Towards Storytelling from Visual Lifelogging: An Overview

Marc Bolaños*, Mariella Dimiccoli*, and Petia Radeva

Abstract—**Visual lifelogging consists of acquiring images that capture the daily experiences of the user by wearing a camera over a long period of time. The pictures taken offer considerable potential for knowledge mining concerning how people live their lives, hence, they open up new opportunities for many potential applications in fields including healthcare, security, leisure and the quantified self. However, automatically building a story from a huge collection of unstructured egocentric data presents major challenges. This paper provides a thorough review of advances made so far in egocentric data analysis, and in view of the current state of the art, indicates new lines of research to move us towards storytelling from visual lifelogging.**

Index Terms—**visual lifelogging, egocentric vision, storytelling**

I. Introduction

Lifelogging consists of a user continuously recording their everyday experiences, typically via wearable sensors including accelerometers and cameras, among others. When the visual signal is the only one recorded, typically by a wearable camera, it is referred to as *visual lifelogging*. This is a trend that is rapidly increasing thanks to advances in wearable technologies over recent years. Nowadays, wearable cameras are very small devices that can be worn all-day long and automatically record the everyday activities of the wearer in a passive fashion, from a first-person point of view. As an example, Fig. 1 shows pictures taken by a person walking down a street while wearing such a camera.

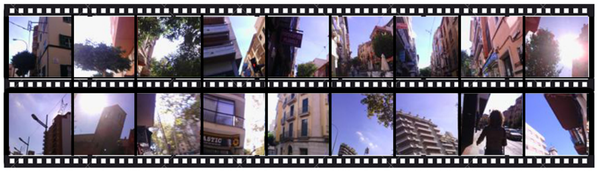

Fig. 1. Example of a sequence acquired by the Narrative Clip wearable camera while the user is walking down a street. The temporal leaps between neighbouring pictures produced by photographic cameras are common in dynamic environments and make the extraction of information from closely spaced images very difficult.

Fig. 2. Examples of wearable cameras on the market: (a) GoPro (2002). (b) SenseCam (2005). (c) Looxcie (2011). (d) Narrative Clip (2013).

Most wearable cameras on the market like GoPro, MeCam, Looxcie or Google Glass (see Fig. 2 (a) and (c)) are video cameras, which have relatively High Temporal Resolution (HTR) (e.g. from 25 up to 60 frames per second) and are more suitable to record specific moments, such as cooking or doing sports. A limited number of wearable cameras, such as Narrative Clip and SenseCam (see Fig. 2 (b) and (d)) are photographic cameras, which have Low Temporal Resolution (LTR) (2-3 frames per minute), and hence are more suitable for acquiring data over long periods of time. On the one hand, data recorded at specific moments with video cameras offer potential for in-depth analysis of daily or special activities, allowing to capture even *how* something happened. On the other hand, data acquired over long periods of time, commonly called *visual lifelogs*, offer considerable potential for inferring knowledge about e.g. behaviour patterns, and hence enable many applications that would not be possible with HTR cameras. As shown by Doherty et al. [32], visual lifelogs captured through a SenseCam, which as opposed to video cameras can capture the whole day, could be used to prevent non-communicable diseases associated with unhealthy trends and risky profiles (such as obesity or depression, among others). Additionally, they could also help prevent cognitive and functional decline in elderly people [29], [44], [57]. However, visual lifelogs present a significant challenge for automatic visual analysis. Indeed, due to the free motion of the camera and to its LTR, abrupt changes in lighting conditions and image content are very frequent (see Fig. 1). In such situations, computer vision techniques based on temporal coherence and motion estimation become unreliable. Recognition algorithms have to cope with the huge variety of objects that appear. In addition, due to the non-intentional nature of the pictures captured, they generally contain severely occluded objects, artefacts such as blurring or light saturation [89] and a large number of non-informative images that capture non-meaningful information such as walls, the sky, parts of objects, etc. Furthermore, the sheer number of data that a visual lifelog consists of and the rate at which they increase (up to 2,000 images per day or around 800,000 images every year) imposes a need for efficient methods to extract and locate relevant content concerning the wearer from the photo stream. Regarding HTR cameras, if they were employed for a lifelog analysis, the problem of the amount of data would be even more acute, and would additionally imply the need of huge computational resources.

In response to the challenges and opportunities introduced by analysis of visual lifelogs, and more generally, by wearable cameras, computer vision scientists have rapidly become more

M. Bolaños, marc.bolanos@ub.edu, M. Dimiccoli, mariella.dimiccoli@cvc.uab.es, P. Radeva, petia.ivanova@ub.edu - Universitat de Barcelona, Barcelona, Spain and Computer Vision Center, Bellaterra, Spain. *The first two authors contributed equally to this work.

interested in the subject over recent years. By searching for the keywords *egocentric vision*, *first person vision*, *ego vision* and *visual lifelogging*, using *Google Scholar*, *DBLP* and *visionbib.com*, we found 274 papers in total devoted to visual lifelogging. For each of them, we annotated the type of camera used in the study and generated the plot in Fig. 3, which represents all the papers related to egocentric vision up to November 2015. As can be seen, interest grew very fast in the last years and the number of papers published increased by over 50% in 2014 alone. Dotted lines show the comparatively small amount of work devoted to the analysis of image streams captured by photo cameras. This trend seemed to temporally change from 2007 to 2010, when the popularity of SenseCam resulted in a growth in the use of photo streams.

An additional indication of the interest in this emerging field is the fact that in the last years, four surveys of wearable cameras and egocentric vision have been published. One, written by Doherty et al. [32], focuses on explaining the *ethical and data management issues* that must be taken into account when developing some health-related application using wearable cameras. The second one, by Betancourt et al. [12], provides a general perspective on egocentric vision and devotes most of its analysis to the *egocentric camera hardware*, *egocentric datasets*, *augmented reality*, *algorithm types* and *feature types* used in the literature from 1997 to 2014. This analysis is focused on providing a historical perspective of egocentric devices and their algorithms in addition to several ways of categorizing the existent papers in this field. The third one, which is a book by Gurrin et al. [41], focuses on data management and distinguishes between *data storage*, *organisation* and *visualisation*; while also provides an overview of potential *applications*. The fourth study, by Harvey et al. [42], the authors present their work from the perspective of providing an aid to human memory. They analyse the human memory mechanisms from a psychological perspective and propose a pipeline for enhancing it based on *segmentation*, *context enhancement (recognising objects and people)* and *image retrieval*.

This paper focuses on addressing the question: *How far are we from being able to automatically tell our stories using egocentric photo streams?* The process of fully understanding the story behind the pictures is fundamental towards enabling a wide range of applications [27] and user cases [45], especially related to health. As we explained, since these applications require observations over long periods of time, data should be acquired by photographic cameras (e.g. SenseCam, Narrative, etc.) instead of video cameras (e.g. GoPro, GoogleGlass, Looxcie, etc.). To this end, a thorough review of the published advances in egocentric data analysis is presented and research insights are provided. In contrast to previous surveys, we review and give details of studies that focus on both photographic and video cameras, considering which aspects should be reformulated and modified for their applicability in the LTR domain, and thus for egocentric storytelling.

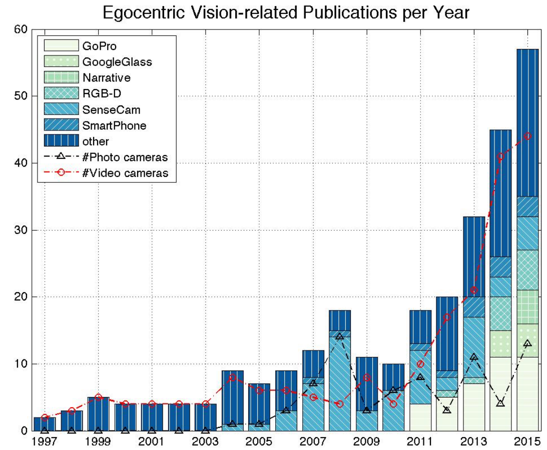

Fig. 3. Histogram of the number of research papers published per year related to egocentric vision. The different colours indicate how many papers used each kind of camera. The dashed blue and black lines make a less specific distinction, showing the number of studies that used photo (LTR) or video (HTR) cameras, respectively.

To summarize, our contributions are as follows:

- Review of methods for acquiring, organizing, summarizing and browsing large collections of unstructured data.
- Organization of the available literature around the central questions necessary to address the storytelling problem: *Was the user interacting with somebody? How?*, *Where* is he/she?, *When* did the event occur? and *What* is the person wearing the camera doing?.
- Highlights of the weaknesses and strengths of the reviewed techniques with respect to their applicability to the LTR domain (at the end of each subsection).
- Extensive analysis of the available datasets and source code related to the storytelling problems.
- Open problems and challenges in the field of egocentric vision with the final goal of storytelling.

The rest of the paper is organised as follows. In Section II, we review the most important papers devoted to the task of acquiring, organising, summarizing and browsing large and unstructured collections of egocentric data. The solutions to these problems provide a basis to further analyse the data content, as in Section III, where we review papers that claim to construct semantic building blocks for storytelling. Concluding remarks about applicability to the LTR domain are given at the end of each subsection. In Section IV, we summarize the available egocentric datasets with the corresponding annotations, as well as the egocentric vision software. Finally, in Section V, we draw our conclusions and give some possible future directions for the research necessary to fill the gap between raw egocentric data analysis and visual storytelling.

II. Visual lifelogging acquisition, segmentation and summarization

This section reviews the literature concerning acquiring, structuring and summarizing visual lifelogging data, which is summarized in Table I.

A. Data Acquisition

The positioning of a wearable camera is of crucial importance for lifelogging data acquisition from the point of view of its later application. Mayol-Cuevas et al. [66] evaluated, partially through simulations on a 3D facet model of the human body, four attributes of optical devices with respect to their position on the wearer's body: social acceptability,

TABLE I
SUMMARY OF ALL THE VISUAL LIFELOGGING PAPERS REVIEWED IN THIS SURVEY RELATED TO ACQUIRING, ORGANIZING, SUMMARIZING AND BROWSING LARGE COLLECTIONS OF UNSTRUCTURED DATA.

