Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa

Does our social life influence our nutritional behaviour? Understanding nutritional habits from egocentric photo-streams

Andreea Glavan[a,1], Alina Matei[a,1], Petia Radeva[b,c], Estefania Talavera[a,c,*]

[a] Bernoulli Institute, University of Groningen, Nijenborgh 9, 9747 AG Groningen, Netherlands
[b] University of Barcelona, Gran Via de les Corts Catalanes, 585, 08007 Barcelona, Spain
[c] Computer Vision Center, Autonomous University of Barcelona, Edifici O, 08193 Barcelona, Spain

ARTICLE INFO

Keywords:
Pattern discovery
Egocentric vision
Nutrition
Behaviour understanding
Lifelogging
Deep learning

ABSTRACT

Nutrition and social interactions are both key aspects of the daily lives of humans. In this work, we propose a system to evaluate the influence of social interaction in the nutritional habits of a person from a first-person perspective. In order to detect the routine of an individual, we construct a nutritional behaviour pattern discovery model, which outputs routines over a number of days. Our method evaluates similarity of routines with respect to visited food-related scenes over the collected days, making use of Dynamic Time Warping, as well as considering social engagement and its correlation with food-related activities. The nutritional and social descriptors of the collected days are evaluated and encoded using an LSTM Autoencoder. Later, the obtained latent space is clustered to find similar days unaffected by outliers using the Isolation Forest method. Moreover, we introduce a new score metric to evaluate the performance of the proposed algorithm. We validate our method on 104 days and more than 100 k egocentric images gathered by 7 users. Several different visualizations are evaluated for the understanding of the findings. Our results demonstrate good performance and applicability of our proposed model for social-related nutritional behaviour understanding. At the end, relevant applications of the model are discussed by analysing the discovered routine of particular individuals.

1. Introduction

Nutrition plays an important role in our daily routine. Recent research (Hamrick, Andrews, Guthrie, Hopkins, & McClelland, 2011) has shown that an average of 2.5 h a day is spent eating or drinking by American people, out of which 78 min drinking or eating while doing other primary activities, such as working, driving, or preparing meals. Food behaviour has been generally regarded as *what* people eat, with a focus on healthy versus unhealthy meals. However, recent studies (Laska, Hearst, Lust, Lytle, & Story, 2015) have shown that *how* and *where* people eat have a direct impact on health, being associated to diseases like diabetes, obesity (Stalonas & Kirschenbaum, 1985), cancer (Hopkinson, Wright, McDonald, & Corner, 2006) and even mental illnesses (Donini, Savina, & Cannella, 2003). Another key question related to eating habits is *how much* people eat; the social aspect of eating (i.e. social eating), which implies the act of two or more people eating together, exercises a considerable influence on this matter. For instance, in Higgs and Thomas (2016), the authors concluded that people are inclined to eat more and, consequently, spend more time in food-related environments, when joined by someone else. Moreover, social eating has proven to be a very powerful facilitator for establishing humans bonds (Dunbar, 2017). What attracts people towards social eating is an opportunity for overeating, which is facilitated by the tendency of people to order and take larger quantities of food while being in a social group setting (Herman, 2017). All of these factors involuntarily shape the way people eat.

Human behaviour when related to nutrition and social has been previously studied with computer vision and machine learning models. For instance, food balance estimation has been analysed from images intentionally collected by the user (Aizawa, Maruyama, Li, & Morikawa, 2013). At the same time, the study of behaviour in our society has been addressed from crowds (Li, 2018) and individuals (Talavera, Wuerich, Petkov, & Radeva, 2020) through the analysis of images. When studying behaviour, the collection of data describing the daily life of people is

* Corresponding author.
E-mail addresses: a.i.glavan@student.rug.nl (A. Glavan), a.matei@student.rug.nl (A. Matei), petia.ivanova@ub.edu (P. Radeva), estefania.tama@gmail.com (E. Talavera).

[1] Both authors contributed equally to this study.

https://doi.org/10.1016/j.eswa.2020.114506
Received 28 September 2020; Received in revised form 24 November 2020; Accepted 14 December 2020
Available online 24 December 2020

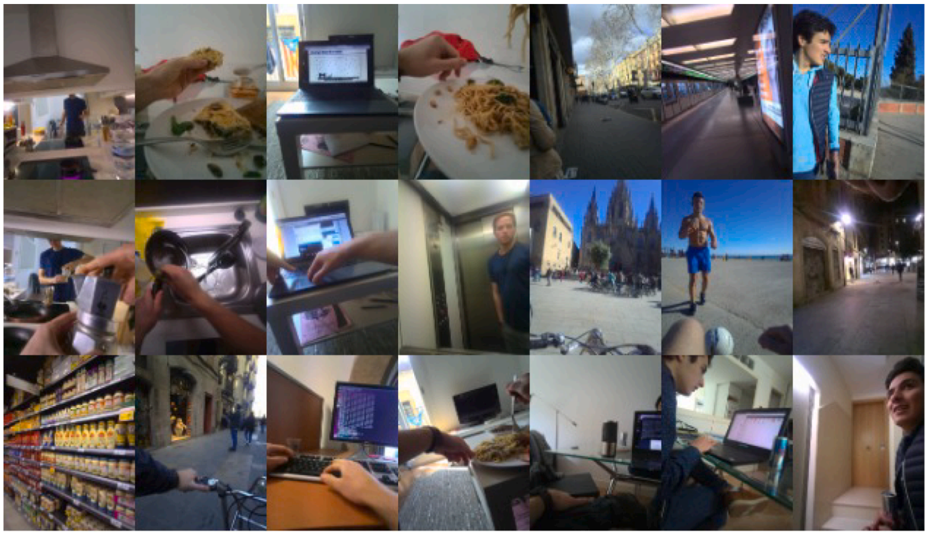


Fig. 1. Examples of egocentric images describing daily activities including both food-related and non-food related activities.

needed. One way through which individual behaviour can be captured is through the means of *lifelogging* (Gurrin, Smeaton, & Doherty, 2014; Bolanos, Dimiccoli, & Radeva, 2016), an emerging mainstream activity in which day-to-day life activities are tracked using various techniques. Our hypothesis is that the type of lifelogging that entails tracking activities in the form of photo-streams recorded by wearable cameras can be of help for the objective understanding of people's behaviour. Such egocentric photo-streams are recorded at a set time interval by portable cameras worn by the user as a necklace. These images provide an objective first-person perspective of the activities conducted by the camera wearer. Some examples of sampled images from egocentric photo-streams can be seen in Fig. 1. The data is extremely meaningful for inferring and extracting patterns related to human behaviour as shown in Clarkson (2002), Varini, Serra, and Cucchiara (2017). However, there is a lack of automated tools that can process egocentric photo-streams with the aim of providing insight into the nutritional routine. What is more, to the best of our knowledge, the effect of the correlation of nutritional habits and social activity on behaviour has not been studied from images before.

This work is motivated by the above-drawn conclusions by several studies on the effects of social interaction on the food-related human behaviour. Hence, the focal research questions are: *Can social- and food-related descriptors help us discover nutritional habits from visual information through analysis of egocentric photo-streams?*. And if yes, how does people's social daily life impact their nutritional behaviour? And what is the extent of this influence? Can we quantify it? To answer these questions, we propose an automatic system for understanding people's social-nutritional habits from egocentric photo-stream. We address the automatic extraction of the main influencing factors, as well as how the respective behaviour manifests. The understanding of these factors can help people towards becoming self-aware of their habits. This is especially beneficial for comprehending the social psychology underlying long-term nutritional routines. Our proposed system analyses the social characteristic of eating and its impact on food-related habits. The approach we propose is able to analyse social behaviour by identifying social interaction and the regularity of appearance of people with whom the camera wearer socially engages. We employ the pipeline previously proposed in Talavera, Glavan, Matei, and Radeva (2020) to extract nutritional information from egocentric timelines corresponding to whole recorded days. We account for the impact of social activity on eating habits by identifying and quantifying the instants of social eating. To do so, we create social-nutritional descriptors for the recorded days of the camera wearer. Days are compared based on these social-nutritional descriptors, which capture the routine behaviour of the day. Finally, to discriminate between routine and non-routine related days, we apply anomaly detection seeking non-routine related days by means of the Isolation Forest algorithm (Liu, Ting, & Zhou, 2008). Moreover, for a broader reach of the applicability of our model, we introduce several ways of visualizing the obtained results and correlations. We study the performance of the proposed tool on data collected by 7 users included in the EgoRoutine dataset (Talavera et al., 2020), who have visually lifelogged their daily lives for an average of approximately 15 days each. This case study analysis displays the relevancy of the proposed model for applications in the field of behaviour interpretation.

The contributions of this work are twofold:

- *Routine discovery captured by social and nutritional indicators*: To the best of our knowledge, this is the first work that addresses the automatic discovery of nutritional habits in relation to social interactions from collections of egocentric photo-streams. A novel routine discovery pipeline is proposed for the personalized discovery of routine and non-routine nutritional behaviour from multiple unseen egocentric photo-stream timelines in relation to the interrelation among food-related and social interactions of the user.
- *Social-eating metrics* are introduced for the quantification and analysis of the nutritional behaviour of the camera wearer. These metrics are used for the later identification of similarity among days, i.e. the discovery of routine vs. non-routine related habits at the level of daily life.

The rest of the paper is organized as follows: in Section 2 we address relevant literature to our research, followed by Section 3 which provides an overview of the methods and techniques used for the construction of the proposed pipeline model. The experimental setup of the work is described in Section 4. In Section 5 and Section 6, we present the experimental results and discussions, respectively. Finally, Section 7 draws our final conclusions.

2. Related works

Food detection and recognition from images have been widely studied. For example, in Kagaya, Aizawa, and Ogawa (2014), the authors focused on the recognition and tracking of food-intake in images where food occupies a significant part of the image. This approach is not suitable for our research since it only focuses on classifying single images. In the case at hand, we are working with continuous sets of labels that describe the lifestyle of the user, which has been addressed in the literature (Talavera et al., 2020). The work proposed in Talavera et al.

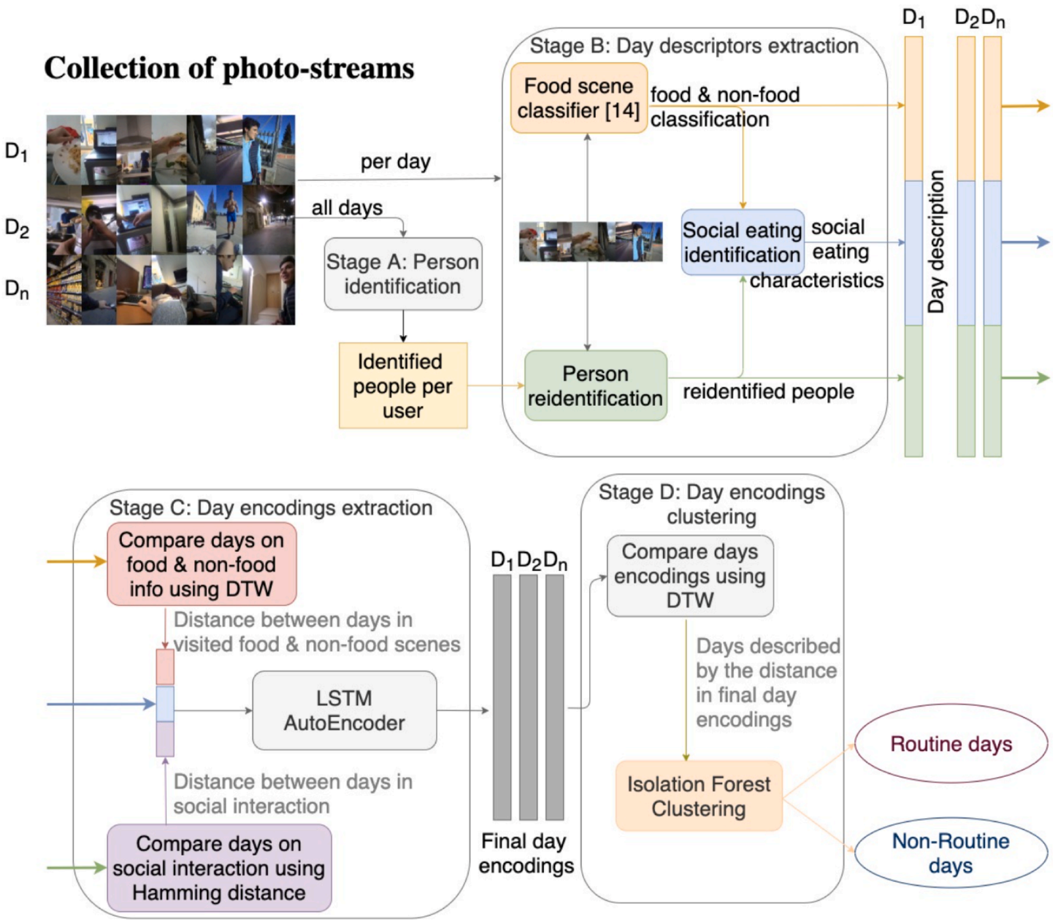

Collection of photo-streams
D1
D2
Dn
per day
all days
Stage A: Person identification
Identified people per user
Stage B: Day descriptors extraction
Food scene classifier [14]
food & non-food classification
Social eating identification
social eating characteristics
Person reidentification
reidentified people
Day description
D1
D2
Dn
Stage C: Day encodings extraction
Compare days on food & non-food info using DTW
Distance between days in visited food & non-food scenes
LSTM AutoEncoder
Distance between days in social interaction
Compare days on social interaction using Hamming distance
D1 D2 Dn
Final day encodings
Stage D: Day encodings clustering
Compare days encodings using DTW
Days described by the distance in final day encodings
Isolation Forest Clustering
Routine days
Non-Routine days

Fig. 2. Overview of the proposed pipeline for identifying eating routine.

