Pervasive and Mobile Computing 95 (2023) 101846

Contents lists available at ScienceDirect

Pervasive and Mobile Computing

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

Behavioural patterns discovery for lifestyle analysis from egocentric photo-streams

Martín Menchón [a,*], Estefania Talavera [b], José Massa [c], Petia Radeva [d,1]

[a] INTIA, UNCPBA, CONICET, Argentina
[b] University of Twente, Enschede, The Netherlands
[c] INTIA, UNCPBA, Argentina
[d] University of Barcelona, Barcelona, Spain

ARTICLE INFO

Keywords:
Behaviour analysis
Routine discovery
Pattern discovery
Data mining
Egocentric vision
Lifelogging

ABSTRACT

Automatic tools for the analysis of human behaviour are very important when aiming to understand the lifestyle of people. Egocentric wearable cameras allow the capture of images during long periods of time and in this way bring objective evidence of the experiences of the user.

In this paper, we propose a novel framework to discover behavioural patterns following an unsupervised greedy approach based on extracted image descriptors. The method collects and constructs time-frames to extract the semantics of user behaviour in terms of contextual information, such as places, activity, present objects, and others. Later, the similarity among the user time-frames is computed to assess correlations and thus obtain the user's routine descriptors. To evaluate the performance of our method, we present several score metrics and compare them to state-of-the-art works in the field. We validated our method on 315 days and more than 390,000 images extracted from 14 users. Results show that behavioural patterns can be successfully discovered and that they are able to characterize the routine of people bringing important information about their lifestyle and behaviour change.

1. Introduction

Understanding human behaviour is increasingly gaining attention from the computer science community [1–3]. Aiming at developing automatic tools for the analysis of people behaviour, researchers rely on the use of wearable sensors. These devices allow the automatic capture of behavioural information throughout the day. This information tends to be represented in the form of a time series. These data can localize a person at a certain time [4], display the activity happening at a certain moment in time [5], deduce the emotional state of a person [6], identify social interactions [7] or help monitoring health conditions [8].

Different wearable sensors are available nowadays, the most popular being the GPS and accelerometers. These devices are able to acquire information during long periods of time. However, their gathered data lack semantics, they are not readable by the user and thus do not allow for a deep understanding of the user experiences. In contrast, visual information does allow the comprehension of the context, leading to a user-friendly display of the behaviour of a camera wearer. Wearable cameras are devices that are worn

* Correspondence to: INTIA, Facultad de Ciencias Exactas, Universidad Nacional del Centro de la Provincia de Buenos Aires, Paraje Arroyo Seco, Campus Universitario (7000), Tandil, Argentina.

E-mail addresses: mmenchon@exa.unicen.edu.ar (M. Menchón), e.talaveramartinez@utwente.nl (E. Talavera), jmassa@exa.unicen.edu.ar (J. Massa), petia.ivanova@ub.edu (P. Radeva).

1 Fellow IAPR.

https://doi.org/10.1016/j.pmcj.2023.101846
Received 27 March 2023; Received in revised form 23 August 2023; Accepted 21 September 2023
Available online 27 September 2023

Fig. 1. Visual timeline with collected images by one of the users in the *EgoRoutine* dataset. Rows represent the different collected days and columns indicate time. White spaces correspond to not recorded moments.

attached to clothing, and capture a first-person view of the life of the wearer. The collection of images gathered with a wearable camera is known as egocentric photo-streams. The frame rate of the camera limits the amount of information that one can gather. Low frame rate cameras allow collection of images throughout long periods of time (e.g. the whole day), while high frame rate cameras allow a detailed description of the activity at hand (e.g. sport activity).

Until today, many approaches have been introduced in the literature to address different tasks within computer vision that allow a better understanding of the activities performed by the camera wearer. Just to mention some examples, such tasks include: recognition of daily life activities [9], prediction of next action in an activity [10], detection of places [2,11], social interactions [12], routine-related days identification [1], among others.

In this work, we propose a step forward presenting a robust approach for the discovery of patterns of behaviour from egocentric photo-streams. Instead of considering a fixed, predetermined (usually, small) number of actions [3], we are interested in discovering the actions that characterize the behaviour of a user in an unsupervised framework. Our method extracts and constructs descriptors from images in form of semantic concepts automatically. The method follows a greedy fashion that looks for common descriptive structures, i.e. beginning from a seed it extracts structures seeking similar user behaviour throughout the time periods of different days. The result of our learning approach is a set of routine patterns characterizing the behaviour of a user. This differs from our previous work in [1] in that routine was described as a collection of patterns of behaviour that provide a summary of the lifestyle of the user. Here, the concept *Pattern* relates to the sequence of activities that a person follows in her or his daily routine. We understand that patterns differ in complexity, i.e. number of sequential activities a person follows daily. These patterns represent a solid and quantified base to draw a summary of the lifestyle of the person (see Fig. 1).

This work builds on top of our preliminary work towards the discovery of patterns in egocentric photo-sequences [13], where we presented an unsupervised greedy method to identify a set of behavioural patterns from a semantic clustering methodology based on the similarity among the time-frames that reflect the days that have been gathered for a user.

In this paper, the contributions are five-fold:

1. We present a novel and robust methodology for behavioural pattern discovery from egocentric photo-streams based on extracted semantic information from the images.
2. We introduce the use of n-grams of different order to describe sequence of behavioural patterns.
3. We provide an extensive analysis for the performance of our model by testing on four different publicly available egocentric datasets. An ablation study with respect to hyperparameters of our proposed pipeline shows the best configuration of our proposal.
4. We propose a novel metrics, we call Representativity in order to evaluate the quality of the extracted patterns.
5. We designed a framework for qualitative evaluation based on a questionnaire for user assessment. Moreover, we propose a proper visualization scheme to illustrate the extraction of routine discovery.

The rest of the paper is structured as follows: In Section 2, we describe related works. In Section 3, we define our approach for the discovery of behavioural patterns from collections of egocentric photo-streams. In Section 4, we give the experimental setup and results. There, we present a detailed discussion on the obtained results in Section 5. Finally, we conclude this work in Section 6.

2. Related works

Human behavioural pattern discovery is a challenging task. One can agree the richness of all human activities characterizing the human behaviour. Most likely, on one hand a user would do more than a prefixed number, because of the high diversity among the activities that an individual can follow [14]. Nowadays, there is a lack of common tools and devices to gather information about the behaviour of individuals.

Behaviour and wearable sensors. In [15], a review of different strategies that rely on different sources of data for human behaviour is described. The major part of the proposed approaches addresses the task of human behaviour analysis as a supervised learning problem, i.e. learning to identify a small set of pre-defined human activities (e.g. 12 activities like running, walking, etc.). However, most likely, we can expect that a user would do more than a prefixed number (e.g. 12) of activities. Moreover, different users tend to perform different activities. For example, we expect different activities when thinking of a child or an elderly. Another example is the work proposed in [16], where the authors present a tool for classification of a limited set of activities based on the analysis of collected values by sensors. In this work, the term *pattern discovery* is described by the authors as "extracted information from low-level sensor data without any predefined assumptions or models". This work was based mainly on sensor data and cannot be extended for the analysis of unknown activity classes. However, the method is not capable of analysing the variety of scenarios that can appear in the daily life of a person, since this kind of devices are unable to give information capturing the semantic meaning of the activities or the surrounding environment. We think it is key for behaviour understanding, that the problem is addressed from a discovery perspective following an unsupervised approach. This means that we should aim at learning to discover patterns of activities with no pre-defined activity labels. Creating such a robust model, without the restriction of a predefined and limited set of human activities to recognize, is still an open issue in the literature.

The discovery of daily routine can be seen as the identification of patterns of behaviour. Patterns have appeared in the literature when addressing other tasks. For instance, in order to detect thematic patterns in time series, [17] introduced a model for the discovery of the theme in a video. To that end, the authors proposed to address the task as a cohesive sub-graph mining problem and define as binary quadratic programming problem the route to reach the solution. Although the work shows robustness for thematic pattern discovery, their proposed approach is not able to find common patterns in different time-sequences. In our work, this is relevant since our goal is to analyse time series that describe the days of camera wearers. Furthermore, one theme or topic is not sufficient to describe the behaviour of an individual. In human behaviour analysis, multiple pattern discovery is needed followed by undefined tasks so that it can adapt to individual patterns and situations.

Human behaviour and computer vision. Human behaviour is a broad concept addressed in the computer vision literature. The field includes activity recognition in images [18] and videos [19], social relationship recognition [20], sentiment analysis [21], crime recognition [22,23], crowd behaviour analysis [24], among others. Most of this work focus on the recognition of a certain set of classes following an unsupervised approach. They tend to stay within the defined frame of their training dataset. There are few works in the literature that could be used in the discovery of behavioural patterns. From the state-of-the-art in the field, we gain insight into the best way of describing images. For instance, using global descriptors extracted using pre-trained convolutional neural networks as in [25,26], or using detected objects and activities as semantic descriptors as in [1].

Human behaviour through egocentric vision. Through the analysis of daily experiences described from a first-person view, i.e. egocentric perspective, we can get understanding the behaviour of a person. This concept has grown interest in the computer vision community, more specifically in the field of computer vision. Previous works have addressed the task of temporal segmentation by detecting strong changes in the temporal line [27,28]. In [27], extracted semantics in the form of bag-of-words in combination with global descriptors extracted from the images are analysed. In [28], the authors identify the boundaries between events by relying on an LSTM-based model that compares the visual context generated from a set of frames in the past to the predicted visual context in the future. The authors in [29] focused on the location depicted in the image to segment videos into different time-slots. Despite identifying time-slots, these works segment the video, but do not describe what these segments represent and do not study how the segments are related throughout the data collection. Furthermore, in *Organizing egocentric videos of daily living activities* [30], the authors proposed a framework to automatically segment and organize a set of egocentric videos of daily living scenes. Through an unsupervised temporal segmentation, each egocentric video is divided in scenes by considering the extracted features of a neural network. The segmented videos can provide rich descriptors for video analysis that are useful for query, retrieval and lifestyle description. However, they do not claim to discover user routine as a concept of repeating events typical for a user.

Routine analysis through egocentric vision. In the field of egocentric vision, there are several works that focus on the analysis of camera wearer behaviour. In [1], the authors proposed an unsupervised framework for photo-stream (day) classification into routine and non-routine. Topics are discovered based on extracted semantics from the images. Later, a representation of the day is built based on the relevance of the topics. A clustering method is used to identify similar days, which define the routine-related cluster. They describe a day as routine or non-routine while here we discover routine patterns as a combination of repeated events, that do not necessarily cover the whole day. In this way, we provide a more fine-grained definition of the descriptors of the identified routine patterns.