TOWARDS STORYTELLING FROM VISUAL LIFELOGGING							
II-B **Informative Images Detection**							
[96]	[61]						
II-C **Temporal Segmentation**							
[60]	[30]	[28]	[62]	[86]	[64]	[18]	[73]
[88]	[23]						
II-D **Egocentric Summarization**							
[83]	[49]	[40]	[64]	[15]	[61]		
II-E **Content-Based Search and Retrieval**							
[94]	[24]	[68]	[4]	[93]			

absolute field of view (FOV), resilience to body motion, and view of the handling space region. That study concluded that wearable cameras placed on the chest are the most socially acceptable and therefore offer the advantage of not interfering with social interactions. In addition, they are relatively resilient to the disturbances introduced by the wearer's own motion and are closely linked to the user's workspace, since they allow visualisation of the manipulative space in front of the wearer's chest. However, the FOV is quite narrow and does not allow the focus of the wearer's attention to be modelled. In contrast, cameras worn on the head have a wider FOV and do allow this attention to be modelled, but they are the most sensitive to the wearer's motion and suffer from low social acceptability. A compromise between the size of the FOV, accessibility to the handling regions, sensitivity to ego-motion and social acceptability is offered by wearable cameras placed on the shoulder. The authors also considered the possibility of wearing multiple devices on different parts of the body so that their FOVs would be complementary, with the joint FOV computed as the union of the individual FOVs.

Remarks: Since for long-term image acquisition social acceptability is crucial, placement on the chest is usually considered the best choice. In addition, it has the advantage of offering access to the handling space and the manipulation of objects can be focused.

B. Informative Image Detection

Once images have been acquired, before proceeding with any structuring, analysis and summarization, proper cleaning of the images is necessary. This need stems from the fact that egocentric images are non-intentional images, that is, nobody decides when and of what to take a picture. As a result, a significant number of images can be blurred, can be dark, or can capture non-informative data (the sky, the ground, walls, etc.). In Xiong and Grauman [96], informative images are defined as "intentional" images, obtained once those with undesired artefacts, such as light saturation, blurred images, or useless information (the sky, walls, etc.) have been removed. Lidon et al. [61] define as informative any image that includes objects and/or people, and which is of reasonable quality, assuming that it does not include any undesired artefacts (e.g. blurring, darkness or occlusions). With this definition, they trained a binary CNN to make this distinction.

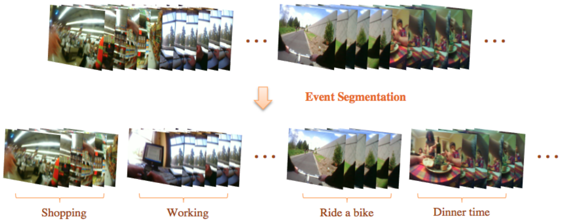

Fig. 4. Example of the desired event segmentation applied to lifelogging data. The goal is to group with respect to their main event, considering the activities, objects or people involved.

C. Temporal Segmentation

Lifelogging data typically consist of long unstructured videos or photo streams. Organising and structuring them into homogeneous temporal segments, corresponding to different events and/or environments (see Fig. 4), are very important to facilitate browsing and analysis of the images. State-of-the-art methods for egocentric data segmentation can be classified into two broad classes depending on whether the homogeneous segments represent what the *wearer sees* or *does*.

The former class uses features that can capture the characteristics of the environment around the wearer as image representation. Early work aiming at segmenting the sequences into visually homogeneous segments was based on low-level features. Li et al. [60] have proven that it is possible to distinguish different events simply by treating SenseCam images as time-series data and calculating the eigenvalue peaks in consecutive windows of images. Doherty et al. [30], [28] used different descriptors for image representation and the metadata available from the camera sensors. Lin and Hauptmann [62] proposed a simple approach based on using colour features in a time-constrained K-means clustering algorithm, capable of maintaining temporal coherence on the splitting of events. Spriggs et al. in [86] proposed a method for simultaneous temporal segmentation and recognition of activity related to cooking. They captured videos at the same time from a single wearable video camera and multiple other static cameras, sensors, microphones, etc., and used both sensor data and visual GIST descriptors to describe the frames. For the unsupervised scene segmentation, they applied a Gaussian mixture model. More recently, Talavera et al. [88] proposed the use of CNNs computed on the whole image using AlexNet as a fixed feature extractor for image representation. That work, designed for egocentric photo streams, uses a graph-cut algorithm to temporally segment the photo streams and includes an agglomerative clustering approach with concept drifting methodology, called ADWIN.

Methods focusing on what the camera wearer does mostly use motion information as image representation. Usually, optical flow is used to distinguish between *static*, *moving the head/camera* and *in-transit* frames [18], [64] (see Fig. 5). To focus on long-term ego-activities, Poleg et al. [73] proposed the use of so-called integral motion, which is closely related

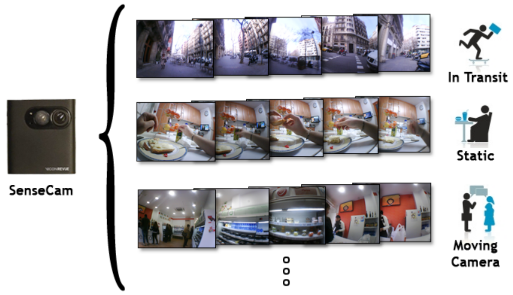

Fig. 5. Motion-based segmentation framework proposed in [18]. By including motion features to describe egocentric pictures they can separate the events considering the dynamism of the activities performed.

to the wearer's activity. By integrating the instantaneous displacements at fixed image patches, the variations due to head rotation are eliminated, since their mean is practically zero, leaving only the consistent displacement caused by forward motion. A different approach, based on CNNs, is adopted by Castro et al. [23]. They gathered a large egocentric dataset from a single user and fine-tuned a CNN pre-trained on ImageNet for activity classification. They proved that the network trained on the data of a single user can be re-trained to generalise to new users. The main problem with this approach is that a new set (several thousands of images) must be labelled from scratch whenever it is necessary to predict the events affecting a new user with the model.

- *Remarks:* The applicability of motion as a feature, though relevant when dealing with videos, has proven to be rather limited for photo streams. In the latter case, the use of richer representations, such as global CNN-based features, seems crucial to compensate this limitation. The use of time-dependent methods for egocentric segmentation is also a must considering the nature of the data. A promising approach to improve the results of the segmentation of egocentric sequences is the addition of semantic-level features (scenes, objects, social interaction, actions, etc.). This additional information would be an important step to bring machine segmentation closer to the way humans segment unconstrained streams of images.

D. *Egocentric Summarization*

Summarization is the process of generating a proper, compact and meaningful representation [90] of a given sequence through a subset of representative frames or segments. This step is crucial to help manage and browse large volumes of lifelogging video content efficiently. Basically, there are two kinds of summaries that can be produced: a static video story board, which is composed of a set of salient images extracted or synthesised from the original sequence, and dynamic video skimming: a shorter version of the original video made up of several shots, comprised of a series of frames. To fully exploit the potential of visual lifelogs in a variety of applications, an egocentric summarization method should be designed to aid in the visualisation, indexing and browsing of autobiographical events, with the least possible semantic loss.

Story board summarization has been traditionally formulated as grouping images into coherent collections by relying on low-level spatio-temporal features and then selecting the most representative image (or set of images) from each collection [83]. Based on this classical approach, Jinda-Apiraksa et al. [49] and Chowdhury et al. [25], developed similar techniques for keyframe selection in egocentric sequences based on quality measures [49], [25] and both quality and diversity measures [25]. More complex features for grouping were used by Bolaños et al. [15]. Their methodology, adapted for photo cameras, uses the AlexNet CNN as a feature extractor to characterise each frame. Then, using those features, they apply event segmentation using a hierarchical clustering algorithm and a posterior single keyframe selection by applying the Random Walk algorithm to each of the segments.

While these methods rely solely on low-level features, some recent work has introduced a semantic level in the keyframe selection process. Ghosh et al. [40] suggested that video summarization should be driven by the presence of important people and objects. Following this idea, they proposed a method that reveals salient people and objects based on their interaction time with the camera wearer and then selected keyframes according to keyobject event occurrences. Lu and Grauman [64], following on from their previous work, suggested that video summarization should preserve the narrative character of a visual lifelog and proposed a shot selection consisting of three terms: 1) a term that models *story coherence* by favouring shots capable of following the inherent story; 2) a term that models *importance*, to choose only shots that show some important aspect of the day; and 3) a term that models *diversity* and avoids repeating similar events.

Summarization that considers semantic topics was recently proposed by Varini et al. [91] and Schinasi et al. [81]. In [91], it is assumed that interesting scenes in a cultural experience, such as visiting a museum, are those associated with certain patterns of behaviour of the camera wearer that are learned and used for classification. Taking into account the topic of interest of the user, different summaries can be generated from the same video. In [81], topics are revealed from a set of social media messages as highly connected messages in a graph, whose nodes encode messages and whose edges encode their similarities. Finally, the images that best represent the topic are selected based on their relevance and diversity. Lidon et al. [61], also working on photo sequences, proposed an event keyframe ranking method based on a trade-off between image relevance and diversity after removing non-informative images (containing undesired artefacts, e.g. blurring, darkness or occlusion, or showing the sky, walls or object parts) by using a new binary CNN-based filter. Their relevance criteria took into consideration several semantic measurements, including whether faces and/or objects were present, as well as whether the images had a high saliency value.

- *Remarks:* A semantic-oriented approach to egocentric summarization seems to be the most suitable for lifelogging data. Indeed, users would ideally search for complex autobiographical events that encompass simpler human actions and may not be directly correlated with their visual appearance. When dealing with photographic cameras, and due to the nature of their data, the only possible way to tackle the summarization problem is through the keyframe selection approach. Taking this into account, methods like [64] should be reformulated, either considering the video sub-shots as single frames, or developing a fine-grained segmentation procedure. This procedure should separate the data into a large number of events to have enough segments to apply the sub-shot selection correctly.