(2020) introduces a model which employs a Convolutional Neural Network pipeline for classification of unseen chronological sets of images based on the food scene depicted by each image. The model is able to recognize 16 distinct classes out of which 15 are food-related classes (i.e. ranging from 'bar' to 'market indoor' and 'restaurant'), the 16th class represents the 'non-food' label.

The field of food-related scene recognition in egocentric vision is fairly new, with limited research available on this specific topic. The classification of food-related scenes from egocentric images was first presented in Talavera et al. (2019). The authors developed a hierarchical classification model and introduced a food-scene taxonomy as well as a new egocentric dataset, called *EgoFoodPlaces*. The dataset is split semantically in meta-classes corresponding to nutrition-related activities (i.e. eating, preparing, acquiring). Each meta-class is in turn split into sub-classes, until a three-level taxonomy is reached. A deep Convolutional Neural Network (CNN) is applied to each level of the taxonomy, resembling a DECOC classifier (Pujol, Radeva, & Vitria, 2006), which decomposes a multi-class classification problem into multiple classification problems organized hierarchically. The model proposes not only the classification into 15 food-related scene classes corresponding to the lowest level of the taxonomy, but also the recognition of meta-classes at different levels (e.g. cooking, shopping, eating), which provides a more general view on the nutritional behaviour.

Person re-identification refers to the task of identifying known people in a set of images, based on facial features. It has been approached from various angles, ranging from probabilistic approaches to deep learning approaches (Zheng, Gong, & Xiang, 2011; Yi, Lei, Liao, & Li, 2014). In egocentric images, this task was addressed in a supervised fashion with a limited amount of data and people appearing in the frames in Talavera, Cola, Petkov, and Radeva (2019). Therefore, we do not consider their work as relevant for our study and instead, we address person re-identification as an unsupervised task, i.e. relying on clustering techniques for the identification of people throughout the photo-sequences. Social interactions and food-related scenes were previously used for image classification in Herruzo, Portell, Soto, and Remeseiro (2017). The authors label images as 'Eating', 'Socializing' or 'Sedentary' and the possible combinations of these three labels, resulting in twelve in total. Their focus was on image classification given a limited set of general scenarios. Moreover their proposed model did not look for patterns throughout time or on how the different labels interact.

Previous works on routine discovery are based on the main idea of clustering similar days and disregarding non-routine days considered as outliers. Hence, anomaly detection methods were able to identify days considered as non-routine related (Talavera, Petkov, & Radeva, 2019). In their work, the authors analyzed a day as the aggregation of the feature vectors extracted of the images of that day. Even though they achieved good performance, their work loses time-related behaviour by aggregating the feature vectors. Moreover, they only study context information without higher semantics such as activity, object in the scene, or others.

The study presented in Talavera et al. (2020) addressed the discovery of nutritional habits. Their proposed classification of images does not provide a level of understanding with respect to the social interaction aspect in the behaviour of the individual. Therefore, we go a step further and incorporate the study of routine by combining different daily behaviour descriptors such as social interaction and food-related scenes occurrence. Moreover, we do not simply aggregate day descriptors as in Talavera et al. (2020), but study their accumulated information through the use of Long-Short-Term-Memory deep neural networks. Moreover, their experiments reinforced the suitability of the Isolation Forest clustering algorithm, as an anomaly detection technique, for discerning between days which follow a routine pattern and days which exhibit irregular behaviour. From the above-mentioned works, and based on our own perception of this unsupervised task, we continue addressing the discovery of routine-related behaviour through anomaly detection.

3. Methodology

This section presents in detail the constituent components of the proposed pipeline for routine discovery in relation to social interactions. Fig. 2 depicts a broad overview of the entire pipeline, starting from the input of the photo-stream timelines, corresponding to all the recorded days of a particular user, to the final outcome, represented by the two clusters of days: routine and non-routine related when it comes to eating and social habits, respectively. We have decomposed our proposed method into four stages which are described closely, for a better comprehension, in the following subsections. Moreover, the pipeline design employs high levels of modularity, which implies that some stages can be decoupled and used individually.

From this point on, we will use the terms ‘camera wearer’ and ‘user’ interchangeably, since the design of the proposed pipeline has been highly inspired by the EgoRoutine dataset (Talavera et al., 2020), which includes 7 users, playing the role of camera wearers.

In Algorithm 1 we provide an overview of the workings of our proposed pipeline in pseudocode format. The information illustrated in this algorithm coincides with the graphical depiction in Fig. 2, albeit it provides a more detailed set of steps for each of the described stages. Each step is in turn discussed in the following sub-Sections to a greater detail.

3.1. Individualized person identification per user

In the first stage of the pipeline (i.e. stage A), we aim to identify and group all individuals present in the collection of egocentric photo-streams recorded by the same user. Since egocentric photos imply a first person perspective, we assume that the individuals that can be identified by analysing the photos are engaged, to some degree, in a social activity with the camera wearer. A visual representation of stage A is presented in Fig. 3.

Stage A has been divided in two steps, see Fig. 3. In the first one, stage A.1, each photo in the user’s collection of photo-streams is analysed using the face recognition functionality of the OpenCV library (Bradski, 2000). This process leads to the identification and extraction of all faces available in the photo-streams. An extracted face is then represented as an embedding, meaning an 128-d feature vector, as resulting from the face recognition process. Subsequently to the extraction of all

Algorithm 1: Pseudocode of the proposed system.

```
Result: Routine and Non-Routine day identification.
images = load photostream of user’s days;
all_faces = [];
for image in images do
    img_faces = identify faces in image;
    all_faces += img_faces;
end
nr_people, people = DBSCAN clustering of all_faces;
for each daily photostream do
    daily_descriptors = [];
    food_classification = [];
    reidentified_faces = [];
    for image in photostream do
        food_classification += Food classifier result of image;
        reidentified_people += Identify person based on (people, image);
    end
    social_eating = Compute eating times based on (food_classification, reidentified_people);
    daily_descriptors += [food_classification, reidentified_people, social_eating];
end
distance_matrix = DTW distance of daily_descriptors;
social_distance = Hamming distance of social_eating;
day_description = [distance_matrix, daily_descriptors, social_distance];
encodings = LSTM(day_description);
encoded_matrix = DTW distance of encodings;
routine, non_routine = Isolation Forest clustering of encoded_matrix;
return routine, non_routine
```

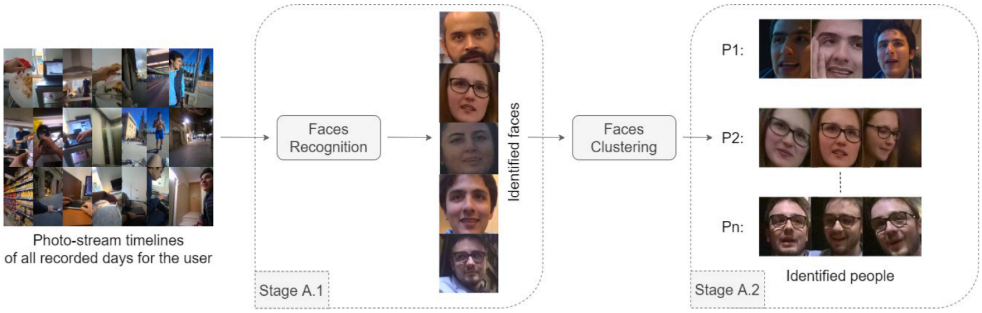


Fig. 3. Overview of stage A of the pipeline which entails the identification of people which recurrently appear in the user’s recorded days, indicating social interaction. Stage A.1 implies the identification and characterization of all human faces which are then clustered in stage A.2 in order to identify the people with which the user had recurrent interaction over the recorded days.

the faces, stage A.2 cluster them. Since the collection of recorded photos spans a period of multiple days (the exact number of days might vary depending on the user), we expect that people who are related to the user's routine to reappear. Moreover, the number of reappearing people is unknown, therefore we opted for the DBScan clustering with Euclidean distance approach (Birant & Kut, 2007), which does not require any prior knowledge of the number of clusters, in this context specifically, the number of reappearing people. DBScan is a density based clustering approach, therefore it groups together core sample points, in our case, faces, which lie in dense regions. The final vector space contains a collection of points, each corresponding to a 128-d vector describing an identified face, with the format:

$$identified\_face = [m_1 \quad m_2 \cdots \quad m_{128}], \tag{1}$$

where m_i represents the ith measurement of the detected face, resulting from the application of OpenCV.

The density approach accounts for the reappearing persons: the more a person is socially engaged in the routine of the user, the more instances of their face will be present in the recorded photo-streams. This creates a dense region of reappearing face instances, thus allowing the clustering algorithm to group them as an individual. On the other hand, points which have a limited number of neighbours in the extracted faces space (i.e. lie in a low density area) will be deemed as outliers. Since we want to eliminate the cases of randomly appearing, unrelated people (for example, people on the public transport or faces in ads) being captured by the egocentric camera, we discard the outliers group altogether.

The outcome of stage A is therefore a series of identified individuals as represented by a cluster of their faces, extracted from all the photos recorded by the user. This outcome will flow into stage B of the pipeline; nevertheless stage A can be also used individually for other applications which employ unsupervised person re-idetification.

3.2. Nutritional and social day descriptors extraction

The second stage of the pipeline operates on individual timeline photo-streams, corresponding to the recorded days. The aim of this stage is to analyse the day photo-stream in order to create a detailed day descriptor vector consisting of nutritional, social eating characteristic and social interaction information. Each of the three components of the day descriptor vector are computed by a corresponding sub-stage and concatenated together in order to obtain the final day descriptor, as seen in Fig. 4.

With the input being a photo-stream consisting of a single recorded day from the user, each image is processed independently by the three sub-stages. For stage B.1, the food scene classifier proposed by Talavera et al. (2020) is applied. The classifier has the following behaviour: given an image, it decides whether the scene depicted is not related to any food environment. If this holds, the image is labeled as 'non-food', otherwise the classification process continues. The classifier's decision making process entails discriminating between 15 food scene classes ranging from 'bakery shop' to 'kitchen' and 'restaurant'. A vector containing the classification likelihood of each food scene is obtained, the final classification is the food scene class with the highest percentage in likelihood. Based on the food scene classification, the pipeline has the ability to compute the top 5 most appearing food scenes in a day with their corresponding percentages of appearance.

For stage B.2, person re-identification is performed using the same photo-stream. A similar procedure as before is employed for face recognition. The OpenCV library is utilized to identify faces in the image. In the case in which at least a face is found, this result is cross-referenced with the respective user's identified individual clusters from stage A. If a match is found, the appearance of the previously identified person in the image at hand is stored in a binary vector of dimension 1×P, where P is the total number of identified individuals as per stage A of the pipeline. For example, if person N has been re-identified in the image at hand, the binary vector corresponding to the image will have a positive bit at position 1×N, where N ⩽P. An overall metric accounting for the total time spent in social interactions is computed as $f_i^p = n_i$ where n_i is the number of occurrences of face i. Quantifying the amount of time a person appears within the photo sequences is based on the images metadata and is computed separately.

Gathering this information for all images corresponding to a day, the nutritional information is cross-referenced with the social re-identification in stage B.3. This results into a series of social eating characteristics for the day:

- Total Eating Time (TET),
- Time Eating Alone (TEA),
- Time Eating with One other person (TEO),
- Time Eating with a Group (more than one other person) (TEG).