Table 1
Explanation of the used notations throughout the paper.

Notation	Explanation
$n(i, j)$	Node corresponding to day i and timeslot j
S_i, S_j	Scene labels of the ith and jth time-slots
A_i, A_j	Activity labels of the ith and jth time-slots
O_i, O_j	Array of objects corresponding to the ith and jth time-slots
$DH(S_i, S_j)$	Hamming distance between scene labels S_i and S_j
$DH(A_i, A_j)$	Hamming distance between activity labels A_i and A_j
$DJ(O_i, O_j)$	Jaccard distance between array of objects O_i and O_j
T_i, T_j	Time slots of the nodes expressed in seconds normalized by the day length
p	Constant defining the importance of the different terms in the distance
q	Constant defining the importance of the time difference term in the distance
P	Aggregated set or cluster of nodes
S	Set of nodes
neighbours(_)	Neighbours of a node
Entropy(P)	Entropy of the cluster P
x_i	Newly added node to set P
$F(x_i)$	Occurrence frequency of the distance between x_i and the centroid of P
DV	Set of all distance values between existing nodes of P
dx_i	Distance between x_i and P
N	Total number of days
K	Total number of time-slots

Clustering egocentric images in passive dietary monitoring with self-supervised learning. In [31], the authors introduced a self-supervised learning framework for clustering large volumes of egocentric images into separate events. The goal is to simplify the post-processing and annotation tasks associated with analysing a large volume of captured images. The framework utilizes a two-stream structure with multi-task learning, combining a masked autoencoder (MAE) branch and a contrastive learning branch. The MAE branch reconstructs the original images and computes the mean squared error loss, while the contrastive learning branch focuses on learning discriminative image features. The two branches are combined using a joint loss function.

Our first steps towards routine pattern discovery. Our preliminary work in *Behavioural pattern discovery from collections of egocentric photo-streams* [13] presented a case study of our proposed greedy approach for behavioural pattern discovery. This work represents an extension of that research line by extending the greedy framework to detect patterns of different orders defining relations of actions in the daily routine. We perform an ablation study at a hyperparameter level and by benchmarking on publicly available egocentric datasets in order to illustrate the performance of our proposal. Furthermore, we show that by introducing an entropy-based measure to form and analyse the clusters of images, a more consistent output is achieved. We compare the performance of our framework to the state-of-the-art approaches, such as *Organizing egocentric videos of daily living activities* [30] and show that our proposal outperforms the state-of-the-art as well as provides an efficient computational method for behavioural pattern discovery.

3. Behavioural routine pattern discovery

In this section, we describe our proposed pipeline for pattern discovery from egocentric photo-streams for lifestyle analysis. We refer to the following terms:

- *Day*: sequence of images collected by a wearable camera describing the activities performed by the camera wearer.
- *Time-slot*: set of consecutive images that are recorded within two defined time stamps.
- *Patterns/clusters*: a collection of time-slots that share descriptors and conform a group with a maximum acceptable level of entropy. These time-slots can be sequenced in time and through days.
- *n-gram*: repeated sequence of n patterns.
- *Routine*: a collection of patterns that describe the lifestyle of the camera wearer.

We also provide the notations in Table 1 to facilitate understanding throughout the paper.

Given a set of days described by photo-streams, the latter are first characterized by the detected concepts with a pre-trained deep network (see Section 3.1). Later, days are translated into graphs (see Section 3.2). To do this, photo-streams are divided into time-slots, where each time-slot represents a node in the graph (see Section 3.3). The created time-slots are compared to each other,

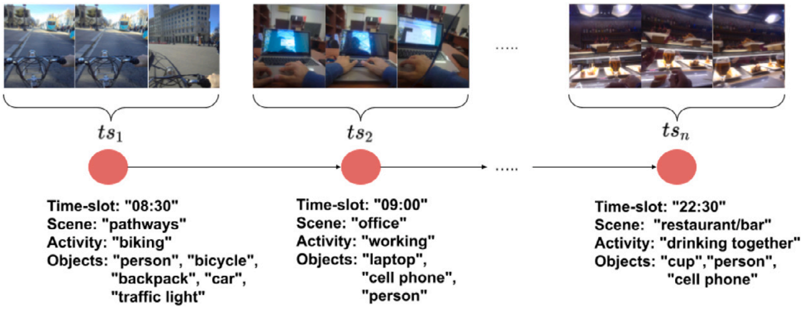

Fig. 2. Illustration of time-slots created for image sequences corresponding to 30' within the image time-series. We indicate the automatically extracted and selected labels to describe the time-slot.

our greedy algorithm defines patterns based on the computed similarities (see Section 3.4). Finally, the sequences of repetition of these patterns are extracted throughout the different days (see Section 3.5). In the following subsections, we describe in detail the different steps.

3.1. Days characterization by semantic features extraction

Most people in society tend to follow a life based on a routine. Routine has been described as the sequence of recurrent activities that a person follows in daily life [1]. A typical work-related routine day could be described by the following example of a sequence of activities: commuting, office, eating, office and commuting. If we pay attention, we can observe in this example that location and activity play an important role when describing human behaviour (i.e. commuting as an activity and office as a location). Keeping this in mind, we represent the images that compose the collected photo-streams with extracted elements in the form of labels. We extract the following labels:

- location,
- activity, and
- objects detected in the image

in order to construct a description of the visual depicted scene.

3.2. Graph representation of the day

The collected days by the users are divided into time-slots of half-hour. A time-slot is later represented by a node in a linear graph, see illustration in Fig. 2. Images are characterized by the detected concepts with pre-trained deep networks where we rely on networks pre-trained to identify scene [32], activity [9] and objects [33] that the images depict. Therefore, a node, representing a group of images, is characterized by the scene and activity with the highest occurrence, and the set of objects that appear more than K times (in our case, $K = 10$) in the group of images of each time-slot, where we select labels with high occurrence to ensure that they are representative for the time-slot. Fig. 2 depicts the process of forming time-slots and the labels that describe them.

3.3. Detecting similar time-slots

After building the linear graphs that describe the daily user behaviour, we build a metric that allows us to compare different days. The metric compares nodes representing time-slots to reflect the inherent time feature of a potential pattern. With this, we can differentiate activities such as eating that can happen at different times of the day (e.g. having breakfast, lunch or dinner). The distance d among nodes m_i, m_j is computed as follows:

$$d((S_i, A_i, O_i), (S_j, A_j, O_j)) = p\,(DH(S_i, S_j) + DH(A_i, A_j) + DJ(O_i, O_j)) + q\,(|T_i - T_j|) \quad (1)$$

where S_i, S_j are the scene labels of ith and jth time-slots, A_i, A_j are their activity labels and $O_i O_j$ are arrays containing the corresponding objects detected by Yolov4. $DH(S_i, S_j)$ is the Hamming distance that is $DH(S_i, S_j) = 1$ if S_i is different of S_j and 0 if they coincide, $DH(A_i, A_j)$ is the Hamming distance between nodes activities, $DJ(O_i, O_j)$ is the Jaccard distance [34] between nodes objects, where T_i and T_j are the time slots of the nodes expressed in seconds normalized by the day length in seconds, p

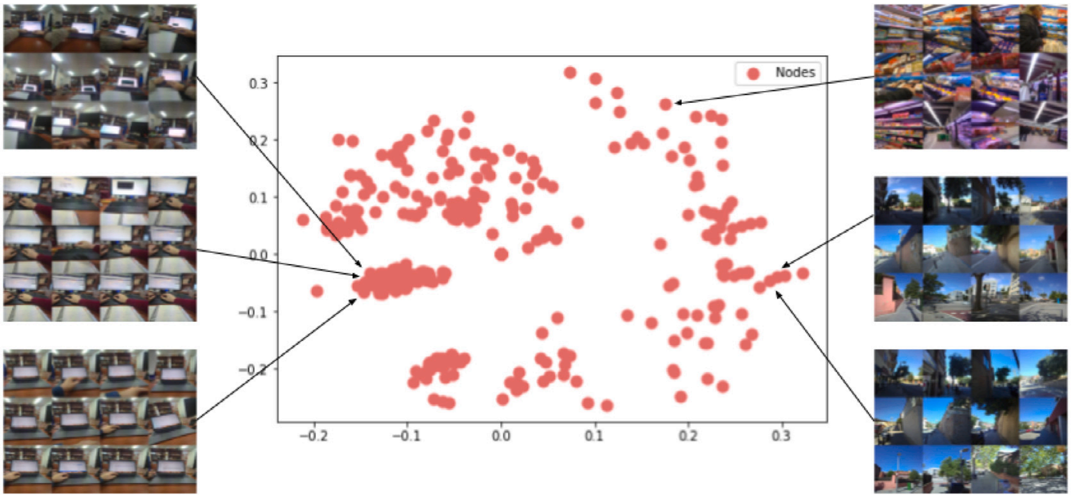

Fig. 3. Mapping time-slots corresponding to 30' by using MDS [35] on the pairwise distances computed by Eq. (1).

and q are constants defining the importance of the different terms in the distance (Eq. (1)). We found empirically that the weights p equal to 0.85 and q equal to 0.15 provide the optimal trade-off between both terms.

In order to illustrate the similarity between time-slots, we compute the pair-wise distance among the nodes of the same time-slot and apply multi-dimensional scaling (MDS) [35] to map the nodes in a two-dimensional space. Fig. 3 presents the spatial distribution of the nodes after applying the MDS. In this example, we can observe the suitability of MDS for the representation of the samples. Different nodes are separated in space, while similar nodes are placed closer. We can conclude that the spatial relationship among nodes with highly similar images is maintained.

3.4. Aggregating similar time-slots to create events across days and time

The next step is to detect routine patterns consisting of repeated similar time-slots. Our proposed approach for behaviour pattern identification follows an heuristic greedy algorithm. The method relies on a seed node that aggregates neighbouring nodes based on the defined distance (see Eq. (1)). We consider neighbours ($neighbours(\_)$) as adjacent nodes in time, as well as nodes on different days that occur at the same time. Let us consider a node $n(i,j)$ corresponding to the ith day and jth time-slot. We initialize the aggregated set P (cluster) with a seed composed of two time-slots/nodes that are the closest and belong to different days but that happen at the same time frame. Then, we determine the nodes of minimal distance to P (any element of P) from their neighbours.