E. Content-Based Search and Retrieval

Retrieving images from a large personal database allows us to browse, search and find images of previously seen objects or places and thereby has the potential to solve a broad range of problems in egocentric vision, such as:

- searching for elements (Have I seen this before?);
- navigating (How often do I visit this place?);
- understanding the environment (Where am I right now?);
- efficiently organising huge amounts of data.

Following these premises, in [94] Wang et al. built a system for content-based searching and browsing that starts by splitting the stored data into segments and extracting three kinds of information: 1) time and other relevant attributes, 2) low visual features, and 3) audio features. Then, in the retrieval step, they applied time-based filtering by comparing the time attributes of the images in the database with the query introduced by the user. A clustering step then extracts a representative clip from each cluster; and finally, the user can provide one or more query images for the system to refine the search based on visual features and improve the query result. Still, several open issues remain: in many situations it is difficult to recall the time and where the photo we are looking at was taken; visual features are too simple to capture real object shape and texture differences; and furthermore, audio features are not provided by all wearable devices. Aghazadeh et al. [4] proposed to retrieve novel scenes and actions with respect to a previously acquired egocentric dataset by using a set of ”alignment” sequences, and matching them with a new ”query” sequence by using dynamic time warping.

Assuming that searching, browsing or summarization in visual lifelogging would largely benefit from semantic concept representation, Wang and Smeaton [93] investigated the selection of the most appropriate combination of concepts for event representation. Their strategy basically consists of reasoning on semantic networks using a density-based approach. Min et al. [24], [68] represented millions of egocentric images on a sparse graph. They represented each image as a node in the graph, and added an edge between two nodes, when they belonged to the same bag in a BoW representation. Relying on this representation, they showed that local density clustering is more suitable than global clustering methods, considering the high redundancy that lifelogging data inherently possess.

• *Remarks:* Many issues remain regarding content-based retrieval techniques, for instance: How can we make use of the basic building blocks extracted from lifelogging (actions, people and environments)? The usage of a multi-level and multi-modal descriptions based on the recognition of actions, people, objects and environments could provide a detailed image description close to text-level, which could allow high retrieval accuracy.

In methods such as [24], [68], new challenges would arise when dealing with photo data, considering the higher variability of consecutive images compared to video sequences.

III. Visual Lifelogging Analysis

We present an overview of the most important papers on visual lifelogging analysis and the problems they tackled, organised around four basic questions: *Is the user interacting? How? Where* is the user? *When are the events occurring?* and *What* is the user doing?. Table II lists the papers and related information.

TABLE II
Summary of all the visual lifelogging analysis-related papers reviewed.

VISUAL LIFELOGGING ANALYSIS								
III-A **Interacting? How?: Social Interactions**								
[31]	[5]	[6]	[35]	[9]	[1]	[3]	[2]	[72]
[85]								
III-B **Where?: Scene Understanding**								
Concept Recognition	[21]							
III-B1 Object Recognition	[75]	[74]	[37]	[17]	[14]			
III-B1 Object Discovery	[52]	[19]	[16]					
III-B2 Spatial Localisation	[55]	[13]	[95]					
III-C **When?: Time-Based Localisation**								
[62]	[88]	[23]						
III-D **What?: Action Recognition**								
III-D1 Body movements	[73]	[56]						
III-D2 Object-hand interaction	[34]	[87]	[10]	[71]				
	[59]	[58]	[77]	[76]				
III-D3 Attention	[36]	[65]	[79]					
III-D4 Other Approaches	[97]	[84]	[51]					

A. Interacting? How?: Social Interactions

Following the definition by Rummel [78], *social interactions* are all acts, actions or practices of two or more people mutually oriented towards each other. Given the powerful social nature of humans, the analysis of social interactions in lifelogging data is of fundamental importance to understanding human behaviour. Furthermore, the presence of people and social interactions are consistently associated with event memorability [47] and therefore, their detection is also potentially useful for keyframe extraction or to estimate the importance of events in a lifelog [31]. From the perspective of computer vision, social interactions can be characterised by patterns of attention between individuals. Analysing attention patterns requires the detection, tracking and locating of people in 3D environments. Indeed, when interacting with others, we naturally tend to place ourselves in certain positions so as to stand close to those we interact with and avoid occlusions. F-formations [53] have been demonstrated to be a suitable formalism for modelling social interaction behaviour. Following the original definition by Kendon [54]:

> *An F-formation arises whenever two or more people sustain a spatial and orientational relationship in which the space between them is one to which they have equal, direct, and exclusive access.*

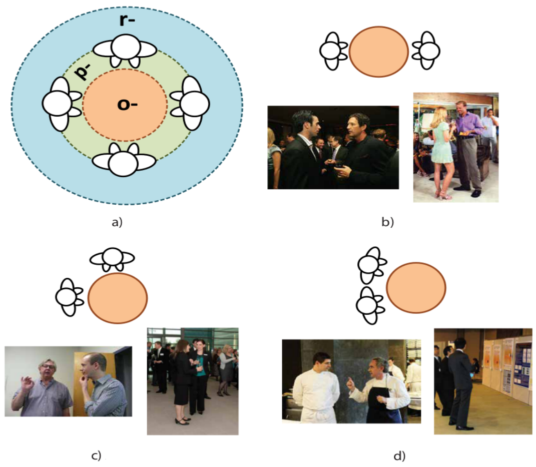

Fig. 6. Different arrangements of F-formations that are useful for social interaction analysis: (a) circular arrangement. (b) vis-a-vis arrangement. (c) L-arrangement. (d) side by side arrangement. Image adapted from [82].

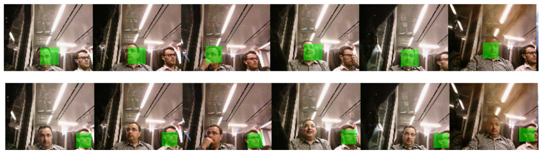

Fig. 7. Example of multi-face tracking obtained by applying the method in [1] to track multiple faces in LTR sequences captured by a wearable camera. Each row represents the track of a different person.

Examples of F-formations are given in Fig. 6. The F-formations theory has been successfully applied in social interaction analysis [46] using classical videos or still images, and more recently to egocentric videos [6]. Head estimation and 3D location are crucial for the detection of F-formations. Indeed, a rough estimate of someone's head pose allows us to understand with a certain precision what the person is looking at; while it is important to estimate the distance people have from the camera wearer and other people if there is interaction.

In sequences captured through a wearable camera, pose estimation is a challenging task due to the continuous changes of aspect ratio, scale and orientation. A common way to address this problem [5], [6], [9], [72] is to assume that where a group interacts in a discussion, the head of each person will be oriented for a while towards the person who is speaking, and to use a model to capture this behaviour over time. Generally, in video sequences, this is achieved through a hidden Markov model or Markov random fields, where the latent variable corresponds to the head pose and the observed variables to the results of a multiple person tracker, applied to the input images. The only works devoted to the analysis of photo sequences are [1], [3], [2]. In this context, tracking people is very challenging due to the abrupt and very frequent changes of view. The proposed approach basically consists of computing backward and forward correspondences for each face detected in the sequence and of grouping similar tracklets into bags, which should correspond to different people (see Fig. 7). A combination of first-person and third-person views is considered by Soo and Shi [85] to predict social saliency, considered as the likelihood of joint attention, in real-world scenes with multiple social groups. This is basically achieved by modelling social formation features that encode the geometric relation between the joint attention and spatial distribution of the members of a social group.

- *Remarks:* In general, there is common agreement about the need to track people, head orientation and 3D locations to detect F-formations that represent social groups in egocentric sequences; however, two fundamental problems arise. First, since in different social scenarios, distances and poses can assume different degrees of significance, clearly a need emerges for an algorithm to be able to adapt to different situations and learn how to treat distance and orientation features depending on the context. As a consequence, the choice of which data to use for training is crucial. Second, distances and poses strongly depend on where the camera is worn (eyeglasses, on the head, on the neck, etc.). Except [1], [3], all the methods mentioned above rely strongly on temporal coherence, since they were conceived for video sequences. Further advances in the analysis of social interactions through photographic cameras would require us to focus on features that are less sensitive to changes over time, such as people's body movements, which are consistently associated with emotional experiences [67] and could, therefore, be considered cues of social interactions.

B. *Where? Scene Understanding*

To answer the question ”Where is the user?”, we require a semantic understanding of the elements that surround the camera wearer, such as objects, people and environments, since they represent the cues available to recognise his/her surroundings. In this section, we provide an overview of computer vision tasks related to scene understanding, such as *object recognition*, *spatial localisation*, *scene parsing* and *scene recognition*. All of them share the goal of determining what the most promising techniques are for understanding scenes in lifelogging data.

1) Object Recognition and Object Discovery: Scenes can also be characterised by a vocabulary of concepts that can be found in them. With this aim, we consider the following problems: *object recognition*, which intends to identify the category that a given object belongs to; and *object discovery*, which detects, recognises and reveals new objects in images that possibly have never been seen before by the algorithm in the previous images. Due to the free motion of the camera and to the passive acquisition of lifelogging data, objects are frequently occluded and their appearance may vary broadly. Thus, the *object recognition* problem in egocentric data is becoming a challenging and active research field. The first work on object recognition in the domain of lifelogging is by Byrne et al. [21], who successfully validated supervised concept recognition, referring to relevant objects or scenes as concepts. Furthermore, using the output of the detector, they showed that the images that compose a lifelog collection tend to be temporally consistent in their visual properties, as well as in the concepts they contain. Because of this concept consistency, they suggested that an efficient automatic extraction and inference of higher-level semantic concepts based on co-occurrences and known relationships would be feasible.

Bolaños et al. [17] developed an active labelling method to generate a sufficiently large number of training examples to train an efficient supervised classifier. The method, based on a combination of hierarchical clustering trees, uses an unsupervised learning algorithm to organise the data, selecting the most informative part, asking the user for their labels, and using the feedback provided to improve the classification in a semi-supervised way. Ren et al. in [75], [74] and Fathi et al. in [37] used head-mounted cameras and proposed methods that recognise objects held in the user's hand. They segmented the background from the foreground (hands and objects) using optical flow features and relying on the fact that foreground objects will usually move in a more dynamic way while the background is more static.

