Neurocomputing 472 (2022) 175–197

Contents lists available at ScienceDirect

Neurocomputing

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

Survey paper

Egocentric Vision-based Action Recognition: A survey

Adrián Núñez-Marcos [a,*], Gorka Azkune [b], Ignacio Arganda-Carreras [c,d,e]

[a] Deustotech Institute, University of Deusto, Avenida de las Universidades, No. 24, Bilbao 48007, Basque Country, Spain
[b] IXA NLP Group, Faculty of Computer Science, Euskal Herriko Unibertsitatea (EHU/UPV), M. Lardizabal 1, Donostia 20008, Basque Country, Spain
[c] Donostia International Physics Center (DIPC), Manuel Lardizabal 4, Donostia 20018, Basque Country, Spain
[d] Ikerbasque, Basque Foundation for Science, Plaza Euskadi 5, Bilbao 48009, Basque Country, Spain
[e] Department of Computer Science and Artificial Intelligence, University of the Basque Country, M. Lardizabal 1, Donostia 20008, Basque Country, Spain

ARTICLE INFO

Article history:
Received 6 May 2021
Revised 8 November 2021
Accepted 21 November 2021
Available online 8 December 2021

Keywords:
Deep learning
Computer vision
Human action recognition
Egocentric vision
Few-shot learning

ABSTRACT

The egocentric action recognition EAR field has recently increased its popularity due to the affordable and lightweight wearable cameras available nowadays such as GoPro and similars. Therefore, the amount of egocentric data generated has increased, triggering the interest in the understanding of egocentric videos. More specifically, the recognition of actions in egocentric videos has gained popularity due to the challenge that it poses: the wild movement of the camera and the lack of context make it hard to recognise actions with a performance similar to that of third-person vision solutions. This has ignited the research interest on the field and, nowadays, many public datasets and competitions can be found in both the machine learning and the computer vision communities. In this survey, we aim to analyse the literature on egocentric vision methods and algorithms. For that, we propose a taxonomy to divide the literature into various categories with subcategories, contributing a more fine-grained classification of the available methods. We also provide a review of the zero-shot approaches used by the EAR community, a methodology that could help to transfer EAR algorithms to real-world applications. Finally, we summarise the datasets used by researchers in the literature.

1. Introduction

Since the introduction of the first wearable camera [122], commercial and lightweight cameras such as GoPro and similars have become widely used, producing a vast amount of first-person or egocentric videos to analyse. These videos are recorded from the point of view of the wearer of the camera, producing videos with large, non-linear and unpredictable head and body motion and a lack of global context, which pose a challenge from a machine learning standpoint. Hence, the increasing amount of data and the interesting setting of these types of videos have attracted the computer vision and the machine learning communities towards the vision-based EAR research field.

In contrast to third-person or exocentric videos, first-person or egocentric videos contain rich intrinsic features, motivating their use for novel approaches, i.e. without relying exclusively on approaches from the exocentric vision literature. For example, these features include the occlusion-free interactions with objects, the focus on the manipulation of objects, the gaze movement and so forth, which have been identified in the literature [106] and are helpful to discern actions. These cues make the first-person or egocentric action recognition a research field on its own, apart from the third-person vision research. In fact, exploiting the intrinsic features of this type of vision seems to be crucial to correctly recognise the content of videos [139].

Nowadays, the egocentric vision research line has been adopted by various research groups and several solutions have been proposed. Even new features such as the use of sound are being leveraged in recent works [9,31], as some actions cannot be distinguished using only visual cues. Even though the field is advancing, it still has to become as large as the third-person one. In addition, the results are still far from being acceptable. In fact, the majority of the research is focused on the supervised learning setting in which labels are provided in the training stage. This requires large annotated datasets, which is a laborious task. There are, however, works that have analysed the use of few-shot [208] and zero-shot [205] learning frameworks. These require a few annotated samples at most, being more suitable for real-world applications than the classic supervised settings. Nevertheless, more research is required in order to steer new solutions in the correct direction.

* Corresponding author.
E-mail addresses: a.nunez@opendeusto.es (A. Núñez-Marcos), gorka.azcune@ehu.eus (G. Azkune), ignacio.arganda@ehu.eus (I. Arganda-Carreras).

https://doi.org/10.1016/j.neucom.2021.11.081
0925-2312/© 2021 The Author(s). Published by Elsevier B.V.

1.1. Vision-based Exocentric Action Recognition

In order to settle the basis for the action recognition field (later focused only on EAR), we briefly describe the evolution of the exocentric (third-person) vision-based action recognition field over the last few years.

Before the success of Deep Learning, hand-engineered features were used for action recognition; for instance, extracting the foreground (optionally), computing features from the inputs (using, e.g. traditional algorithms such as LBP [143,144], SIFT [116] and SURF [18]) and applying a classifier to obtain an action prediction. The foreground extraction can be done, for example, to segment hands and objects in egocentric-vision frames. The other two steps can be applied for both types of action recognition approaches. Other approaches include computing Optical Flow OF features, computing the skeleton and joints, trajectory-based recognition and so forth. These solutions are also seen in the EAR literature with small adjustments to fit better the features that can be found in egocentric videos (e.g. hands and objects).

With Deep Learning, features are automatically extracted, instead of manually. The exocentric action recognition field switched to three main approaches [232]: multi-stream Convolutional Neural NetworkCNNs (being the two-stream network the most used one [174]), 3D CNN [81] and those based on Recurrent Neural NetworkRNN, e.g. the Long-Short Term Memory LSTM [73]. There are other methods such as those using graphs (e.g. [13]) which can also be found within one of these categories. Moreover, thanks to the use of Neural Network NN architectures, transfer learning could be applied, allowing large models to be trained with huge datasets (Imagenet [50] for static images, UCF101 [180] for third-person videos and so forth) before being finetuned on specific tasks and/or smaller datasets (egocentric datasets, for example).