Our proposed model relies on the entropy of the formed groups of nodes, i.e. clusters (P), to determine if a given node will be added to a cluster. This criterion is based on the idea that the entropy of the nodes of a cluster should remain low and that the inclusion of a node very different to the cluster nodes would increase the entropy of the cluster. Shannon entropy [36] was calculated by the discretization of the values from the multi-dimensional scaling with a precision defined by a number of bytes (equal to 2). Entropy has been used not only as a measure to compare nodes, but also as a global goodness metric [37] based on the idea that each cluster should be a subset of nodes with minimal variation i.e. minimal degree of internal inhomogeneity, considering the features node values from the point of view of the information theory.

Hence, the distance of a neighbour node to the set P is defined in terms of the entropy between them. We employ the entropy of $P' = P \bigcup x_i$ as follows (see Eq. (2)):

$$Entropy(P') = -\sum_{i=1}^{n} F(x_i) log F(x_i) \tag{2}$$

Where x_i represents the newly added node to the set P, and $F(x_i)$ is the occurrence frequency of the distance between x_i and the centroid of P. Let us consider the set of all distance values $DV = \{d_1, d_2, d_3, \dots, d_n\}$ between existing nodes of $P = \{x_1, x_2, x_3, \dots, x_m\}$ to P, and dx_i the distance between x_i and P, $F(x_i)$ is the amount of times that dx_i, discretized with a resolution of 2 bytes, appears in the set of DV.

The entropy of including the different nodes is computed and used to rank them. The node to be included in the cluster corresponds to the one that leads to the lower entropy increase. We set a value threshold to assess the maximum entropy increase allowed, above which no nodes are further included. We present the formalization of our introduced method with Algorithm 1:

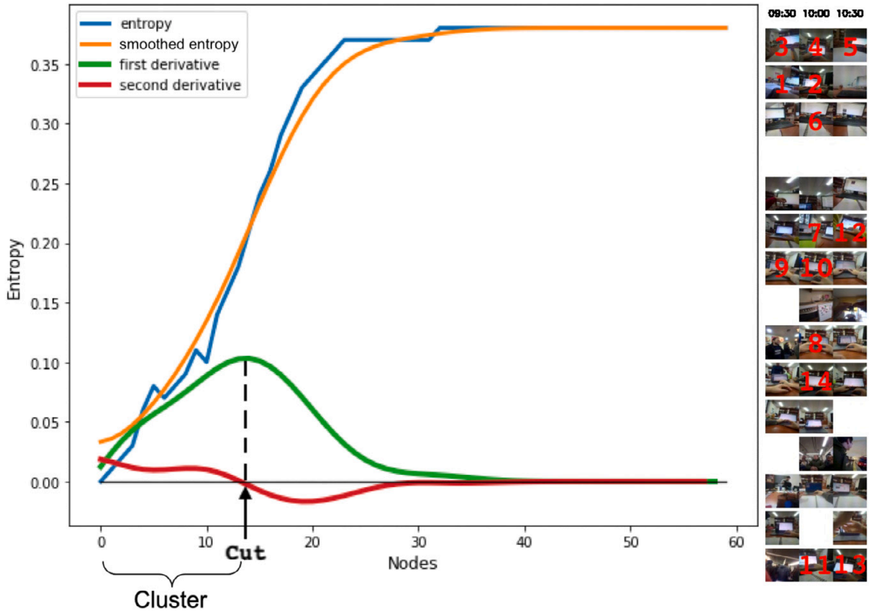

Fig. 4. Example of a cluster entropy (blue), filtered entropy signal with a Gaussian filter (orange), first (green) and second (red) derivatives of the smoothed entropy as functions of the iteration of cluster growing. The maximum value of the first derivative indicates the cut of the cluster growing, i.e. when the process stops adding time-slots. On the right, we visualize the time-slots that are sequentially added to the cluster (i.e. pattern). The red numbers show the order of the aggregated time-frames.

Initialization: Given all nodes $S_0 = S = \{n(i,j)\},\ i = 1,\dots N$ corresponding for days from 1 to N, and j=1,..K corresponding for time-slots from 1 to K, these nodes are described by the semantic set of concepts detected in the images
Output: Set of clusters P_t;

→ Let $node_0 = n(i_0, j_0)$ and $node_1 = n(i_1, j_0)$ are the closest nodes from S_0 so that $d(n(i_0,j_0), n(i_1,j_0)) = min_{w,l,z}(d(n(i_w,j_l), n(i_z,j_l))$ all over the days i_w and i_z and time-slots j_l;

→ Let $v_1 = Entropy(P_1)$ where $P_1 = \{n(i_0,j_0), n(i_1,j_0)\},\ S_1 = S_0 \setminus \{n(i_0,j_0), n(i_1,j_0)\},\ t = 1$;

while $S_t \neq \emptyset$ *or* v_t *does not fulfil the stopping criterion (see Section ??)* **do**

`// now we evaluate all neighbours from the aggregated sets`
→ Let $node_t = argmin_{node_t \in neighbours(S_{t-1})} Entropy(P_{t-1} \bigcup node_t)$;

→ Update set of nodes $P_t = P_{t-1} \bigcup \{node_t\},\ S_t = S_{t-1} \setminus \{node_t\},\ t = t+1$;

`// compute entropy of updated set of nodes` P_t
→ $v_t = Entropy(P_t)$;

end

Algorithm 1: Our proposed heuristic greedy algorithms for the discovery of behavioural patterns in collections of egocentric photo-streams.

Complexity: In the worst case, the presented algorithm performs for each candidate node, an iteration over all the nodes minus the nodes that have been already considered. Taking into account that the maximum number of nodes is $|S|$, the computational complexity is $O(|S|^2)$, where $|S| = \sum_{i=0}^{N} t_i$ indicates the number of nodes, the value N describes the number of days under analysis, and the value t_i represents the number of time-slots in day i.

Discovering routine patterns: One of the important questions in the Algorithm 1 is to define a good stopping criterion in order to extract a cluster of nodes with semantic similarity between them. We consider that similar nodes are related to small time-slot similarity criteria, and by vice-versa, if we add dissimilar nodes, the entropy will increase rapidly. Both the aggregation process and the evolution of the criterion are monotonically increasing functions and in this way, a change can be detected by means of the maximum of first derivative of v_t and zero-crossing of its second derivative. We find the cluster with similar nodes by smoothing v_t with a Gaussian filter and calculating the first and second derivatives to define a stopping criterion for the clustering process.

Stopping criterion: We define a single pattern (we call pattern of order 1) by stopping the growing of a cluster P_t at iteration t when the entropy achieves an inflection point that is when the value of the first derivative v_t' is over a defined threshold k and the second derivative v_t'' crosses zero i.e.

$$\text{Stopping criterion}(v_t, t, k, \epsilon) : v_t' > k \quad |v_t''| < \epsilon.$$

In this case, the constructed cluster gives us a **single routine pattern**. We visualize the process of finding patterns in Fig. 4. We show entropy as a function which is proportional to cluster growth at its first and second derivatives. On the right we show which of the frames from different days and in which order are aggregated to the cluster according to the minimal entropy. The columns of the frames correspond to different time-slots and the rows correspond to different days. Each cluster of frames together with the extracted semantic concepts form a single pattern. For example, the cluster in Fig. 4 corresponds to the pattern of working at the office in the morning.

Note that during the process of pattern discovery, each time-slot is only considered once. Therefore, when a pattern is detected, the time-slots that compose it are discarded from the rest of the process. In the next steps of the algorithm, the rest time-slots are analysed using the same procedure in order to find the rest of the behavioural patterns.

3.5. Pattern discovery of higher order extracted by n-grams

A pattern of behaviour can be composed of a repeated sequence of simple patterns (e.g. activities like each morning the user takes breakfast, gets the metro and goes to the work office). In order to extract a composition of repeated single patterns, here, we rely on the well-known probabilistic n-grams of different order to identify patterns of distinct order (length). The algorithm of n-grams is shown in Algorithm 2:

Input: Set of clusters P_t ($t = 1$ to N, where N is the maximum number of identified clusters) as a result of the Algorithm 1.
The individual members of this set are considered patterns of order 1 or $(1grams)$
Output: Patterns corresponding of $(2grams)$ and $(3grams)$;

→ Initialize an empty list $2G$ and $3G$ to store the 2-grams and 3-grams;

for *t=1 to N-2* **do**
→ Append the consecutive pair P_t, $P_t + 1$ to the list $2G$;

→ Append the consecutive tuple P_t, $P_t + 1$, $P_t + 2$ to the list $3G$
end

Algorithm 2: N-grams extraction from the identified clusters

We consider each identified pattern as a single element. An n-gram is defined as a sequence of n elements that occur in the photo-stream collection. An n-gram of order 1, i.e. $n = 1$ describes the simplest scenario where the behaviour is described by a single pattern. A higher degree of n-grams, i.e. high values of n, describes more complex behaviour composed of a set of n single patterns. For example, an n-gram of $n = 3$ would detect the repetitive behaviour of a person for 3 sequential activities such as commuting, working with a laptop and work-meeting that used to repeat throughout the days within a similar time interval. Fig. 5 shows an example of clusters and how to identify different n-grams by their order.

4. Experimental framework

In order to validate our proposal, in this section, we describe the datasets on which we test our proposed framework, the metrics used for the validation, and the experimental setup of our method.

4.1. Datasets

We evaluate the robustness of our model for the discovery of behavioural patterns by analysing four different egocentric datasets that describe the daily life of people in the wild. Table 2 quantitatively presents these datasets. They were collected with different wearable cameras, namely, Narrative Clip [41] and OMG Autographer [42], thus, the visual perspective slightly differs. Fig. 6 shows some sample images per dataset. The following is a brief description of the used datasets:

- *The NTCIR-12* [38], *The NTCIR-13* [39] and *The NTCIR-14* [40] are available egocentric datasets released in the Lifelog: Personal Lifelog Search challenge. The datasets are composed of 79, 91 and 43 days respectively, and were collected by a total of 7 different users.
- *Egoroutine:* dataset introduced in [1] consists of unlabelled egocentric photo-streams. A total of 114, 225 images were collected during 102 days by 7 different users capturing the daily lives of the users.