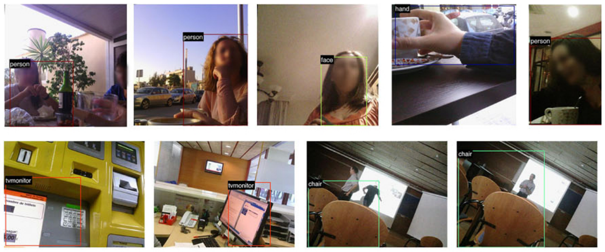

Fig. 8. Examples of objects revealed by the ego-object discovery methodology [16] for two different subjects (one per row). Better viewed in digital format.

Focusing on the task of *object discovery* in lifelogging data (see example in Fig. 8), Kang et al. [52] proposed a method, starting from an initial segmentation, that clusters only samples with higher correlation that should belong to the same object type. To this end, starting from the initial segmentation, they provide a merging strategy for segments that closely co-occur in most images. In this way, they complete objects that might be composed of different, but clearly defined parts (e.g. a laptop composed by a screen and keyboard). With the same goal, Bolaños et al. in [19], [16] proposed the use of a state-of-the-art objectness detector and a pre-trained CNN specialised in object recognition to extract a set of rich features for each object candidate followed by clustering them. The clustering integrates a "Bag of Refill" strategy of previously discovered object instances as a knowledge reuse methodology.

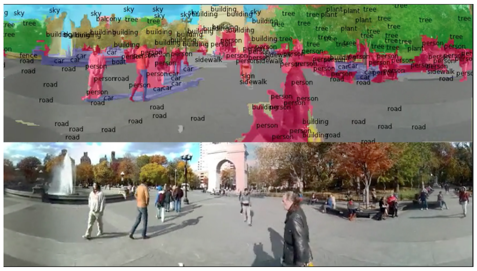

Fig. 9. Example of the result obtained (top) by applying a scene parsing algorithm to a conventional non-egocentric image (bottom). We can see the different segments found (separated by different colours) and the classes assigned to each of them. Picture adapted from [33].

2) Spatial Localisation: Bettadapura et al. in [13], proposed a method called FOV localisation that combines localisation techniques with egocentric images to localise the user(s) in the environment. To do so, they used a reference dataset, which can be images from Google Street View or pre-recorded videos from fixed cameras, and matched them to the data acquired by the user's photographic or video camera to obtain his/her localisation. They tested the system on multiple datasets captured indoors and outdoors. Additionally, they proposed a combined FOV localisation system for simultaneous localisation of multiple users of wearable devices. Wannous et al. in [95] also proposed a methodology for localisation and action-related event recognition. They used a shoulder-mounted video camera to acquire images of daily indoor living (e.g. kitchen, office, library, etc.) and built a 3D model of the different scenes. In their work, they proved that their models were more powerful than simpler 2D ones and were able to recover information from previously seen scenes with query images.

• *Remarks:* Another interesting approach that egocentric vision could benefit from is scene parsing. This is based on image segmentation; that is, separating out all the regions in an image that belong to different objects or regions. Furthermore, these kinds of techniques classically consist of providing pixel-level segmentation of the whole image and at the same time assigning an object class to each of the pixels (see the example of scene parsing in Fig. 9). To do this, most of the methods use pixel-level classifiers to achieve an initial segmentation and then a graphical model is applied to smooth and correct the boundaries of the segments [33], [98]. A limited amount of work in this field can be found in the literature but none of it was specifically designed or tested on egocentric and lifelogging datasets. Considering the differences we could find in an egocentric dataset (and more precisely in lifelogs) with respect to those typically used in scene parsing, we can enumerate some clear points to take into account when working on scene parsing:

- Scene parsing datasets are usually composed of natural and urban scenes (in general, outdoors) and their corresponding class distributions have a high percentage of training samples related to those environments, that is, the egocentric lifelogging datasets for scene parsing would be very different considering the indoor and routine settings where people usually spend most of their time.
- Also taking into account the fact that egocentric vision datasets are composed of routine and redundant scenes, scene parsing methods focusing on lifelogging images should provide some higher context and knowledge-reuse mechanisms to take advantage of the previously parsed images in the egocentric sequence.

Related to scene parsing, it would also be useful to be able to recognise the scene the user is in. Although no work has been presented with this purpose using egocentric data, a good example with conventional images is the dataset Places205 [100]. This information could help when deciding, for instance, how we should segment the day into events or use this information to exploit the environment-object relationships.

Although good methodologies have been proposed for object recognition and object discovery using egocentric and lifelogging images, there is still a lot of work to do to semantically describe the camera wearer's environment at a high level. The development of object detection methods specifically designed for egocentric images could not only improve existent recognition and discovery methods, but also set a more robust basis for the future appearance of scene parsing of lifelogging images. To achieve these goals, new computer vision techniques able to cope with blurring, light saturation and the occlusion of objects have to be developed. Hence, new techniques for gathering huge labelled datasets not only for object detection, but most importantly for scene parsing, must be developed. Furthermore, the addition of GPS or visual localisation techniques to scene parsing could clearly improve understanding of the environment. The most promising technique applicable to scene parsing is using Fully Convolutional Networks [63], which are able to infer the classes of each pixel treating the image as a whole instead of the current pixel-level centred classifications.

Finally, note that all the work on object recognition relies on the user-like focus and point of view that head-mounted cameras offer. This approach would not be feasible for real applications, where neck hanging cameras are usually used because they are considered less obtrusive and more user-friendly [43], despite not always being able to show what the user is doing. Moreover, these algorithms, which rely heavily on temporally close video frames and motion information, would not be applicable to LTR photographic cameras either.

C. When? Time-Based Localisation

Time information is particularly important to determine the causal relations in human behaviour. For instance, it could be useful in understanding which factors determine crises in people affected by bipolar disorder. The most common annotation tool used for keeping a record of the time in lifelogging data is the time stamp provided by cameras. By using this information, one can easily establish the temporal placement of the data in the long term, the order of the images, and their temporal distance for photographic cameras in the short term or daily. Some works have studied incorporating temporal information as a complementary feature indicator for achieving an indirect prediction. As an example, in [62], [88] the authors have treated the data acquired as a time-series to properly segment the different events present in a day. In [23], both the day of the week and the time of the day have been used for training a classifier with the ability to categorise different events. Naaman et al. [69] studied the role of the time stamp as a memory cue in a psychological experiment on conventional images and concluded that people are unable to retrieve their memories when only given the time and date; consequently, additional information is needed for retrieval methods to be effective.

D. What? Action Recognition

Inferring what the camera wearer is doing from a visual lifelog basically requires the categorisation of everyday activities. The categories to focus on depend on the kind of application. For instance, in healthcare and well-being applications, occupational therapy research may guide the selection of the target activities and related concepts (see Fig. 10 as an example of sports category recognition). For diet monitoring applications, eating actions will be the focus; whereas in applications related to the diagnosis of dementia, the focus will be on daily life activities such as dressing, making coffee and cooking. In quantified-self applications, activities like housework, watching TV, working/studying, eating/drinking, etc. are the most prevalent activities.

Traditional action recognition methods can be broadly classified depending on the kind of features they use to represent actions; with body movement analysis and the use of the objects involved in the action being the most common choices. Only very recently has the scene context been used to improve action recognition. Still, the choice of the representation strongly depends on the kind of actions to be classified.

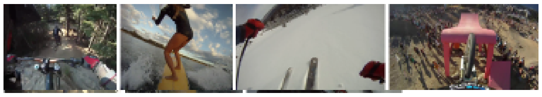

Fig. 10. Examples of first-person point of view images performing various sports. Image adapted from [56].

1) Body movement-based methods: In an egocentric setting, general body movements such as running, walking, moving the head/camera or staying still are usually estimated relying on motion features (when this is possible with the temporal resolution of the camera). Usually, based on such features, the ego-action classification can also be used for event segmentation. Typically, video cameras like GoPro, which capture around 30 fps, are used to gather data. Poleg et al. [73] proposed integrating instantaneous displacements of fixed image patches over a long period of time to remove the zero mean variations due to head rotation. By applying this process, they leave only the consistent displacement caused by forward motion. The cumulative displacement curves show different patterns for ego-motion activities, so that activities become easy to classify. Instead of focusing on the goal of building discriminative motion features, Kitani et al. [56] used several modifications of classical motion-based feature vectors and built a complex Bayesian model for clustering.

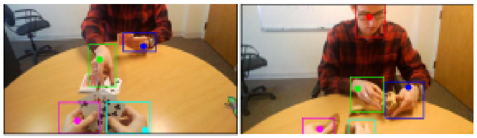

Fig. 11. Examples of first-person point of view images for recognising activities involving hands. The algorithm is capable of detecting the left and right hand of the user, in pink and light blue respectively; and the left and right hand of the person he/she is interacting with, in dark blue and green, respectively. Image adapted from [58], devoted to hand disambiguation.

2) Object-hand interaction-based methods: A first-person point of view offers an ideal perspective from which to analyse hand-object manipulation or hand-eye coordination (see Fig. 11). The main idea, introduced by Fathi et al. [34] and further improved by [10], [71], [87] is that objects are correlated with actions (e.g. dish and nibbling) and actions with activities, and these correlations can be exploited to build robust object models. However, the challenges come from additional occlusions (from manipulated objects, or self-occlusions of fingers by the palm) and the fact that hands interact with the environment and often leave the camera FOV.

Others have focused on different problems related to hand-object manipulation such as capturing the variability of hand appearance over a diverse set of imaging conditions and hand poses [59], disambiguating and tracking the observers hands and those of social partners [58], improving robustness against camera motion [74], [76], [77], or capturing the appearance of visual composites of humans and objects in interaction [71].

3) Attention-based methods: The use of manipulation-based approaches is restricted to scenes and objects where the user's hands present significant information. Attention-based approaches aim to identify objects to which the user pays particular attention, even in the absence of manipulation, since they could be key factors in self-behaviour recognition. In general, these methods are applicable to data acquired by head, eyeglass or ear-mounted cameras only. Attention can be used to find salient objects as in Matsuo et al. [65], or to capture the relationship between action and gaze, as in [36].

