ORIGINAL RESEARCH

A deep learning based wearable system for food and drink intake recognition

Dario Ortega Anderez[1] · Ahmad Lotfi[1] · Amir Pourabdollah[1]

Abstract

Eating difficulties and the subsequent need for eating assistance are a prevalent issue within the elderly population. Besides, a poor diet is considered a confounding factor for developing chronic diseases and functional limitations. Driven by the above issues, this paper proposes a wrist-worn tri-axial accelerometer based food and drink intake recognition system. First, an adaptive segmentation technique is employed to identify potential eating and drinking gestures from the continuous accelerometer readings. A posteriori, a study upon the use of Convolutional Neural Networks for the recognition of eating and drinking gestures is carried out. This includes the employment of three time series to image encoding frameworks, namely the signal spectrogram, the Markov Transition Field and the Gramian Angular Field, as well as the development of various multi-input multi-domain networks. The recognition of the gestures is then tackled as a 3-class classification problem ('Eat', 'Drink' and 'Null'), where the 'Null' class is composed of all the irrelevant gestures included in the post-segmentation gesture set. An average per-class classification accuracy of 97.10% was achieved by the proposed system. When compared to similar work, such accurate classification performance signifies a great contribution to the field of assisted living.

Keywords Gesture recognition · Accelerometer · Deep learning

1 Introduction

Eating difficulties are those that alone or combined, hamper the preparation or the intake of food and/or beverages (Westergren 2001), with major causes including cognitive impairment, poor appetite or feeding dependency. Incidentally, a poor diet can contribute to weight loss and malnutrition, leading to potential functional limitations, metabolic abnormalities and diminished immunity (Payette and Shatenstein 2005). Recent statistics outline eating difficulties as a prevalent issue among the elderly population. For instance, the survey conducted in Westergren et al. (2002) with 520 elderly patients in hospital rehabilitation, reveals 82% of

the patients exhibit some form of eating difficulty. The survey conducted in Lohrmann et al. (2003), including 3000 patients from 11 different hospitals, acknowledge 21.1% of the patients younger than 80, and 36.4% of those aged 80 or older require eating assistance.

As of now, dietary behaviour is generally monitored by the use of self-assessment questionnaires. Though, two major shortcomings are found on the use of these conventional approaches. First, the data entry process may result cumbersome, since normally the questionnaires have to be filled manually by the subjects. Second, various studies indicate self-reported estimates of daily activities are subjective and variable (Smith et al. 2005; Rush et al. 2008).

The rapid technological development in ubiquitous computing seen in recent years is translating into an increasing research attention towards Human Activity Recognition (HAR) (Lee et al. 2011; Gayathri et al. 2015; Adama et al. 2018; Ortega-Anderez et al. 2019; Anderez et al. 2020; Casella et al. 2020). Current portable or wearable devices such as smart-phones and smart-watches already integrate a broad array of sensors (i.e. accelerometers, magnetometers, gyroscopes), allowing for human behaviour analysis and monitoring in applications such as health care and

✉ Dario Ortega Anderez
   dario.ortegaanderez2013@my.ntu.ac.uk

   Ahmad Lotfi
   ahmad.lotfi@ntu.ac.uk

   Amir Pourabdollah
   amir.pourabdollah@ntu.ac.uk

[1] School of Science and Technology, Nottingham Trent University, Clifton Campus, College Drive, Nottingham NG11 8NS, UK

well-being monitoring. In line with this, this paper proposes a two-fold wrist-worn accelerometer based food and drink intake recognition system in support of elderly well-being.

First, the Crossings-based Adaptive Segmentation Technique (CAST) (Anderez et al. 2018b, 2020) is employed for spotting potential eating and drinking gestures from the continuous accelerometer readings. Despite the sparse occurrence of food and drink intake gestures, this segmentation technique has shown the capability of retrieving 100% of the eating and drinking gestures embedded within accelerometer signals. Once the segment set containing potential eating and drinking gestures is built, a 1D CNN fed with the raw accelerometer data is parametrically optimised and proposed as a benchmark classification model. A posteriori, various multi-input-multi-domain networks are proposed on top of the benchmark model for classification performance comparison purposes. These include the employment of three different time series to image encoding frameworks, namely the signal spectrogram, the Markov Transition Field (MTF) and the Gramian Angular Field (GAF), as well as a 31-dimensional hand crafted feature vector. The problem is ultimately tackled as a 3-class supervised classification problem ('Drink', 'Eat' or 'Null'), where the 'Null' class embodies all the irrelevant gestures retrieved by the signal segmentation technique. That is, any gesture which is not an eating or a drinking gesture.

The main contributions of this paper are as follows. First, we provide a thorough investigation upon feature extraction and the optimisation of Convolutional Neural Networks for food and drink intake gesture recognition which can serve as a point of direction for future research in the field. Such investigation includes:

- Hyper-parameter optimisation (the number of layers, the number of filters and the filter size) of a 1D CNN fed with raw accelerometer data.
- Employment of three time series to image encoding frameworks for feature extraction, namely the signal spectrogram, the Markov Transition Field (MTF) and the Gramian Angular Field (GAF). Despite the good performance exhibited in other applications, to the best of our knowledge, the MTF and the GAF have not been employed by any gesture or activity recognition research work.
- Development and comparison of a wide range of novel multi-input multi-domain deep learning frameworks for gesture recognition.

Ultimately, this paper provides an accurate unobtrusive solution for the recognition of eating and drinking gestures which outperforms most previous similar work. Given the unobtrusiveness of the solution and the recognition performance achieved, this work signifies a great step towards the field of Ambient Intelligence in the form of a monitoring system for the analysis of personal dietary behaviours.

The rest of the paper is organised as follows. Section 2 provides a critical analysis upon the use of CNNs for gesture and activity recognition purposes, as well as a review of previous research work on the recognition of eating and drinking gestures using wearable sensors. Section 3 describes the methodology employed to implement the food and drink intake recognition system. Section 4 presents the results achieved by the use of the various CNN-based classification frameworks proposed. Section 5 draws the conclusions from the results obtained.

2 Previous work

This section is divided into three different parts. In Sect. 2.1, previous literature on the use of convolutional neural networks for activity and gesture recognition is discussed. Section 2.2 reviews previous work on eating and drinking recognition with the use of wearable sensors. Ultimately, the motivation for undertaking this work is presented in Sect. 2.3.

2.1 The use of convolutional neural networks for activity recognition

The use of deep learning (Zeng et al. 2018, 2019, 2020), and specially that of Convolutional Neural Networks (CNNs), has revolutionised the state-of-the-art of challenging problems such as speech recognition and image classification (Ronao and Cho 2015). Likewise, CNNs are gaining increasing attention within the field of HAR due to the numerous advantages they provide as compared to traditional state-of-the-art HAR feature extraction and classification methods. First, conventional HAR solutions typically require the computation of hand-crafted or self-engineered features, thus relying on human domain knowledge. Second, according to human expertise, only shallow features, such as basic signal statistics, can be learned through the use of conventional hand-crafted feature extraction methods (Yang et al. 2015). Despite the good performance exhibited by the use of shallow features on the recognition of low-level activities such as walking, sitting or jogging, gaining insights into context-aware activities such as using the toilet or having lunch, may require more complex computations (Wang et al. 2019). Third, in contrast to traditional HAR approaches, CNNs are able to exploit the translation invariant nature of human gestures/activities as well as the local dependency attribute of temporal sequences (Ronao and Cho 2015).