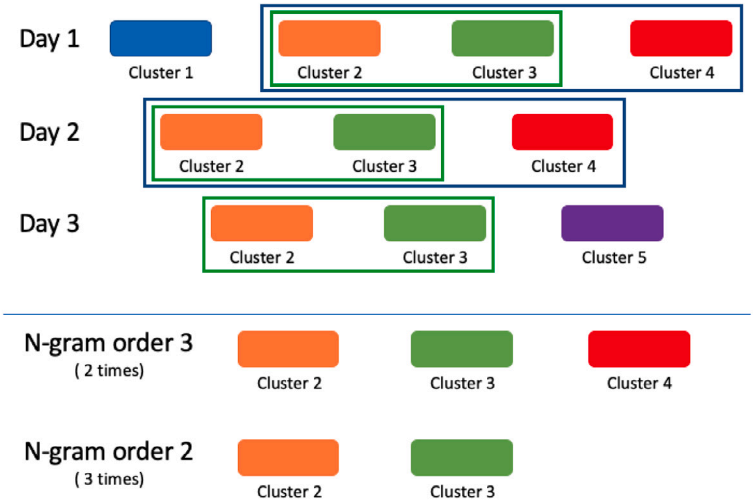

Fig. 5. Example of n-grams of orders 3 and 2. At the top, there are three days conformed by different identified clusters. The combination of clusters 2, 3, and 4 define an n-gram of order 3, which appears on days 1 and 2. The combination of clusters 2 and 3 detail an n-gram of order 2 that appears in the three days.

Fig. 6. Sample images from the different egocentric datasets evaluated in this work namely, EgoRoutine [1], NTCIR-12 [38], NTCIR-13 [39], and NTCIR-14 [40].

Table 2
Distribution of the datasets on which we evaluate our proposed method for behavioural pattern discovery.

Dataset	#Users	#Days	#Images	Avg images/day
NTCIR-12 [38]	3	79	86915	1100,19
NTCIR-13 [39]	2	91	112186	1232,81
NTCIR-14 [40]	2	43	81210	1888,60
EgoRoutine [1]	7	102	114225	1119,85
Total	14	315	394536	1252,49

Recently, several challenging egocentric vision datasets appeared like: *Ego4D* [43], the *EPIC-Fusion* [44] and the *EgoCOM* [45]. Note that no one of them covers the whole day and several weeks that are critical in order to discover routine patterns. Ego4D [43] is composed of a diverse range of daily life activities recorded from a first-person perspective. Still, despite being the largest field data set, the data collected do not describe a long and continuous collection on the time axis, being unsuitable for the task of routine analysis. Therefore, we cannot identify patterns from its collected set of different individual activities. The same happens for the *EPIC-Fusion* dataset [44], a set of videos describing food-related preparation activities in the kitchen. In the case of *EgoCOM* [45], it presents a dataset of videos from an egocentric perspective describing a conversation. Likewise the previously mentioned datasets, it is not applicable because of the lack of sequences of activities.

4.2. Validation metrics

In this work, we evaluate the performance of our proposed model and the obtained discovered routine in a quantitative and qualitative manner.

Quantitative evaluation: The routine discovery is an unsupervised learning problem in which validation cannot be evaluated in terms of accuracy. Instead, patterns are quantitatively evaluated with metrics that assess the goodness of the discovered patterns (created clusters). We follow the proposed validation metrics in [13]. Let us consider the whole set of patterns of a user $\mathbf{P} = \{P_1, P_2, \dots, P_n\}$.

Note that in the daily life of people, a pattern tends to appear on multiple days. We consider that an ideal routine pattern would correspond to a pattern that repeats in as many days as possible. Hence, we consider the occurrence of the discovered patterns and formally define it as Occ, which gives the number of days per cluster divided by the amount of clusters normalized by the number of days for the user, as follows:

$$Occ = \frac{1}{|\mathbf{P}| * N} \sum_{P=1}^{|\mathbf{P}|} ds_P. \tag{3}$$

where ds_P is the number of days a pattern P covers, N is the total number of days of the user and $|\mathbf{P}|$ is the total number of patterns found in all days of a user.

Let the duration of a pattern P within a day be defined by its starting time $t_{i,P}^{first}$ and end time $t_{i,P}^{last}$, respectively, being i a day. We consider the set of occurrences of the pattern P, corresponding to the pattern duration for each day i, defined as $t_{i,P} = [t_{i,P}^{last} - t_{i,P}^{first}]$. We compute the mean $\mu_P = \text{mean}_i(t_{i,P})$ and the standard deviation $\sigma_P = \text{std}_i(t_{i,P})$ of the durations of the pattern along the different days. We define the compactness Cmp as a function of the standard deviation of the durations as follows:

$$Cmp = max\{0, 1 - \frac{1}{|\mathbf{P}|} \sum_{P=1}^{|\mathbf{P}|} \sigma_P/\mu_P\}. \tag{4}$$

Thus, the Cmp expresses of all patterns appear with a similar duration along different days.

Note that both the Occ and Cmp vary between 0 and 1. The more days are covered by a routine, both Occ and Cmp will tend to 1. When time-frames are less covered by routine patterns, both magnitudes tend to 0. Both values are defined when we have discovered at least 1 pattern in the dataset of a user.

We define the representativity (R) of a pattern with regard to the daily life of the user as the average of its frame occurrence (Occ) and compactness (Cmp), namely:

$$R = \frac{1}{2}(Occ + Cmp). \tag{5}$$

The Occ component allows us to evaluate the number of repetitions of a pattern; the more repetitions of the pattern over the days, the higher its value. The aim is to reward the algorithm that finds the same pattern over more days.

The Cmp allows us to know the duration of the patterns found. If the value is closer to 1, the cluster is more compact. If the value is close to 0, the cluster is not very compact. The objective is to highlight the algorithm that finds more compact patterns. The best result is the one with the higher (R) value.

Silhouette score is a standard measure to evaluate the compactness of clusters. In this work, we address the discovery of behavioural patterns as an unsupervised learning problem. Therefore, as an additional validation metric we rely on the silhouette score to compare the performance of our discovered patterns and the ones identified by the other baseline approaches.

We implemented the silhouette score function available from [46], which is based on the work proposed in [47]. The silhouette function of feature i is defined as,

$$s(i) = \frac{b(i) - a(i)}{max\,(a(i), b(i))} \tag{6}$$

where:

- i represents the extracted descriptor by pretrained models. These descriptors represent either the global descriptors [30] or extracted semantics (as we propose in this work),
- $a(i)$ is the average distance between feature i and all other features in the same cluster,
- and (b) is the average distance between feature i and all features in the nearest neighbouring cluster to which it does not belong.

The silhouette value ranges between −1 and 1. A high silhouette value indicates that the feature is well-fitted within its own cluster, while a low value suggests that the feature may be closer to features in another cluster. A value close to 0 suggests that the feature is near the boundary between two clusters. The average silhouette of all features in a cluster provides an overall measure of cluster quality.

Table 3
Results for the ablation study on our method for different values of threshold k over different datasets. We present results as average for all the users. The highest representativity values for each dataset are highlighted in bold font.

Dataset		Model	Evaluation			
Name	#Users	k	Silhouette	Occ	Cmp	R
EgoRoutine [1]	All	0.02	0.603 ± 0.024	0.411 ± 0.084	0.966 ± 0.013	0.689 ± 0.042
		0.03	0.743 ± 0.021	0.427 ± 0.086	0.954 ± 0.021	**0.691** ± 0.044
		0.035	0.732 ± 0.016	0.411 ± 0.079	0.960 ± 0.016	0.687 ± 0.045
		0.04	0.640 ± 0.049	0.411 ± 0.085	0.850 ± 0.121	0.630 ± 0.071
		0.05	0.602 ± 0.000	0.440 ± 0.085	0.555 ± 0.120	0.500 ± 0.014
NTCIR-12[38]	All	0.02	0.746 ± 0.010	0.283 ± 0.015	0.997 ± 0.006	**0.640** ± 0.010
		0.03	0.777 ± 0.023	0.300 ± 0.044	0.977 ± 0.006	0.637 ± 0.021
		0.035	0.740 ± 0.012	0.287 ± 0.015	0.950 ± 0.000	0.617 ± 0.006
		0.04	0.721 ± 0.028	0.280 ± 0.042	0.850 ± 0.000	0.565 ± 0.021
		0.05	–	–	–	–
NTCIR-13[39]	All	0.02	0.778 ± 0.028	0.205 ± 0.035	0.990 ± 0.000	0.595 ± 0.021
		0.03	0.821 ± 0.014	0.215 ± 0.049	0.980 ± 0.000	**0.600** ± 0.028
		0.035	0.799 ± 0.007	0.210 ± 0.057	0.965 ± 0.021	0.590 ± 0.014
		0.04	0.762 ± 0.014	0.210 ± 0.028	0.920 ± 0.071	0.565 ± 0.021
		0.05	–	–	–	–
NTCIR-14[40]	All	0.02	0.656 ± 0.007	0.335 ± 0.078	0.970 ± 0.028	0.655 ± 0.021
		0.03	0.687 ± 0.007	0.355 ± 0.092	0.975 ± 0.021	0.665 ± 0.035
		0.035	0.650 ± 0.014	0.375 ± 0.106	0.970 ± 0.028	**0.675** ± 0.035
		0.04	0.647 ± 0.021	0.385 ± 0.134	0.950 ± 0.014	0.665 ± 0.078
		0.05	0.643 ± nan	0.300 ± nan	0.790 ± nan	0.540 ± nan
All datasets	All	0.02	0.692 ± 0.019	0.344 ± 0.101	0.976 ± 0.018	0.660 ± 0.046
		0.03	0.755 ± 0.020	0.359 ± 0.105	0.966 ± 0.019	**0.663** ± 0.048
		0.035	0.728 ± 0.017	0.351 ± 0.100	0.960 ± 0.016	0.656 ± 0.052
		0.04	0.704 ± 0.043	0.356 ± 0.108	0.876 ± 0.097	0.615 ± 0.067
		0.05	0.621 ± 0.040	0.393 ± 0.101	0.633 ± 0.160	0.513 ± 0.025

4.3. Experimental setup

Image characterization: We characterize the depicted environment and context in images to represent the egocentric photo-sequences and daily experiences as the combination of detected concepts describing location, activity and detected objects. To do so, we rely on pre-trained networks as follows:

- *Places*: For recognizing the scene in the single images from the photo-stream, we applied the *VGG16-Places365* [32] CNN trained with the "Places-365" dataset consisting of 365 classes.
- *Activities*: For the characterization of the activities, we use the network proposed in [9]. This model allows us to classify a given image as belonging to one of 21 activities of Daily Living.
- *Object detection*: We use the Yolov4 [33] CNN model for the recognition of objects. Detected objects with a given class probability > 0.5 are considered.