4) Other approaches: To detect activities that cannot be fully characterised by body movement, object-hand manipulation or object-gaze relationships, motion has been the most commonly used feature. Instead of trying to compute ego-motion, these approaches describe the frames that compose the actions, they use a set of motion and visual word features in a local (on a single frame) and global (on a set of consecutive frames) manner and create a specific structure for obtaining a temporally and spatially consistent representation of the action. Song et al. [84] obtained an accuracy rate of activity recognition of about 80% using the dataset they published (LENa dataset), by adopting the dense trajectory approach. In [79], the authors used a wearable video camera to capture and recognise a diverse set of actions (e.g. throwing, hand shaking, hugging or waving) which, in this case, is made by other people towards the camera user. Recently, a newer approach for action recognition was proposed by the same authors in [80]. On this occasion, they used CNN features to describe the frames of an HTR video. To obtain a rich and motion-like representation, they then proposed the use of a temporal pooling operator (PoT). An interesting alternative to motion was proposed by Yan et al. [97], who exploited the fact that typically people tend to perform the same actions in the same environment (e.g. people at work typically have a coffee break) and their results show the advantage of sharing information between tasks. Kanade et al. [51] explored the problem of activity recognition from a deeper perspective. They proposed several methods for activity recognition, some based on object and scene understanding, which are specifically adapted to their eye-glass mounted wearable device.

• *Remarks:* In essence, the most common cues on which activity recognition in egocentric videos relies on are body movement, object-hand interaction and patterns of attention. Body movement-based methods rely on motion estimation and therefore are not directly applicable to data acquired by photographic cameras. Object-hand interaction and patterns of attention are feasible for data acquired by wearable cameras attached to the head or somewhere near the person's eyes that could follow his/her gaze. However, when the camera is worn as a necklace or attached to the clothes, attention-based methods fail, making it impossible to see what the user is manipulating and making it very difficult to estimate the centre of attention. Similarly, object-hand methods can be very difficult to apply considering the free motion of the camera and the difficulty in regularly showing the hands of the user. To the best of our knowledge, there is no published work on recognition of egocentric activities recorded by freely worn cameras. In this context, it would be a requirement for robust activity recognition to take into account information concerning whether the camera wearer is stationary or moving.

IV. Availability of datasets and software

A. Egocentric Vision Datasets

As egocentric vision is a relatively new research field, the creation of standardised and rich enough datasets and annotations to test and compare the new algorithms is crucial to boost the development of the field. In Table III, we provide a summary of currently available public egocentric datasets, specifying, for each of them, the following information: the name and the reference paper where the datasets were presented or were used for the first time (where data can be found); a short description; the kind of annotated data they contain; and the camera used to acquire the data.

Only two of the publicly available egocentric datasets, EDUB [16] and AIHS [50] use photographic cameras, and thus, are useful to test and compare algorithms for visual lifelogging. Most of them are acquired using video (HTR cameras), making the analysis of long periods of time difficult. Although nearly all of them show scenes of daily living and some of them record many continuous hours of video [64], [71], [73], there is a strong need to create rich datasets with detailed annotations to ensure the robustness, applicability and usability of the algorithms for visual storytelling construction.

Following, we enumerate the available datasets (referenced by their main citation) for each of the relevant tasks applicable for analysing the main building blocks of lifelogging data:

- Social interaction analysis: [35], [6], [70]
- Object recognition/detection/discovery: [16], [75], [37], [71], [10], [26], [20]
- Gaze prediction: [36]
- Hand detection/segmentation: [37], [8], [71], [59], [7], [11]
- Gesture recognition: [8], [20], [22]
- Activity recognition: [36], [35], [73], [71], [79], [86], [84], [10], [48], [56]
- Novelty or informative region detection: [64], [4]

This analysis reveals the lack of well-established and widely accepted datasets.

B. Egocentric Vision Software

The publication of the source code is crucial to guarantee the reproducibility of research results and to allow quantitative comparisons on different datasets. To divulge available egocentric vision-related software, we present a list of the most relevant repositories, including source code for object recognition, object discovery, activity recognition, event segmentation, keyframe-based summarization and informative image detection in Table IV.

V. Conclusions and Future Directions

This review summarized the state of the art of visual lifelogging analysis from a storytelling perspective, focusing on the progresses made so far in this context in the field of computer vision. In the first part of this survey we reviewed several techniques for acquiring, organizing, summarizing and browsing large collections of unstructured data. In the second part, we organized the available literature around the central questions necessary to address the storytelling problem: *Was the user interacting with somebody? How?*, *Where* is he/she?, *When* did the event occur? and *What* is the person wearing the camera doing?. For each research question we highlighted the weaknesses and strengths of available methods with respect to their applicability to the LTR domain. Additionally, we reviewed all the available datasets and source code.

Generally, from this review, we can draw some conclusions regarding the crucial points that must be followed in short-term research into egocentric vision. *First*, there is a need to develop more algorithms suited to data acquired through photo cameras, in particular for social interaction detection and analysis, as well as for activity and context recognition. *Second*, in view of the large number of datasets made publicly available in the last few years, it would be useful to foster cooperation within the lifelogging scientific community to elaborate richer lifelogging datasets. By doing this, researchers could validate their algorithms and promote competition. *Third*, considering that visual storytelling has to preserve semantics, a promising direction is to continue leveraging semantic information for both egocentric data analysis and summarization. Given the wide variety of settings in which lifelogging cameras are being deployed, visual recognition could largely benefit from the use of ontologies. Moreover, this paper showed that the interest in analysis from the computer vision community over the last few years has increased considerably. In parallel, we witnessed a burst in the study and applicability of convolutional neural networks, suggesting that expectations for making progress in the coming years are growing fast. This progress should be accompanied by the creation of larger and more consolidated datasets that will compensate the enormous data demand of CNNs. In particular, research efforts should focus on the problems of 1) developing more sophisticated transfer learning strategies able to reduce the need of large annotated datasets and 2) exploiting temporal coherence of concepts that characterize visual lifelogs. However, given the current limitations of CNNs in terms of computational cost and resources, the analysis would be limited to post-processing. Finally, a promising area of research that has not been explored for storytelling via ego-vision yet, is *text description generation* from images. This problem, tackled for instance in [99], [92], consists of rendering a visual to text translation of what is happening in the images. The development of these new kinds of multi-modal techniques could open up a new area, full of potential for egocentric storytelling, in which we could provide a human-like description of what happened in a precise scene or event. The application of these algorithms to the medical field, and more precisely to people with dementia, could help provide patients with a richer context to understand better what happened to them in a given situation.

Acknowledgments

This work was partially funded by research projects TIN2012-38187-C03-01, TIN2015-66951-C2-1-R (MICINN), SGR 1219 (AGAUR) and 20141510 to Maite Garolera (Fundaci Marat TV3). M. Dimiccoli is supported by a *Beatriu de Pinòs* grant (Marie-Curie COFUND action), and P. Radeva by an *ICREA Academia* grant.

References

[1] M. Aghaei, M. Dimiccoli, and P. Radeva. Multi-face tracking by extended bag-of-tracklets in egocentric photo-streams. *To appear in Computer Vision and Image Understanding*, 2015.

[2] M. Aghaei, M. Dimiccoli, and P. Radeva. Towards social interaction detection in egocentric photo streams. In *Machine Vision, International Conference on*, 2015.

[3] M. Aghaei and P. Radeva. Bag-of-tracklets for person tracking in life-logging data. *Artificial Intelligence Research and Development: Recent Advances and Applications*, 269:35, 2014.

[4] O. Aghazadeh, J. Sullivan, and S. Carlsson. Novelty detection from an ego-centric perspective. In *Computer Vision and Pattern Recognition, IEEE Conference on*, pages 3297–3304, 2011.

[5] G. Alletto, S.and Serra, S. Calderara, and R. Cucchiara. Head pose estimation in first-person camera views. In *Pattern Recognition 22nd International Conference on*, pages 4188–4193. IEEE, 2014.

[6] S. Alletto, G. Serra, S. Calderara, F. Solera, and R. Cucchiara. From ego to nos-vision: Detecting social relationships in first-person views. In *Computer Vision and Pattern Recognition Workshops, IEEE Conference on*, pages 594–599, 2014.

[7] Sven Bambach, Stefan Lee, David Crandall, and Chen Yu. Lending a hand: Detecting hands and recognizing activities in complex egocentric interactions. In *Computer Vision, IEEE International Conference on*, 2015.

[8] L. Baraldi, F. Paci, G. Serra, Luca Benini, and R. Cucchiara. Gesture recognition in ego-centric videos using dense trajectories and hand segmentation. In *Computer Vision and Pattern Recognition Workshops, IEEE Conference on*, pages 702–707, 2014.

[9] L. Bazzani, M. Cristani, D. Tosato, M. Farenzena, G. Paggetti, G. Menegaz, and V. Murino. Social interactions by visual focus of attention in a three-dimensional environment. *Expert Systems*, 30(2):115–127, 2013.

[10] A. Behera, D. C? Hogg, and A. G. Cohn. Egocentric activity monitoring and recovery. In *Asian Conference on Computer Vision*, pages 519–532. Springer, 2013.

[11] A. Betancourt, P. Morerio, E. I Barakova, L. Marcenaro, M. Rauterberg, and Carlo S. Regazzoni. A dynamic approach and a new dataset for hand-detection in first person vision. In *Computer Analysis of Images and Patterns*, pages 274–287. Springer, 2015.

[12] A. Betancourt, P. Morerio, C. S. Regazzoni, and M. Rauterberg. The evolution of first person vision methods: A survey. *IEEE Transactions on Circuits and Systems for Video Technology*, 25(5):744–760, 2015.

[13] V. Bettadapura, I. Essa, and C. Pantofaru. Egocentric field-of-view localization using first-person point-of-view devices. In *Applications of Computer Vision, IEEE Winter Conference on*, pages 626–633, 2015.

[14] V. Bettadapura, E. Thomaz, A. Parnami, G. D. Abowd, and I. Essa. Leveraging context to support automated food recognition in restaurants. In *Applications of Computer Vision, IEEE Winter Conference on*, pages 580–587, 2015.