The advantages presented above have recently deviated the attention of human activity/gesture recognition research work towards the implementation of CNN

frameworks, which as proven by recent work in the field (Duffner et al. 2014; Yang et al. 2015; Ignatov 2018), can outperform traditional state-of-the-art approaches such as Random Forest, Support Vector Machines or K-Nearest Neighbours. However, despite the good performance exhibited by CNNs, major discrepancies are found among the literature.

One of such discrepancies is found on the segmentation of the sensory signals, which is mainly due to the differing duration of different gestures or activity cycles. While excessively short segments would miss fundamental characteristics of a gesture/activity, long sequences may retrieve characteristics from multiple gestures/activities, thus lowering the ultimate classification performance. Generally, the length of the segments is either roughly estimated based on the characteristics of the gesture or activity set studied (Ronao and Cho 2015, 2016), or calculated as a hyper-parameter of the classification problem itself (Lee et al. 2017; Ignatov 2018).

Different approaches are also found on the pre processing of the signals. Typically, 1D filters are directly used on the raw sensor data (Ronao and Cho 2015; Yang et al. 2015; Ronao and Cho 2016; Ignatov 2018; Anderez et al. 2019). However, alternative solutions have also been proposed. In Lee et al. (2017), the accelerometer signals are unified into the magnitude of the tri-dimensional vector. While this approach can leverage the computational cost of the network, a poor performance (classification accuracy = 92.95%) at recognising a basic set of three high-level activities, suggests that crucial information is thrown away at such unification step. Various studies employing multiple sensor nodes for data collection (Jiang and Yin 2015; Ha et al. 2016), propose time series to image encoding frameworks to capture the spatial dependency between the different sensors, as well as the local dependency over time. A posteriori 2D CNNs are used for feature learning and classification. As proven in Ha et al. (2016), 2D CNNs can outperform 1D CNNs on time series classification tasks, however, the exhibited improvement is considerably low.

Ultimately, the network architecture has also varied considerably between different HAR works. While some studies propose shallow networks with only one convolutional layer (Lee et al. 2017; Ignatov 2018), other studies have opted for the employment of networks with 2 convolutional layers (Jiang and Yin 2015; Ha et al. 2016) or yet deeper architectures (Yang et al. 2015; Ronao and Cho 2015). In theory, increasing the number of convolutional layers allows for the computation of more complex features, which as shown in Ronao and Cho (2015), can lead to a better classification performance. However, employing deep architectures may also lead to network overfitting and consequently to a worse classification performance (Ignatov 2018).

2.2 Eating and drinking recognition with the use of wearable sensors

Eating and drinking recognition can be considered alone a research area within the human activity recognition field. This is mainly due to most of the activities studied by HAR work exhibit a quasi-periodic nature (e.g. walking, jogging), whereas eating and drinking are composed of sparsely occurring gestures embedded in continuous data streams.

Various solutions for gesture recognition have been proposed in the past years. The work in Chen et al. (2017) proposes a sliding-window segmentation approach alongside a hand-crafted feature vector and a SVM classification model to recognise drinking gestures from signals collected by a single wrist-worn inertial sensor. A classification recall of 91.3% is claimed by this method. However, despite the good classification performance achieved, the experiments are run under a extremely constrained scenario where the chairs are height adjusted to each individual and the experimental data set lacks of a 'Null' class. The work in Schiboni and Amft (2018) proposes a Gaussian Mixture Hidden Markov Model (GMM-HMMs) network for the recognition of drinking gestures. The experimental data is collected from seven participants following their usual daily routines while wearing a single wrist-worn inertial sensor which embodies a tri-axial accelerometer, a tri-axial magnetometer and a tri-axial gyroscope. A classification precision of 75.2% and a classification recall of 76.1% are reported in this work. Another drinking recognition solution is proposed in Amft et al. (2010). The experimental data is collected from six participants wearing a single wrist-worn inertial unit containing a tri-axial accelerometer, a tri-axial compass and a tri-axial gyroscope while performing a set of various free-living scenarios. A Feature Similarity Search (FSS) is a posteriori used for the classification of the gestures, achieving a classification recall of 84.0% .

The work in Dong et al. (2014) presents a two-fold approach for the recognition of meal periods using data from a single wrist-worn inertial sensor. A wrist motion energy-based custom-peak segmentation technique is proposed to identify potential time windows containing meal periods. Once the potential eating periods are identified, a 4-dimensional feature vector alongside a Naive Bayes classifier are used for the ultimate classification. A classification recall of 81.0% is achieved by this work. In Junker et al. (2008), a gesture recognition system to identify a set of four dietary gestures ('drink', 'cutlery', 'spoon' and 'hand-held') from data collected from five inertial units (one on the trunk and two on each arm) is proposed. First, a two-fold gesture spotting approach based on the sliding-window and bottom-up segmentation technique (Keogh et al. 2004) and a FSS, is used to identify potential eating and drinking gestures. A posteriori a Hidden Markov Model (HMM) is used for

classification, achieving a classification precision of 73.0% and a classification recall of 79.0%. The work in Anderez et al. (2020) proposes an accelerometer-based adaptive segmentation technique (CAST) to identify potential eating and drinking gestures embedded in the continuous sensor readings. A posteriori, a soft Dynamic Time Warping (DTW)-based gesture discrepancy measure alongside a hand-crafted feature vector are used to train a range of different classifiers. The best results are obtained using a Deep Neural Network, which exhibits an average per-class classification accuracy of 98.2%, a classification precision of 95.7% and a classification recall of 95.0%.

2.3 Research motivation

Different limitations are found within the above reviewed works. First, various eating and drinking recognition systems still rely on the use of several sensor units (Junker et al. 2008; Anderez et al. 2018a), making such solutions excessively intrusive for their daily use. Second, some studies rely on experimental work undertaken in extremely constrained environments. For instance, in Chen et al. (2017), the chairs are height-adjusted to individuals. Besides, the individuals are instructed as to how to perform the drinking gestures and the experimental dataset is only composed of drinking gestures. In addition, the performance of gesture recognition systems under unconstrained environments still lies far away from that achieved by HAR systems. The sparse occurrence of gestures and the subsequent difficulty to develop adaptive segmentation techniques to accurately spot such gestures, generally translates into true positive missing at this preliminary segmentation (spotting) step, which then further propagates to the final classification step. For instance, the results in Amft et al. (2010) show an 84% recall at spotting drinking gestures. The work in Junker et al. (2008) obtains an 80% spotting recall. Besides, accurate eating and drinking recognition systems still rely on specific domain knowledge (Anderez et al. 2020).

To our view, gesture spotting and recognition experimental work should be undertaken in realistic scenarios where the participants can freely perform the proposed activities/gestures. Moreover, the resultant experimental data sets should include a reasonable 'Null' class so that the implemented systems face the challenges one would expect to encounter in real life.

Overall, the different drawbacks found within different reviewed works suggests there are still many open challenges on the implementation of systems for eating and drinking gesture recognition, as well as on the deployment of suitable CNN architectures for activity/gesture recognition. In line with this, this paper presents a CNN-based eating and drinking gesture recognition system which aims at overcoming the above-mentioned drawbacks. To do so, first an adaptive

segmentation technique is proposed to mitigate the problem present on sliding window-based segmentation approaches. Second, the study presented here aims to unravel a great array of unanswered questions with regards to the use of CNNs for gesture recognition, as well as to propose an accurate domain knowledge independent eating and drinking recognition system.

3 Methods

This section presents the methodology employed to develop the proposed fluid and food intake recognition system. The section is divided regarding the different methodology phases as follows. Section 3.1 presents the experimental setup, Sect. 3.2 describes the signal pre-processing steps employed to accommodate the raw signals for network fitting, Sect. 3.3 defines the time series to image encoding frameworks employed, Sect. 3.4 describes the single-input and the multi-input multi-domain CNN-based frameworks proposed for gesture classification.