We would like to highlight the fact that the networks VGG16-Places365 and Yolov4 used for feature extraction were not previously trained on egocentric images, but on conventional ones, which can lead to some misdetections in the images. As we will see, our model tolerates suboptimally detected labels being able to discover efficiently routine patterns.

The code of this work is available at.[2] We also developed a tool for the visualization of the discovered behavioural patterns by the different algorithms, which is available at.[3]

In this platform, the user can define different values for the parameters evaluated and observe the obtained clusters and behavioural patterns. We hope this tool encourages research in the area of behaviour analysis.

5. Results and discussion

In this section, we describe quantitatively and qualitatively the obtained results.

5.1. Quantitative results on behavioural pattern discovery

The evaluation metrics Silhouette, Occ, Cmp, and R are used to assess the performance of the model on the four datasets: EgoRoutine [1], NTCIR-12 [38], NTCIR-13 [39], and NTCIR-14 [40]. Table 3 shows the study of our algorithm on different datasets with different entropy thresholds k.

[2] Link to repository https://github.com/martinmenchon/Behavioural_patterns_discovery_for_lifestyle_analysis
[3] Link to results and visualizations https://martinmenchon-routinepatterns.streamlit.app

Table 4
Results for the best setup of our model when using as entropy threshold the global best k = 0.03 (as shown in Table 3), and compared to the performance of DBSCAN. The highest silhouette and representativity values for each dataset are highlighted in bold font.

Dataset	Model	Evaluation			
Name		Silhouette	Occ	Cmp	R
EgoRoutine [1]	Our model(k = 0.03)	**0.74** ± 0.02	0.43 ± 0.09	0.95 ± 0.02	**0.69** ± 0.04
	DBSCAN	0.62 ± 0.38	0.67 ± 0.31	0.21 ± 0.22	0.44 ± 0.23
NTCIR-12[38]	Our model(k = 0.03)	**0.78** ± 0.02	0.30 ± 0.04	0.98 ± 0.01	**0.64** ± 0.02
	DBSCAN	0.71 ± 0.10	0.54 ± 0.26	0.05 ± 0.08	0.30 ± 0.17
NTCIR-13[39]	Our model(k = 0.03)	**0.82** ± 0.01	0.22 ± 0.05	0.98 ± 0.00	**0.60** ± 0.03
	DBSCAN	0.76 ± 0.11	0.53 ± 0.05	0.00 ± 0.00	0.27 ± 0.03
NTCIR-14[40]	Our model(k = 0.03)	**0.69** ± 0.01	0.35 ± 0.09	0.98 ± 0.02	**0.66** ± 0.03
	DBSCAN	0.59 ± 0.18	0.72 ± 0.08	0.05 ± 0.04	0.38 ± 0.06

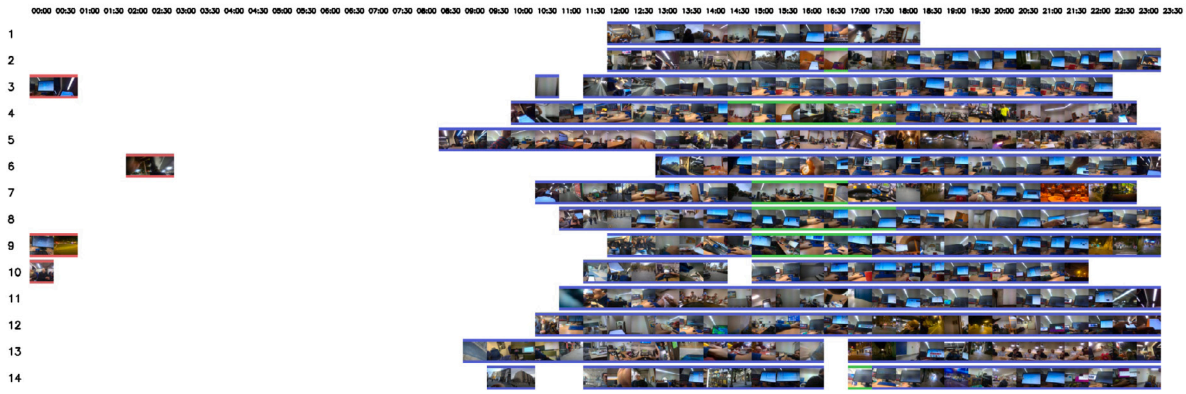

Fig. 7. Visual timeline with collected images of User 1 in the EgoRoutine dataset. Each colour indicates a cluster detected by DBSCAN. We can see that most of the time-slots belong to only one cluster (labelled in blue). The other two clusters contain less time-slots (labelled in red and green).

On the EgoRoutine [1] dataset, users appear to get the best results using the threshold value of $k = 0.03$ for all assessment measures. The Silhouette, Occ, and Cmp evaluation measures perform best for the NTCIR-12 [38] dataset when the threshold value is $k = 0.02$, whereas the R evaluation metric performs best when the threshold value is $k = 0.04$.

For all evaluation metrics on the NTCIR-13 [39] dataset, the threshold value of $k = 0.03$ produces the best results, with the exception of the R evaluation metric, which produces the best results with a threshold value of $k = 0.02$. For NTCIR-12 [38] and NTCIR-13 [39] dataset, the proposed method found no patterns for a $k = 0.05$.

The threshold value of $k = 0.03$ produces the best results for the NTCIR-14 dataset [40].

Overall, the proposed method has a robust performance on all datasets and evaluation metrics, and the choice of the threshold parameter k is critical for obtaining a good performance. We also observe that the performance of our method decreases as the entropy threshold value increases. This is because a high threshold value leads to a smaller number of patterns, which may not capture a significant part of the user's routine. On the other hand, a low threshold value may result in a high number of patterns, which may introduce noise in the pattern recognition process. Considering the highest value of average(R) in the four evaluated datasets, our algorithm with ($k = 0.03$) obtained better results in two of them (EgoRoutine [1] and NTCIR-13 [39]). The same situation is repeated if we consider the average and standard deviation of the 14 users of all datasets together with 0.663 and 0.048, respectively.

Comparison against related works:

We evaluate the task of pattern discovery addressed as the grouping of similar time-frames using the DBSCAN clustering technique [48]. The time characteristic is balanced with respect to the remaining 365 characteristics following a weighting procedure applied to the input feature vector, as described in [49]. The resulting clusters are considered patterns.

The results of our method vs. DBSCAN for all datasets can be seen in Table 4. The results obtained by our proposed algorithm result in a higher representativity value for all datasets. Also, the silhouette of the clusters formed by our proposed solution is better than the clusters formed by DBSCAN in all cases.

It can be observed that the DBSCAN finds fewer clusters than the rest of the methods. The problem is that these clusters tend to contain almost all the time-slots and all the user's days, resulting in clusters without a routine sense, as it is shown in Fig. 7. This is reflected in higher Occ values than our proposed method but significantly lower Cmp values.

Also, we compare against our baseline work on behavioural pattern recognition presented in *Behavioural pattern discovery from collections of egocentric photo-streams* [13], which was tested only on the EgoRoutine dataset. In addition, we implemented the

Table 5
Detailed results per user for the best set up of our model when using as entropy threshold the global best k (k) value, and compared to the performance of the models proposed in *Behavioural pattern discovery from collections of egocentric photo-streams* [13], in *Organizing egocentric videos of daily living activities* [30] and DBSCAN. The highest silhouette and representativity values for each dataset are highlighted in bold font.

Dataset	Model	Evaluation						
Name	Method	#Clusters	AVG #ImagesClust	#DaysCluster	Silhouette	Occ	Cmp	R
EgoRoutine [1]	Our method (k = 0.03)	13.14 ± 4.18	723.17 ± 127.98	6.05 ± 0.74	**0.74** ± 0.02	0.43 ± 0.09	0.95 ± 0.02	**0.69** ± 0.04
	Approach in [13]	15.57 ± 5.38	740.03 ± 126.34	4.84 ± 0.67	0.65 ± 0.03	0.34 ± 0.08	0.96 ± 0.02	0.65 ± 0.04
	DBSCAN	2.43 ± 1.51	700.06 ± 1049.13	9.63 ± 5.22	0.62 ± 0.38	0.67 ± 0.31	0.21 ± 0.22	0.44 ± 0.23
	Approach in [30]	5.86 ± 4.74	690.99 ± 567.02	2.31 ± 1.12	0.00 ± 0.01	0.17 ± 0.10	0.15 ± 0.17	0.16 ± 0.12
NTCIR-12[38]	Our method (k = 0.03)	21.33 ± 6.35	736.20 ± 183.20	7.84 ± 1.05	**0.78** ± 0.02	0.30 ± 0.04	0.98 ± 0.01	**0.64** ± 0.02
	Approach in [13]	18.33 ± 11.37	638.54 ± 95.93	5.75 ± 0.28	0.69 ± 0.07	0.22 ± 0.02	0.97 ± 0.03	0.59 ± 0.03
	DBSCAN	7.33 ± 0.58	256.78 ± 114.86	14.06 ± 6.30	0.71 ± 0.10	0.54 ± 0.26	0.05 ± 0.08	0.30 ± 0.17
	Approach in [30]	15.67 ± 11.59	826.04 ± 48.28	3.22 ± 0.41	−0.01 ± 0.03	0.13 ± 0.02	0.16 ± 0.12	0.14 ± 0.06
NTCIR-13[39]	Our method (k = 0.03)	26.50 ± 2.12	639.53 ± 281.47	9.26 ± 1.78	**0.82** ± 0.01	0.22 ± 0.05	0.98 ± 0.00	**0.60** ± 0.03
	Approach in [13]	23.00 ± 7.07	584.63 ± 288.94	7.64 ± 2.32	0.76 ± 0.06	0.18 ± 0.02	0.96 ± 0.01	0.56 ± 0.01
	DBSCAN	14.00 ± 15.56	269.86 ± 250.58	24.92 ± 12.61	0.76 ± 0.11	0.53 ± 0.05	0.00 ± 0.00	0.27 ± 0.03
	Approach in [30]	17.00 ± 9.90	916.17 ± 320.84	3.74 ± 0.06	−0.02 ± 0.01	0.09 ± 0.04	0.06 ± 0.09	0.08 ± 0.03
NTCIR-14[40]	Our method (k = 0.03)	28.50 ± 16.26	833.03 ± 272.65	7.21 ± 1.80	**0.69** ± 0.01	0.35 ± 0.09	0.98 ± 0.03	**0.66** ± 0.04
	Approach in [13]	29.00 ± 15.56	892.60 ± 337.74	6.37 ± 1.85	0.71 ± 0.01	0.31 ± 0.07	0.98 ± 0.02	0.64 ± 0.02
	DBSCAN	8.50 ± 6.36	412.05 ± 313.67	14.88 ± 5.83	0.59 ± 0.18	0.72 ± 0.08	0.05 ± 0.04	0.38 ± 0.06
	Approach in [30]	9.00 ± 11.31	2157.44 ± 941.24	2.44 ± 0.62	−0.02 ± 0.02	0.12 ± 0.03	0.52 ± 0.65	0.32 ± 0.34

Fig. 8. Visual timeline with collected images of User 5 in the EgoRoutine dataset. Each colour indicates a cluster detected by approach in *Organizing egocentric videos of daily living activities*.

approach introduced in *Organizing egocentric videos of daily living activities* [30] and evaluated their discovered clusters, which we treat as patterns. We consider both works as the state of the art.