[15] M. Bolaños, R. Mestre, E. Talavera, X. Giró-i Nieto, and P. Radeva. Visual summary of egocentric photostreams by representative keyframes. In *Multimedia & Expo Workshops, IEEE International Conference on*, pages 1–6, 2015.

[16] M. Bolaños and P. Radeva. Ego-object discovery. *arXiv preprint arXiv:1504.01639*, 2015.

[17] M. Bolaños, M. Garolera, and P. Radeva. Active labeling application applied to food-related object recognition. In *ACM International workshop on Multimedia for cooking & eating activities*, pages 45–50, 2013.

[18] M. Bolaños, M. Garolera, and P. Radeva. Video segmentation of lifelogging videos. In *Articulated Motion and Deformable Objects*, pages 1–9. Springer, 2014.

[19] M. Bolaños, M. Garolera, and P. Radeva. Object discovery using cnn features in egocentric videos. In *Pattern Recognition and Image Analysis*, pages 67–74. Springer, 2015.

[20] I. M. Bullock, T. Feix, and A. M. Dollar. The yale human grasping dataset: Grasp, object, and task data in household and machine shop environments. *The International Journal of Robotics Research*, 34(3):251–255, 2015.

[21] D. Byrne, A. R. Doherty, C. G.M. Snoek, G. J.F. Jones, and A. F. Smeaton. Everyday concept detection in visual lifelogs: validation, relationships and trends. *Multimedia Tools and Applications*, 49(1):119–144, 2010.

[22] Minjie Cai, Kris M Kitani, and Yoichi Sato. A scalable approach for understanding the visual structures of hand grasps. In *Robotics and Automation (ICRA), 2015 IEEE International Conference on*, pages 1360–1366. IEEE, 2015.

[23] D. Castro, S. Hickson, V. Bettadapura, E. Thomaz, G. Abowd, H. Christensen, and I. Essa. Predicting daily activities from egocentric images using deep learning. In *ACM International Symposium on Wearable Computers*, pages 75–82, 2015.

[24] V. Chandrasekhar, C. Tan, W. Min, L. Liyuan, L. Xiaoli, and L. J. Hwee. Incremental graph clustering for efficient retrieval from streaming egocentric video data. In *Pattern Recognition, IEEE International Conference on*, pages 2631–2636, 2014.

[25] S. Chowdhury, P. J. McParlane, S. Ferdous, and J. Jose. My day in review: Visually summarising noisy lifelog data. In *ACM International Conference on Multimedia information Retrieval*, 2015.

[26] D. Damen, O. Haines, T. Leelasawassk, A. Calway, and W. Mayol-Cuevas. Multi-user egocentric online system for unsupervised assistance on object usage. In *Computer Vision Workshops, European Conference on*, pages 481–492. Springer, 2014.

[27] M. Dimiccoli and P. Radeva. Visual lifelogging in the era of outstanding digitization. *Digital Presentation and Preservation of Cultural and Scientific Heritage*, (V):59–64, 2015.

[28] A. R. Doherty, C. Ó Conaire, M. Blighe, A. F. Smeaton, and N. E. O'Connor. Combining image descriptors to effectively retrieve events from visual lifelogs. In *ACM international conference on Multimedia information retrieval*, pages 10–17. ACM, 2008.

[29] A. R. Doherty, K. Pauly-Takacs, N. Caprani, C. Gurrin, C. J.A. Moulin, N. E. O'Connor, and A. F. Smeaton. Experiences of aiding autobiographical memory using the sensecam. *Human–Computer Interaction*, 27(1-2):151–174, 2012.

[30] A. R. Doherty and A. F. Smeaton. Automatically segmenting lifelog data into events. In *International Workshop on Image and Audio Analysis for Multimedia Interactive Services*, pages 20–23, 2008.

[31] A. R. Doherty and A. F. Smeaton. Combining face detection and novelty to identify important events in a visual lifelog. In *Computer and Information Technology Workshops, IEEE International Conference on*, pages 348–353, 2008.

[32] A.R. Doherty, S.E. Hodges, Abby C King, A.F Smeaton, E. Berry, C. J.A. Moulin, S. Lindley, P. Kelly, and C. Foster. Wearable cameras in health: the state of the art and future possibilities. *American journal of preventive medicine*, 44(3):320–323, 2013.

[33] C. Farabet, C. Couprie, L. Najman, and Y. LeCun. Learning hierarchical features for scene labeling. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 35(8):1915–1929, 2013.

[34] A. Fathi, A. Farhadi, and J. M. Rehg. Understanding egocentric activities. In *Computer Vision, IEEE International Conference on*, pages 407–414, 2011.

[35] A. Fathi, J. K. Hodgins, and J. M. Rehg. Social interactions: A first-person perspective. In *Computer Vision and Pattern Recognition,IEEE Conference on*, pages 1226–1233, 2012.

[36] A. Fathi, Y. Li, and J. M. Rehg. Learning to recognize daily actions using gaze. In *Computer Vision, European Conference on*, pages 314–327. Springer, 2012.

[37] A. Fathi, X. Ren, and J. M. Rehg. Learning to recognize objects in egocentric activities. In *Computer Vision and Pattern Recognition, IEEE Conference On*, pages 3281–3288, 2011.

[38] T. Feix, R. Pawlik, H.-B. Schmiedmayer, J. Romero, and D. Kragic. A comprehensive grasp taxonomy. In *Robotics, Science and Systems: Workshop on Understanding the Human Hand for Advancing Robotic Manipulation*, pages 2–3, 2009.

[39] Axel Furlan, Stephen Miller, Domenico G Sorrenti, Li Fei-Fei, and Silvio Savarese. Free your camera: 3d indoor scene understanding from arbitrary camera motion. In *British Machine Vision Conference*, 2013.

[40] J. Ghosh, Y. J. Lee, and K. Grauman. Discovering important people and objects for egocentric video summarization. In *Computer Vision and Pattern Recognition, IEEE Conference on*, pages 1346–1353, 2012.

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

[42] Morgan Harvey, Marc Langheinrich, and Geoff Ward. Remembering through lifelogging: A survey of human memory augmentation. *Pervasive and Mobile Computing*, 2015.

[43] D. S. Hayden, C. Vondrick, S. X. Jia, Y. Landa, R. C. Miller, A. Torralba, and S. Teller. The accuracy-obtrusiveness tradeoff for wearable vision platforms. In *Computer Vision and Pattern Recognition Workshop on Egocentric Vision, IEEE Conference On*, 2012.

[44] S. Hodges, L. Williams, E. Berry, S. Izadi, J. Srinivasan, A. Butler, G; Smyth, N. Kapur, and K. Wood. Sensecam: A retrospective memory aid. In *Ubiquitous Computing*, pages 177–193. Springer, 2006.

[45] F. Hopfgartner, Y. Yang, L. M. Zhou, and C. Gurrin. User interaction templates for the design of lifelogging systems. In *Semantic Models for Adaptive Interactive Systems*, pages 187–204. Springer, 2013.

[46] H. Hung and B. Kröse. Detecting f-formations as dominant sets. In *ACM international conference on multimodal interfaces*, pages 231–238, 2011.

[47] P. Isola, J. Xiao, A. Torralba, and A. Oliva. What makes an image memorable? In *Computer Vision and Pattern Recognition IEEE Conference on*, pages 145–152, 2011.

[48] Y. Iwashita, A. Takamine, R. Kurazume, and M.S. Ryoo. First-person animal activity recognition from egocentric videos. In *Pattern Recognition, IEEE International Conference on*, pages 4310–4315, 2014.

[49] A. Jinda-Apiraksa, J. Machajdik, and R. Sablatnig. A keyframe selection of lifelog image sequences. *Erasmus Mundus M. Sc. in Visions and Robotics thesis, Vienna University of Technology*, 2012.

[50] N. Jojic, A. Perina, and V. Murino. Structural epitome: a way to summarize ones visual experience. In *Advances in Neural Information Processing Systems*, pages 1027–1035. 2010.

[51] Takeo Kanade and Martial Hebert. First-person vision. *Proceedings of the IEEE*, 100(8):2442–2453, 2012.

[52] H. Kang, M. Hebert, and T. Kanade. Discovering object instances from scenes of daily living. In *Computer Vision, IEEE International Conference on*, pages 762–769, 2011.

[53] A. Kendon. *Studies in the behavior of social interaction*, volume 6. Humanities Press Intl, 1977.

[54] A. Kendon. *Conducting interaction: Patterns of behavior in focused encounters*, volume 7. CUP Archive, 1990.

[55] B. Kikhia, A.y Boytsov, J. Hallberg, H. Jonsson, and K. Synnes. Structuring and presenting lifelogs based on location data. In *Pervasive Computing Paradigms for Mental Health*, pages 133–144. Springer, 2014.

[56] K. M. Kitani, T. Okabe, Y. Sato, and A. Sugimoto. Fast unsupervised ego-action learning for first-person sports videos. In *Computer Vision and Pattern Recognition, IEEE Conference on*, pages 3241–3248, 2011.

[57] M. L. Lee and A. K Dey. Lifelogging memory appliance for people with episodic memory impairment. In *Proceedings of the 10th international conference on Ubiquitous computing*, pages 44–53. ACM, 2008.

[58] S. Lee, S. Bambach, D. J Crandall, J. M Franchak, and C. Yu. This hand is my hand: A probabilistic approach to hand disambiguation in egocentric video. In *Computer Vision and Pattern Recognition Workshops, IEEE Conference on*, pages 557–564, 2014.

[59] C. Li and K. M. Kitani. Pixel-level hand detection in ego-centric videos. In *Computer Vision and Pattern Recognition, IEEE Conference on*, pages 3570–3577, 2013.

[60] N. Li, M. Crane, and H. J. Ruskin. Automatically detecting "significant events" on sensecam. *International Journal of Wavelets, Multiresolution and Information Processing*, 11(06):1350050, 2013.

[61] A. Lidon, M. Bolaños, M. Dimiccoli, P. Radeva, M. Garolera, and X. Giró-i Nieto. Semantic summarization of egocentric photo stream events. *arXiv preprint arXiv:1511.00438*, 2015.