3.3. Day descriptors manipulation and final day encodings extraction

In stage C of the pipeline, see Fig. 5, the day descriptor vectors are contextualized with respect to all the days recorded by the user. Routines imply consistency both over habits, but also over time. Having this general characteristic of routine in mind, we compare the day descriptors in order to investigate how much distinction exists between

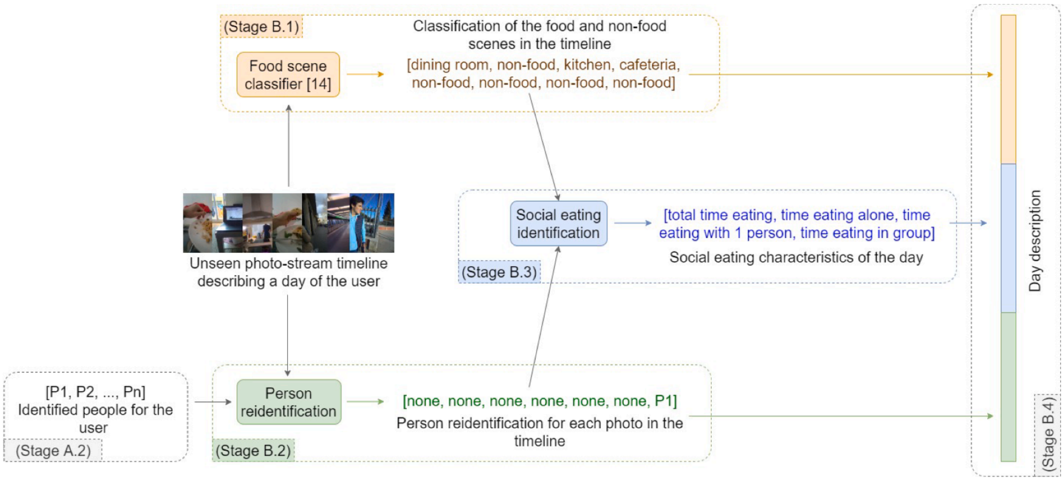

(Stage B.1)
Food scene classifier [14]
Classification of the food and non-food scenes in the timeline
[dining room, non-food, kitchen, cafeteria, non-food, non-food, non-food, non-food]
Unseen photo-stream timeline describing a day of the user
Social eating identification
[total time eating, time eating alone, time eating with 1 person, time eating in group]
Social eating characteristics of the day
(Stage B.3)
[P1, P2, ..., Pn]
Identified people for the user
(Stage A.2)
Person reidentification
(Stage B.2)
[none, none, none, none, none, none, P1]
Person reidentification for each photo in the timeline
Day description
(Stage B.4)

Fig. 4. Stage B: The extraction and assembly of day descriptions occurs for every recorded day of the user. Stage B.1 provides nutritional descriptors, which offer information about the visited food scenes as extracted by the classifier proposed in Talavera et al. (2020). Stage B.2 provides details into the social aspect of the day.

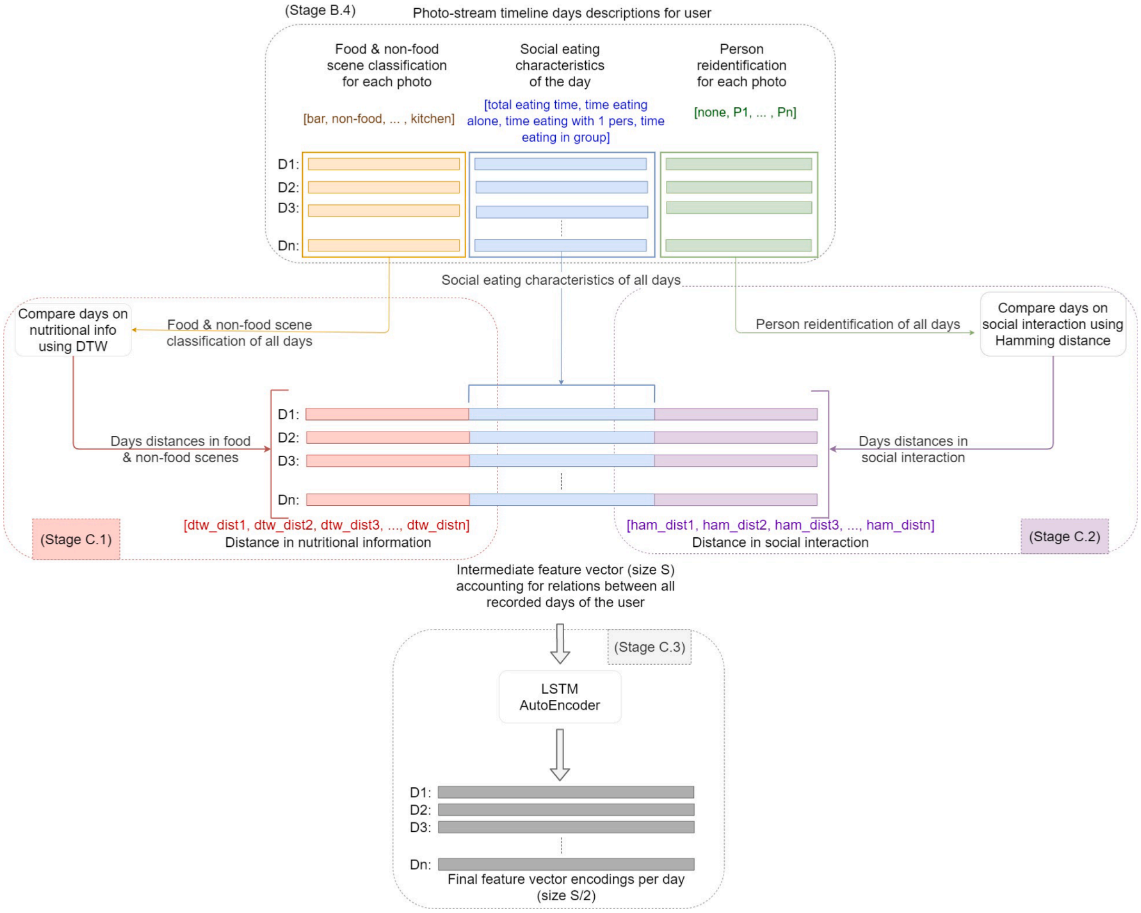

(Stage B.4)
Photo-stream timeline days descriptions for user
Food & non-food scene classification for each photo
Social eating characteristics of the day
Person reidentification for each photo
[bar, non-food, ... , kitchen]
[total eating time, time eating alone, time eating with 1 pers, time eating in group]
[none, P1, ... , Pn]
D1:
D2:
D3:
Dn:
Social eating characteristics of all days
Compare days on nutritional info using DTW
Food & non-food scene classification of all days
Person reidentification of all days
Compare days on social interaction using Hamming distance
Days distances in food & non-food scenes
Days distances in social interaction
D1:
D2:
D3:
Dn:
[dtw_dist1, dtw_dist2, dtw_dist3, ..., dtw_distn]
Distance in nutritional information
[ham_dist1, ham_dist2, ham_dist3, ..., ham_distn]
Distance in social interaction
(Stage C.1)
(Stage C.2)
Intermediate feature vector (size S) accounting for relations between all recorded days of the user
(Stage C.3)
LSTM AutoEncoder
D1:
D2:
D3:
Dn:
Final feature vector encodings per day (size S/2)

Fig. 5. Stage C: The day descriptors given by the previous stage, stage B, are manipulated and contextualized with respect to all the days recorded for the user. Stage C.1 compares all the days, using dynamic time warping (DTW). Later, Stage C.2 compares all the days with respect to the social activity using Hamming distance. Finally, in stage C.3 the feature space is scoped down by using an LSTM AutoEncoder to extract encodings from all the recorded days.

days. We define two steps (i.e. C.1 and C.2), which simultaneously compute the difference in nutritional activities and the social pairwise interaction distance between/among all the days of the user, respectively.

To compare days with respect to the different nutritional and social activities, we can not directly compare image to image since activities of the days can have different duration. Instead, in Stage C.1 we apply Dynamic Time Warping (DTW) (Müller, 2007) on the days, as represented only as their nutritional activities information extracted from the day descriptor vectors. For each day, the nutritional information is presented in a timeline manner, each image in the timeline being represented by its food scene category or 'non-food' label. We define a day as the vector containing the food and non-food classes ordered by their temporal appearance throughout the day. Given two such vectors, s' and s'', each representing a day, we can compute the path warping $w = (w_0, w_1, \ldots, w_Q)$, where Q is the length of the path. Each element in the path w_q is a tuple $(w_q[1], w_q[2])$ which indicates the mapping between the two timelines (i.e. element $w_q[1]$ in s' corresponds to element $w_q[2]$ in s''). Eq. (2) describes the formula behind the DTW algorithm which computes the optimal path warping (i.e. minimal distance path) which defines the best correspondence between the sequences:

$$DTW\_dist(s', s'') = \sum_{q=0}^{Q} dist(s'_{w_q[1]}, s''_{w_q[2]}). \tag{2}$$

DTW is a suitable approach for this problem, since it is designed to measure similarity between time series, by stretching or shifting one of the series time wise in order to match the second series. DTW also accounts for small differences in time, which is desirable since daily events take place within a certain time margin.

Stage C.2 compares the days with respect to their social interaction. We judge social activity by the (re) appearance of the identified people (from the person identification phase of the pipeline) in the day photo-stream. Socially, a day is described by a chronological sequence of 24 binary vectors which describe the people which are socially engaged with the user each hour of the day. For social comparison of the days, we propose Hamming distance: this computes how many of the bits differ between the two binary sequences representing the compared days. Given two sequences of binary vectors, b' and b'', corresponding to two distinct days, we compute the Hamming social distance as per: 3:

$$Social\_dist = \sum_{i=0}^{23} hamming\_dist(b'[i], b''[i]). \tag{3}$$

where i indicates the hour and $hamming\_dist(b'[i], b''[i])$ computes the number of bit differences between two social interactions $b'[i]$ and $b''[i]$.

Prior to this point, the pipeline is employed to extract features from the days that can capture the correlation between social activity and nutritional routine. An overview of all the extracted features is given as follows, elements 3–9 being computed by the pipeline phase described in Section 3.2; we note that not all of the features the pipeline has the capacity of extracting are employed as day descriptors for the routine discovery clustering proposed in our model.

1. DTW nutritional distances between the days (as given by Eq. (2))

2. Hamming social interaction distances between the days (as given by Eq. (2))
3. Total time eating per day (TET)
4. Time eating alone per day (TEA)
5. Time eating with one person per day (TEO)
6. Time eating in a group per day (TEG)
7. Top 5 food scenes per day
8. Top 5 food scenes percentages per day
9. Social interaction time per day.

An intermediate feature vector is created for each day. The vector is the result of the concatenation of the nutritional distances (as given by Eq. (2)), the social eating characteristics of the day (TET, TEA, TEO and TEG as computed by the pipeline phase described in Section 3.2) and, finally, the social interaction distances (as given by Eq. 3). The social eating characteristics are time based metrics describing the day. The final vector has the following aspect:

$$Day = concat(dist_{DTW}(s', s''), [TET \quad TEA \quad TEO \quad TEG], Social\_dist). \tag{4}$$

The size of the intermediate feature vector is dependent on the number of days recorded by the user. In order to capture routine, at least 7 recorded days are required (i.e. by analysing one week which includes 5 working days, routine should become apparent). Therefore, the size of a feature vector is expected to be at least 1 × 18 (7 nutritional distances, 7 social interaction distances and 4 social eating metrics) and increases significantly with the increase in the number of recorded days.

In order to compress the information that appears in the time series, we pass the intermediate feature vectors through stage C.3, during which an Long-Short-Term-Memory (LSTM) auto-encoder is applied.

Auto-encoders (Baldi, 2012Bengio, 2009) are machine learning tools used to reduce the feature size of the input data by encoding it to a feature vector of a set size, based on the input features. The resulting encoding vector can be extracted and used for data compression or machine learning tasks. In the context of autoencoders, the compression and decompression of data are done using functions based on neural networks; these functions are learned by the network based on examples, as opposed to the user-defined ones (Park, Marco, Shin, & Bang, 2019). In this case, Long-Short-Term-Memory (LSTM) networks (Greff, Srivastava, Koutník, Steunebrink, & Schmidhuber, 2016Hochreiter and Schmidhuber, 1997) are used, due to the strength of their recurrent and persistent memory.

Long-Short-Term-Memory networks are a type of networks based on Recurrent Neural Networks (RNN) tailored towards extracting features from data with a time component (i.e. time series data). RNNs capture time dependencies of sequence data due to their recurrent connections between the neural units. However, RNNs fail for the task of capturing long term connections from sequences with significant temporal lags because of their short-term memory. This issue is overcome by the LSTM networks which have a more sophisticated update equation. LSTMs utilize memory cells that combine multiplicative interactions between logistic and linear units with input and output gates.