Multi-stream networks started with two branches (the two-stream network by [174]), taking RGB and OF frames, to extract spatial and temporal features and a classifier on top to make the classification. They later evolved to include more information (such as gaze [114] or visual rhythm images [40]) or even to add streams with varying information ([170] includes bones, joints and their motion as input to their multi-stream setting). Later, the computation of OF was alleviated by the proposal of [237], which had a two-stream network learning motion features in an end-to-end fashion, allowing a real-time processing.3D CNN (e.g. C3D [199]) learn spatio-temporal features using the 3D convolution operation. They are computationally heavier than the multi-stream approaches; there are even works that aimed to divide the 3D operation into a 2D and a 1D operation (as in the Xception network [38]). Furthermore, an approach called Two-Stream Inflated 3D ConvNet or I3D [30] mixed these last two ideas (two-stream network and the 3D CNN) and became an standard for extracting spatio-temporal features.

In fact, regarding feature extraction, it was usual to have a network such as an I3D extracting spatio-temporal features in a short-term span while having an RNN such as the LSTM extracting temporal features in a longer temporal span. This type of architecture was popularised by [56] and, afterwards, many works started applying it, e.g. [200,65,229].

The majority of these architectures can be directly applied to egocentric videos. However, as seen in Section 2, there are better ways to deal with egocentric videos.

1.2. Contributions and Arrangement

In this paper, we contribute the following:

- A taxonomy to classify EAR methods into categories and subcategories.
- A review of the EAR proposals using this taxonomy.

The rest of the paper is arranged as follows: Section 2 presents the aforementioned taxonomy and reviews the fine-grained classified literature; Section 3 presents the EAR methods that use or have the potential to be used within the the zero-shot paradigm; Section 4 summarises the egocentric video datasets and, finally, Section 6 provides the final conclusions.

2. Egocentric Action Recognition

The idea of using egocentric videos has only started to be exploited in the last decade thanks to novel, lightweight and affordable devices such as GoPro and similars. In fact, lifelogging has become widely used. Indeed, the number of datasets in the state of the art of the EAR field has progressively increased during this decade, with releases such as the large EPIC Kitchens dataset [42]. This has also motivated the research on the topic [84,14,21,49,139,11,12], being mainly divided into three areas: (i) activity recognition/classification, (ii) video summarisation and (iii) object detection. In this section, we aim to provide an extensive review on the action recognition subfield, referred to as EAR throughout the document. Examples of egocentric actions are shown in Fig. 1.

First of all, it should be noted that the literature presents two conflicting terms: actions and activities. [139] discussed that both terms are semantically different: an action is a short event such as ”opening a jar” while an activity is a semantically higher event in which various actions are combined, lasting from several minutes to hours. Nonetheless, part of the literature does not take this difference into account and uses the word ”activity” instead of action. Moreover, some works even denote the motion using the word ”action”, i.e. the movement generated when something is being cut would be called an action, regardless of the objects present in the scene. In this survey, we will differentiate between actions and activities and between actions and motion, being the motion for us the movement generated from an action independently of the object.

Reviewing the literature on EAR, it is noticeable that there are various special cues intrinsic to egocentric videos that drive the type of approach that researchers use to tackle the EAR challenge. For example, [106] used (i) the hand pose and its movement [17], (ii) the head motion and (iii) the gaze direction as egocentric cues in their work. In addition, they also stressed the importance of objects in the egocentric setting. In general, from the literature, we can extract the main egocentric features or cues used, summarised in Fig. 2. Hence, we can split these characteristics into two groups: those related to the appearance or objects and those related to the movement or motion.

Therefore, in this chapter, we split the literature into four sections depending on the type of modality driving the approaches: (i) object- or appearance-based approaches, (ii) motion-based approaches, (iii) hybrid approaches (combining appearance and motion) and (iv) other approaches that consider other modalities such as the sound or that are making a contribution not related to these modalities. The proposed taxonomy used for this section is illustrated in Fig. 3 and all the references following this categorisation can be found in Table 1.

We believe that having a taxonomy to divide the literature allows researchers to have a better perspective of the kinds of works that have been published or the research lines that are currently active. For a beginner, this makes it easy to find works of interest and to explore similar ones. The possible disadvantage that