[62] W.-H. Lin and A. Hauptmann. Structuring continuous video recordings of everyday life using time-constrained clustering. In *Electronic Imaging 2006*, pages 60730D–60730D. International Society for Optics and Photonics, 2006.

[63] J. Long, E. Shelhamer, and T. Darrell. Fully convolutional networks for semantic segmentation. In *Computer Vision and Pattern Recognition, IEEE Conference on*, pages 3431–3440, 2015.

[64] Z. Lu and K. Grauman. Story-driven summarization for egocentric video. In *Computer Vision and Pattern Recognition, IEEE Conference on*, pages 2714–2721, 2013.

[65] K. Matsuo, K. Yamada, S. Ueno, and S. Naito. An attention-based activity recognition for egocentric video. In *Computer Vision and Pattern Recognition Workshops, IEEE Conference on*, pages 565–570, 2014.

[66] W. W. Mayol-Cuevas, B. J. Tordoff, and D. W. Murray. On the choice and placement of wearable vision sensors. *Systems, Man and Cybernetics, Part A: Systems and Humans, IEEE Transactions on*, 39(2):414–425, 2009.

[67] A. Mehrabian. Significance of posture and position in the communication of attitude and status relationships. *Psychological Bulletin*, 71(5):359, 1969.

[68] W. Min, X. Li, C. Tan, B. Mandal, L. Li, and J.-H. Lim. Efficient retrieval from large-scale egocentric visual data using a sparse graph representation. In *Computer Vision and Pattern Recognition Workshops, IEEE Conference on*, pages 541–548, 2014.

[69] M. Naaman, S. Harada, Q. Wang, H. Garcia-Molina, and A. Paepcke. Context data in geo-referenced digital photo collections. In *ACM international conference on Multimedia*, pages 196–203, 2004.

[70] S. Narayan, M. S. Kankanhalli, and K. R. Ramakrishnan. Action and interaction recognition in first-person videos. In *Computer Vision and Pattern Recognition Workshops IEEE Conference on*, pages 526–532, 2014.

[71] H. Pirsiavash and D. Ramanan. Detecting activities of daily living in first-person camera views. In *Computer Vision and Pattern Recognition, IEEE Conference on*, pages 2847–2854, 2012.

[72] F. Poiesi and A. Cavallaro. Predicting and recognizing human interactions in public spaces. *Journal of Real-Time Image Processing*, pages 1–19, 2014.

[73] Y. Poleg, C. Arora, and S. Peleg. Temporal segmentation of egocentric videos. In *Computer Vision and Pattern Recognition, IEEE Conference on*, pages 2537–2544, 2014.

[74] X. Ren and C. Gu. Figure-ground segmentation improves handled object recognition in egocentric video. In *Computer Vision and Pattern Recognition, IEEE Conference on*, pages 3137–3144, 2010.

[75] X. Ren and M. Philipose. Egocentric recognition of handled objects: Benchmark and analysis. In *Computer Vision and Pattern Recognition Workshops, IEEE Conference on*, pages 1–8, 2009.

[76] G. Rogez, M. Khademi, J.S. Supančič, J.-M. M. Montiel, and D. Ramanan. 3d hand pose detection in egocentric rgb-d images. In *Computer Vision Workshops, European Conference on*, pages 356–371. Springer, 2014.

[77] G. Rogez, J. S. Supancic III, and D. Ramanan. Egocentric pose recognition in four lines of code. *arXiv preprint arXiv:1412.0060*, 2014.

[78] R. J. Rummel. Social behavior and interaction–chapter 9. *Understanding Conflict and War–The Conflict. John Wiley*, 1976.

[79] M. S. Ryoo and L. Matthies. First-person activity recognition: What are they doing to me? In *Computer Vision and Pattern Recognition, IEEE Conference on*, pages 2730–2737, 2013.

[80] M. S. Ryoo, B. Rothrock, and L. Matthies. Pooled motion features for first-person videos. In *Computer Vision and Pattern Recognition, IEEE Conference on*, pages 896–904, 2015.

[81] M. Schinas, S. Papadopoulos, Y. Kompatsiaris, and P. A. Mitkas. Visual event summarization on social media using topic modelling and graph-based ranking algorithms. In *ACM International Conference on Multimedia information Retrieval*, 2015.

[82] F. Setti, C. Russell, C. Bassetti, and M. Cristani. F-formation detection: Individuating free-standing conversational groups in images. *PloS one*, 10(5):e0123783, 2015.

[83] A. F. Smeaton, P. Over, and A. R. Doherty. Video shot boundary detection: Seven years of trecvid activity. *Computer Vision and Image Understanding*, 114(4):411–418, 2010.

[84] S. Song, V. Chandrasekhar, N.-M. Cheung, S. Narayan, L. Li, and J.-H. Lim. Activity recognition in egocentric life-logging videos. In *Computer Vision - ACCV Workshops*, pages 445–458. Springer, 2014.

[85] H. Soo Park and J. Shi. Social saliency prediction. In *Computer Vision and Pattern Recognition, IEEE Conference on*, pages 4777–4785, 2015.

[86] E. H. Spriggs, F. De La Torre, and M. Hebert. Temporal segmentation and activity classification from first-person sensing. In *Computer Vision and Pattern Recognition Workshops, 2009. IEEE Computer Society Conference On*, pages 17–24, 2009.

[87] S. Sundaram and W. W. Mayol-Cuevas. Egocentric visual event classification with location-based priors. In *Advances in Visual Computing*, pages 596–605. Springer, 2010.

[88] E. Talavera, M. Dimiccoli, M. Bolaños, M. Aghaei, and P. Radeva. R-clustering for egocentric video segmentation. In *Pattern Recognition and Image Analysis*, pages 327–336. Springer, 2015.

[89] C. Tan, H. Goh, V. Chandrasekhar, L. Li, and J.-H. Lim. Understanding the nature of first-person videos: Characterization and classification using low-level features. In *Computer Vision and Pattern Recognition Workshops IEEE Conference on*, pages 549–556. IEEE, 2014.

[90] B. T. Truong and S. Venkatesh. Video abstraction: A systematic review and classification. *ACM Transactions on Multimedia Computing, Communications and Applications*, 3(1):3, 2007.

[91] P. Varini and R. Serra, G.and Cucchiara. Personalized egocentric video summarization for cultural experience. In *ACM International Conference on Multimedia information Retrieval*, 2015.

[92] S. Venugopalan, M. Rohrbach, J. Donahue, R. Mooney, T. Darrell, and K. Saenko. Sequence to sequence-video to text. In *Computer Vision IEEE Conference on*, pages 4534–4542, 2015.

[93] P. Wang and A. F. Smeaton. Semantics-based selection of everyday concepts in visual lifelogging. *International Journal of Multimedia Information Retrieval*, 1(2):87–101, 2012.

[94] Z. Wang, M. D. Hoffman, P. R. Cook, and K. Li. Vferret: content-based similarity search tool for continuous archived video. In *ACM workshop on Continuous archival and retrieval of personal experiences*, pages 19–26, 2006.

[95] H. Wannous, V. Dovgalecs, R. Mégret, and M. Daoudi. *Place recognition via 3d modeling for personal activity lifelog using wearable camera*. Springer, 2012.

[96] B. Xiong and K. Grauman. Detecting snap points in egocentric video with a web photo prior. In *Computer Vision, European Conference on*, pages 282–298. Springer, 2014.

[97] Y. Yan, E. Ricci, G. Liu, and N. Sebe. Recognizing daily activities from first-person videos with multi-task clustering. In *Computer Vision - ACCV*, pages 522–537. Springer, 2014.

[98] J. Yang, B. Price, S. Cohen, and M.-H. Yang. Context driven scene parsing with attention to rare classes. In *Computer Vision and Pattern Recognition, IEEE Conference on*, pages 3294–3301, 2014.

[99] L. Yao, A. Torabi, N. Cho, K.and Ballas, C. Pal, H. Larochelle, and A. Courville. Describing videos by exploiting temporal structure. *stat*, 1050:25, 2015.

[100] B. Zhou, A. Lapedriza, J. Xiao, A. Torralba, and A. Oliva. Learning deep features for scene recognition using places database. In *Advances in Neural Information Processing Systems*, pages 487–495, 2014.

Marc Bolaños received his BSc in Computer Science from the Universitat de Barcelona (UB), Spain, in 2013. He received his interuniversity MSc in Artificial Intelligence from the Universitat Politècnica de Catalunyain 2015 and is currently a PhD candidate at the UB. His research interests are in the area of deep neural networks and egocentric vision for health improvement.

Mariella Dimiccoli is a Beatriu de Pinòs fellow (Marie-Curie COFUND action) at the Computer Vision Center and associate professor at UB. She received a Computer Engineering degree (Cum Laude) from the Technical University of Bari in 2004 and a PhD on Image Processing from the Universitat Politécnica de Catalunya (Cum Laude and European Honours) in 2009. Her current research interests are the automatic analysis and organisation of lifelogging data acquired by a wearable camera and their use in health applications.

Petia Radeva did her undergraduate study at the University of Sofia, Bulgaria, in 1989. In 1996, she received her Ph.D. from the Universitat Autnoma de Barcelona in Computer Vision applied to Medical imaging. Currently, she is a tenure associate professor at UB. She is the head of the Consolidated Research Group ”Computer Vision at the UB” and the head of MiLab at the Computer Vision Center (www.cvc.uab.es). Her research interests are in the development of learning-based approaches, especially in egocentric vision applied to health. She is associate editor of the Journals Pattern Recognition and Visual Communication and Image Representation. She received the ICREA Academia 2015 award from Catalonia given to the best 30 researchers of the year.

TABLE III: Summary of currently available public egocentric datasets.