The hidden layers of the network include memory cells that are fully interconnected. The memory cell corresponds to input and output gates which, being fed inputs from other memory cells, decide how to update the memory cell at hand: which new information needs to be capture or discarded. The update follows Eq. 5 with c_i being the i^{th} memory cell and its output at time t being $y^{c_i}(t)$.

$$y^{c_i}(t) = y^{out_i}h(s_{c_i}(t)), \tag{5}$$

where the internal state $s_{c_i}(t)$ is defined as per Eq. (6):

$$\begin{aligned} s_{c_i}(0) &= 0, \\ s_{c_i}(t) &= s_{c_i}(t-1) + y^{in_i}g(net_{c_i}(t)), \quad t > 0, \end{aligned} \tag{6}$$

with in_i and out_i being the corresponding input and output gates of memory cell c_i; functions g and h are differentiable functions and have the role to compress net_{c_i} and to scale the cell's output, respectively.

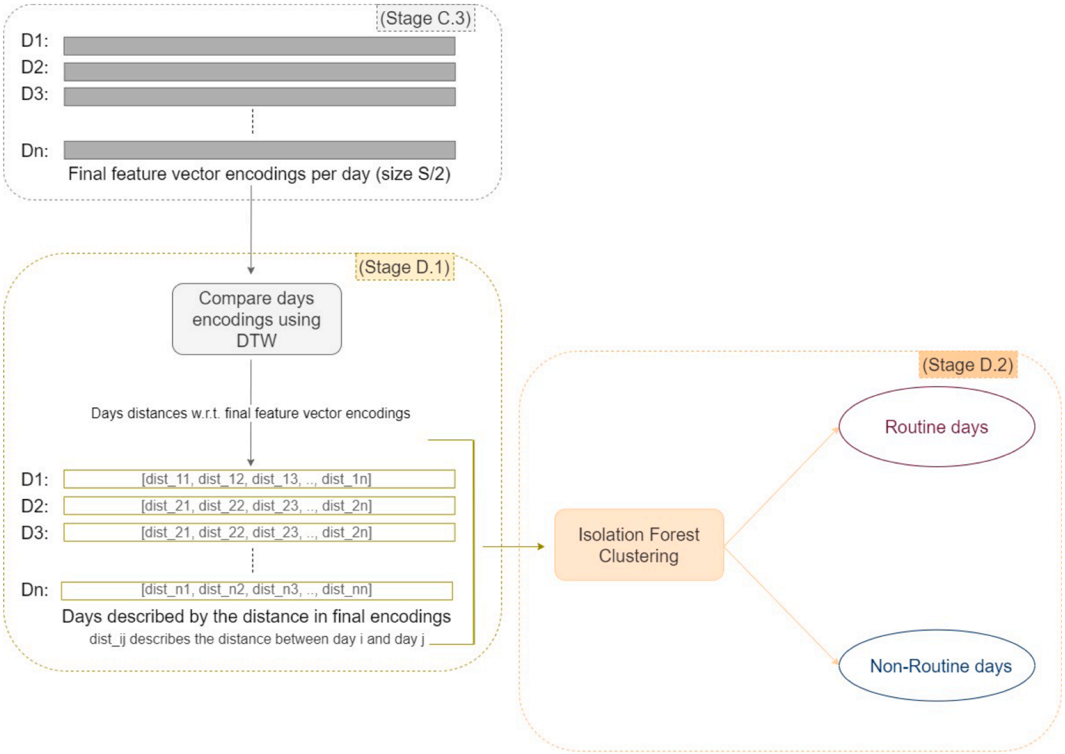

(Stage C.3)
D1:
D2:
D3:
Dn:
Final feature vector encodings per day (size S/2)
(Stage D.1)
Compare days encodings using DTW
Days distances w.r.t. final feature vector encodings
D1:
[dist_11, dist_12, dist_13, .., dist_1n]
D2:
[dist_21, dist_22, dist_23, .., dist_2n]
D3:
[dist_21, dist_22, dist_23, .., dist_2n]
Dn:
[dist_n1, dist_n2, dist_n3, .., dist_nn]
Days described by the distance in final encodings
dist_ij describes the distance between day i and day j
(Stage D.2)
Isolation Forest Clustering
Routine days
Non-Routine days

Fig. 6. Stage D: The extracted day encodings are compared against each other in stage D.1 using Dynamic Time Warping (DTW). A day is finally described as how distinct it is compared to the rest of the recorded days. This information is passed to stage D.2 which produces the routine vs. non-routine clustering of all the recorded days of the user using the Isolation Forest approach.

Table 1
Distribution of *EgoRoutine* dataset including the number of recorded days and the total corresponding number of images for each of the 7 users in the dataset, appended with the distribution of routine and non-routine days as annotated by the users.

	User 1	User 2	User 3	User 4	User 5	User 6	User 7	Total
#Days	14	10	16	20	13	18	13	104
#Images	20,543	11,815	21,727	18,977	17,046	16,592	11,207	117,907
#Routine days	11	7	12	7	6	15	7	65
#Non-routine days	3	3	4	13	7	3	6	39

In this way, the LSTMs have the capacity to 'unlearn' information that is no longer relevant and overcome the exponential error decay of RNNs, which improves the ability of the LSTM to adapt to time changes and gaps in the data and provide better results in the sequence mapping.

The LSTM autoencoder can be formalized as a tool consisting of two LSTM neural networks, one with the purpose of encoding (i.e. the encoder) and the other with the purpose of decoding (i.e. the decoder). The two functions are related to one another as such:

$$\begin{aligned} a &= Encoder(b), \\ b^{'} &= Decoder(a). \end{aligned} \tag{7}$$

The compressed encoding vector a is used as input for the decoder, which predicts the original input feature vector b. The goal of the model is to minimize the loss, which we define as the Euclidean distance between the predicted $b^{'}$ and the original b. To achieve this, as much information as possible is preserved from the original vector b, while ensuring that the encoding a captures sufficient information for the reconstruction of the original vector.

3.4. Day encodings clustering for identification of nutritional routine and non-routine days

For the final stage of the pipeline, the day encodings from stage C are encoded using an LSTM and finally clustered for the identification of eating-social routine-related days. An overview of the final stage can be found in Fig. 6.

The encoding representation of the days are further compared using DTW in stage D.1. Even though the encoding are not given in a timeline format, they still preserve the temporal characteristic of the previous representation of the days from which they have been derived. The results of the DTW algorithm provide a comprehensive view on how distinctive the days are. Each day is therefore represented by a sequence of distances corresponding to the pairwise DTW distance applied between the respective day and the rest of the recorded days.

Finally, stage D.2 identifies the routine and non-routine days. The day representations given by stage D.1 are clustered using the Isolation Forest method (Liu et al., 2008). Moreover, the authors in Talavera et al. (2020) have shown that Isolation Forest outperformed other clustering approaches for the specific task of nutritional routine identification on the EgoRoutine dataset, also employed in this work. The Isolation Forest algorithm is an anomaly detection method based on tree ensamble: the method chooses a feature randomly for which it selects a random value within the maximum and a minimum margins; using recursive partitions chosen at random, it builds a tree-like structure; it decides on anomalies based on the length of the paths: a shorter path (i.e. closer to the root of the tree) is more likely to describe an anomaly given that outlier data is known to be more sparse. These path lengths are used in the anomaly score, based on the normalized and average distance of the paths; we compute the anomaly score $a(x,n)$ for observation x, given a set of n samples as follows:

$$a(x,n) = 2^{\frac{-E(h(x))}{c(n)}}. \tag{8}$$

where $h(x)$ is the path length of point x from the root node to the last external node; $E(h(x))$ identifies the average of $h(x)$ from a collection of isolation trees. The average path length is denoted by $c(n)$ as follows:

$$\begin{aligned} c(n) &= 2H(n-1) - \frac{2(n-1)}{n}. \\ H(n) &= ln(n) + \gamma \ (Euler's\ constant). \end{aligned} \tag{9}$$

The anomaly outcome of the clustering represents the non-routine days, while the non-anomaly days are considered to be the routine days. Finally, based on these clustering results, visualizations are created in order to better understand the detected routines and their components. The visualizations will be presented and qualitatively analysed in the following section.

4. Experimental framework

This section presents the dataset used for measuring the performance and applicability of our model. The experimental setup is also presented, as well as the metrics applied for the evaluation of the conducted experiments.

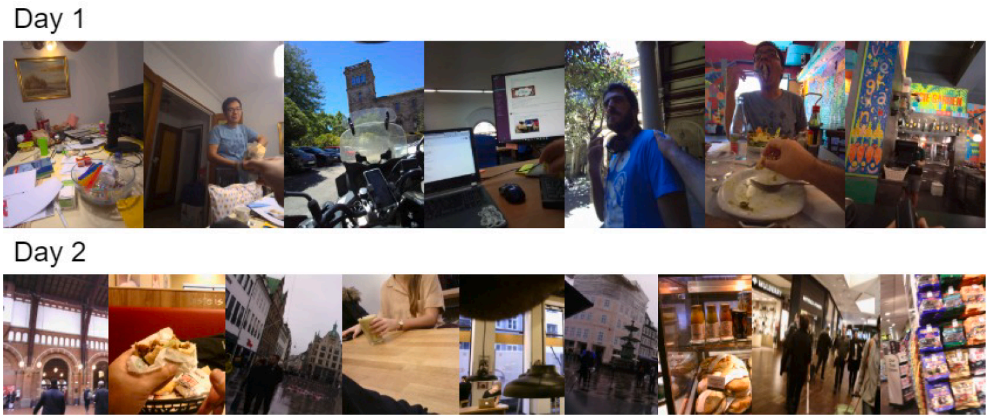


Fig. 7. Examples of images in the *EgoRoutine* dataset extracted from two random days in the dataset.

Table 2
Overview of all the conducted experiments for the ensemble of the day representations to be used for the routine discovery clustering. The features used per experiment are indicated with *x*. * indicates that DTW was applied. Time eating (TE).

		Exp. 1*	Exp. 2*	Exp. 3	Exp. 4	Exp. 5	Exp. 6*	Exp. 7*	Exp. 8*
Extracted features	Nutritional dist	x	x	x	x	x	x	x	
	Social dist	x	x	x	x	x	x	x	
	TE	x	x	x	x		x	x	x
	TE alone	x	x	x	x		x	x	x
	TE one pers	x	x	x	x		x	x	x
	TE group	x	x	x	x		x	x	x
	Top 5 Loc		x		x			x	
	Top 5 %		x					x	
	Social time					x			x
Day representations before clustering	Directly on Features	x	x						
	Custom distance			x	x	x			
	Directly on encodings						x	x	x

4.1. *Dataset*

In this work, we employ the *EgoRoutine* dataset proposed in Talavera et al. (2020) to evaluate our proposed system. This dataset consists of unlabeled egocentric data captured by 7 distinct users over different time spans (measured in days). The collections of egocentric photo-streams capture the daily lives of the users such as working with a PC, walking in street, eating alone, shopping at the supermarket, having lunch with colleagues, among others. An overview of the dataset distribution over the 7 users can be seen in Table 1. The gathered data describes a real-life scenario, i.e. the dataset predominantly capture non-food related environments since these tend to represent a relatively small fraction of a person's day. Nonetheless, we observed that the collections also includes significant instances of the users in food and social environments, as can be seen in Fig. 7. The dataset also includes user annotations of the recorded days. The users have been asked to label a day in terms of routine behaviour which spans both the social and eating aspects. This resulted into a binary categorization of the days into 'routine' and 'non-routine' days. Table 1 shows an overview of the distribution of routine and non-routine days for all the 7 users.

4.2. *Experimental setup*

Due to the modularity of our proposed pipeline, we evaluate the performance of the different stages that compose it.

- *Individualized person identification per user:* In our proposed method, we empirically observed that the best grouping was obtained when the values *Eps*(ϵ) and *MinPts* are set to 0.5 and 5, respectively. The distance function used is the Euclidean distance.
- *Day descriptors manipulation and final day encoding extraction:* We experimented on how to optimally describe the days for the most effective routine clustering. We experimented with the combinations of multiple day descriptors, both nutritional and social, in order to gain insight about which of these proposed features are most representative of the routine.

  The auto-encoder proposed for our pipeline is based on a ReLu activation layer (Ramachandran, Zoph, & Le, 2017) and an Adam optimizer (Kingma & Ba, 2014) architecture for the LSTM encoding and decoding layers. Our implementation is built on top of the existing models proposed in Chollet et al. (2015), for the Keras deep learning framework. In terms of encoding sizes, we experimented with various values in order to preserve as much of the meaningful features of the data as possible while ensuring a size reduction. The size of the encoding is half the size of the intermediate feature vector, which has been determined empirically.
- *Routine vs non-routine related days discovery:* We evaluate the performance when applying clustering directly on the extracted features, on the computed distance between combinations of extracted features or on the encoding obtained from the extracted features. All clustering parameters remain the same for the different experiments, specifically the number of Isolation Forest estimators used is 1000. The random seed state used is 0, which determines the random factor. These parameter values were determined to perform best empirically, as proved in Talavera et al. (2020).