In Table 5, we present the results of each model evaluated for each dataset. If we compare the models based on this metric, we can observe that our method is consistently obtaining better results than DBSCAN [48], the work in [13], and the adapted work from [30].

Based on the representativity value, [13] obtains similar performance for mainly all datasets, most of the times 0, 04 lower than our proposed method. Also, the silhouette of the clusters are slightly worse on all datasets therefore our proposed solution could be considered an improvement of this method.

The [30] method generates elongated clusters with little repetition, not very compact. Probably because when finding the clusters, it uses features of neural networks that are not trained for egocentric images, resulting, in most cases, a low Cmp value. In Fig. 8, we can see clusters labelled in green and turquoise tend to cover many hours with little repetition.

From a dataset user perspective, our method obtains a better R value in 13 of the 14 total users. [13] obtained the same R score as our method for only one user, and DBSCAN scored slightly better in only one user. The approach in [30] was consistently outperformed by other methods. These results can be found in Table 6. These results match what we expected since we consider the method proposed in this work an improvement of the one presented in [13]. We also believe that the case where the DBSCAN was superior is due to the inclusion in our proposed method of the timing of each time-slot in the pattern search algorithm. In the case where the same activity is displaced by many hours, but it is always the same activity, it would be penalized by the algorithm. In Fig. 9 is shown the DBSCAN result for user 3, their most recurrent activity is working in the office and is divided over many timeslots during the day. DBSCAN does not distinguish whether this activity is performed in the morning, afternoon or evening, considering everything as the same pattern.

Table 6
Best performing cases per user when relying on the proposed Representative metric that combines the occurrence and compactness of the found patterns.

Model	Avg. representativity	#Best performing cases
Our method (k = 0.03)	**0.66**	**13**
Behavioural pattern discovery [13]	0.62	1
DBSCAN	0.37	1
Organizing egocentric videos of daily living activities [30]	0.16	0

Fig. 9. DBSCAN results for User 3. The algorithm identified two patterns; the major one (labelled in yellow) shows work activities in the office throughout the day. This pattern is observed on most of the user's days and the main object that stands out in the images is a laptop computer.

5.2. Qualitative results on behavioural pattern discovery

The run of experiments of our ablation study derived in multiple visualizations. In Fig. 10, we present several samples of the discovered patterns for User 2. Per pattern, we present sample images that appear on different days, including their time frame. Moreover, we add the set of labels on which the model relied for the identification of the pattern. As an example, we focus on the first pattern. It describes a working activity where a laptop and tv/monitor are the main objects. Likewise, the second and third patterns describe walking outdoor and having breakfast/working home office activities, respectively.

With these examples, we can observe that the found patterns make sense and follow the daily activities that the user performs. These results suggest that our approach is highly effective in discovering meaningful patterns in daily-life images, and that these patterns can be used to provide useful insights into the daily activities of users.

Word Clouds: We can also build word clouds considering all the most relevant places, activities, and objects present in the clusters as shown in Fig. 11. The font size of the different words represents the times this word was found in the set of descriptors that composed the cluster. Colours do not give meaning, are simply associated to the different words through the figures. This visualization help deriving meaning from the grouped words. For instance, we could think that clusters 3, 4, 5 and 7 describe "a person working at the office" given that words such as "working", "office", "mouse", "TV/monitor" and "laptop" compose them. In contrast, the words "walking outdoor", "transportation", "pathways", "car", "truck" of Cluster 2 seem to indicate the user is walking on the street. Besides, we can join all relevant words of all clusters found to build a single cloud, this can give us a quick idea of the life of a person. In Fig. 11, words like "office", "person", "laptop", and "working" seem to indicate most users' life consisting of working at the office.

Fig. 10. Several discovered patterns in the collected days of User 2. We can observe images describing the different time-slots that compose the pattern, together with the locations, activities and objects that were detected in the group of images.

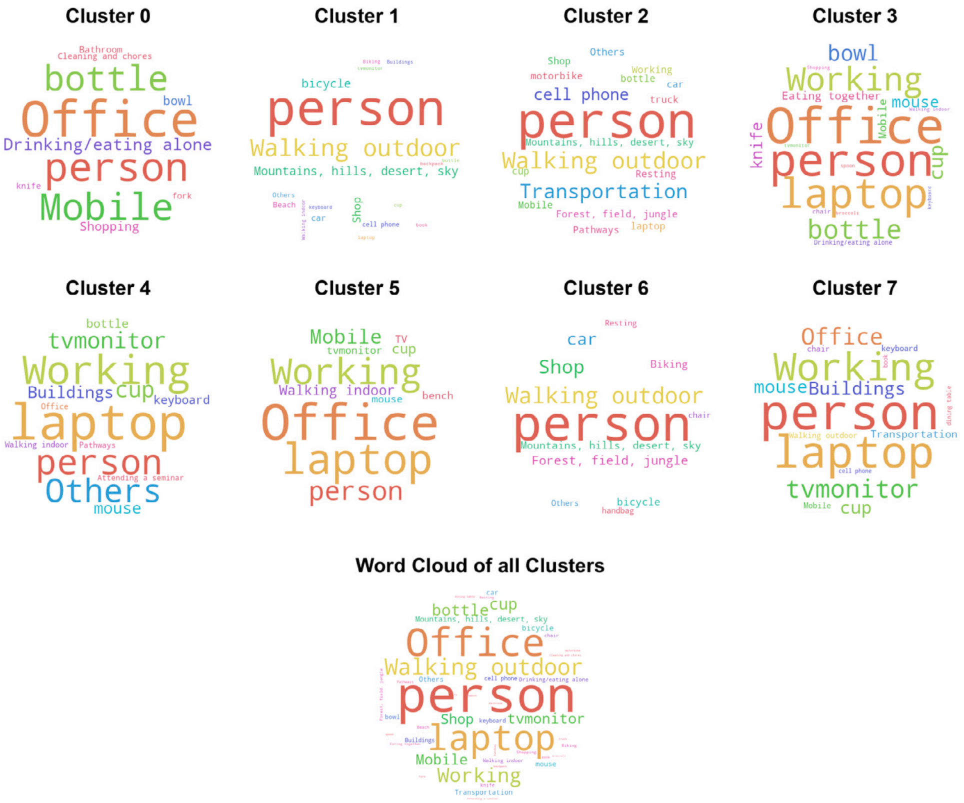

Fig. 11. Word cloud of each cluster of User 2. The size of the word indicates its relevance and occurrence within the cluster. The cluster at the bottom is the world of clouds of all clusters. Colours are used for visualization purposes but do not have any associated meaning.

In this paper, we present a set of visualizations that allows us to describe the insights and understanding of the discovered routine patterns. In Fig. 12, we present the identified patterns for User 2 when setting time-slots of 30 min. Every image appearing in the survey corresponds to the one in the median time of the time-slot.

As we can observe, behavioural patterns are identified throughout the 10 collected days by User 2. Just to mention some of the patterns, we can observe a person preparing breakfast within the time slots 8–9 h. This activity is highlighted in blue. From 11:00 h to 12:30 h and 14:00 h to 15:30 h a laptop is the main visual descriptors of the images, meaning the user is working. These patterns are highlighted in green and violet respectively. These two patterns do not describe the same pattern, because of the time difference, one in the morning and one in the afternoon. We find that this is important to note because a further analysis of working patterns can be done by differentiating morning and afternoon behaviour. Global patterns are clearly identified, having breakfast, working, leaving the house, working, and walking on the street. However, some patterns describing working with the laptop are not correctly grouped together by the proposed framework, creating some visual noise. This is something that we plan to investigate further, but that can be due to the wide variety within the collected frames in terms of viewpoints, lighting and others.

Qualitative evaluation: We propose to validate the routine discovery by involving several humans evaluators. Given the wide diversity that comes from human judgment, we prepared a questionnaire that includes a set of items asking the interviewed users to evaluate the achieved clustering quality and to compare our results to the other methods. This survey shows, on one side, the set of collected days and, on the other side, the found patterns to 25 interviewed.

Individuals who visually evaluated the obtained patterns were asked the same questions proposed in *Behavioural pattern discovery from collections of egocentric photo-streams* [13]:

1. What patterns do you see in this mosaic of images showing the days of user #ID? *Note that "a pattern is defined as a recurring habit of the camera wearer".*
2. Given the following collected egocentric photo-streams, reflecting the life of the camera wearer, can you see patterns of behaviour?
3. Given the recorded days for user #ID, do you think the found patterns adequately represent the user's behaviour?.
4. Do you consider that the found patterns are predominantly: (a) no patterns, (b) sub-patterns, (c) merged patterns, or (d) correct patterns? *We introduced the following terms: "Sub-patterns: they are sub-parts of a bigger pattern. No patterns: they do not represent any pattern. Merged patterns: they showed merged patterns. Correct patterns: they correctly described a pattern".*
5. Given these two discovered sets of patterns, which pattern set do you think better represents the user's routine behaviour?

Fig. 12. Discovered patterns for User 2 between 8:00 h and 19:30 h. Every row corresponds to one day. We can see patterns of breakfast from 8:00 h to 9:00 h (blue). Working from 11:00 h to 12:30 h (green). Then walking on the beach from 12:30 h to 15:00 h (light blue) and later continued working again (violet). Finally, walking on the street from 17:00 h to 19:30 h (turquoise).