Name	Description	Type of Annotations	Camera
Egocentric Dataset of the University of Barcelona (EDUB) [16]	With 4912 images acquired by the wearable camera Narrative; divided into 8 different days which capture daily life activities like shopping, eating, riding a bike, working, etc. It was acquired by 4 different subjects, 2 days each; and with 11,294 different object segmented instances from 21 different classes (TV, hand, person, car, sign, etc.).	Object labels and segmentations	Narrative
All I Have Seen (AIHS) [50]	Contains 19 days with a total of 45,612 images of 640 x 480 resolution, containing around 15 recurrent places/scenes appearing like home rooms, work office, work building, supermarkets, playgrounds, campus, biking trails, etc.	Not available	SenseCam
Intel Egocentric Object Dataset [75]	Has 10 video sequences (100,000 frames) from 2 subjects manipulating 42 different types of everyday object instances.	Object labels and foreground and background segmentations	PointGrey
GeorgiaTech Egocentric Activities (GTEA) [37]	The videos captured by a cap-worn camera show 7 types of daily activities, such as making a sandwich/coffee/tea, each performed by 4 different subjects. Each activity video is labelled with the list of objects involved; each frame has left hand, right hand, and background segmentation marks	Objects list and hands and background segmentations	GoPro
GTEA Gaze+ Dataset [36]	With video and audio recordings of 7 meal-preparation activities such as making pizza/pasta/salad collected using eye-tracking glasses. Each activity was performed by 5 different subjects. Each frame has eye-gaze fixation data, and different activities such as opening fridge are annotated.	Gaze and actions performed	Tobii
First-Person Social Interactions Dataset [35]	Day-long videos of 8 subjects spending their day at Disney World. The cameras are mounted on a cap worn by the subjects. Elan annotations containing the number of active participants in the scene, and the type of activity: walking, waiting, gathering, sitting, buying something, eating, etc.	Actions performed and social interactions at each time period	GoPro
Huji EgoSeg Dataset [73]	With 29 videos captured by an egocentric camera annotated in Elan format. The videos (some from YouTube and others recorded by Hebrew University of Jerusalem researchers) contain various daily activities.	Actions performed at each time period	GoPro
UT Ego Dataset [64]	Has 4 videos captured by a Looxcie wearable camera (head-mounted). Each video is about 3-5 hours long, captured in a natural, uncontrolled setting. The videos capture a variety of daily activities.	Important regions annotation	Looxcie
Interactive Museum Dataset [8]	A gesture recognition dataset taken from an egocentric perspective in a virtual museum environment. It has 5 different users who performed 7 hand gestures.	Hand gestures	No Information

VINST - Visual Diaries [4]	With 31 videos capturing the visual experience of a subject walking from a metro station to work. It consists of 7236 images in total. Each image is annotated with a location ID which covers 9 unique labels in total. Temporal segments corresponding to novel ego motions are annotated as well.	Location and ”novel ego-motions” annotations per frame	No Information
UCI Activities of Daily Living Dataset (ADL) [71]	Has 1 million frames of dozens of people performing 18 daily indoor activities such as brushing their teeth, washing dishes, or watching television, each performed by 20 different subjects. It includes annotations of 42 object classes.	Activities, object bounding boxes and classes, hand positions and interaction events	GoPro
EGO-HPE [6]	A set of egocentric videos with different subjects for head pose estimation. Each video is annotated at the frame level for five yaw angle orientations (-75, -45, 0, 45, 75) with respect to the subject wearing the camera.	Face orientation	Vuzix Smart Glass
EGO-GROUP [6]	A social group detector dataset for egocentric vision, which consists of 10 videos collected in different situations: a laboratory, a coffee break, a conference room and an outdoor scenario.	People group composition	Vuzix Smart Glass
JPL First-Person Interaction Dataset [79]	Human activity videos taken from a first-person viewpoint. The dataset specifically aims to provide first-person videos of interaction-level activities, recording how things look from the perspective of a person/robot participating in physical interactions.	Actions performed in each time period	GoPro
NUS First-person Interaction Dataset [70]	Dataset for interaction recognition with 8 interactions in 2 perspectives (first-person and third-person) resulting in 16 classes in total. The dataset will be made publicly available at a later date. It contains 2 human-human interactions, 2 human-object-human interactions and 4 human-object interaction classes. It contains 260 videos with at least 15 samples in each class.	Interaction type	GoPro
CMU Multi-Modal Activity Database (CMU-MMAC) [86]	Multimodal dataset of 18 subjects cooking 5 different recipes (brownies, pizza, etc.); also contains audio, body motion capture, and IMU data.	Frame-level action	No Information
CMU EDSH (hands under varying illuminations) [59]	Dataset of over 600 hand images taken under various illumination conditions and different backgrounds. Each image is segmented at the pixel level.	Hand segmentation	GoPro
EgoHands Dataset [7]	Contains 48 Google Glass videos of complex, first-person interactions between two people. The main intention of this dataset is to enable better, data-driven approaches to understand hands in first-person computer vision.	Hand segmentation	Google Glass
Unige-Hands Dataset [11]	Videos recorded in 5 different locations (office, street, bench, kitchen and coffee bar) intended for hand detection.	Hand/No Hand label per frame	GoPro
Yale Human Grasp Dataset [20]	Dataset with 27.7 hours of tagged video recorded by two housekeepers and two machinists during their regular work activities. It includes the tagged grasp type with its time information, objects manipulated and parameters of the performed task.	Grasp tagging, and interval and object labels	RageCams

UT Grasp Data Set [22]	Dataset under controlled environment performed by four different subjects. They were asked to grasp a set of objects placed on a desktop with specific types of grasps. The most common subset of 17 grasp types from Feix's Taxonomy [38] were selected to perform these everyday activities.	Hand grasp type and start/end frame number	GoPro
Life-logging EgoceNtric Activities (LENA) [84]	Egocentric video database containing 13 categories of activities relevant to lifelogging applications performed by 10 different subjects. Each subject recorded 2 clips for one activity (20 clips per activity). Each clip has a duration of 30 seconds.	Activities performed.	Google Glass
COGNITO [10]	Non-periodic manipulative tasks in an industrial context. All the video sequences were captured with on-body sensors consisting of IMUs, a backpack-mounted RGB-D camera for top-view and a chest-mounted fish-eye camera for the front view of the workbench.	Activity labels and objects and wrist tracklets	RGB-D and others
Michigan-Milan Indoor Dataset [39]	With 10 video sequences collected with common smartphones in a variety of environments, including offices, corridors and large rooms, where the observer moves freely (6 DoF) around the scene.	Image segmentations with the labels ”ceiling”, ”floor” or ”wall”	Smartphone
Bristol Egocentric Object Interactions Dataset [26]	Dataset captured with wearable gaze tracker software containing various pre-defined actions of daily living in different indoor locations (kitchen, workspace, gym, laser printer, corridor and weight-lifting machine). The videos in each sequence are recorded by 3-5 different users.	3D maps and 3D objects GT	ASL Mobile Eye XG
DogCentric Activity Dataset [48]	DogCentric Activity Dataset is composed of dog activity videos taken from a first-person animal viewpoint. The dataset contains 10 different types of activities, including activities performed by the dog itself, interactions between people and the dog, and activities performed by people or cars. The videos are in 320x240 image resolution, 48 frames per second.	Activity performed	GoPro
UEC EgoAction Dataset [56]	A set of videos (acquired by the researchers or public from YouTube) recording different sports (skiing, mountain biking, etc.). Each video is several minutes long and contains a wide set of actions performed by the user.	Activities performed	GoPro

TABLE IV: List of the most relevant public software related to egocentric vision.

Alireza Fathi's Egocentric Vision Toolbox [36], [37], [74]	Toolbox including functions for applying different data processing to egocentric videos, including motion estimation, image segmentation, object classification and action classification among others.
OpenCV and CUDA	http://ai.stanford.edu/~alireza/GTEA_Gaze_Website/Code/index.html
Ego-Object Discovery [16], [19]	Object Discovery Algorithm on Egocentric Images. Semi-supervised algorithm that uses initial object proposal generation, a CNN-based feature representation, false positive filtering, and an interactive object discovery with Refill strategy.
Matlab and Caffe	https://github.com/MarcBS/Ego-Object_Discovery
Detecting Activities of Daily Living in First-person Camera Views [71]	Train and test code for the problem of detecting activities of daily living (ADL). It applies novel representations including temporal pyramids to approximate temporal correspondences, and composite object models that exploit the differences between the objects when being interacted with.
Matlab	http://people.csail.mit.edu/hpirsiav/codes/ADLdataset/adl.html
Temporal Pooling of CNN Vectors [80]	It includes the pooled time series (PoT) representation framework as well as basic per-frame descriptor extractions including a histogram of optical flows (HOF) and histogram of oriented gradients (HOG).
Java and OpenCV [exec. only]	https://github.com/mryoo/pooled_time_series/
Temporal Segmentation of Egocentric Videos [73]	Software for segmentation and event classification of egocentric HTR videos. It applies a hierarchical classification using cumulative displacement curves.
Matlab and C++	http://www.vision.huji.ac.il/egoseg/
Doherty Wearable Camera Browser [30]	Application for data segmentation annotation and browsing. It supports analysis of images from the following photographic cameras: Vicon Autographer, Revue, or SenseCam.
[exec. only]	http://sensecambrowser.codeplex.com/
R-Clustering for Event Segmentation [88]	Segmentation of events in egocentric lifelogging photo streams. It uses convolutional neural network features and an energy minimisation (Graph-Cut) technique to segment photo sequences.
Matlab and Caffe	https://github.com/MarcBS/SR-Clustering
Motion-Based Egocentric Segmentation [18]	It applies a robust SIFT-Flow motion estimation suitable for photo sequences to perform photo stream segmentation in motion-related events.
Matlab	https://github.com/MarcBS/Motion_Video_Segmentation
Egocentric Vision Keyframe Summarization [15]	The code extracts a visual summary of a set of egocentric images captured by a photo camera. The result is a collage with one image summarizing every event in the image set. It uses a frame representation by means of a convolutional neural network followed by an event segmentation based on agglomerative clustering and keyframe selection based on Random Walk.
Matlab and Caffe	https://github.com/MarcBS/Egocentric-Visual-Keyframes-Summary
Egocentric Snap Points Detection [96]	Automatic prediction of snap points in unedited egocentric video that is, those frames that look as if they could be photos taken intentionally. It makes use of a generative model for snap points that rely on a photo prior to intentional (conventional) images together with domain-adapted features.
Matlab and C	https://github.com/bxiong1202/snap-points