3.1 Experimental setup

The experiment conducted embodied 6 volunteers (5 male and 1 female) having a meal which included crisps, soup, chicken breast and cake. The participants wore a wrist-worn tri-axial accelerometer (sample frequency 25 Hz) on their dominant hand while having the meal. The data provided by the accelerometer therefore embodied three different time series, namely a_x, a_y, a_z, which correspond respectively to the medio-lateral, antero-posterior and vertical acceleration inputs of the dominant wrist of the participants as these move about in space during the experiment. Before the meal took place, the participants were asked to move and act freely around the house for unlimited time. This ensured the dataset embodied a wide 'Null' class composed of irrelevant gestures from a variety of other quotidian activities, so that the system development accounts for the challenges one would expect to face in real life. Given the food provided, the experiment included the use of various utensils. Moreover, the utensils provided differed between different participants (i.e. various participants used a mug to drink water while others used a glass), therefore incorporating inter-utensil variability. Furthermore, one left-handed person took part on the experiment, thus adding extra variability to the dataset. The labelling of the gestures was a posteriori aided by the use of video recordings, whereby a gesture was classified regarding the type of gesture that had caused the peak on the acceleration on the y-axis. With this, a total of 587 "Null" gestures, 59 "Drinking" gestures and 167 "Eating" gestures were retrieved for further feature extraction and classification, with an average segment length of 1.22 seconds

for the "Null" class, 4.51 seconds for the "Drinking" class and 3.12 seconds for the "Eating" class.

3.2 Signal pre-processing

The signal pre-processing process is divided into three different stages: signal shift, signal segmentation and segment padding.

3.2.1 Signal shift

Eating and drinking gestures generally require a movement of the dominant hand towards the mouth. Given the orientation shift of the y-axis when the wrist-worn accelerometer is worn by a left handed person, a 180° shift is applied to the signal corresponding to the accelerometer y-axis from the data collected from the left-handed participant.

3.2.2 Signal segmentation

The aim of signal segmentation is to either break down the signal into segments that share a common characteristic or to filter out unwanted segments of the signal. In this case, an adaptive segmentation technique, namely the CAST (Anderez et al. 2020), is employed to identify potential segments containing an eating or a drinking gesture. This technique makes use of the crossings between two moving averages $\bar{y}_1[t]$ and $\bar{y}_2[t]$ (fast and slow respectively) to identify those potential eating and drinking gestures. The moving averages are calculated over the accelerometer y-axis signal as:

$$\bar{y}[t] = \frac{1}{n} \sum_{i=0}^{n-1} y[t-i] \tag{1}$$

where n is the number of data samples over which the moving average is calculated.

In this case, after the experimental work undertaken in Anderez et al. (2020), $n = 25$ (1 s) and $n = 150$ (6 s) are used for the calculation of $\bar{y}_1[t]$ and $\bar{y}_2[t]$ respectively.

The intuition behind this technique is the sequence of movements which compose a complete eating or drinking gesture. First, one has to grasp the corresponding piece of food or utensil (i.e. spoon), then such food or utensil is taken to the mouth and ultimately, the hand is returned to the rest position. The presented sequence of movements leads to a cross-over of the fast moving average $y[1]$ to the slow moving average $y[2]$ when the hand is moving towards the mouth, followed by the subsequent cross-down when the hand is moved back to the rest position. This can be observed on the example of the performance of the CAST at spotting a drinking gesture and an eating gesture depicted in Fig. 1.

3.2.3 Segment padding

Contrary to traditional sliding-window approaches, the CAST adapts to the duration of the gestures themselves, leading to a gesture set of signal segments with varying lengths. The segments are a posteriori padded to the length of the longest segment retrieved by the CAST ($n = 394$) to allow for network batch training on the 1D CNN. The GAF and the MTF time series to image encoding frameworks utilise such padded segments of length ($n = 394$). In the case of the signal spectrogram framework, n is rounded up to the nearest higher power of 2 ($n = 512$).

Fig. 1 Example of the performance of CAST at spotting a drinking gesture and an eating gesture
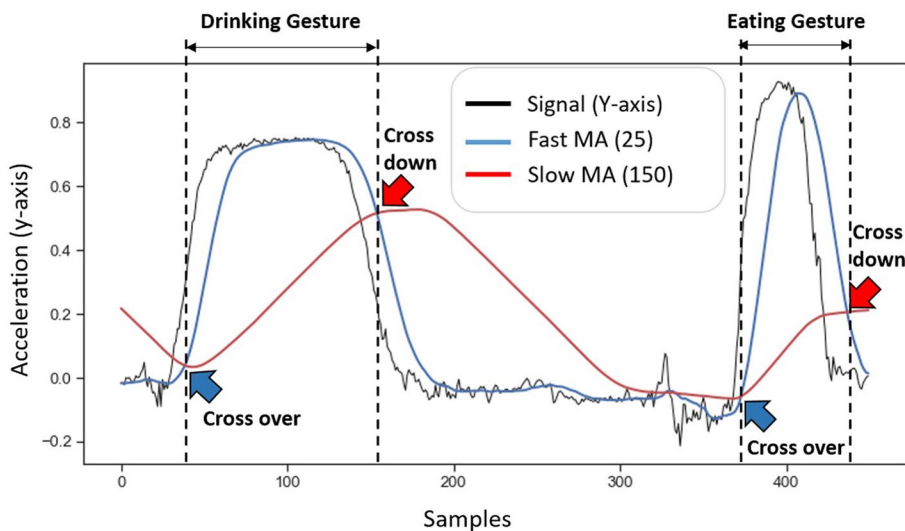
3.3 Time series imaging

Inspired by the work in Wang and Oates (2015); Lawal and Bano (2019), three different frameworks are employed for encoding the accelerometer signal segments into images, namely the signal spectrogram, the Markov Transition Field (MTF) and the Gramian Angular Field (GAF). In this work, the image encoding is independently applied to the magnitude of the 3-dimensional accelerometer signal as well as to the y-axis signal (previously employed for signal segmentation). Different pictorial examples of the time series to image encoding frameworks employed on the different gesture classes are shown in Fig. 2.

3.3.1 Signal spectrogram

The signal spectrogram is a visual representation which depicts the strength spectrum of frequencies of a signal as it varies with time. Given a time series $X = \{x_1, x_2, \ldots, x_n\}$,

X is first converted into the frequency domain using the Fast Fourier Transform (FFT) as follows:

$$FFT(X) = \frac{\sum_{k=1}^{W_l} |a_k|^2}{W_l} \tag{2}$$

where $a1, a2, \ldots, a_{W_l}$ are the FFT components of the corresponding window of length W_l. In this case, a window length W_l of 32 samples with 50% overlapping is used across the padded segments ($N = 512$).

A posteriori, the signal spectrogram is calculated as follows:

$$spectrogram\{x(t)\}(\tau, \omega) = |X(\tau, \omega)|^2 \tag{3}$$

Eventually, the resulting signal spectrogram is encoded into a 2-dimensional (time and frequency) graph, with a third dimension (signal amplitude of a particular frequency at a specific time) represented by a colour scale.