Table 2 presents a comprehensive overview of all the performed experiments of our ablation study. The column titles of the extracted features correspond to the features extracted by the pipeline as per the overview in Section 3.3. The column titles of the ensemble of day representation refer to the data the clustering was applied on: directly on the selected day features, on the custom distance of the days (see Eq. (10)), directly on the extracted LSTM encodings from the days described by the selected features. The selected features for the day descriptors and the ensemble method for each of the experiments are indicated by an x in the corresponding column. The * indicates that these experiments also include the variant of applying DTW on the day representations before clustering (we denote these experiments as Exp. N, where N is the number of the original experiment). DH stands for the Hamming distance and DE for the Euclidean distance, the distance formula is given by:

$$\begin{aligned}
&Custom\_dist(d1, d2) = w_0 * DTW(d1\_N, d2\_N)+\\
&w_1 * DH(d1\_S, d2\_S) + w_2 * DE(d1\_TET, d2\_TET)+\\
&w_3 * DE(d1\_TEA, d2_TEA) + w_4 * DE(d1\_TEO, d2\_TEO)+\\
&w_5 * DE(d1\_TEG, d2\_TEG) + w_6 * DTW(d1\_top5, d2\_top5)+\\
&w_7 * DE(d1\_ST, d2\_ST)
\end{aligned} \tag{10}$$

where the day features correspond to the features extracted by the pipeline (see Section 3.3) for day *d*.

A weight vector *w* is defined for each experiment employing the custom distance, in order to signal which features of the day will be enabled in the computation of the custom distance. We have the following weights vectors for experiments 3, 4 and 5, respectively: (1, 1, 1, 1, 1, 1, 0, 0), (1, 1, 1, 1, 1, 1, 1, 0), (1, 1, 0, 0, 0, 0, 0, 0, 1).

4.3. *Implementation details*

Our proposed system works sequentially. Therefore, its complexity can be estimated by computing the aggregated complexities of the modules that compose it. On one side, given an image, pre-trained convolutional neural networks are applied in a fast-forward manner, i.e. no training involved, for food scene classification, which has a complexity of $\mathcal{O}(n)$ (He & Sun, 2015), where n is the number of pixels of the input image. This complexity is the result of adding the complexity of the different layers that compose the network namely, convolutional, ReLu and max-pooling layers, all with a complexity of $\mathcal{O}(n)$. On the other side, person re-identification is performed using DBSCAN. This algorithm visits each sample in the set, i.e. detected face, only once and has an average run-time complexity of $\mathcal{O}(m\,logm)$, where m is the number of sample images in the set (Birant & Kut, 2007). Afterwards, an

Table 3
Silhouette score of the clustering of routine and non-routine days for different experiments; the experiments are defined in Table 2. The NA results were obtained for the cases in which all days of the user were clustered as part of the same group which made the silhouette score inapplicable. The results greater then 0.5 are highlighted in the table.

	User 1	User 2	User 3	User 4	User 5	User 6	User 7	All Users
OurSystem*(Exp. 6.1)*	**0.545**	**0.613**	**0.576**	**0.614**	0.428	**0.553**	0.436	**0.538 ± 0.077**
Exp. 1	0.341	0.488	0.292	0.314	0.237	0.143	0.218	0.290 ± 0.109
Exp. 1.1	**0.546**	**0.555**	**0.523**	**0.503**	0.316	0.166	0.404	0.431 ± 0.145
Exp. 2	0.336	0.488	0.373	0.107	NA	0.172	0.218	0.282 ± 0.141
Exp. 2.1	**0.545**	**0.555**	**0.524**	**0.503**	0.321	0.165	0.394	0.430 ± 0.145
Exp. 3	0.433	0.472	0.293	**0.560**	**0.649**	0.220	0.277	0.415 ± 0.158
Exp. 4	0.475	0.470	0.293	**0.556**	**0.646**	0.221	0.275	0.419 ± 0.159
Exp. 5	0.309	0.171	0.268	0.484	0.258	0.145	0.155	0.256 ± 0.118
Exp. 6	**0.516**	0.410	0.287	**0.547**	0.260	0.186	0.443	0.379 ± 0.136
Exp. 7	0.330	**0.636**	0.429	0.121	0.238	0.372	**0.839**	0.424 ± 0.243
Exp. 7.1	0.458	**0.533**	0.491	0.290	0.192	**0.685**	**0.527**	0.454 ± 0.164
Exp. 8	0.150	NA	0.291	NA	0.255	0.352	NA	0.262 ± 0.084
Exp. 8.1	0.359	NA	0.356	NA	0.272	**0.571**	NA	0.390 ± 0.127

Table 4
Silhouette score of the discovered routine and non-routine related clusters by our system against the method proposed in Talavera et al. (2020).

	User 1	User 2	User 3	User 4	User 5	User 6	User 7	All Users
OurSystem	**0.545**	**0.613**	**0.576**	**0.614**	0.428	**0.553**	0.436	**0.538 ± 0.077**
Method in Talavera et al. (2020)	0.477	0.478	0.272	0.341	0.156	0.137	0.222	0.298 ± 0.141

LSTM network is applied for the encoding of temporal descriptors in a fast-forward manner. In this setting, LSTMs are linearly affected by the size of the input and have a time complexity per weight of $\mathscr{O}(n)$ (Hochreiter & Schmidhuber, 1997). Days are compared using DTW (Salvador & Chan, 2007), with a complexity of $\mathscr{O}(d^2)$, where d represents the encoding vector of a day. Finally, routine discovery with Isolation Forest has employed that has complexity of $\mathscr{O}(m\,log\psi(n))$ (Liu et al., 2008), where m, and $\psi(n)$ represent the size of the testing data and the sub-sampling size for the training, respectively. Therefore, we can conclude that the complexity of our routine discovery system is $\mathscr{O}(n + m\,logm + d^2)$.

4.4. Evaluation metrics

We use different evaluation metrics for the different stages that compose our eating-social behavioural patterns pipeline.

- *Face clustering/ Stage A:* For the evaluation of our proposed pipeline, we use several metrics depending on the different stages of the pipeline. For the evaluation of the faces clustering employed in stage A of the pipeline, we use three metrics: Silhouette score (Rousseeuw, 1987) and the Structural Similarity Index (SSIM) (Hore & Ziou, 2010; Wang, Bovik, Sheikh, & Simoncelli, 2004). These metrics are used for the evaluation of the internal cohesion of the clusters.

  The Silhouette score metric describes the relatedness of each point with respect to the cluster group it has been assigned to, and it is described by the following equation:

$$Silhouette_{score} = \frac{b(i) - a(i)}{max(a(i), b(i))}. \tag{11}$$

  where $a(i)$ is the average distance between point i and points withing the same cluster, and $b(i)$ is the minimum average distance from i to points in the other clusters.

  In contrast, SSIM is a metric used to compare two images and measure their similarity; SSIM is associated with the human visual system with respect to quality perception of images. It is a full reference metric, meaning that one of the compared images is considered to be the original. SSIM accounts for image distortions based on three factors: loss of correlation, luminance distortion and contrast distortion.
- *Day encoding/Stage C:* For the evaluation of the auto-encoder in step C of the pipeline, we use Mean Squared Error (MSE) (Wang & Bovik, 2009) to compare the quality of the encoding against the original feature vector. We do this by using the decoding layer which decodes the produced encoding. The decoder should produce a feature vector with a high similarity to the original feature vector for a good estimated performance of the auto-encoder.

The Mean Square Error (MSE) estimator is given by Eq. (12), where the two instances to be compared are represented by f and g, having the same dimensions MxN. In our work, MSE is used to compare two images represented as matrices for evaluating the face clustering and for the evaluation of the encoding given by the LSTM autoencoder:

$$MSE(f, g) = \frac{1}{MN}\sum_{i=1}^{M}\sum_{j=1}^{N}(f_{ij} - g_{ij})^2. \tag{12}$$

We also compute the Mean Absolute Error (MAE) in order to analyse the performance quality of the encoder. MAE represents the average of the absolute errors and is computed according to Eq. (13). Similarly to MSE, two instances of the same dimensions are compared element wise. This metric accounts for the absolute difference between the elements, unlike the MSE.

$$MAE(f, g) = \frac{1}{MN}\sum_{i=1}^{M}\sum_{j=1}^{N}|f_{ij} - g_{ij}|. \tag{13}$$

Moreover, we compute the Mean Euclidean (MED) distance, as per Eq. (14), between the original feature vector and the feature vector resulting from the decoding. A small MSE index and Euclidean distance correspond to higher accuracy of the encoding with respect to the original feature vector:

$$MED(f, g) = \frac{1}{N}\sqrt{\sum_{i=1}^{N}(f_i - g_i)^2}. \tag{14}$$

Table 5
Evaluation of the DBScan facial clustering in stage A of the proposed pipeline.

Metric	User 1	User 2	User 3	User 4	User 5	User 6	User 7	All Users
Silhouette Score	**0.319**	0.243	**0.358**	0.275	**0.301**	0.252	0.291	0.291 ± 0.039
Mean SSIM	0.622	0.585	**0.778**	0.576	0.603	**0.693**	**0.679**	0.648 ± 0.072

Table 6
Evaluation of the accuracy of the day encodings obtained in stage C of the proposed timeline. The mean MSE and Euclidean distance scores give the average difference between the intermediate feature vector (i.e. the original feature vector) describing the day (see Fig. 5 in the Appendix) and its decoded final day encoding given by the decoder layer of the autoencoder. The lowest scores are highlighted in the table.

	User 1	User 2	User 3	User 4	User 5	User 6	User 7	All Users
Mean MSE	**3,349.58**	9,924.15	9,504.83	19,664.31	7,735.90	15,333.18	6,806.21	10,331.16 ± 5,083.29
MAE	299.46	423.27	515.49	208.31	297.08	389.57	144.26	325.34 ± 127.68
Mean Euclidean dist	**1,143.50**	**1,487.22**	1,992.29	3,819.56	1,636.51	3,270.30	**1,496.90**	2120.89 ± 941.42
Size original feature vector	32	24	36	44	30	40	30	33.71 ± 6.27
Size encoding	16	12	13	22	15	20	15	16.14 ± 3.35

- *Eating-social routine related days/Stage D:* The output of the clustering in stage D of the pipeline is evaluated by the Silhouette score of the outcome clusters. As such, we gain knowledge with respect to the quality and consistency of the identified clusters. We mainly focus on evaluating the obtained routine clusters qualitatively, however we also compare the obtained results with the annotations that indicate the user-defined routine versus non-routine days. We evaluate the the performance of the clustering w.r.t. the given annotations with Accuracy, Precision, Recall, and F-Measure metrics.

5. Results

This section presents the results obtained for the conducted experiments described in Section 4. We evaluate and present the achieved performance for some specific elements in our proposed pipeline, specifically the face clustering employed in stage A, as well as the LSTM autoencoder in stage C (see Fig. 2). Finally, we analyse the discovered routine for one of the users in the *EgoRoutine* dataset, through means of meaningful visualisations of the results given by our proposed method for the specific user.

- *Overall results of the proposed pipeline.* Based on the various types of feature vectors representing a recorded day from a user, as obtained from the experiments in Table 2, we applied the routine identification Isolation Forest clustering method. Table 3 illustrates the Silhouette scores per user, for each of the attempted experiments. Experiment 6.1, corresponding to our proposed pipeline, obtains an average Silhouette score of 0.538 over all the users. Our method is significantly superior, since no other experiments achieve an average Silhouette score over the 0.5 mark. We observe that the performance of the experiments is highly user dependent, for example Experiment 7 achieves a Silhouette score of 0.839 for user 7. However, it does not generalize well for the rest of the users (i.e. user 4 has only a 0.121 score). Moreover, some experiments (i.e. experiment 2, 8 and 8.1) do not measure any Silhouette score for some users, since all the days attributed to the respective user are assigned to the same routine or non-routine cluster. This indicates that the features included in the day representation are not descriptive for the routine of the user. Comparing the second best ranking experiment (i.e. experiment 7.1) with our proposed approach, it becomes evident that considering more information about the days does not translate into higher performance (i.e. as per Table 2, experiment 7.1 includes information about the top 5 food-related locations identified throughout a day in addition to the features considered by our model in experiment 6.1).

  Our method generalizes well for the users in the dataset, 5 out of 7 users obtaining a Silhouette score higher than 0.5, with no user having a score lower than 0.4. This indicates that the collection of features and processing steps included in our proposed method are representative for the recorded days and capture the nutritional routine. Given that no users are identical, an alternative approach would imply customizing the routine discovery user-wise: choosing the experiment which achieves the highest score for each particular user.