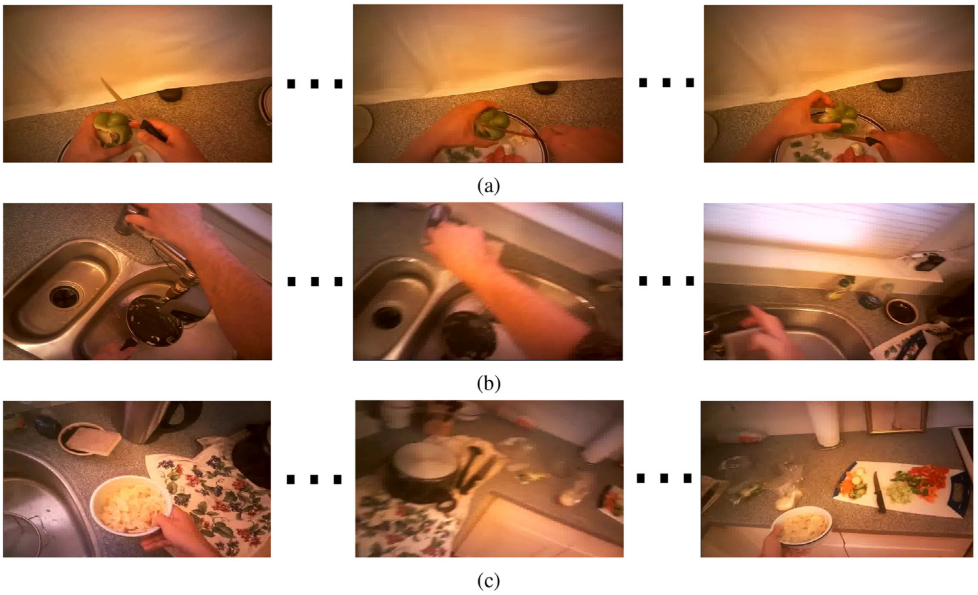

Fig. 1. Examples of egocentric actions (subsampled frames) from the Extended GTEA Gaze + dataset: (a) "cut bell pepper" action, (b) "wash pan" action and (c) "move bowl" action.

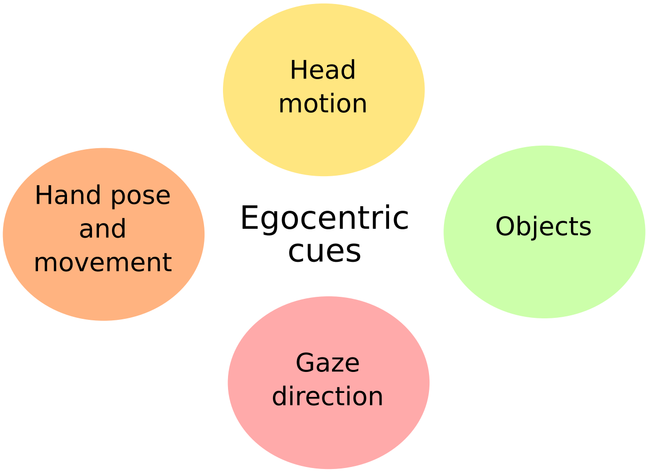

Fig. 2. Intrinsic egocentric cues.

these kinds of taxonomies may have is that, unless they are very fine-grained (which is not practical), for some works there are overlaps between categories, i.e. a specific research may fall into various categories. There might be better representations for such a taxonomy (e.g. a graph) that cannot be represented here but could be beneficial for the EAR community. We hope this taxonomy proposal enriches the research and motivates researchers to propose new ways to divide the literature.

2.1. Object-driven Action Recognition

The current literature is highly dominated by works that believe that objects present in the scene and, specially, objects related to tasks are the main cues in the recognition of actions. That is, analysing objects in videos can become a critical hint towards recognising actions. In fact, [59] argued that the egocentric paradigm is specially beneficial to analyse actions that involve objects due to three reasons: (i) object occlusions are minimised, as the space where these are manipulated is always present; (ii) objects are often seen at consistent viewing directions with respect to the egocentric camera, as poses and the displacement of the manipulated objects are also consistent in workspace coordinates; and (iii) the camera is usually focusing on objects and actions, that are usually in the centre of the image or video, thus obtaining high quality image measurements.

Regarding the classification of objects, there are various ways in the literature to categorise them. [192], for example, opted for defining objects by the type of space they are in. That is, the space observed by the subject (the one wearing the camera) is known as the *observable space*. Then, any object that is graspable or can be reached using the hands is contained within the *manipulation space*. Lastly, an object that is grabbed by the subject is said to be a manipulated object.

In a complementary way, [139] stated that four types of objects can be observed:

- Active and passive objects: active objects are those relevant for actions and passive objects are background or non-important items.
- Salient and non-salient objects: the former are those that are fixated by the gaze or those in which the focus is put on while the latter can be considered background or non-attended objects.
- Manipulated objects: objects that are in the hands are said to be manipulated.
- Multi-state objects: those that have changes in terms of colour or shape.

It is specially important to stress that active objects are considered important to estimate the action [161], but recognising them is also a challenging task due to hand occlusion or background clutter. To diminish the effect of the background clutter, [60,59] proposed to first detect a Region of Interest ROI before localising objects. In fact, there are authors that aim to detect active objects in an unsupervised way (without categorising them). Namely, [85] generated a pool of segmentations, individually searching for instances of specific objects (one at a time) by enforcing constraints such as geometric consistency. [44] used a gaze tracker to infer the most important objects and analysed the interactions with them. [129] made a segmentation process in two steps: first, they generated a probabilistic