Fig. 2 Examples of the employed imaging techniques for each of the classes ('Drink', 'Eat', 'Null'). In the examples provided, the plot and the corresponding spectrogram, MTF and GAF, are visual representations of the y-axis of the accelerometer signal
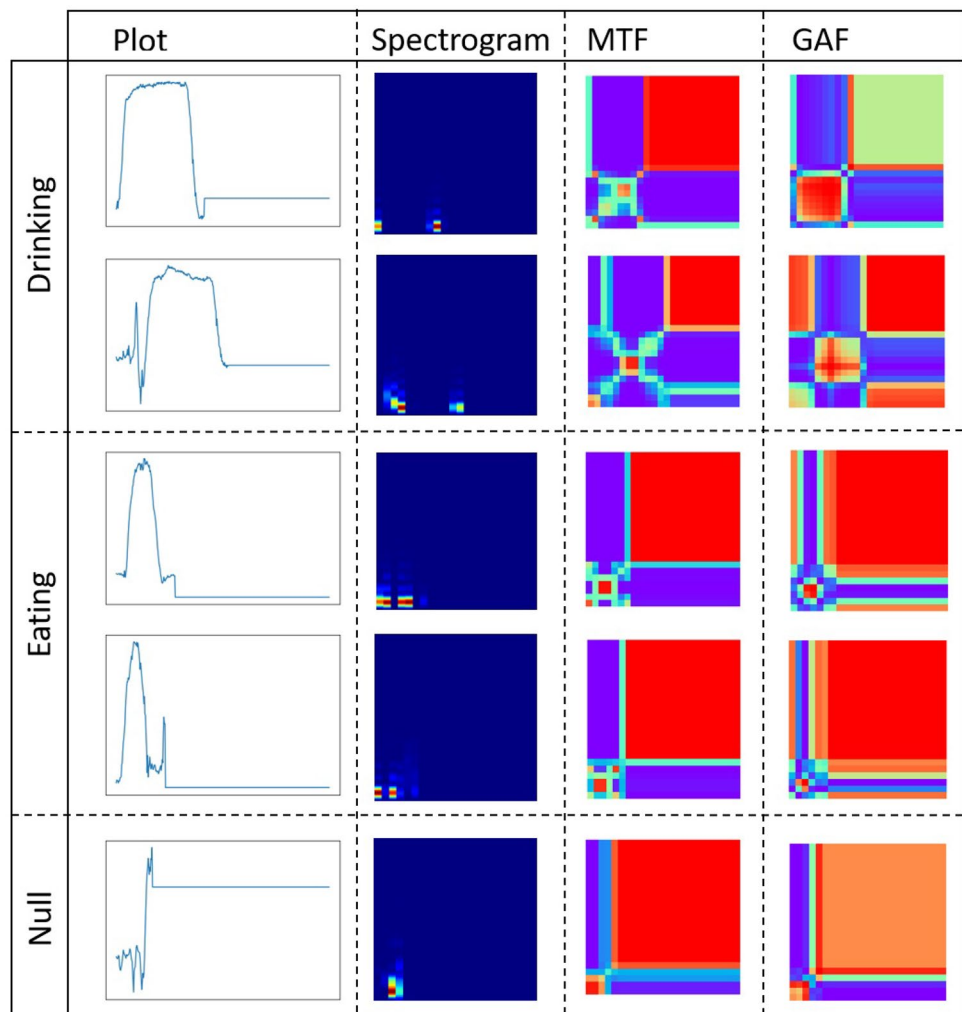
3.3.2 Markov Transition Field

The Markov Transition Field (MTF) framework is employed to encode dynamical transition statistics of the signal. To preserve the sequential information enclosed within the signal, the framework proposed by Wang and Oates (2015) is employed, whereby the Markov transition probabilities are represented sequentially, thus preserving information in the time domain. Given a time series $X = \{x_1, x_2, ..., x_n\}$, Q quantile bins are identified and each x_i is assigned to the corresponding bins q_j ($j \in [1, Q]$). A posteriori a $Q \times Q$ weighted adjacency matrix W is constructed with the count of the transitions among quantile bins in the form of a first order Markov chain along the time axis. $w_{i,j}$ is then estimated as the frequency at which a point in the quantile q_j is followed by a point in the quantile q_i. This, after normalisation $\sum_j w_{i,j} = 1$ gives as a result the Markov transition matrix W. However, W is insensitive to the distribution of X and the temporal dependency on time steps t_i.

To overcome the loss of the temporal dependency, the Markov Transition Field (MTF) matrix M is defined as follows:

$$M = \begin{bmatrix} w_{ij|x_1 \in q_i, x_1 \in q_j} & \cdots & w_{ij|x_1 \in q_i, x_n \in q_j} \\ w_{ij|x_2 \in q_i, x_1 \in q_j} & \cdots & w_{ij|x_2 \in q_i, x_n \in q_j} \\ \vdots & \ddots & \vdots \\ w_{ij|x_n \in q_i, x_1 \in q_j} & \cdots & w_{ij|x_n \in q_i, x_n \in q_j} \end{bmatrix} \quad (4)$$

The $Q \times Q$ Markov transition matrix (W) is computed by dividing the data into Q quantile bins, where the quantile bins that contain the data at time stamp i and j are q_i and q_j respectively ($q \in [1, Q]$). M_{ij} in MTF denotes the transition probability of $q_i \rightarrow q_j$. That is, the matrix W is spread out into the MTF matrix M by considering temporal position.

3.3.3 Gramian Angular Field

Given a time series $X = \{x_1, x_2, ..., x_n\}$ where each x_i is normalised as:

$$\tilde{x}_i = \frac{(x_i - max(X) + (x_i - min(X))}{max(X) - min(X)} \quad (5)$$

$\tilde{X}$ can be represented in polar coordinates as follows:

$$\begin{cases} \phi = \arccos(\tilde{x}_i), -1 \leq \tilde{x}_i \geq 1, \tilde{x}_i \in \tilde{X} \\ r = \frac{t_i}{N}, t_i \in N \end{cases} \quad (6)$$

where t_i is the time stamp and N is a constant regularisation factor of the polar coordinate system.

The above encoding offer two major advantages. First, the function is bijective. That is, each value in the original signal correspond to one value in the polar coordinate representation

and vice versa. Second, the absolute temporal relations are preserved through the r coordinate.

Further to the conversion, the angular perspective can be easily exploited by considering the trigonometric sum between each pair of points. Thusly, the GAF is defined as:

$$G = \begin{bmatrix} cos(\phi_1 + \phi_1) & \dots & cos(\phi_1 + \phi_n) \\ cos(\phi_2 + \phi_1) & \dots & cos(\phi_2 + \phi_n) \\ \vdots & \ddots & \vdots \\ cos(\phi_n + \phi_1) & \dots & cos(\phi_n + \phi_n) \end{bmatrix} \quad (7)$$

Taken the definition of the inner product of two vectors x and y as:

$$< x, y > = x \cdot y - \sqrt{1 - \tilde{X}^2} \cdot \sqrt{1 - \tilde{X}^2} \quad (8)$$

G is therefore a Gramian matrix as shown in Equation 9:

$$G = \begin{bmatrix} < \tilde{x}_1, \tilde{x}_1 > & \dots & < \tilde{x}_1, \tilde{x}_n > \\ < \tilde{x}_2, \tilde{x}_1 > & \dots & < \tilde{x}_2, \tilde{x}_n > \\ \vdots & \ddots & \vdots \\ < \tilde{x}_n, \tilde{x}_1 > & \dots & < \tilde{x}_n, \tilde{x}_n > \end{bmatrix} \quad (9)$$

3.4 Network Architectures

This work proposes a range of single-input and multi-input multi-domain CNNs for the recognition of eating and drinking gestures from continuous accelerometer readings. The intuition behind, is the great potential of CNNs to identify the relevant patterns from accelerometer temporal sequences given the translation invariant nature of gestures. In addition, CNNs are knowledge domain independent since the features are automatically learned through the training step. Such feature learning takes place following a hierarchical structure, whereby the most elementary patterns are captured at the left-most layers and more complex patterns are learned at the right-most ones.