  Moreover, in Table 4, we compare the performance of our proposed model and the one proposed in Talavera et al. (2020) due to their similarity. We can observe how even though both methods find coherent clusters, with a score >0, the clusters discovered by our model are of higher quality.
- *Face clustering:* Table 5 showcases the clustering results for the identified people for all users using DBSCAN with Euclidean distance. We measure the Silhouette score and Structural Similarity Index Measure (SSIM) within each identified cluster, and report the average values per user. In terms of Silhouette score, the obtained results do not exceed the 0.3 mark, however, since the clustering is applied on images depicting human faces, the Mean SSIM index offers better insight into the cohesion within the face clusters. The obtained SSIM scores are larger than the 0.5 mark for all users, user 6 obtaining the highest score of 0.778. This indicates that images of faces belonging to the same cluster are structurally similar, which accounts for the fact that the depictions of the face of the same individual can be captured from different angles or in different lighting, depending on the image instance. This reinforces our assumption of face clusters representing particular individuals.
- *Evaluation of autoencoder*

  For the evaluation of the autoencoder employed during stage C of the pipeline (see Fig. 5), we measure the accuracy of the given final encoding. Using the decoder layer, we obtain the reconstructed feature vectors from the final encoding. The reconstructed feature vectors are subsequently compared to the original feature vectors (i.e. the input to the auto-encoder) using MSE, MAE, and mean Euclidean distance. The obtained results per user are shown in Table 6. As mentioned in Section 3.3, the size of the encoding differ per user, being influenced by the number of days recorded for the user. The number of recorded days increases the size of the intermediate feature vector in stage C of the pipeline, which is the original feature vector passed as input to the auto-encoder.

  Since a most accurate decoding is desired, low MSE, MAE, and Euclidean distance scores are sought. We observe that the lowest average MSE score is obtained for user 1. There is a tendency of high increase in MSE score for the encoding of large dimensions such for

Table 7
Quantitative metrics of the routine vs. non-routine clustering for our proposed model and the model proposed in Talavera et al. (2020).

	Our System			Method in Talavera et al. (2020)		
	Precision	Recall	F1-Score	Precision	Recall	F1-Score
Non-Routine	0.41	0.28	0.33	0.47	0.18	0.26
Routine	0.64	0.75	0.69	0.64	0.88	0.74
Average (Avg)	0.52	0.52	**0.51**	0.55	0.53	**0.50**
Weighted Avg	0.55	0.58	**0.56**	0.58	0.62	**0.56**
Accuracy (Acc)		0.58			0.62	
Weighted Acc		**0.52**			**0.52**	

user 4 and user 6, which obtained the highest score. This implies that the auto-encoder might have faced some challenges in preserving all the meaningful information in the original feature vector consisting of a large number of features. A larger encoding might have preserved more information. However, a balance must be reached since the size of encoding have a direct influence on the clustering results. Even with a MSE score of 19,664.31 denoting the accuracy of the user's day encoding, the corresponding Silhouette score of the routine clustering is an adequate 0.614. The MAE values indicate small differences with respect to the original – encoded vectors. The lowest value belongs to user 7, 144.26, but overall the standard deviation of 127.68 indicates consistency across all users. In terms of mean Euclidean distance between the original vectors and the decoding, the distance is more or less consistent (no more than 2000) for the users with a lower size of the original feature vector, however, as seen with MSE for users 4 and 6, the distance spikes. This indicates that there is room for improvement in terms of finding the optimal encoding length, personalized for each user.

Table 8
Mean nutritional and social metrics (in minutes) per user in the EgoRoutine dataset. The metrics are reported over all the recorded days, over the routine and, respectively, non-routine days.

	Mean times	Time eating (TE)	TE alone	TE w/ one pers	TE in group	Social interaction time
User 1	overall	65.00	20.20	5.90	38.90	157.01
	Routine	42.81	22.46	4.65	15.70	161.59
	Non-routine	104.96	16.15	8.16	80.65	94.00
User 2	overall	62.70	39.76	9.50	13.45	94.00
	Routine	41.95	33.30	6.87	1.78	71.55
	Non-routine	111.11	54.83	15.62	40.66	146.39
User 3	overall	62.34	24.74	22.09	15.51	86.04
	Routine	66.55	29.49	19.48	17.58	74.91
	Non-routine	49.70	10.50	29.92	9.28	119.44
User 4	overall	101.39	60.41	19.35	21.63	62.37
	Routine	81.50	33.96	15.91	31.63	85.01
	Non-routine	157.86	122.00	29.90	5.96	12.67
User 5	overall	131.37	64.69	31.49	35.19	70.95
	Routine	161.04	79.37	40.97	40.70	76.25
	Non-routine	44.62	31.68	4.07	8.87	50.56
User 6	overall	40.33	11.66	11.26	17.40	72.54
	Routine	27.95	9.23	8.93	9.79	51.41
	Non-routine	83.65	20.18	19.42	44.04	146.50
User 7	overall	72.51	32.36	15.86	24.28	71.63
	Routine	67.63	33.24	12.22	22.16	68.86
	Non-routine	88.77	29.42	28.00	31.34	80.86

The experimental analysis indicates that, given a collection of egocentric photo-streams, the best set of descriptors is the day representation using the LSTM encoding of the concatenated Nutritional distance and Social distance, together with the computed TE, TEA, TEO, TEG. As for the set of parameters that allow the best performance of the methods in our system, we obtained the following setting: DBSCAN with 0.5 Eps and 5 MinPts for face clustering, i.e. people re-identification; Isolation Forest with 1000 estimators and random seed of 0 for non-routine anomaly detection; and LSTM with ReLu activation Layer, an Adam Optimizer and N/2 encoding size, where N is the size of the intermediate feature vector.

5.1. Routine vs. non-routine clustering results

The *EgoRoutine* dataset was made publicly available together with 'routine' and 'non-routine' labels, which were assigned by the camera wearer. These labels accounted for the occurrence and relation of the whole set of activities that users perform throughout the days and therefore are not only focused on social-nutritional habits. However, we find it interesting to evaluate the goodness of our model for the task of 'general' routine discovery given a collection of days. We compare the performance of our model against the labels obtained in Talavera et al. (2020). Results are shown in Table 7.

We can observe that our system obtains an overall 0.58 and 0.52 score in accuracy and weighted accuracy, respectively. The highest metric score obtained is the recall of the 'routine' class (i.e. a score of 0.75) which shows that the model can successfully identify and correctly cluster the relevant data instances (i.e. the recorded days) that have been categorised as 'routine' days by the users. The somewhat reduced performance of the model w.r.t. the 'non-routine' category is justified by the reduced number of non-routine days in the data set (i.e. 39 non-routine days compared to the 65 routine days). Moreover, non-routine behaviour is more diverse and inconsistent: for example, 2 non-routine days recorded by the same user are not necessarily expected to follow the same behavioural eating pattern since this is highly dependent on the undertaken activities. The performance of our system is similar to the one of Talavera et al. (2020). We can observe that both obtain the same weighted accuracy and average F1-score. If we relate to the results described in Table 4, our method produces a better representation in the data space of the days since the clusters are of higher quality. In contrast to the model proposed in Talavera et al. (2020), our pipeline also accounts for the social nature of eating. Results show that this added complexity does not represent a drawback for the discovery of routine-related days but a richer description. A detailed account of how the social features are impacting the clustering of the 'routine' and 'non-routine' days is presented through the case study in Section 5.2.

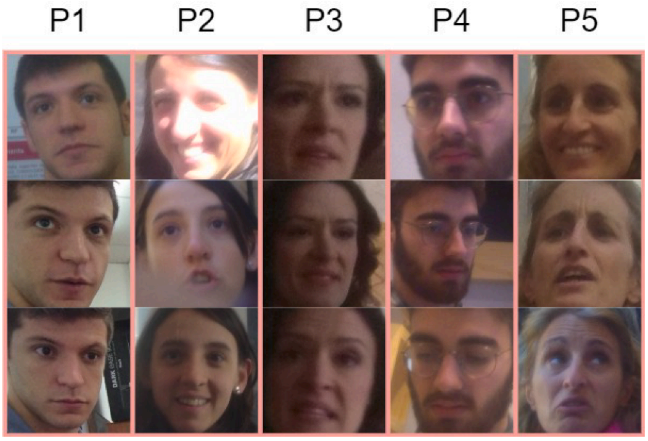


Fig. 8. Example of several people identified in the collection of photo-streams of user 6 based on the face clustering from stage A of the proposed pipeline.

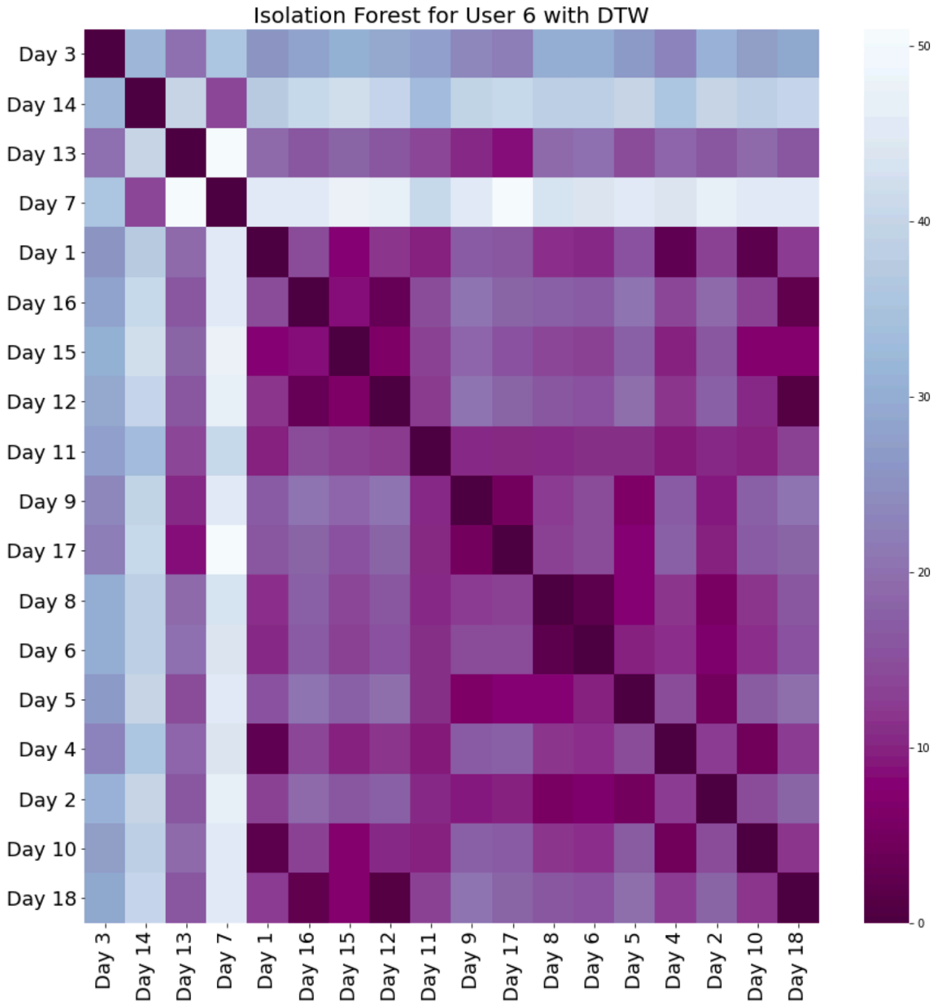


Fig. 9. Routine clustering output resulting from our method applied to user 6.

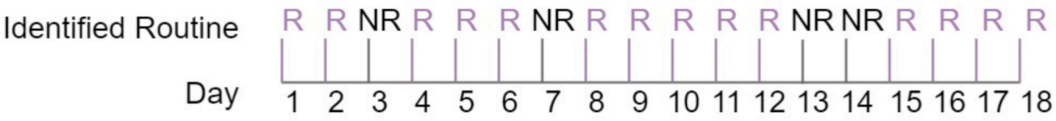


Fig. 10. Identified routine over the chronologically ordered days for user 6 in the EgoRoutine dataset.