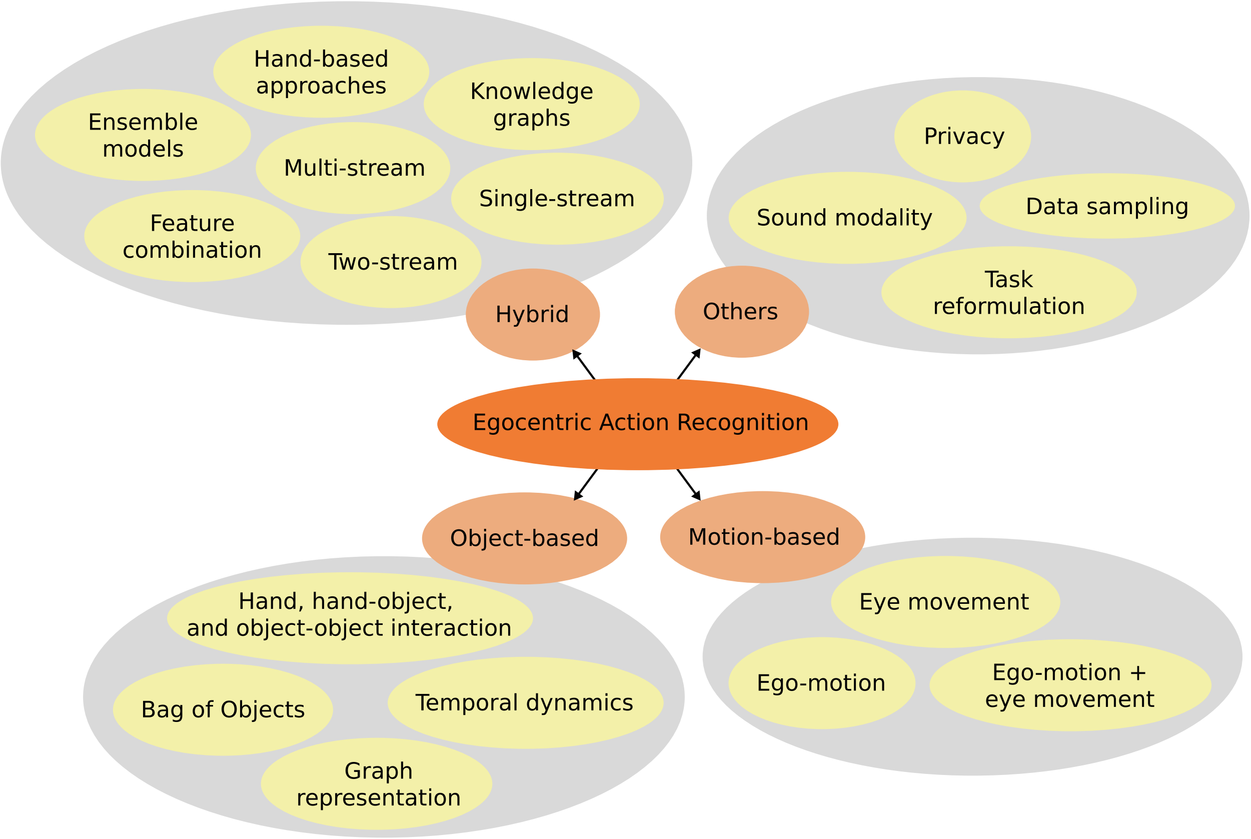

Fig. 3. The proposed taxonomy used to summarise the literature on EAR.

Table 1
Summary of the literature following the taxonomy proposed in Fig. 3.

Category	Sub-category	References
Object-based approaches	Bag of Objects approaches	[192,148,125,61,124,135,4,88]
	Hand-Object and-Hand relations	[20,19,33,67,16,120,196,138]
	Graph representations	[59,133]
	Temporal dynamics	[230]
Motion-based approaches	Eye movement	[224,225]
	Ego-motion	[191,163,178,137,153] [175,154,93,177]
	Eye movement and ego-motion	[142,215]
Hybrid approaches	Two-stream architectures	[121,95,211,189,105,202] [234,209,117,187,118,228]
	Multi-stream architectures	[195,64,74,207,83,128]
	Single-stream, multiple tasks	[176,188,86,149,146,115]
	Combination of multiple features	[182,171,216,35,176,233,134,131] [79,155,52,239,238,87,226,227,94]
	Knowledge graphs	[212,165]
	Hand-based recognition	[236,66,28]
Other approaches	Sound modality	[9,31,32,90]
	Task reformulation	[130,213]
	Privacy	[152,54,183,198]
	Data sampling	[218]

boundary map of the scene and, second, they made use of the fixation point to get the closed contour that included that point. [190] presented *EYEWATCHME*, an integrated vision and state estimation system that, at the same time, tracked, among others, the position of hands and active objects. The approaches using the gaze are specially interesting, as [97,72] showed that the eyes always look directly at the objects that are being manipulated (active objects). In fact, these approaches could be integrated in an action recognition system that aimed to use active objects' information. More recently, [100] stressed the importance of hands for the detection of active objects. They proposed to automatically segment hands first and, then, including this information in an object localisation network, achieved a more precise localisation of objects. This highlights the importance of hands in the active object detection problem.

Bag of Objects approaches. There are several studies in which the bag of objects approach is used (see Fig. 4 for an example). Works such as those of [148,124] made use of bags of active and passive objects to infer actions, being the objects first detected by an object detector and, then, classified into active or passive.

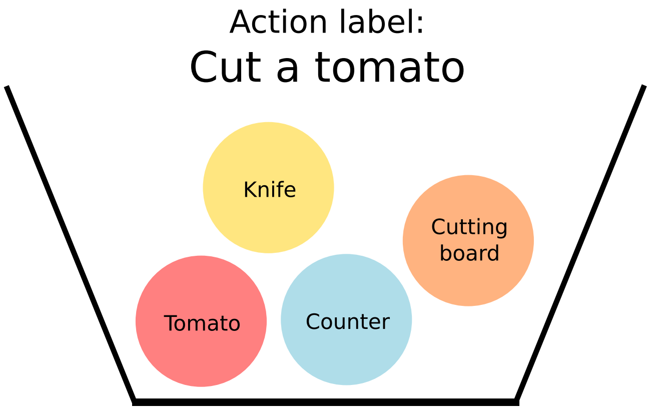

Fig. 4. Bag-of-objects approaches aim at discovering actions using a collection of objects.

[192] used two complementary sets: one of observable objects and another one of manipulable objects.