3.4.1 Benchmark Model - 1D CNN

Based on the exhibited success at similar classification tasks (Anderez et al. 2019), a 1D CNN fed with raw accelerometer data is proposed as a benchmark model. Given the accelerometer time series $x_i^0 = [x_1, ..., x_N]$, where N is the length of the accelerometer segments (in this case, $N = 394$ samples), the output of the convolutional layers is given by:

$$c_i^{l,j} = \sigma \left(b_j^l + \sum_{m=1}^{M} w_m^{l,j} x_{i+m-1}^{l-1,j} \right), \quad (10)$$

where l is the layer index, M is the kernel size, w_m^j is the weight for the jth map and m^{th} filter index, b_j^l is the bias term for the jth filter at layer l, and σ is the activation function. For

clarification, the subscript π of neuron c defines the ith neuron on layer l, while the subscript i in the time series x refers to the ith accelerometer data sample. The mth filter index represents the mth parameter of the convolution filter.

In this case, the activation function employed is the rectified linear unit (ReLU):

$$\sigma(z) = max(0, z) \tag{11}$$

Following the convolutional layer, a pooling layer performs a non-linear down-sampling by retrieving the maximum value among a set of nearby inputs. This is given by:

$$p_i^{l,j} = \max_{r \in R} \left(C_{i \times T + r}^{l,j} \right) \tag{12}$$

where T is the pooling stride and R the pooling size (in this study, 1 and 2 respectively).

Several convolutional and pooling layers can be stacked to form deeper network architectures. The output from the stacked convolutional and pooling layers is flattened to form the feature vector $f^I = [f_1, ..., f_I]$, where I is the number of units in the last pooling layer. f^I is then used as input to the fully-connected layer:

$$h_i^l = \sum_j w_{ji}^{l-1} \left(\sigma(f_i^{l-1}) + b_i^{l-1} \right) \tag{13}$$

where w_{ji}^{l-1} is the connection weight term from the ith node on layer $l-1$ to the jth node on layer l, σ is the activation function (ReLU) and b_i^{l-1} is the bias term.

The output from the fully connected layer is then used as input to the softmax function, by which the gesture classification is computed as:

$$P(c|p) = \underset{c \in C}{argmax} \frac{e^{(f^{l-1}w^L + b^L)}}{\sum_{k=1}^{N_C} e^{(f^{L-1}w_k)}} \tag{14}$$

where L is the index of the last layer, c is the gesture class and N_C is the total number of gesture classes.

The network training is conducted using the Adaptive Moment Estimation (Adam) optimiser on batches of 32 accelerometer segments for a total of 30 epochs. Categorical cross-entropy is used as the loss function. A dropout rate of 0.5 is used on the fully connected layer to mitigate overfitting issues.

3.4.2 Benchmark Network Optimisation

The performance of the 1D CNN is studied across various key network parameters. These include the number of layers (l), the number of filters within a layer (j) and the filter size (M) as follows:

– 1 = [1,2,3]

– j = [16,32,64,128,256]
– M = [6,12,25,50,75,100,125,150]

Given the sample frequency employed for data collection (25 Hz), the filter size ranges from $M = 0.24$ seconds to $M = 6$ seconds.

3.4.3 CNN frameworks description

Once the 1D benchmark network is optimised, various multi-input multi-domain networks are built on top to evaluate whether a further improvement on the classification performance can be achieved. The different proposed CNN frameworks are described below:

– 1. 1D CNN: Optimised 1D CNN benchmark network fed with raw accelerometer data.
– 1.1.1. Spec(Mag): Optimised 1D CNN benchmark network fed with raw accelerometer data combined with a 2-layered 2D CNN fed with spectrogram images of the magnitude of the tri-dimensional accelerometer signal.
– 1.1.2. Spec(y): Optimised 1D CNN benchmark network fed with raw accelerometer data combined with a 2-layered 2D CNN fed with spectrogram images of the y-axis of the accelerometer signal.
– 1.2.1. MTF(Mag): Optimised 1D CNN benchmark network fed with raw accelerometer data combined with a 2-layered 2D CNN fed with MTF images of the magnitude of the tri-dimensional accelerometer signal.
– 1.2.2. MTF(y): Optimised 1D CNN benchmark network fed with raw accelerometer data combined with a 2-layered 2D CNN fed with MTF images of the y-axis of the accelerometer signal.
– 1.3.1. GAF(Mag): Optimised 1D CNN benchmark network fed with raw accelerometer data combined with a 2-layered 2D CNN fed with GAF images of the magnitude of the tri-dimensional accelerometer signal.
– 1.3.2. GAF(y): Optimised 1D CNN benchmark network fed with raw accelerometer data combined with a 2-layered 2D CNN fed with GAF images of the y-axis of the accelerometer signal.
– 1.4. F.V: Optimised 1D CNN benchmark network fed with raw accelerometer data combined with a 2-layered NN fed with a 31-dimensional hand-crafted feature vector.

The architecture of the 2D CNNs employed for the feature learning of the resultant spectrogram, MTF and GAF images is defined by $l = 2$, $j = 5 \times 5$ and $M = 16$. The framework including the hand-crafted feature vector (F.V) employs a 2-layered Neural Network (NN) with 16 neurons on each layer. Such feature vector includes a wide range of signal

descriptive statistics as well as the duration of the different gestures. A visual representation of the different CNN frameworks employed is shown in Fig. 3.

A posteriori, the classification performance of each of the frameworks is evaluated by adopting a Leave-One-Out cross-validation strategy, whereby on each validation step one of the experiment participants is used as the test set and the remaining subjects as the training set. Given that six participants took part in the experiment, the resultant model performance metrics are then reported as the average of the six runs.

4 Experimental results

The results achieved by the different CNN-based frameworks for the recognition of eating and and drinking gestures are presented in this section. The problem has been tackled as a 3-class classification problem, with the classes being 'Drink', 'Eat' and 'Null'. The 'Null' class embodies all the irrelevant gestures retrieved by the segmentation technique. That is, all the gestures which are not an eating or a drinking gesture.

A parametrically optimised 1D CNN fed with raw accelerometer data is first proposed as a benchmark classification model. Such optimisation is achieved by studying the performance of the network across the number of layers l, the number of filters j and the filter size M. This can be better observed in Fig. 4 where the average per-class classification accuracy of the networks is plotted against the different studied parameters. The optimisation process is performed layer by layer. That is, once the values j and M are optimised for the 1-layered CNN, a second convolutional layer is added

to that optimised network. This process is then repeated for the implementation of the 3-layered CNN. From Fig. 4, it can be seen that while an increase on the classification performance is achieved by increasing the number of layers (this is confirmed by an ANOVA-Tukey HSD test), no direct relationship can be observed between the classification performance and the number of filters or the filter size. Despite the improvement seen on the classification performance achieved by the increase made to the number of layers, a further analysis is made by analysing the distribution of the average per-class classification accuracy across the different configurations. As it can be seen on Fig. 5, the performance distribution exhibited by the 1-layered and the 3-layered CNNs exhibit a negative skewness. This indicates the use of a 1-layered network and that of a 3-layered network for this specific problem can lead to underfitting and overfitting issues respectively, therefore a 2-layered network would be recommended as the more conservative architecture for future similar problems where the execution of network optimisation is not possible.

In this case, as shown in Table 1, the best average performance across j and M is achieved by the 3-layered CNN with an average per-class classification accuracy of 96.06%. The best classification performance is also achieved using a 3-layered CNN (the configuration is described in the table). Such network achieves an average per-class classification accuracy of 97.10%, an average per-class classification precision of 93.01% and an average per-class classification recall of 93.96%. The performance achieved on each class are reported in Table 3.