Table 7
Qualitative results from questionnaires. The highest results per question for each dataset are highlighted in bold font.

Dataset	Question 2		Question 3			Question 4				Question 5	
	Yes	No	Yes	No	May be	No-Patterns	Sub-Pattern	Merged patterns	Correct patterns	Our algorithm	Method in [30]
EgoRoutine [1]	**92%**	8%	**76%**	4%	20%	0%	32%	36%	**48%**	**76%**	24%
NTCIR-12[38]	**80%**	20%	**76%**	0%	24%	0%	24%	28%	**56%**	**60%**	40%
NTCIR-13[39]	**88%**	12%	**72%**	8%	20%	4%	16%	20%	**68%**	**76%**	24%
NTCIR-14[40]	**68%**	32%	**48%**	20%	32%	12%	28%	**36%**	28%	**60%**	40%

We rely on the defined questionnaires to collect impressions by different people regarding the quality of the discovered patterns. The results that we obtained are shown in Table 7, where the percentage of answers per question is indicated. From the collected answers, we can discuss several points:

• Regarding question 2, *"Given the following collected egocentric photo-streams, reflecting the life of the camera wearer, can you see patterns of behaviour?"*, almost all the people surveyed recognize that there are patterns of behaviour in the image sequences. From this, we can deduce that the way we understand a pattern of behaviour is a shared feeling among the people who participated in the questionnaire. Moreover, it indicates that the output of the framework is qualitatively accurate.

• When it comes to question 3, *"Given the recorded days for user #ID, do you think the found patterns adequately represent the user's behaviour?"*, the first 3 datasets (EgoRoutine, NTCIR-12, and NTCIR-13) generally all find that the patterns are adequate to represent user behaviour. But in the NTCIR-14 dataset, only half (48%) recognize that the patterns reflect user behaviour, while the other half of those surveyed believe that there are no patterns (20%) or believe that it may be (32%).

• As for question 4, *"Do you consider that the found patterns are predominantly: (a) no patterns, (b) sub-patterns, (c) merged patterns, or (d) correct patterns?"*, the majority of the users consider that the shown patterns are correct. Except in NTCIR-14 where the majority (36%) indicated that there were merged patterns, and only 28% indicated that the patterns were detected correctly. It should be noted that, after visually checking the NTCIR-14 dataset, we felt that it was more challenging to identify patterns in the image sequences. This was because of the heterogeneity of user activities at different times and the characteristics of the images taken (fish-eye) made it difficult to distinguish objects and activities involved. This appreciation is reflected in the 12% of people who declare not to recognize any patterns at all.

• Finally, question 5, *"Given these two discovered sets of patterns, which pattern set do you think better represents the user's behaviour?"*. From the responses, we perceive the respondents' appreciation of the results of our framework. We compared the best performing setting (include specifications and parameters) with the work proposed in [30], and consistently got approved. For some of the dataset we can see that the difference was not that big (NTCIR-12). However, considering the average of the results of our algorithm for all datasets, we are 68% more liked that the state-of-the-art in this topic.

Table 8
N-grams found for User 03 for different order n. The numbers within brackets indicate the n sequence of clusters that compose the n-gram.

1st order	2nd order	3rd order	4th order
(a) [4]	(i) [4, 3]	(r) [4, 3, 2]	(y) [3, 2, 5, 7]
(b) [3]	(j) [3, 2]	(s) [3, 2, 5]	(z) [4, 3, 2, 1]
(c) [2]	(k) [2, 5]	(t) [2, 5, 7]	
(d) [5]	(l) [5, 7]	(u) [3, 2, 1]	
(e) [7]	(m) [2, 1]	(v) [7, 6, 1]	
(f) [0]	(n) [7, 6]	(w) [5, 7, 6]	
(g) [1]	(o) [6, 1]	(x) [2, 5, 6]	
(h) [6]	(p) [6, 0]		
	(q) [5, 6]		

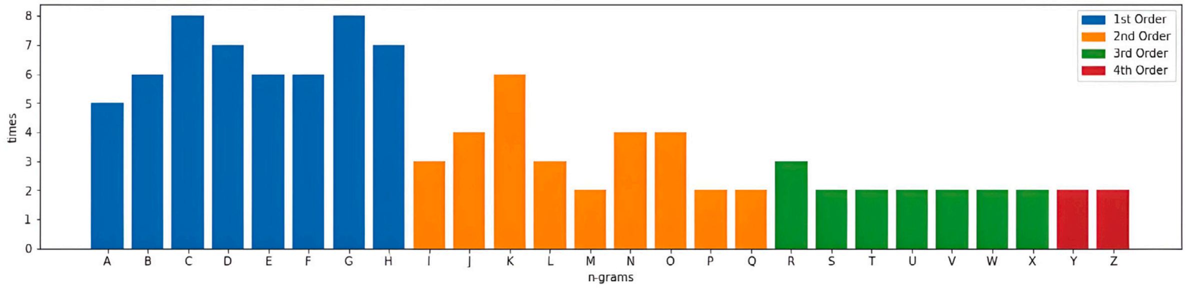

Fig. 13. N-grams plot for User 03. Each character indicates the found n-gram, every vertical bar indicates the frequency of the times these patterns and the colour indicates the order of the found n-grams. We can observe that as the order increases, the number of patterns and their repetitions decreases.

From the above described results, we can indicate the suitability of our proposed model with respect to the baseline approaches for the identification of behavioural patterns. Moreover, given these results we can highlight the generalization capabilities of the framework, which was tested on four different datasets.

5.3. N-grams for complex daily behaviour discovery

We find that it is interesting to analyse n-grams that occur as many times as possible. However, the higher order the n-gram is, the more complex it becomes and the fewer times it is expected to be repeated. We have observed that n-grams of order 2 and 3 have a good balance of repetitions and complexity. In contrast, and as expected, n-grams of order 1 have many repetitions and little complexity and n-grams of order 4 have greater complexity and fewer repetitions.

In order to show the potential of n-grams for the description of behaviour, we present the obtained n-grams when setting different values as number of repetitions. Table 8 shows the list of n-grams and their order throughout the collected image sequences for the User 03. We can observe that if the order of the n-grams increases, fewer are discovered. In Fig. 13, we can also note that n-grams composed of a higher number of elements, i.e., sequence of patterns, lead to fewer found n-grams. This highlights what we could deduce from self-experience, i.e., it is difficult to follow a long sequence of activities in daily life.

An example of identified clusters and n-grams is presented in Fig. 14. In Fig. 14(a), clusters are indicated with different colours. As we can observe, clusters are formed by a set of sequential images describing an environment or scene through different days at a similar time frame. Combination of clusters that maintain a consecutive order can be identified and by counting the number of times the pattern appears we can identify daily behaviour. Fig. 14(b) presents the list of n-grams of different orders that describe the days of the user. For instance, we can count up to 3 times the sequence of cluster 6 followed by cluster 2 for this user. Moreover, when we observe the listed n-grams and their occurrence, the higher the order of the n-gram the less they appear in the day collection. N-grams with a high number of elements will describe the life of a person with a well-defined routine, i.e. a set of activities that occur in the same sequence through the days of the week. However, we observed that it is not likely, people tend to vary their daily activities leading to days rich in n-grams of low order where with low order, we refer to 1, 2, or 3 sequential clusters.

5.4. Limitations of the proposed method

The proposed method exhibits the following limitations. Firstly, by defining the algorithm's granularity through timeslots (in our case, a value of 30 min), there is a risk of overlooking small patterns that occur within those 30-minute intervals. Similarly, if the value is set to a small duration, such as 5 min, the total number of discovered patterns increases significantly, potentially breaking down longer patterns into multiple smaller ones, thereby losing the conceptual sense of routine and transforming it into a mere event-based division. Secondly, the execution time of the algorithm significantly increases as the number of days and images in the user dataset grows. This is due to the algorithm having to test all possible nodes in the greedy algorithm until finding the one that

(a) Clusters found

```
                          Order 1:                        Order 2:

Day 1: [0, 6, 2]          a) [0]: 4 times    h) [6, 2]: 3 times    o) [4, 3]: 1 time
Day 2: [0, 1]             b) [1]: 3 times    i) [3, 1]: 2 times    p) [1, 7]: 1 time
Day 3: [0, 3, 1, 6, 2]    c) [2]: 3 times    j) [0, 1]: 1 time     q) [7, 6]: 1 time
Day 4: [0, 4, 3, 1, 7, 6, 2]  d) [6]: 3 times  k) [0, 3]: 1 time   r) [4, 3]: 1 time
                          e) [3]: 2 times    l) [0, 4]: 1 time
                          f) [4]: 1 time     m) [0, 6]: 1 time
                          g) [7]: 1 time     n) [1, 6]: 1 time
```

(b) N-grams found

Fig. 14. Identified clusters and n-grams over a simplified collection of photo-streams of User 02. 14(a) each row represents a day of a user and each cluster is indicated by a different number and colour. 14(b) these days are converted to lists of clusters ids, preserving the temporal order of the clusters. This information is used to calculate the n-grams, including the number of times they appears. N-grams of order 1 and 2 are presented in this image.

introduces the least entropy into the cluster. As can be expected, the method needs enough data in order to discover behavioural patterns (repeated events during different days).

6. Conclusions

In this work, we address the task of unsupervised behavioural pattern discovery from egocentric photo streams. We rely on extracted semantics of the images in the form of activity labels and detected objects to depict what happens in the life of the camera wearer. Our proposed model studies the variation of the entropy of defined clusters across days' temporal adjacent images to assess the quality of the grouping. In order to assess the goodness of the discovered cluster of related behaviour, we present and rely on custom-made metrics for this specific task. Patterns are later formed by seeking n-grams with different degrees of complexity, where the parameter n indicated the number of clusters. The quantitative and qualitative assessment of the obtained results indicates the robustness of our approach against state-of-the-art experiments.

Future research lines will aim to include other descriptors in the analysis so more complex situations can be identified. We believe this research will support healthcare by monitoring the lifestyle of people, extracting their routines, and providing objective and precise computational tools for behaviour change and improvement. In addition, the analysis of n-grams can be improved, for example, by performing a classification or clustering of the histograms of n-grams of each user to classify the type of routine in comparison with other users.

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.

Data availability

Data used in this work is public and code repository will be published after the acceptance of this paper.

