entire sequence of input features. A fusion of the two-streams of spatiotemporal features is performed in the end using a fully connected layer that combines the information from the hands, head and mouth and computes the probability of an input video sequence to describe a bite instance.

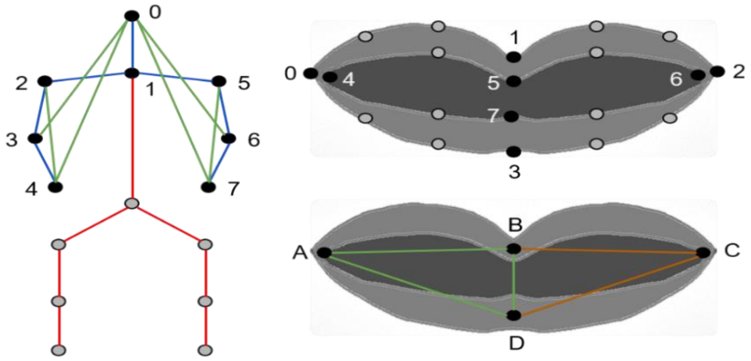

Fig. 2. Upper body (left) and mouth (right and down) human motion features that are used as input to the proposed deep network.

4 Experimental Evaluation

In this section, we evaluate our proposed network in the task of bite detection. Initially, we describe the dataset used for the evaluation of the proposed automatic bite detection method and then we present the preprocessing of features before their introduction to the proposed deep network. Moreover, we present the experiments for the optimization of the hyper-parameters of the proposed network, prior to the evaluation of our bite detector in the task of continuous bite detection.

4.1 Dataset

The evaluation dataset consists of 85 videos that depict people eating from a side view and it is part of the one used in [5]. It is a challenging dataset as there are occlusions of hands and mouth and features a variety of eating patterns and different subjects, types of meals (i.e., soup, breakfast and dinner) and kitchen utensils.

The dataset is manually annotated by human experts (i.e., nutritionists) and ground truth annotations of bite instances are provided. To train and validate our proposed network, we cropped a total of 12121 video clips with a duration of 2 seconds or 50 frames each and formed an isolated dataset. In this dataset, there are 4149 positive samples (i.e., annotated bite instances) and 7972 negative samples (i.e., randomly cropped non-bite instances). Additionally, this dataset is randomly split in a training set that contains 90% of samples and is used for training the proposed deep network and a validation set that contains the remaining 10% of the samples and is used for the optimization of the hyper-parameters of the proposed network.

4.2 Data preprocessing and augmentation

There are 25 body and 70 face features extracted from each video frame. We select only the 8 most relevant body (i.e., nose, neck, left and right hands) features and the 8 most relevant face (i.e., mouth) features as shown in Fig. 2. Furthermore, due to body parts' occlusions, there are features that are not detected or have abnormal values (i.e., outliers) and even frames with no detected features at all. To overcome this problem we apply two preprocessing stages to the extracted features. The first stage fills the empty values with values from the previous temporal instance (i.e., frame), while the second stage performs spline interpolation to the temporal sequences of features in order to remove outliers and smooth feature values across time. Moreover, in order to diminish the influence of the location of people in the videos, we perform normalization by transforming the coordinates of the selected features to a local coordinate system. More specifically, we assume as local origins the neck and nose for the upper body and mouth features respectively.

Although the size of the isolated bite detection dataset may seem quite large, it is not sufficient to properly train a deep network, like the one proposed in this work. To overcome this problem, we propose a data augmentation technique that is based on the manipulation of the temporal sequences of features in a way that adds variation to the input of the network and assists it in identifying the real difference between bite and non-bite instances and achieving higher accuracy in the task of bite detection.

More specifically, we propose two augmentation operations to the input temporal sequences of features. The first operation concerns the addition of a small displacement in the input, thus affecting the global location of features in the videos, while the second operation concerns the circular wrapping of the input by a certain small value in order to add variation to the temporal order of the feature values.

4.3 Hyper-parameter optimization

The optimization of the hyper-parameters of the proposed deep network is performed based on the performance of the network on the validation set. In our case, the parameters that are optimized are the number of CNN and LSTM layers since all the other parameters (i.e., number and size of kernels and filters and dropout percentages) are kept fixed to optimal values chosen based on our knowledge on the task and after initial experimentation.