[125] argued that, as an extension of the traditional bag of objects, spatio-temporal binning approaches could capture space–time relations and, to solve the issue of inflexible predefined schemes, they first proposed to learn the spatio-temporal partitions that were most discriminative. For that, they generated a pool of randomly generated candidates and used a boosting approach to select the best ones. Second, to further improve the first contribution, they aimed to create object-centric partitions, i.e. regions of videos where active objects are supposed to appear, by creating a histogram of active objects for each video. For the classification, they computed features from each proposal in the pool and applied the boosting operation to get the best proposals that were used to train the final classifier.

One aspect related to this bag of objects are the object fluents, i.e. a time-varying attribute of objects or groups of objects whose values are the specific states of the attribute [113,62,132]. For example, for a mug, the states or fluents can be *empty* and *full* (binary fluents). Specifically, [113] proposed to represent an action as concurrent and sequential object fluents. Given an egocentric video, beam search was used to recognise the fluents per frame and then infer actions. The bag of objects used in the work of [61] was composed of sequences of visual patches of objects (a sequence represented the changes of an object during a video). [4] also modelled object state transitions as a means of inferring actions. In their model, a CNN extracted visual features from a set of frames selected from *K* segments (uniformly sampled across each video), one per segment. The network was later divided into two branches by means of a point-wise convolution: the first one was in charge of learning nouns, while the second one took care of learning states. A global average pooling was applied to obtain a feature vector from each branch, one per frame. For the noun vectors, a point-wise convolution led to a single feature vector while, for the states, two channels were left after the same operation. The two channels of the state branch represented the verb (the type of change applied from the pre-state to the post-state), learnt using a Fully-Connected FC layer. For the action classification task, another FC layer was used. [88] analysed the use of object detections from YOLO [159] as a tool to detect indoor actions and to experiment with various detection parameters. They observed that the presence of certain objects was highly correlated with some actions and that the lack in the detection of those relations hampered the detection of actions. Thus, they compensated this using the temporal information of objects, i.e. they gathered detections of various frames to get a more complete picture of the scene. More specifically, they trained a NN with a per-frame bag of objects to infer the location (physical place), they also did the same using a lSTM network to infer the location using the whole video. Finally, for action recognition, another lSTMm was used, including in the input the location and shape of the bounding boxes of the detected objects apart from the presence vectors.

New methodologies to represent the bag of objects approach are also arising, such as that of [135]. They presented a preliminary work on object-based action recognition in which they detected objects using a pre-trained CNN and they recognised the action without training any other model. Specifically, to estimate the action, they exploited web data to compute the semantic similarity between the detected object names and the names of the action classes.

Hands, Hand-Objects and Object-Object interactions. The interaction between humans (using hands mainly) and objects and also between objects is also a quite analysed topic in the EAR field. [20] presented their bag of relations, which extended the idea of the bag of objects including, not only the object itself, but also the part of the body that interacted with the object (object-body) and also the object-object relations. With the same idea of the ”bag of interactions”, [19] proposed a Histogram of Oriented Pairwise Relations in which the spatial relations (distances, orientations and alignments) between visual-words were represented. Similarly, [33] also aimed to capture hands and the objects that were being manipulated. For that, they leveraged the R*CNN presented by [67] to detect the primary region (hands) and the secondary regions (objects). The output of that module was given to an lSTM to process the evolution of the video. Going one step further, [196] presented a unified model which, given a single RGB image, in a single feed-forward pass, estimated the 3D hand and object poses, their interactions and the object and action classes. They extracted features using a Fully Convolutional NetworkFCN in which each output cell predicted 3D hand poses and object bounding box coordinates. Then, these cells were associated with a vector that contained target values for the hand and object pose, the object and action class and the overall confidence value. Those predictions with the highest confidence were passed to their *interaction RNN*.

In contrast, without the need to include interactions, there is research about the sole use of the shape and pose of hands to determine actions. [16] argued that they could infer actions in their dataset using only that information. To test their hypothesis, they masked out the region where there were no hands and used a CNN to infer actions. Even though the results were not perfect, they showed that there is a high correlation between hands and actions. Taking into account the temporal domain by applying a simple majority voting, they concluded that their results improved as a consequence of the importance that certain hand poses may have, being more distinctive than others.

While the interactions between hand and objects are important, the relation between different objects is also a central element of actions, i.e. in a given scenario, only a subset of objects may be relevant to the task. That is why [120] proposed a way to model arbitrary relations between arbitrary subgroups of objects. Their method was first divided into two parts: (i) in the coarse-grained part, a CNN extracted features from each frame, these were passed through a Multi-Layer Perceptron MLP and, to join all the features, the Scale Dot-Product Attention (SDP-Attention) of [201] was applied to them; and (ii) in the fine-grained part, the Region Proposal Network RPN proposed by [160] was used to extract object ROI, which were fed to the Recurrent Higher-Order Interaction (Recurrent HOI) module they contributed. This module employed a learnable attention mechanism to decide the set of candidate objects that were relevant for each action. Finally, the output of both streams were concatenated and a FC layer with a softmax activation was used.