After the optimisation of the 1D CNN, the different frameworks proposed in Sect. 3.4.3 are evaluated. The classification performance achieved by each of the frameworks

Fig. 3 Diagram showing the different single-input and multi-input multi-domain networks proposed. It should be noticed that the top part (1D CNN) is a common factor on all the proposed networks. The rest of the models are built on top of that one by combining the respective learned features at a common fully connected layer
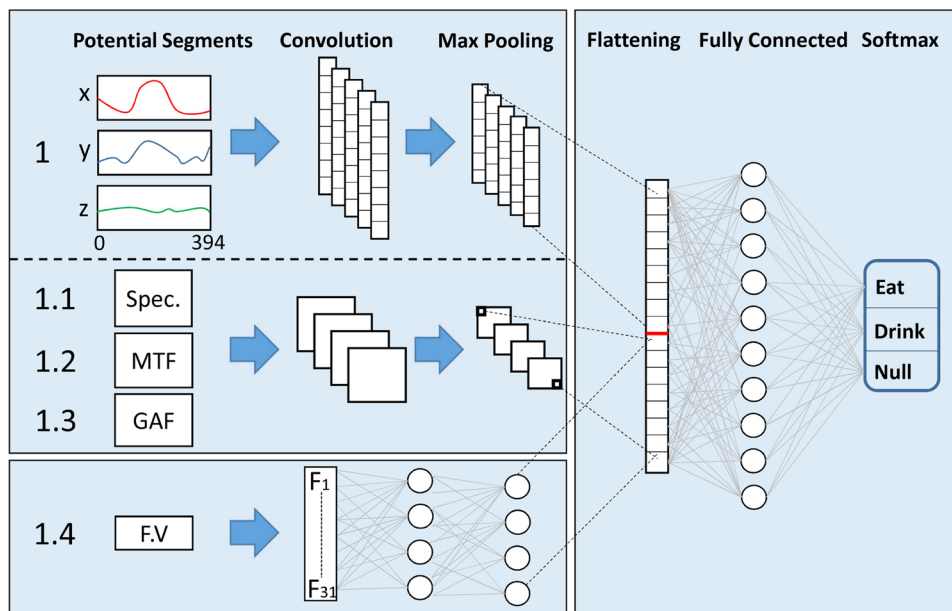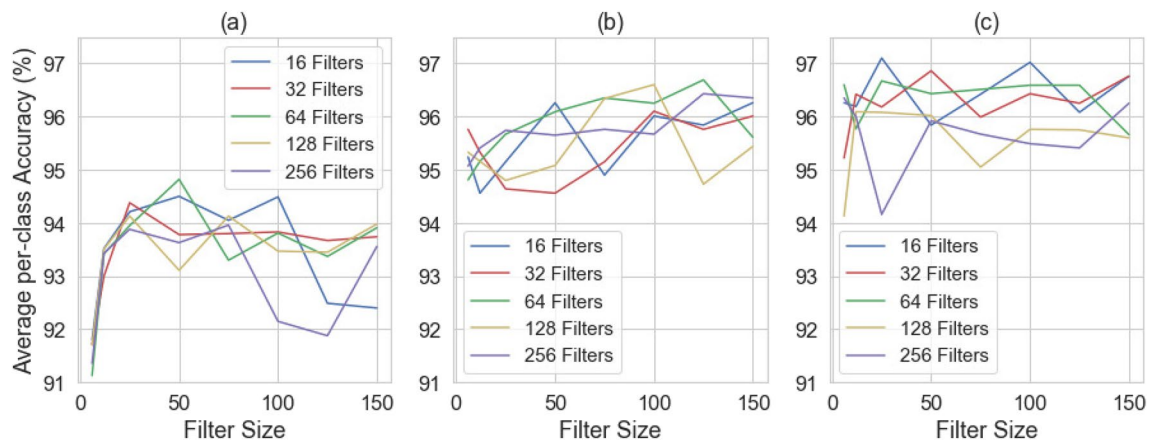
Fig. 4 Classification performance of the 1D CNN across the parameters *l*, *j* and *M*, where **a** depicts the average per-class classification accuracy of the 1-layered CNN, **b** of the 2-layered CNN and **c** of the 3-layered CNN
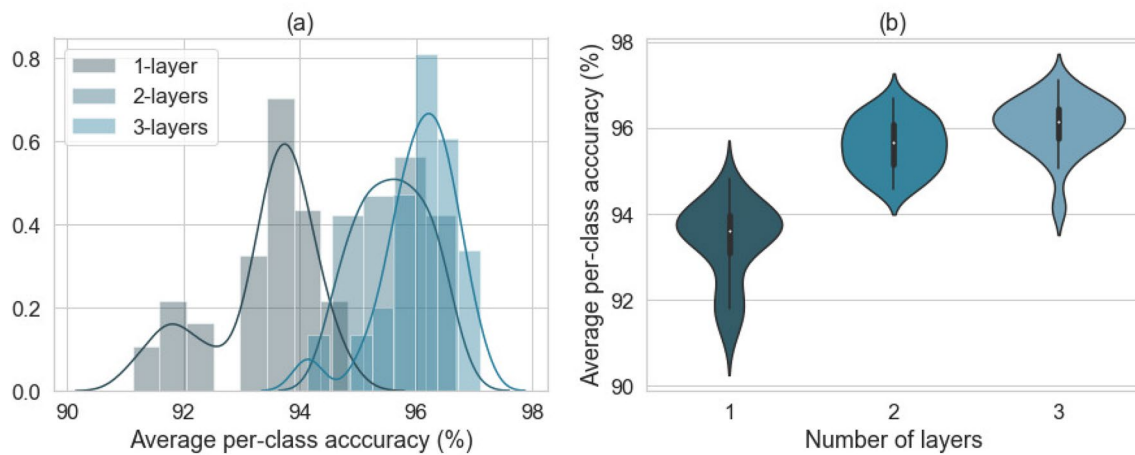

Fig. 5 Study upon network architecture (number of layers). **a** The distribution of the classification accuracies achieved by the 1-layered, 2-layered and 3-layered CNNs. **b** The corresponding violin plot

Table 1 Summary of results. The Avg. perform. (%) column reports the mean of the average per-class classification accuracy across *j* and *M*

1D CNN	Avg. perform. (%)	Best Configuration	Acc. (%)	Prec. (%)	Rec. (%)
1 Layer	93.36	$j^1 = 64$ filters, $M^1 = 50$	94.82	86.46	90.23
2 Layers	95.59	$j^2 = 64$ filters, $M^2 = 125$	96.69	91.40	94.28
3 Layers	96.06	$j^3 = 16$ filters, $M^3 = 25$	97.10	93.01	93.96

Acc. (%), Prec. (%) and Rec. (%) report the respective values achieved by the best network configurations described in the Best Configuration column

can be seen in Fig. 6. The results indicate the benchmark 1D CNN model outperforms the rest of the proposed frameworks, with only the F.V framework obtaining comparable results. Despite the implicit additional information provided by the rest of the frameworks, the required additional complexity of the network led to overfitting issues.