The optimization of the number of layers is performed using the following approach. We consider a maximum number of 3 stacked layers and start by introducing one-by-one the recurrent layers (LSTMs) of block 3 and then add convolutional layers for blocks 1 and 2 as long as the performance of the network is increased. Finally, we add the shortcut connections of blocks 1 and 2. The experiments are performed on the initial isolated dataset (i.e., no data augmentation) as the experiments with data augmentation are performed on the optimized network. The performance of the proposed deep network on the validation set for different number of convolutional (Conv) and LSTM layers is presented in Table 1.

Table 1. Experiments with respect to the proposed network architecture.

Proposed network architecture			Performance on validation set
Block 1	Block 2	Block 3	
-	-	1 LSTM	0.81
-	-	2 LSTMs	0.807
-	-	3 LSTMs	0.847
1 Conv	-	3 LSTMs	0.857
2 Conv	-	3 LSTMs	0.864
3 Conv	-	3 LSTMs	0.875
3 Conv	1 Conv	3 LSTMs	0.882
3 Conv	2 Conv	3 LSTMs	0.894
3 Conv	3 Conv	3 LSTMs	0.922
3 Conv + shortcut	**3 Conv + shortcut**	**3 LSTMs**	**0.927**

From Table 1, we conclude that deeper networks with more parameters and thus processing capabilities and with the ability to extract both spatial and temporal information from their inputs achieve higher performance in the task of bite detection.

Furthermore, the shortcut connections are beneficial to the proposed network by slightly improving both its robustness and its classification accuracy.

As far as data augmentation is concerned, we test both data augmentation operations (i.e., displacement and circular wrapping). For the displacement operation, we compute a random displacement in the range [0, 0.1] for both x and y coordinates and add it to all features across time, thus not affecting their relative position. On the other hand, for the circular wrapping operation, we consider either 1 value (i.e., original sequence of features) or 3 values, namely {–5, 0, 5}, where 0 corresponds to the original sequence, while -5 and 5 corresponds to a circular wrapping of the temporal sequence by 5 positions (i.e., frames) to the left and right respectively. Table 2 presents the results of the experimentation with the data augmentation operations, where we can see that augmenting data with the displacement operation improves the performance of the proposed network. However, there is a threshold over which further increase of the displacement augmentation factor leads to a drop in the performance of the proposed network. This can be attributed to the fact that additional samples do not improve the requested variation of the input and the network has already learned the differentiation between bites and non-bites. On the other hand, the combination of the displacement and circular wrapping augmentation operations leads to a significant increase in the performance of the proposed network in the task of bite detection.

Table 2. Experiments with respect to the data augmentation operations.

Data augmentation operations			Performance on the validation set
Displacement factor	Circular wrapping factor	Combined factor	
1	1	1	0.927
3	1	3	0.939
5	1	5	0.937
3	**3**	**9**	**0.947**

4.4 Results on continuous bite detection

To evaluate the ability of the proposed methodology to identify bite instances in a continuous fashion, we test our deep network on the 85 continuous videos of the provided dataset [5]. To achieve this, we employ an overlapping sliding window of 60 frames, which is slightly larger than the clips of 50 frames used for training and as we want smoother output probabilities from our network. Additionally, we employ a step of 1 frame, meaning that a bite detection probability is computed for each frame, taking also into account its neighboring frames.

Since the output of the proposed network is a continuous signal of bite detection probabilities, post-processing should be applied to detect exact locations of bite instances and remove false alarms. Initially a n^{th}-order median filter is applied to smooth signal and remove small and abrupt changes (i.e., sawtooth effect and outliers). Afterwards, the mean m and standard deviation s of the signal are computed and all predictions below the threshold of $m + s$ are zeroed. Then, all local maxima (i.e., peaks) of the signal are detected and all peaks with width below a threshold T_W are removed. Finally, the distance between peaks is computed and for peaks having distance smaller than a threshold T_D, we preserve only the peak with the highest probability to belong to the bite instance. We set the order of the media filter n to 40 so as to achieve heavy smoothing and the distance threshold T_D to 50 frames as we believe that under normal circumstances a person cannot receive two bites in less than 2 seconds time difference. Finally, after experimentation, we set the width threshold T_W of the peaks to 22 frames, which describes the duration of a clear bite instance and corresponds to almost 1 second.