[138] investigated the acquisition of additional features that modelled the interaction between hands and objects. For that, they followed the bag-of-visual-words (BoVW) approach to model actions. To infer the class of new samples, Dynamic Time Warping DTW was applied to compare the features from a new sample and the ones of the rest of samples. Next, they trained an object detector to recognise left and right hands. With these detections, the distance to any object could be determined. As active objects should be in contact with hands, those objects that were being manipulated (very close to the hands' position) were considered actives and the distance between both hands and each hand and the active object were computed. The addition of these features to the presence of objects boosted the performance on the action recognition. [140] proposed a novel NN based on SPD manifold learning. This approach employed skeleton information for hand gesture (action) recognition and was divided into three stages: (i) a CNN to encode skeletal data; (ii) a Gaussian embedding to encode first- and second-order statistics; and (iii) the learning of the SPD matrix and the mapping of this matrix to an Euclidean space for the classification of actions.3D hand pose and gesture (action) recognition were both the objective in the model proposed by [219]. This started by learning joint-aware features using a ResNet network and then the model branched in (i) the action recognition and (ii) the hand pose estimation parts. These were trained iteratively, as the output from one of them was the input to the other one and vice versa. Within these branches, they proposed to use multi-order multi-stream feature analysis. That is, various features were computed: static, those representing velocity and those representing acceleration. For the latter, they took into account the slow and fast moving joints and proposed to compute them separately. Each of these features were fed to a multi-scale relation module that went from fine-grained hand features to more holistic features and then class scores were computed with a Temporal Convolution Network (TCN). In a similar fashion (even with the same dataset), but not specifically intended for egocentric videos, [110] decoupled hand posture variations and hand movements using a two-stream network. For the first one, a 3D CNN was employed, taking also the fingertips' relative position as an extra cue. The other stream was implemented with another CNN. A FC layer computed the score per stream before fusing them for gesture recognition.

Recently, [46] presented a graph architecture to model hand skeleton data to recognise actions. Specifically, they employed a spatio-temporal graph CNN. In fact, by exploiting the symmetry of hand graphs, they proposed to use various sub-graphs to build separate models for finger movements. In contrast, [111] argued that, even though graph methods achieved good results, they were inherently limited in capturing features of hand interactions. To solve that, they contributed a self-attention based method: the hierarchical self-attention network (HAN). A joint self-attention module extracted local features and a finger self-attention module aggregated them. For temporal reasoning, the temporal self-attention module was in charge of modelling the dynamics of the fingers and the entire hand.

Graph representations. Graphs are also used to represent actions, as in the case of the work of [59], in which they built a hierarchical graph (a tree-shaped graph) for activity recognition in which an activity was composed of action nodes. The latter had some leaf nodes: object and hand nodes. Their goal in inference time was to be able to predict hands, objects, actions and activities. To train the system, they employed an algorithm similar to the Expectation-Conditional Maximization of [127]. Recently, [133] presented a work in which they built a topological map (represented by a graph) of the scene (of the physical space) from egocentric videos. In order to cluster zones, they employed a Siamese network that took pairs of images and was able to find pairs that corresponded to the same zone. Then, the graph they constructed had collections of clips within nodes (representing zones and the clips in which those zones were visited) and edges represented weak spatial connectivity between zones based on how people traversed them. From this graph they could infer the primary places of interactions and the actions related to those spaces. Moreover, they showed how to link zones across multiple related environments (such as kitchens from different datasets). [167] proposed a method to jointly recognise, localise and summarise actions. First, they applied a centre-surround model to detect a central region and its surroundings, obtaining superpixels from which features were extracted using a GoogleNet [193]. These were used to build a graph with the superpixels as nodes. By applying a random walk, all the vertices could be annotated in a single run. Finally, a fractional knapsack-type formulation was adopted to obtain a summary of the actions (given that there may be more than one action occurring at the same time and that many superpixels may be labelled as background). [96] parameterised left and right hands and objects as individual graphs to be then joined in a single multi-graph structure. This allowed their model to learn interactions between both hands and between each hand and objects.

Temporal dynamics. The appearance in a frame, the local features, can be extended to model the whole appearance of the video or, better said, its dynamics and how it evolves. [230] proposed to model the high level dynamics of the sub-events within an action by dynamically pooling features of sub-intervals of time series using a temporal feature pooling function. Specifically, each frame was encoded using a CNN, in which each activation neuron was considered a point in the time series, and features were pooled in determined intervals (sub-events) to model the short-term changes. Then, these sub-event dynamics were temporally aligned and a group of Fourier coefficients were extracted in a temporal pyramid to encode the overall video representation. Transformer layers can also be employed to model this evolution, transforming the problem in a sequence-to-sequence task. For example, [103] presented their *Trear*, a Transformer-based architecture that took RGB and depth images. Each modality was fed to an inter-frame attention encoder (not sharing weights among them), merging later in the mutual-attentional fusion block, allowing them to create cross-modal representations. The latter are fed to a linear layer to obtain a per-frame prediction, averaged at the end across frames for the final action prediction.

2.2. Motion-driven Action Recognition

Apart from the object cues, which have shown to be relevant in egocentric contexts, there are also cues related to the motion: eye movement, hand motion and head motion. There is also a feature called ego-motion, usually referring to the global motion generated from objects in the scene, the movement of the body and the head. Fig. 5 shows an example of the ego-motion of a video.