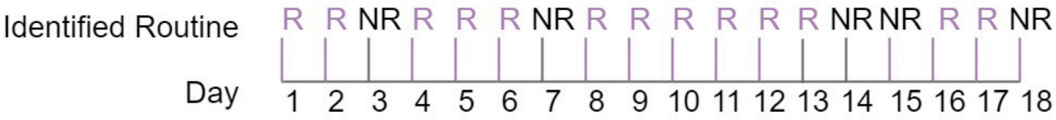


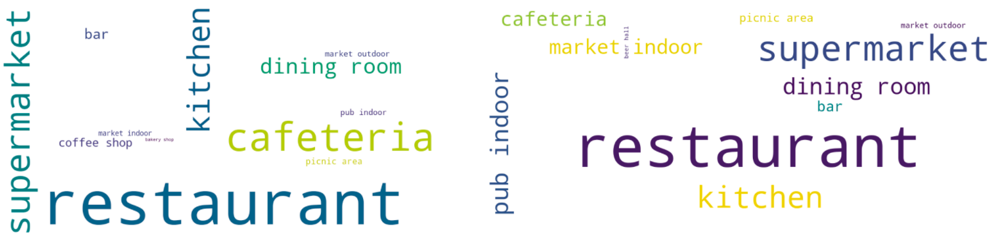
Fig. 11. Identified routine resulting from Talavera et al. (2020) over the chronologically ordered days for user 6 in the EgoRoutine dataset.

5.2. *User specific results: a case study*

We present the results at user level and as average to show the opportunities that the proposed model presents for behaviour understanding from visual data. Later, and due to space limitations, we describe the identified routines for user 6 in the EgoRoutine dataset and treat it as a case study. We analyze this user's eating and social habits in order to showcase the applications of our proposed pipeline model.

Table 8 displays the total eating and social interaction times in minutes, for each user. These times are accumulated over all images recorded by each user and over all the images corresponding to the routine and non-routine days. The instances in which an increase in total eating time also corresponds to an increase in social eating (i.e. time eating with one other person or in a group) are highlighted in the table. The increase is judged in the context of the user.

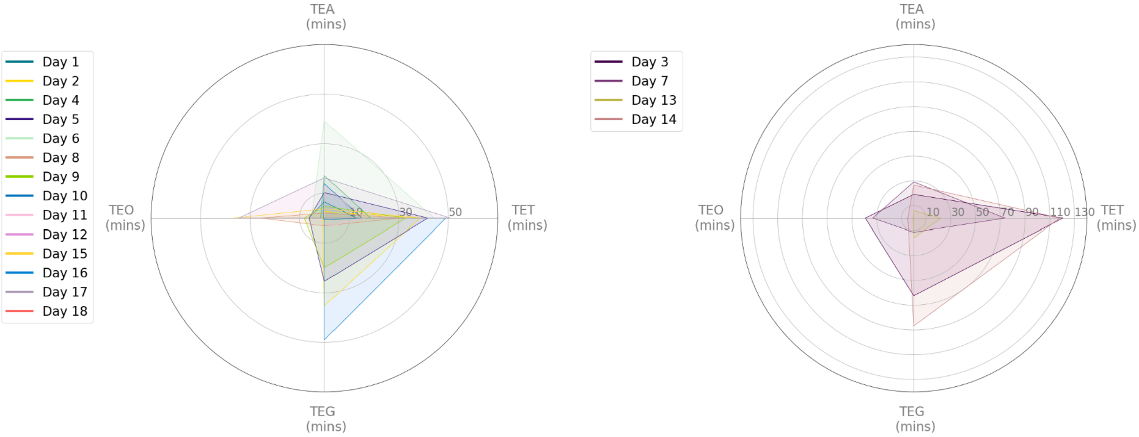
As per Table 8, the correlation between increased total eating and social eating times occur for almost all users (except user 4). In approximately all cases (except user 5), the enhancing influence of social eating on the total eating time is present for the non-routine days. This implies that most users maintain a nutritional and social balance,



(a) Routine

(b) Non-routine

Fig. 12. Word cloud representation of the identified clusters by Isolation Forest for user 6. Word cloud (a) correspond to the identified routine and word cloud (b) to the remaining group of outliers which describe non-routine behaviour. The size of the food scene names indicates the frequency of the food scene within the routine, the colours were only chosen for presentation purposes.



(a) Routine days

(b) Non-routine days

Fig. 13. Social interaction times with respect to food routine. Plots illustrate total time eating (TET), time eating alone (TEA), time eating with one person (TEO) and time eating with more than one person (TEG). All times are illustrated in minutes.

quantitatively speaking, during their routine; contrasting, during non-routine days, they have the tendency to over-indulge themselves in social eating contexts, as hypothesised in the introduction of this work. It is also notable that, for some users (i.e. users 2, 5, 6, 7), an increase in social eating times also corresponds to an increase in general social interaction times. Therefore, it can be inferred that food/meal sharing is a meaningful opportunity for creating and maintaining social bonds. Given the rather limited amount of data at hand, a definite correlation between the social influence on eating habits cannot be decisively determined. However, the results presented in Table 8 are powerful indicators that eating within a social context can lead to more time spent eating and, probably, to larger quantities of food consumed.

A sample of the identified persons with whom user 6 has interacted throughout the recorded days is presented in Fig. 8. The persons have been re-identified in various situations by means of facial recognition; this can be observed in the images included in Fig. 8 through the change in backgrounds of the images. Moreover, the social engagement with the user is visible from the facial expressions captured in the images, i.e. the identified persons are captured smiling, talking, etc.

Fig. 9 showcases the identified routines in the re-ordered daily routine distance matrix for user 6 in the EgoRoutine dataset. The routines are based on the discovered relations among days obtained by our proposed method. We identify two aspects: a routine cluster (i.e. days 1–2, 4–6, 8–18, with the exception of days 13 and 14), and non-routine behaviour, consisting of days 3, 7, 13, and 14. These groups are visualized by day in Fig. 10.

In contrast, Fig. 11 indicates the same group visualization, but using the routine identification method proposed in Talavera et al. (2020). We notice that the user maintains a regular routine, with only small deviations in the form of few non-routine days, which may account for the difference between week days and weekends. For example, non-routine days 13 and 14 could fit a weekend period. The distance matrix in Fig. 9 also shows some similarity between the non routine days 3, 14, and 13, which also form a small cluster, indicating similar habits.

Compared with the results from Talavera et al. (2020), which introduces a solely nutritional approach to identifying eating routine for the EgoRoutine dataset, we notice that similar non-routine days were selected, in the form of days 3, 7, and 14. Our proposed model also identifies day 13 as non-routine, unlike the model in Talavera et al. (2020), which also considers days 15 and 18 non-routine. The overlap of non-routine days indicates that the user strayed from their usual habits significantly during those days, whether nutritionally or socially. The differently selected non-routine days indicate that, perhaps, the user strayed from routine from a nutritional standpoint (less time spent in food-related environments, for example) but behaved in a similar way socially, thus our proposed method may consider them routine, unlike the comparison method of Talavera et al. (2020). The routine days are selected similarly for both models, which solidifies the idea of a regular routine being maintained by this user.

Given the detected routines, we computed the word clouds in Fig. 12 that illustrates the most common food-related scenes seen in the detected routine days and non-routine days, respectively. In terms of common food scenes, there is some overlap between the routine and non-routine, in the form of the 'restaurant' and 'supermarket' classes, which are, overall, the most common for all users (Talavera et al., 2020). We can notice differences in terms of low appearance classes: for instance, the class 'pub indoor' appears only in non-routine, whereas the class 'coffee shop' appears only in the routine cloud. This strengthens the

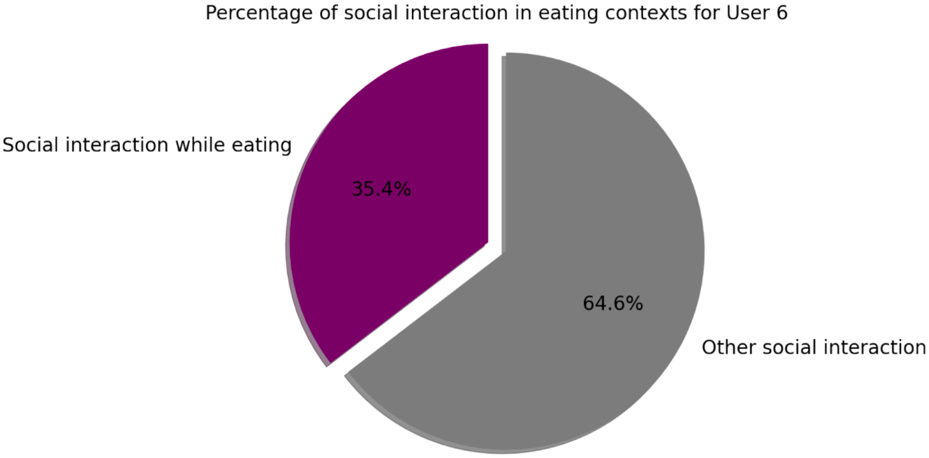


Fig. 14. Social eating interaction with respect to total social interaction identified for user 6.

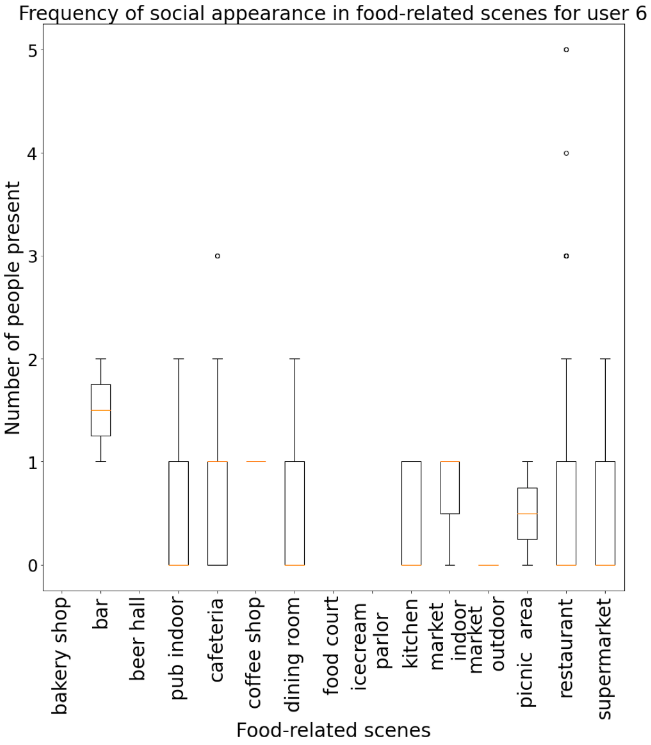


Fig. 15. Distribution of identified people per food-related scene for user 6.

assumption that routine days might correspond to working days, when the user is forced to follow an appointed schedule, while non-routine days to free days, when the user has more freedom of choice in terms of spare time, entertainment activities they can follow.

Based on the identified clusters, we further investigate the social eating aspects with respect to routines. The spider plots in Fig. 13 showcase the amount of time (in minutes) spent in a social eating or interaction context during routine and non-routine days, respectively. What is more, in Fig. 14 we showcase the percentage of social eating in contrast to the total social interaction identified over all days of the user.

For the routine days, a similarity in the shape of the spider plots for the days forming the cluster can be observed. This implies similarity in interaction patterns, although the specific times may differ. The total amount of time dedicated to eating habits daily is relatively low (at most 50 min), out of which the user balances the time eating alone with time eating in the company of other people, with a predilection of group eating. On the other hand, the non-routine spider plot shows various shapes along to time differences, indicating higher variety of behaviour in those days. The total time eating is significantly higher than in the case of routine days (over 100 min spent eating for day 3). What is more, the predisposition of the user for social eating in groups is evident, days 3 and 14 exhibiting high amounts both in total time eating and time eating in a group. The appreciation for social eating is also described by the percentage in Fig. 14, which showcases that almost 40% of the total social interaction of the user is realised in social eating circumstances.

Further investigating the social eating behaviour of the user, Fig. 15 illustrates the mean and standard deviation of people identified in the different food environments for the same user. On average, the user seems to interact with one or sometimes two people, which is consistent with the data shown in Fig. 13, for the routine days, which shows significant periods of time spent eating with one person or in a group (i.e. two or more people). It appears that the user tends to enjoy eating with social company.

Taking into account both aspects previously discussed, food scenes and social interaction, a timeline indicating the daily activity distribution was created. Fig. 16 illustrates the balance of food scenes and social activities within the recorded days classified as routine.

Social interaction is consistent throughout most of the user's recorded days, and although the timing may differ, social eating is present as well. Certain blocks of mixed food activity and social eating may actually represent a lengthier social eating event differing in classification due to the camera position (i.e. people not present in part of the images consisting of food-related scenes). Overall, a routine of socialization and eating towards the second half of the day (social or not) can be seen throughout this user's recordings.

6. Discussions

A broader image of behaviour is captured when investigating the correlation between nutritional and social habits. The knowledge of how nutritional habits are inferred by the social life of people can be applied in practice for discarding unhealthy habits, both from a nutritional and social standpoint, and for establishing new routines for an improved lifestyle. We address this challenge from a computer vision and machine learning point of view, and propose an image-based system for assisted living and well-being monitoring. We believe this can have a relevant societal impact and can have a positive influence in the health care system.