Table 3 presents the overall performance of the proposed method in the task of continuous bite detection. We can observe that our method achieves a bite detection rate (i.e., recall) of 91.71% with a false alarm rate of 8.25%, which means that our method is a very accurate and robust bite detector. Furthermore, Fig. 3 shows a histogram of the distribution of F1-scores (i.e., harmonic averages of recall and precision) for the tested videos. This figure shows that our proposed method achieves superb results on most videos, while there are only a few videos, in which our method achieves mediocre results. Finally, two examples of the predicted bite detection probabilities that our method outputs overlaid on the ground truth are presented in Fig. 4.

Table 3. Experimentation with continuous automatic bite detection.

Recall	Precision	F1-score
0.9171	0.9175	0.9173

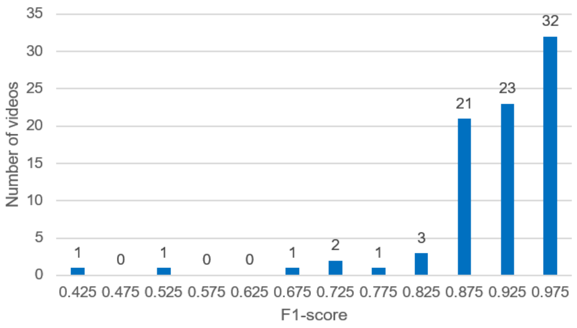

Fig. 3. Distribution of videos based on their F1-score. Total number of videos: 85

5 Conclusions

In conclusion, we present in this work a methodology for accurate bite detection results. Our method is the first one that extracts human body and face motion features from videos and uses them as input to a deep network. Experiments on both isolated and continuous datasets show the superb performance of the proposed bite detector and paves the way for research on additional features and other deep network architectures for the task of bite detection.

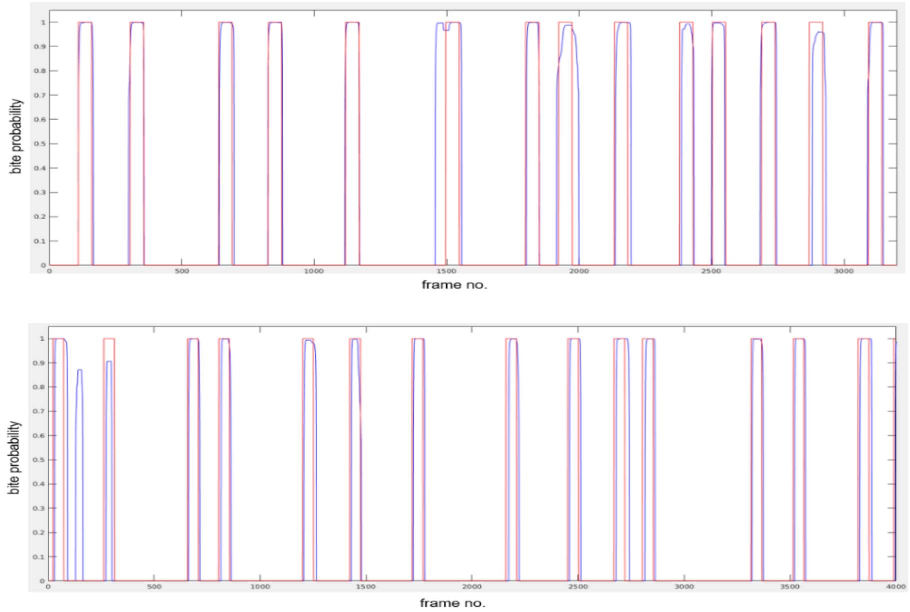

Fig. 4. Bite prediction results (blue signal) overlaid on ground truth (red signal) for two videos of the dataset.

Acknowledgement

This work was supported by the European Project: PROTEIN Grant no. 817732 with the H2020 Research and Innovation Programme.