Eye movement. [98] stated that a person's eye movement is a valuable source of information to recognise actions. In addition, as mentioned by [27], the eye movement can be classified into three types of movements: saccades, fixations and blinks. Saccades are the constant and simultaneous movements of both eyes that are aimed at building a mental "map" of the interesting parts of the scene, fixations are stationary states in which the gaze is fixed on a specific place and blinks are the regular opening and closing movements of the eyelids. [224] limited themselves to actions performed on a table and took hand positions, the locations of the eyes and the head and the recorded ego-videos. Their aim was to be able to segment actions. For that, and based on the fact that eye and head movements are related to the attention as mentioned in the work of [72], they developed a method to detect attention switches. The tracking was done using a head-mounted ISCAN infra-red video based eye tracker. With this, they divided each

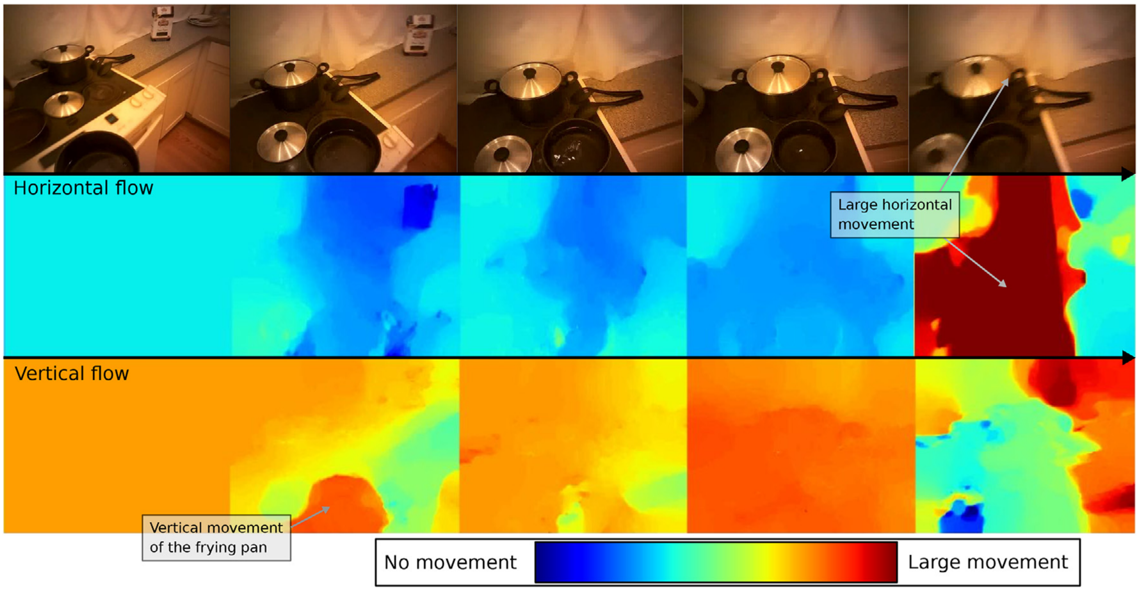
Horizontal flow
Large horizontal movement
Vertical flow
Vertical movement of the frying pan
No movement
Large movement

Fig. 5. Ego-motion example in the EGTEA Gaze + dataset. The top row shows subsampled RGB frames, the middle row has the horizontal optical flow component and the bottom row presents the vertical optical flow component.

video into action segments and used multisensory data to recognise actions. In another work, [225] explored the movement dynamics of some body parts, namely, the eye (gaze), head and hand movements. They integrated and modelled the action using Parallel Hidden Markov Model HMM: body parts were processed in parallel streams and integrated at the end. The benefits were that it allowed different sampling rates and different learnt topologies in each stream and that the noise of a stream was isolated without corrupting the others.

Ego-motion. A large part of the literature aims at capturing the ego-motion or the general motion generated from the head movement and employing it to recognise actions. [92] stated that there are two types of motion: instantaneous motion (directional component) and periodic motion (frequency component). In the first case, actions such as turning one's head have strong directional component while repetitive actions such as walking have strong periodic components. [191] aimed to recognise interactions (each one composed of the manipulation, the object and the location) using low resolution images and temporal templates or motion history images. These templates captured any motion detected in a video, using weights inversely proportional to the temporal distance from the frame in which the motion was detected to the current one. For each class, they computed a mean template and experimented with simple image matching, leading to finding out that normalised cross-correlation performed the best. To infer the location, objects, interactions, events and activities, they proposed a Dynamic Bayesian Network. [163] studied interaction-related actions, i.e. actions that involve interacting with the observer such as ”a person hugging the observer” or ”throwing objects to the observer”. They went one step beyond the work of [92] and explored multi-channel kernels to integrate global and local motion information. They also introduced a methodology that took into account the temporal structure of egocentric videos. Specifically, their global descriptors were histograms extracted from OF data and the local descriptors were composed of 3-D XYT data, i.e. computing salient motion in the video and summarising the gradient values of the detected motion patches. Moreover, they clusterised the motion descriptors and used the visual-word approach to represent the video.

Similar to the previous one, [137] made use of first-person dense trajectories in their motion pyramidal structure. The relative strengths of motion along the trajectories were then used to create various bag-of-words descriptors that were later combined into a single descriptor of the action. A non-linear Support Vector MachineSVM was fed with these descriptors to classify actions. [153] presented their Cumulative Displacement Curves, a method based on the assumption that, over a long period of time, the average displacement caused by the head rotation is practically zero. Therefore, they divided the frames with a fixed grid and accumulated the displacement up to a certain point within each cell (Cumulative Displacement or CD). Analysing trends in these displacements allowed them to focus on long-term actions and to avoid small perturbations due to the head motion. Moreover, for long-term trends, they convolved the CDs with a gaussian kernel to smooth them. For classification, they obtained various features and statistics computed from these motion vectors and applied an SVM. [178] contributed a new dataset called LENA and provided several experiments on it with various feature descriptors for trajectories, namely, Histogram of Oriented Gradients (HOG), Histogram of Optical Flow (HOF) and Motion Boundary Histogram (MBH); Fisher Vector encoding; Principal Component AnalysisPCA for dimensionality reduction; and a linear SVM for the classification step. [175] argued that a method for both short-term (take, put and so forth) and long-term actions (walking, driving and so on) did not exist and proposed a way to solve the task. Their solution was based on OF, in which they aimed to identify the dominant motion, i.e. motion generated by objects and the hands. They compensated the camera motion using a RANSAC-based homography [63] and applied an extension of a Histogram of Optical Flow HOF. Their classification goal was solely to infer if a video showed a short-term or a long-term action, but this could be applied in an EAR system.