4.1 Discussion

The CNN-based system proposed addressed the problem of spotting and recognising eating and drinking gestures with the use of a single wrist-worn tri-axial accelerometer. As demonstrated in previous work (Anderez et al. 2020), the

Table 2 Comparison of the proposed system to previous work on the recognition of drinking gestures

Method	Sensor units	Spot.	Recog.	Accuracy	Precision (%)	Recall (%)
Anderez et al. (2020)	1	✓	✓	99.0	93.3	93.3
Proposed System	1	✓	✓	98.73	92.98	89.83
Junker et al. (2008)	5	✓	✓	–	88.0	83.0
Amft et al. (2010)	1	✓	✓	–	84.0	90.0
Serrano et al. (2017)	4	✓	✓	–	82.28	84.42
Chen et al. (2017)	1	X	✓	–	96.5	91.3
Ramos-Garcia et al. (2015)	1	✓	✓	86.5	–	–

Table 3 Classification performance per class

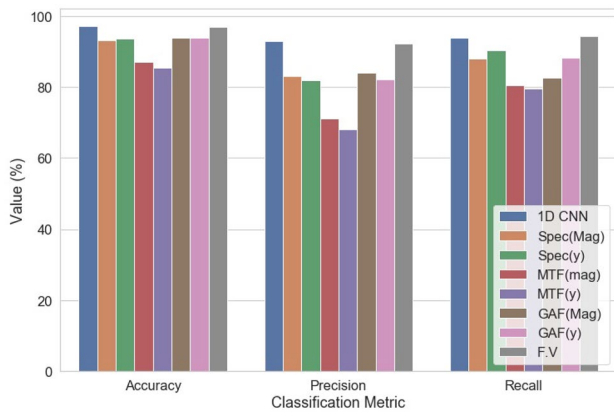|	Accuracy (%)	Precision (%)	Recall (%)
Drink	98.73	92.98	89.83
Eat	96.29	87.43	95.81
Null	96.29	98.60	96.25
Average per class	97.10	93.01	93.96

Fig. 6 Classification performance achieved by the proposed CNN-based frameworks

adaptive segmentation technique employed (CAST), correctly spotted all the eating and drinking gestures embedded in the accelerometer readings. This overcomes the drawback found in previous work at trying to estimate a suitable segment length for the specific classification problem (Lee et al. 2017; Ignatov 2018).

Despite the efforts given to improve the classification performance of the 1D CNN fed with raw accelerometer data, these mostly led to overfitting issues. However, the satisfactory results achieved in this work not only outline the suitability of CNNs for gesture recognition problems, but also signify a great contribution in the field, as supported by the outperformance of most similar work. Given the complexity of an eating activity, previous research has varied the way of tackling its recognition, with some works tackling the recognition of a complete meal period (Dong et al. 2014),

and others aiming at the recognition of specific eating gestures (Junker et al. 2008; Anderez et al. 2020). To fairly evaluate the proposed system against previous similar work, the recognition of drinking gestures in semi-controlled and controlled lab settings is considered. As it can be observed in Table 2, from the research works undertaking both the spotting and recognition phases, only the work in Anderez et al. (2020) exhibit a slightly better performance. However, the system presented here exhibits two major advantages as compared to the work in Anderez et al. (2020). First, the CNN-based system is domain knowledge independent. Second, the presented system only makes use of accelerometer data, whereas the system proposed in Anderez et al. (2020) makes use of both accelerometer and gyroscope data. As stated in Dong et al. (2014), a gyroscope consumes approximately ten times more power than an accelerometer, making the use of the former excessively power consuming for continuous monitoring.

5 Conclusions and future work

This paper has presented a system to address gesture recognition with a case study on eating and drinking. First, an adaptive segmentation technique, namely the CAST, was employed for spotting potential eating and drinking gestures within the continuous accelerometer readings. This technique exhibits a 100% spotting recall, therefore overcoming the drawbacks found in previous literature, where true positives are missing at this preliminary step. This is crucial since the errors taking place at this step propagate to the classification step, therefore affecting the overall performance of the system.

A thorough study on CNNs for eating and drinking gesture recognition was undertaken. A 1D CNN fed with raw accelerometer data was parametrically optimised and proposed as a benchmark classification model. The best classification results were obtained with a network architecture composed of 3 convolutional layers with an overall per-class classification accuracy of 97.10%. However, certain architectural configurations of the 3-layered CNN, show symptoms of model overfitting. Thus, it is crucial not to

assume complex networks will perform better and keep an adequate balance between the complexity of the network, the data available and the complexity of the classification problem itself.

Further to defining a 1D CNN benchmark classification model, various efforts were made to enrich the feature learning process performed through such benchmark model. These included the use of various 2D CNNs fed with the resultant images obtained by the employment of three different time series to image encoding frameworks, as well as a Neural Network (NN) fed with a 31-dimensional hand-crafted feature vector. A posteriori, the above feature learning techniques were combined with the resultant features of the benchmark network at a common fully connected layer. Despite the good performance exhibited by the employed time series to image encoding frameworks in different applications such as audio analysis (Yu and Slotine 2009) or EEG-based sentiment classification (Wang and Oates 2015), in this case, their use did not lead to a better classification performance when added to the 1D benchmark network. The model incorporating the 31-dimensional feature vector did not improve the classification performance of the benchmark model either. Problems of model overfitting were observed in all the cases. Thus, it can be concluded that raw accelerometer data alongside the use of a 1D CNN is the preferred solution, since it offers an adequate balance between underfitting and overfitting, leading to a better classification performance when unseen data is fed into the network.

Overall, the results obtained suggest the eating and drinking gesture recognition proposed is accurate and reliable. In addition, as opposed to comparable systems in terms of the gesture recognition performance (Anderez et al. 2020), the system presented here offers two major advantage in the sense that it does not require domain-specific knowledge and only makes use of accelerometer data. We believe, the results achieved are a great contribution towards unobtrusive diet monitoring, and thus towards the independence and well-being management of elderly people living independently.

Future efforts will be directed towards the development of a system for the recognition of meal periods based on the distribution of eating gestures across time. Further to this, we aim to develop trend analysis techniques to identify irregularities or changes on personal dietary patterns so that cases in which eating assistance is required are accurately identified.

Compliance with ethical standards

Conflict of interest On behalf of all authors, the corresponding author states that there is no conflict of interest.

References

Adama DA, Lotfi A, Langensiepen C, Lee K, Trindade P (2018) Human activity learning for assistive robotics using a classifier ensemble. Soft Comput 22(21):7027–7039. https://doi.org/10.1007/s00500-018-3364-x

Amft O, Bannach D, Pirkl G, Kreil M, Lukowicz P (2010) In: 8th IEEE international conference on pervasive computing and communications workshops, PERCOM Workshops 2010. https://doi.org/10.1109/PERCOMW.2010.5470653

Anderez DO, Lotfi A, Langensiepen C (2018a) A hierarchical approach in food and drink intake recognition using wearable inertial sensors. In: Proceedings of the 11th pervasive technologies related to assistive environments conference, ACM, pp 552–557. https://doi.org/10.1145/3197768.3201542

Anderez DO, Lotfi A, Langensiepen C (2018b) A novel crossings-based segmentation approach for gesture recognition. In: UK workshop on computational intelligence, Springer, pp 383–391

Anderez DO, Lotfi A, Pourabdollah A (2019) Temporal convolution neural network for food and drink intake recognition. In: Proceedings of the 12th ACM international conference on pervasive technologies related to assistive environments, ACM, pp 580–586. https://doi.org/10.1145/3316782.3322784

Anderez DO, Lotfi A, Pourabdollah A (2020) Eating and drinking gesture spotting and recognition using a novel adaptive segmentation technique and a gesture discrepancy measure. Expert Syst Appl 140:112888. https://doi.org/10.1016/j.eswa.2019.112888

Casella E, Ortolani M, Silvestri S, Das SK (2020) Hierarchical syntactic models for human activity recognition through mobility traces. Pers Ubiquit Comput 24(4):451–464. https://doi.org/10.1007/s00779-019-01319-9