Acknowledgements

This work was partially funded by the Horizon EU project MUSAE (No. 01070421), 2021-SGR-01094 (AGAUR), Icrea Academia'2022 (Generalitat de Catalunya), Robo STEAM (2022-1-BG01-KA220-VET-000089434, Erasmus+ EU), DeepSense (ACE053/22/000029, ACCIÓ), DeepFoodVol (AEI-MICINN, PDC2022-133642-I00), PID2022-141566NB-I00 (AEI-MICINN), and CERCA Programme/Generalitat de Catalunya.

References

[1] E. Talavera, C. Wuerich, N. Petkov, P. Radeva, Topic modelling for routine discovery from egocentric photo-streams, Pattern Recognit. (2020) 107330.
[2] A. Furnari, G.M. Farinella, S. Battiato, Recognizing personal locations from egocentric videos, IEEE Trans. Hum.-Mach. Syst. 47 (1) (2016) 6–18.
[3] A. Cartas, E. Talavera, P. Radeva, M. Dimiccoli, Understanding event boundaries for egocentric activity recognition from photo-streams, in: Conference on Pattern Recognition, Springer, 2021, pp. 334–347.
[4] K. Siła-Nowicka, J. Vandrol, T. Oshan, J.A. Long, et al., Analysis of human mobility patterns from GPS trajectories and contextual information, J. Geogr. Inf. Sci. 30 (5) (2016) 881–906.
[5] O.D. Lara, M.A. Labrador, A survey on human activity recognition using wearable sensors, IEEE Commun. Surv. Tutor. 15 (3) (2012) 1192–1209.
[6] Y. Huang, et al., Assessing social anxiety using GPS trajectories and point-of-interest data, in: Conference on Pervasive and Ubiquitous Computing, 2016, pp. 898–903.
[7] M. Atzmueller, L. Thiele, G. Stumme, S. Kauffeld, Analyzing group interaction on networks of face-to-face proximity using wearable sensors, in: Conference on Future IoT Technologies, 2018, pp. 1–10.
[8] A. Pantelopoulos, N.G. Bourbakis, A survey on wearable sensor-based systems for health monitoring and prognosis, IEEE Trans. Syst. Man Cybern. C 40 (1) (2009) 1–12.
[9] A. Cartas, J. Marín, P. Radeva, M. Dimiccoli, Recognizing activities of daily living from egocentric images, in: Iberian Conference on Pattern Recognition and Image Analysis, Springer, 2017, pp. 87–95.
[10] D. Damen, et al., Scaling egocentric vision: The epic-kitchens dataset, in: European Conference on Computer Vision, 2018, pp. 720–736.
[11] A. Matei, A. Glavan, E. Talavera, Deep learning for scene recognition from visual data: a survey, in: Hybrid Artificial Intelligent Systems: 15th International Conference, HAIS 2020, Gijón, Spain, November 11–13, 2020, Proceedings, Springer, 2020, pp. 763–773.
[12] E.S. Aimar, P. Radeva, M. Dimiccoli, Social relation recognition in egocentric photostreams, in: Conference on Image Processing, 2019, pp. 3227–3231.
[13] M. Menchón, E. Talavera, J. Massa, P. Radeva, Behavioural pattern discovery from collections of egocentric photo-streams, in: European Conference on Computer Vision, Springer, 2020, pp. 469–484.
[14] G. Benchetrit, Breathing pattern in humans: diversity and individuality, Respir. Physiol. 122 (2–3) (2000) 123–129.
[15] P.V.K. Borges, N. Conci, A. Cavallaro, Video-based human behavior understanding: A survey, IEEE Trans. Circuits Syst. Video Technol. 23 (11) (2013) 1993–2008.
[16] E. Kim, S. Helal, D. Cook, Human activity recognition and pattern discovery, IEEE Pervasive Comput. 9 (1) (2009) 48–53.
[17] G. Zhao, J. Yuan, Discovering thematic patterns in videos via cohesive sub-graph mining, in: Conference on Data Mining, 2011, pp. 1260–1265.
[18] Q. Li, W. Lin, J. Li, Human activity recognition using dynamic representation and matching of skeleton feature sequences from RGB-D images, Signal Process., Image Commun. 68 (2018) 265–272.
[19] A. Piergiovanni, M.S. Ryoo, Fine-grained activity recognition in baseball videos, in: IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2018, pp. 1740–1748.
[20] X. Liu, W. Liu, M. Zhang, J. Chen, L. Gao, C. Yan, T. Mei, Social relation recognition from videos via multi-scale spatial-temporal reasoning, in: IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 3566–3574.
[21] R. Kaur, S. Kautish, Multimodal sentiment analysis: A survey and comparison, in: Research Anthology on Implementing Sentiment Analysis Across Multiple Disciplines, IGI Global, 2022, pp. 1846–1870.
[22] W. Sultani, C. Chen, M. Shah, Real-world anomaly detection in surveillance videos, in: IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 6479–6488.
[23] K. Boekhoudt, A. Matei, M. Aghaei, E. Talavera, HR-crime: Human-related anomaly detection in surveillance videos, in: Conference on Computer Analysis of Images and Patterns, Springer, 2021, pp. 164–174.
[24] H. Jo, K. Chug, R.J. Sethi, A review of physics-based methods for group and crowd analysis in computer vision, J. Postdr. Res. 1 (1) (2013) 4–7.
[25] A. Glavan, A. Matei, P. Radeva, E. Talavera, Does our social life influence our nutritional behaviour? Understanding nutritional habits from egocentric photo-streams, Expert Syst. Appl. 171 (2021) 114506.
[26] S. Khan, M. Naseer, M. Hayat, S.W. Zamir, F.S. Khan, M. Shah, Transformers in vision: A survey, ACM Comput. Surv. 54 (10s) (2022) 1–41.
[27] M. Dimiccoli, M. Bolanos, E. Talavera, M. Aghaei, S.G. Nikolov, P. Radeva, Sr-clustering: Semantic regularized clustering for egocentric photo streams segmentation, Comput. Vis. Image Underst. 155 (2017) 55–69.
[28] A. Garcia del Molino, J.-H. Lim, A.-H. Tan, Predicting visual context for unsupervised event segmentation in continuous photo-streams, in: Conference on Multimedia, 2018, pp. 10–17.
[29] A. Furnari, S. Battiato, G.M. Farinella, Personal-location-based temporal segmentation of egocentric videos for lifelogging applications, J. Vis. Commun. Image Represent. 52 (2018) 1–12.
[30] A. Ortis, G.M. Farinella, V. D'Amico, L. Addesso, G. Torrisi, S. Battiato, Organizing egocentric videos of daily living activities, Pattern Recognit. 72 (2017) 207–218.
[31] J. Peng, P. Shi, J. Qiu, X. Ju, F.P.-W. Lo, X. Gu, W. Jia, T. Baranowski, M. Steiner-Asiedu, A.K. Anderson, et al., Clustering egocentric images in passive dietary monitoring with self-supervised learning, in: 2022 IEEE-EMBS International Conference on Biomedical and Health Informatics, BHI, IEEE, 2022, pp. 01–04.
[32] B. Zhou, A. Lapedriza, A. Khosla, A. Oliva, A. Torralba, Places: A 10 million image database for scene recognition, IEEE Trans. Pattern Anal. Mach. Intell. (2017).
[33] A. Bochkovskiy, C.-Y. Wang, H.-Y.M. Liao, YOLOv4: Optimal speed and accuracy of object detection, 2020, arXiv preprint arXiv:2004.10934.
[34] R. Real, J.M. Vargas, The probabilistic basis of Jaccard's index of similarity, Syst. Biol. 45 (3) (1996) 380–385.
[35] I. Borg, P.J. Groenen, Modern Multidimensional Scaling: Theory and Applications, Springer Science & Business Media, 2005.
[36] C.E. Shannon, A mathematical theory of communication, Bell Syst. Tech. J. 27 (3) (1948) 379–423.

[37] E. Aldana-Bobadilla, A. Kuri-Morales, A clustering method based on the maximum entropy principle, Entropy 17 (1) (2015) 151–180.
[38] N. Li, C. Gurrin, M. Crane, H.J. Ruskin, NTCIR-12 lifelog data analytics, in: Workshop on Lifelogging Tools and Applications, 2016, pp. 27–36.
[39] C. Gurrin, H. Joho, F. Hopfgartner, L. Zhou, R. Gupta, R. Albatal, D.T.D. Nguyen, Overview of NTCIR-13 lifelog-2 task, in: NTCIR Conference on Evaluation of Information Access Technologies, 2017, pp. 5–8.
[40] C. Gurrin, H. Joho, F. Hopfgartner, L. Zhou, V.-T. Ninh, T.-K. Le, R. Albatal, D.-T. Dang-Nguyen, G. Healy, Overview of the NTCIR-14 lifelog-3 task, in: NTCIR Conference, NII, 2019, pp. 14–26.
[41] Narrative clip web page, 2022, http://getnarrative.com/. (Accessed 18 February 2022).
[42] OMG autographer web page, 2022, https://www.autographer.com/. (Accessed 18 February 2022).
[43] K. Grauman, A. Westbury, E. Byrne, Z. Chavis, A. Furnari, R. Girdhar, J. Hamburger, H. Jiang, M. Liu, X. Liu, et al., Ego4d: Around the world in 3000 h of egocentric video, in: IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 18995–19012.
[44] E. Kazakos, A. Nagrani, A. Zisserman, D. Damen, Epic-fusion: Audio-visual temporal binding for egocentric action recognition, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 5492–5501.
[45] C. Northcutt, S. Zha, S. Lovegrove, R. Newcombe, Egocom: A multi-person multi-modal egocentric communications dataset, IEEE Trans. Pattern Anal. Mach. Intell. (2020).
[46] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, et al., Scikit-learn: Machine learning in python, J. Mach. Learn. Res. 12 (2011) 2825–2830.
[47] P.J. Rousseeuw, Silhouettes: a graphical aid to the interpretation and validation of cluster analysis, J. Comput. Appl. Math. 20 (1987) 53–65.
[48] M. Ester, H.-P. Kriegel, J. Sander, X. Xu, et al., A density-based algorithm for discovering clusters in large spatial databases with noise, in: Kdd, Vol. 96, No. 34, 1996, pp. 226–231.
[49] M. Dousthagh, M. Nazari, A. Mosavi, S. Shamshirband, A.T. Chronopoulos, Feature weighting using a clustering approach, J. Model. Optim. 9 (2) (2019) 67–71.