Our proposed model has shown promising results indicating its suitability for the discovery of habits through the analysis of egocentric photo-streams when considering various food-related environments, eating times, as well as social factors. It provides with information regarding the user's lifestyle in the hope of instilling changes and

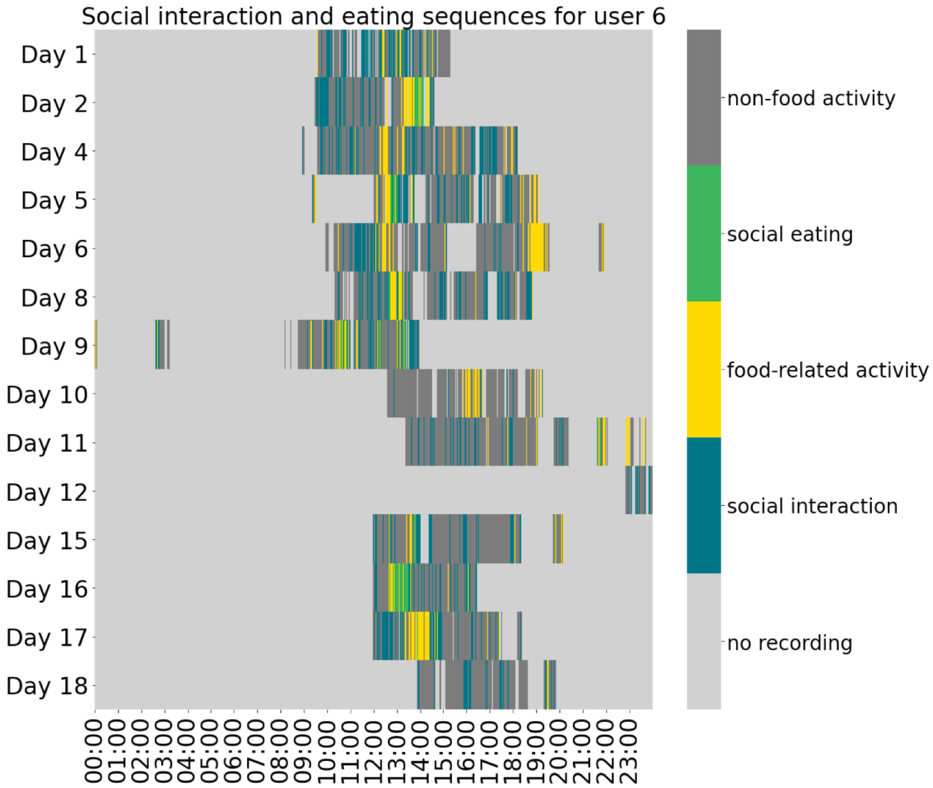


Fig. 16. Timeline of identified routine days with respect to food and social scenes. Showcases both food-related activities as well as social interactions, accounting for social eating moments.

improving overall health.

The design of the proposed pipeline is modular, which implies that the different stages that compose it can be decouple and used individually, as well as replaced by other approaches. For instance, Stage A and Stage B could be decoupled together from the entire pipeline with the aim of computing other various metrics regarding the overall social interaction of the user per day, for example: total interaction time, time spent with one person, time spent in a group. These metrics can offer meaningful insights into the general behaviour of the user.

The proposed visualization aim to provide visual insights into the lifestyle for both, the person who recorded the data and/or the specialist who evaluates the habits of that person. We believe that clear and informative visualization are of importance for an easy interpretation of the obtained results when people with different backgrounds are involved.

Future lines of research will explore the incorporation of different descriptors of the daily activities of the user. What is more, personalized systems could be created for each user where the day description stage is customized based on a case by case basis. For instance, if a user might have a more active or social life, the day descriptors could weigh more these aspects. Within the proposed system, future studies can also focus on finding the optimal trade-off between encoding accuracy and encoding size of the auto-encoder in stage C of the pipeline. We think this could lead to increased performance in the identification of routine-related days. Moreover, based on the obtained results and discovered information, the incorporation of a recommendation system for the improvement of the user's daily routine would lead to healthy living. Therefore, an end-to-end system capable of automatically processing the input data to later recommend actions to the user would be of high relevance for healthcare professionals and the general population.

7. Conclusions

The proposed automated system has successfully discovered social-nutritional routine and non-routine behavioural patterns through the analysis of egocentric photo-streams gathered by wearable cameras. These results provide personally tailored supervision of habits, showing the potential of our tool for innovative applications in the smart industry.

We also propose tools for the visualization of the discovered routines by the proposed unsupervised learning methods. The visualization accounts for both food-related environments and social interactions, thus providing insights for the time as well, which can allow specialists for the evaluation of the lifestyle of the camera wearer.

CRediT authorship contribution statement

Andreea Glavan: Methodology, Software, Data curation, Visualization, Investigation, Formal analysis, Writing - original draft, Writing - review & editing. **Alina Matei:** Methodology, Software, Data curation, Visualization, Investigation, Formal analysis, Writing - original draft, Writing - review & editing. **Petia Radeva:** Conceptualization, Investigation, Resources, Formal analysis, Writing - review & editing. **Estefania Talavera:** Conceptualization, Methodology, Resources, Supervision, Formal analysis, Investigation, Writing - original draft, Project administration, Writing - review & editing.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgment

This work was partially founded by projects RTI2018-095232-B-C2, SGR 1742, CERCA, Nestore Horizon2020 SC1-PM-15-2017 (n° 769643), Validithi EIT Health Program and GreenHabit EIT Digital Program projects. The founders had no role in the study design, data collection, analysis, and preparation of the manuscript.

References

Aizawa, K., Maruyama, Y., Li, H., & Morikawa, C. (2013). Food balance estimation by using personal dietary tendencies in a multimedia food log. *IEEE Transactions on Multimedia, 15*, 2176–2185.

Baldi, P. (2012). Autoencoders, unsupervised learning, and deep architectures. *Proceedings of ICML workshop on unsupervised and transfer learning*, 37–49.

Bengio, Y. (2009). *Learning deep architectures for AI*. Now Publishers Inc.

Birant, D., & Kut, A. (2007). St-dbscan: An algorithm for clustering spatial–temporal data. *Data & Knowledge Engineering, 60*, 208–221.

Bolanos, M., Dimiccoli, M., & Radeva, P. (2016). Toward storytelling from visual lifelogging: An overview. *IEEE Transactions on Human-Machine Systems, 47*, 77–90.

Bradski, G. (2000). The OpenCV library. *Dr. Dobb's Journal of Software Tools*.

Chollet, F. & et al. (2015). Keras. https://keras.io.

Clarkson, B. (2002). Life patterns: Structure from wearable sensors. Ph.D. thesis. Massachusetts Institute of Technology.

Donini, L. M., Savina, C., & Cannella, C. (2003). Eating habits and appetite control in the elderly: The anorexia of aging. *International Psychogeriatrics, 15*, 73–87.

Dunbar, R. (2017). Breaking bread: The functions of social eating. *Adaptive Human Behavior and Physiology, 3*, 198–211.

Greff, K., Srivastava, R. K., Koutník, J., Steunebrink, B. R., & Schmidhuber, J. (2016). Lstm: A search space odyssey. *IEEE Transactions on Neural Networks and Learning Systems, 28*, 2222–2232.

Gurrin, C., Smeaton, A. F., Doherty, A. R., et al. (2014). Lifelogging: Personal big data. *Foundations and Trends in Information Retrieval, 8*, 1–125.

Hamrick, K. S., Andrews, M., Guthrie, J., Hopkins, D., McClelland, K. & et al. (2011). How much time do americans spend on food. US Department of Agriculture, Economic Research Service, 1–58.

He, K., & Sun, J. (2015). Convolutional neural networks at constrained time cost. In *2015 IEEE conference on computer vision and pattern recognition (CVPR)* (pp. 5353–5360). https://doi.org/10.1109/CVPR.2015.7299173

Herman, C. P. (2017). The social facilitation of eating or the facilitation of social eating? *Journal of Eating Disorders, 5*, 1–5.

Herruzo, P., Portell, L., Soto, A. & Remeseiro, B. (2017). Analyzing first-person stories based on socializing, eating and sedentary patterns. In International conference on image analysis and processing (pp. 109–119). Springer.

Higgs, S., & Thomas, J. (2016). Social influences on eating. *Current Opinion in Behavioral Sciences, 9*, 1–6.

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation, 9*, 1735–1780.

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation, 9*, 1735–1780.

Hopkinson, J. B., Wright, D. N., McDonald, J. W., & Corner, J. L. (2006). The prevalence of concern about weight loss and change in eating habits in people with advanced cancer. *Journal of Pain and Symptom Management, 32*, 322–331.

Hore, A. & Ziou, D. (2010). Image quality metrics: Psnr vs. ssim. In 2010 20th international conference on pattern recognition (pp. 2366–2369). IEEE.

Kagaya, H., Aizawa, K. & Ogawa, M. (2014). Food detection and recognition using convolutional neural network. In Proceedings of the 22nd ACM international conference on multimedia (pp. 1085–1088). ACM.

Kingma, D. P. & Ba, J. (2014). Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980.

Laska, M. N., Hearst, M. O., Lust, K., Lytle, L. A., & Story, M. (2015). How we eat what we eat: Identifying meal routines and practices most strongly associated with healthy and unhealthy dietary factors among young adults. *Public Health Nutrition, 18*, 2135–2145.

Li, Y. (2018). A deep spatiotemporal perspective for understanding crowd behavior. *IEEE Transactions on Multimedia, 20*, 3289–3297.

Liu, F. T., Ting, K. M., & Zhou, Z. (2008). Isolation forest. In *2008 Eighth IEEE international conference on data mining* (pp. 413–422). https://doi.org/10.1109/ICDM.2008.17

Müller, M. (2007). Dynamic time warping. *Information Retrieval for Music and Motion*, 69–84.

Park, P., Marco, P. D., Shin, H., & Bang, J. (2019). Fault detection and diagnosis using combined autoencoder and long short-term memory network. *Sensors, 19*, 4612.

Pujol, O., Radeva, P., & Vitria, J. (2006). Discriminant ecoc: A heuristic method for application dependent design of error correcting output codes. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 28*, 1007–1012.

Ramachandran, P., Zoph, B. & Le, Q. V. (2017). Searching for activation functions. arXiv preprint arXiv:1710.05941.

Rousseeuw, P. J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics, 20*, 53–65.

Salvador, S., & Chan, P. (2007). Toward accurate dynamic time warping in linear time and space. *Intelligent Data Analysis, 11*, 561–580.

Stalonas, P. M., & Kirschenbaum, D. S. (1985). Behavioral treatments for obesity: Eating habits revisited. *Behavior Therapy, 16*, 1–14.

Talavera, E., Cola, A., Petkov, N. & Radeva, P. (2019a). Towards egocentric person re-identification and social pattern analysis. arXiv preprint arXiv:1905.04073.

Talavera, E., Glavan, A., Matei, A. & Radeva, P. (2020a). Eating habits discovery in egocentric photo-streams. arXiv:2009.07646.

Talavera, E., Leyva-Vallina, M., Sarker, M., Kamal, M., Puig, D., Petkov, N. & Radeva, P. (2019b). Hierarchical approach to classify food scenes in egocentric photo-streams. arXiv preprint arXiv:1905.04097.

Talavera, E., Petkov, N. & Radeva, P. (2019c). Unsupervised routine discovery in egocentric photo-streams. arXiv preprint arXiv:1905.04076.

Talavera, E., Wuerich, C., Petkov, N., & Radeva, P. (2020). Topic modelling for routine discovery from egocentric photo-streams. *Pattern Recognition, 107330*.

Varini, P., Serra, G., & Cucchiara, R. (2017). Personalized egocentric video summarization of cultural tour on user preferences input. *IEEE Transactions on Multimedia, 19*, 2832–2845.

Wang, Z., & Bovik, A. C. (2009). Mean squared error: Love it or leave it? A new look at signal fidelity measures. *IEEE Signal Processing Magazine, 26*, 98–117.

Wang, Z., Bovik, A. C., Sheikh, H. R., & Simoncelli, E. P. (2004). Image quality assessment: From error visibility to structural similarity. *IEEE Transactions on Image Processing, 13*, 600–612.

Yi, D., Lei, Z., Liao, S. & Li, S. Z. (2014). Deep metric learning for person re-identification. In 2014 22nd international conference on pattern recognition (pp. 34–39). IEEE.

Zheng, W. S., Gong, S. & Xiang, T. (2011). Person re-identification by probabilistic relative distance comparison. In CVPR 2011 (pp. 649–656). IEEE.