Chen LH, Liu KC, Hsieh CY, Chan CT (2017) Drinking gesture spotting and identification using single wrist-worn inertial sensor. In: Proceedings of the 2017 IEEE international conference on applied system innovation: applied system innovation for modern technology, ICASI 2017, pp 299–302. https://doi.org/10.1109/ICASI.2017.7988411

Dong Y, Scisco J, Wilson M, Muth E, Hoover A (2014) Detecting periods of eating during free-living by tracking wrist motion. IEEE J Biomed Health Informat 18(4):1253–1260. https://doi.org/10.1109/JBHI.2013.2282471

Duffner S, Berlemont S, Lefebvre G, Garcia C (2014) 3D gesture classification with convolutional neural networks. In: ICASSP, IEEE international conference on acoustics, speech and signal processing - proceedings, pp 5432–5436. https://doi.org/10.1109/ICASSP.2014.6854641

Gayathri KS, Elias S, Ravindran B (2015) Hierarchical activity recognition for dementia care using Markov Logic Network. Pers Ubiquit Comput 19(2):271–285. https://doi.org/10.1007/s00779-014-0827-7

Ha S, Yun JM, Choi S (2016) Multi-modal Convolutional Neural Networks for Activity Recognition. In: Proceedings - 2015 IEEE international conference on systems, man, and cybernetics, SMC 2015, IEEE, pp 3017–3022. https://doi.org/10.1109/SMC.2015.525

Ignatov A (2018) Real-time human activity recognition from accelerometer data using convolutional neural networks. Appl Soft Comput J 62:915–922. https://doi.org/10.1016/j.asoc.2017.09.027

Jiang W, Yin Z (2015) Human activity recognition using wearable sensors by deep convolutional neural networks. In: Proceedings of the 23rd ACM international conference on Multimedia, ACM, pp 1307–1310. https://doi.org/10.1145/2733373.2806333

Junker H, Amft O, Lukowicz P, Tröster G (2008) Gesture spotting with body-worn inertial sensors to detect user activities. Pattern Recogn 41(6):2010–2024. https://doi.org/10.1016/j.patcog.2007.11.016

Keogh E, Chu S, Hart D, Pazzani M (2004) Segmenting time series: a survey and novel approach. In: Data mining in time series databases, World Scientific, pp 1–21

Lawal IA, Bano S (2019) Deep human activity recognition using wearable sensors. In: Proceedings of the 12th international conference on pervasive technologies related to assistive environments, ACM, pp 45–48. https://doi.org/10.1145/3316782.3321538

Lee MW, Khan AM, Kim TS (2011) A single tri-axial accelerometer-based real-time personal life log system capable of human activity recognition and exercise information generation. Pers Ubiquit Comput 15(8):887–898. https://doi.org/10.1007/s00779-011-0403-3

Lee SM, Yoon SM, Cho H (2017) Human activity recognition from accelerometer data using convolutional neural network. IEEE Int Conf Big Data Smart Comput BigComp 2017:131–134. https://doi.org/10.1109/BIGCOMP.2017.7881728

Lohrmann C, Dijkstra A, Dassen T (2003) The care dependency scale: an assessment instrument for elderly patients in German hospitals. Geriatr Nurs 24(1):40–43. https://doi.org/10.1067/mgn.2003.8

Ortega-Anderez D, Lotfi A, Langensiepen C, Appiah K (2019) A multi-level refinement approach towards the classification of quotidian activities using accelerometer data. J Ambient Intell Hum Comput 10(11):4319–4330. https://doi.org/10.1007/s12652-018-1110-y

Payette H, Shatenstein B (2005) Determinants of healthy eating in community-dwelling elderly people. Can J Public Health 96:S30–S35. https://doi.org/10.1007/BF03405198

Ramos-Garcia RI, Muth ER, Gowdy JN, Hoover AW (2015) Improving the recognition of eating gestures using intergesture sequential dependencies. IEEE J Biomed Health Informat 19(3):825–831. https://doi.org/10.1109/JBHI.2014.2329137

Ronao CA, Cho SB (2015) Evaluation of deep convolutional neural network architectures for human activity recognition with smartphone sensors. In: Proc. of the KIISE Korea computer congress, pp 858–860

Ronao CA, Cho SB (2016) Human activity recognition with smartphone sensors using deep learning neural networks. Expert Syst Appl 59:235–244. https://doi.org/10.1016/j.eswa.2016.04.032

Rush EC, Valencia ME, Plank LD (2008) Validation of a 7-day physical activity diary against doubly-labelled water. Ann Hum Biol. https://doi.org/10.1080/03014460802089825

Schiboni G, Amft O (2018) Sparse natural gesture spotting in free living to monitor drinking with wrist-worn inertial sensors. In: Proceedings of the international symposium on wearable computers, ACM, pp 140–147

Serrano JI, Lambrecht S, del Castillo MD, Romero JP, Benito-León J, Rocon E (2017) Identification of activities of daily living in tremorous patients using inertial sensors. Expert Syst Appl 83:40–48. https://doi.org/10.1016/j.eswa.2017.04.032

Smith BJ, Marshall AL, Huang N (2005) Screening for physical activity in family practice: evaluation of two brief assessment tools. Am J Prev Med 29(4):256–264. https://doi.org/10.1016/j.amepre.2005.07.005

Wang Z, Oates T (2015) Encoding time series as images for visual inspection and classification using tiled convolutional neural networks. In: Workshops at the twenty-ninth AAAI conference on artificial intelligence

Wang J, Chen Y, Hao S, Peng X, Hu L (2019) Deep learning for sensor-based activity recognition: a survey. Pattern Recogn Lett 119:3–11. https://doi.org/10.1016/j.patrec.2018.02.010

Westergren A (2001) Eating difficulties, need for assisted eating, nutritional status and pressure ulcers in patients admitted for stroke rehabilitation. J Clin Nurs 10(2):257–269. https://doi.org/10.1046/j.1365-2702.2001.00479.x

Westergren A, Unosson M, Ohlsson O, Lorefält B, Hallberg IR (2002) Eating difficulties, assisted eating and nutritional status in elderly (65 years) patients in hospital rehabilitation. Int J Nurs Stud 39(3):341–351. https://doi.org/10.1016/S0020-7489(01)00025-6

Yang JB, Nguyen MN, San PP, Li XL, Krishnaswamy S (2015) Deep convolutional neural networks on multichannel time series for human activity recognition. In: IJCAI international joint conference on artificial intelligence, pp 3995–4001

Yu G, Slotine JJ (2009) Audio classification from time-frequency texture. In: ICASSP, IEEE international conference on acoustics, speech and signal processing—proceedings, pp 1677–1680. https://doi.org/10.1109/ICASSP.2009.4959924

Zeng N, Zhang H, Song B, Liu W, Li Y, Dobaie AM (2018) Facial expression recognition via learning deep sparse autoencoders. Neurocomputing 273:643–649. https://doi.org/10.1016/j.neucom.2017.08.043

Zeng N, Wang Z, Zhang H, Kim KE, Li Y, Liu X (2019) An improved particle filter with a novel hybrid proposal distribution for quantitative analysis of gold immunochromatographic strips. IEEE Trans Nanotechnol 18:819–829. https://doi.org/10.1109/TNANO.2019.2932271

Zeng N, Li H, Wang Z, Liu W, Liu S, Alsaadi FE, Liu X (2020) Deep-reinforcement-learning-based images segmentation for quantitative analysis of gold immunochromatographic strip. Neurocomputing. https://doi.org/10.1016/j.neucom.2020.04.001