References

1. Ioakimidis, I., Zandian, M., Eriksson-Marklund, L., Bergh, C., Grigoriadis, A. and Södersten, P: Description of chewing and food intake over the course of a meal, Physiol. Behav., 104:761–769 (2011).
2. Fogel, A., Goh, A.T., Fries, L.R., Sadananthan, S.A., Velan, S.S., Michael, N., Tint, M.T., Fortier, M.V., Chan, M.J., Toh, J.Y., et al: A description of an "obesogenic" eating style that promotes higher energy intake and is associated with greater adiposity in 4.5 year-old children: Results from the GUSTO cohort, Physiol. Behav., 176:107–116 (2017).
3. Ohkuma T, Hirakawa Y, Nakamura U, Kiyohara Y, Kitazono T and Ninomiya T.: Association between eating rate and obesity: a systematic review and meta-analysis, International journal of obesity, 39(11):1589 (2015).
4. Fagerberg, P., Langlet, B., Glossner, A. and Ioakimidis, I.: Food Intake during School Lunch Is Better Explained by Objectively Measured Eating Behaviors than by Subjectively Rated Food Taste and Fullness: A Cross-Sectional Study, Nutrients, 11(3) (2019).
5. Langlet, B., Tang Bach, M., Odegi, D., Fagerberg, P., and Ioakimidis, I.: The Effect of Food Unit Sizes and Meal Serving Occasions on Eating Behaviour Characteristics: Within Person Randomised Crossover Studies on Healthy Women. Nutrients, 10(7), 880, (2018).

6. Hermsen, S., Frost, J.H., Robinson, E., Higgs, S., Mars, M. and Hermans, R.C.J.: Evaluation of a Smart Fork to Decelerate Eating Rate, Journal of the Academy of Nutrition and Dietetics, 116(7):1066-1067 (2016).
7. Theodoridis, T., Solachidis, V., Dimitropoulos, K., Gymnopoulos, L. and Daras, P.: A Survey on AI Nutrition Recommender Systems, 12th International Conference on Pervasive Technologies Related to Assistive Environments Conference, Rhodes, Greece, (2019).
8. Kyritsis, K., Diou, C. and Delopoulos, A.: Food Intake Detection from Inertial Sensors Using LSTM Networks, New Trends in Image Analysis and Processing (ICIAP), Springer International Publishing, pp. 411-418 (2017).
9. Simon, T., Joo, H., Matthews, I. and Sheikh, Y.: Hand keypoint detection in single images using multiview bootstrapping, IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, pp. 4645-4653 (2017).
10. Cao, Z., Simon, T., Wei, S. and Sheikh, Y.: Realtime Multi-person 2D Pose Estimation Using Part Affinity Fields, IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, pp. 1302-1310 (2017).
11. Sazonov, E., and Fontana, J.: A Sensor System for Automatic Detection of Food Intake Through Non-Invasive Monitoring of Chewing, IEEE sensors journal. 12:1340-1348 (2012).
12. Papapanagiotou, V., Diou, C., Langlet, B., Ioakimidis, I. and Delopoulos, A.: A parametric probabilistic context-free grammar for food intake analysis based on continuous meal weight measurements. In Engineering in Medicine and Biology Society (EMBC), 37th Annual International Conference of the IEEE, pp. 7853–7856 (2015).
13. Zhang R. and Amft, O.: Monitoring chewing and eating in free-living using smart eyeglasses, IEEE journal of biomedical and health informatics, 22(1): 23–32 (2018).
14. Kyritsis, K., Diou, C. and Delopoulos, A.: End-to-end Learning for Measuring in-meal Eating Behavior from a Smartwatch, 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), Honolulu, HI, pp. 5511-5514 (2018).
15. Papapanagiotou, V., Diou, C., Zhou, L., Van Den Boer, J., Mars, M. and Delopoulos, A.: A novel chewing detection system based on ppg,audio, and accelerometry, IEEE journal of biomedical and health informatics, 21(3): 607–618 (2017).
16. Mirtchouk, M. Merck, C. and 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).
17. Doulah, A., Farooq, M., Yang, X., Parton, J., McCrory, M.A, Higgins, J.A. and Sazonov, E.: Meal microstructure characterization from sensor-based food intake detection, Frontiers in nutrition, vol. 4, p. 31 (2017).
18. Fontana, J.M., Farooq, M., Sazonov, E.: Automatic ingestion monitor: A novel wearable device for monitoring of ingestive behavior, IEEE Trans. Biomed. Eng. 61:1772–1779 (2014).
19. Zhang, S., Stogin, W. and Alshurafa, N.: I sense overeating: Motif-based machine learning framework to detect overeating using wrist-worn sensing, Information Fusion, vol. 41, pp. 37–47 (2018).
20. Ramos-Garcia, R.I., Muth, E.R., Gowdy, J.N. and Hoover, A.W.: Improving the recognition of eating gestures using intergesture sequential dependencies, IEEE journal of biomedical and health informatics, 19(3): 825–831 (2015).