[154] aimed to recognise long-term activities (helpful to segment long and unstructured videos) with a CNN architecture. They sampled segments of 4 overlapping seconds from videos, spatially divided each frame into a non-overlapping grid of size 32×32 and computed OF features from two corresponding grid cells in consecutive frames. This led to a cube of size $32 \times 32 \times 2$ (due to the x and y components of flow), which was used to create an stack of shape $32 \times 32 \times 120$ from the whole video, finally employed as input to a 3D CNN. [93] extracted features such as Histograms of Oriented Gradients, Motion Boundary Histograms and trajectories, com-

bined all of them and applied PCA to reduce the dimensionality before applying various classifiers: SVM, k-Nearest Neighbors K-NN and the combination of the previous two (SVMkNN). There are some works that provide new ways to arrange motion information, such as that of [164], that presented a new feature representation, called Pooled Time Series POT, based on the time series pooling of feature descriptors, particularly designed for motion information in egocentric videos. However, it could be applied to any feature descriptor such as HOF or CNN features. POT summarised the short- and long-term changes in the descriptors over time, it applied various temporal filters (set of time intervals) that were pooled with various operators and concatenated to obtain a single feature vector.

[177] proposed to combine several features such as dense trajectories (both forward and backward), HOG, HOF (with a compensated head motion), MBH and so on. Temporal pyramids were used to represent features to better capture slow and fast actions. Eventually, each feature vector was used to build a bag of words, which showed an improvement in the performance of the proposed solution. One important conclusion of this work was that, even though the hands and objects are important as egocentric cues, it is not necessary to explicitly segment them. Moreover, the features used in this work were also applied in third-person proposals, creating a bridge between both first- and third-person action recognition.

Combining eye movement and ego-motion. Others, such as the work of [142], combined both approaches and exploited the eye movement and the ego-motion; specifically, [142] analysed the combination of the eye movement taken using an *inside* looking camera and the ego-motion taken using an *outside* looking camera. For the first case, they presented their own encoding method while for the second one they used global OF values. [215] aimed to recognise actions in an unsupervised way in an office and a home environment: they employed encoding saccade information (from an inside camera) and OF encoding obtained from the video frames of an outside camera. They introduced two variants of Multi-Task Clustering, including data from different users in their clusters.

2.3. Hybrid approaches for Action Recognition

So far, the most promising approaches have been the object-driven ones. However, motion-driven methods may add more robustness and, thus, hybrid models are also proposed in the literature. Specially, the Deep Learning approaches dominate the literature due to their advantage in automatically extracting features from different information sources.

Two-stream architectures. A highly popularised approach in the DL community is the two-stream network presented in the work of [174], which employs both RGB and OF information as input. This model was first used for exocentric vision but it was later adapted for egocentric vision [95,189,117,187]. In addition, [174] observed that networks perform better when they do not need to learn to estimate the motion implicitly. Fig. 6 shows an example of a neural two-stream network that takes RGB and OF images as input. [121] proposed an improvement for the appearance stream, dividing it into two modules: one for hand segmentation and the other, that took the output of the first one, for object classification. The hand segmentation part segmented and localised hands, creating a gaussian bump in the region where hands were located (or the space between hands). That part was cropped and fed to the object classification part, which was trained for object recognition. Both this network and the motion stream had their own loss. At the end, both network outputs were concatenated and a FC layer with a softmax activation was used to classify actions. Hence, three different losses were used for training. The fusion of both branches was done with a concatenation operation; however, this fusion was later revisited in the work of [95], in which they contributed a long-term fusion pooling to aggregate the features coming from the two branches and they also analysed the effect of various pooling methods, namely, sum pooling, max pooling and gradient pooling. A combination of all them seemed to provide the best accuracy. An SVM was used as a classifier on top. Instead of employing the standard hard assignment to a single label, [211] used a soft assignment to various motion labels, e.g. {*open, hold, turn, rotate*} can denote the kind of motion used to open a jar or a bottle instead of just using the *open* label. This representation can generalise to unseen actions in which the motion pattern vary in some way, depending on the active object. Later, [209] presented a multi-label verb-only representation for action recognition and action retrieval. Their method allowed for an overlap of labels, removing the ambiguity of previous single label methods. They observed that a multi-verb approach with hard assignment was best suited for recognition tasks while an approach with soft-assignment was better for retrieval tasks.

As the two-stream approaches required an aggregation operation for each clip of the video, [188,189] proposed to extend the architecture in a CNN-RNN fashion using the Convolutional Long-Short Term Memory ConvLSTM network of [214] as the RNN. Moreover, one of the contributions of [189] was a spatial attention layer between the the CNN and the ConvLSTM in the spatial branch: they used Class Activation Maps (CAM) [235] from a pre-trained CNN to encode the video. Following the idea of [189] of adding attention mechanisms, [105] developed a NN that jointly classified actions and learnt attention map distributions using gaze information as supervision during the training. An attention map was sampled from this distribution an applied spatially and temporally to the frames in order to guide the action recognition. At test time, using the received input video, the network could infer both the gaze and the action. The idea of employing the gaze for an attention mechanism was also exploited in the work of [117], who implemented a two-stream network whose spatial branch had an attention mechanism on top. This was composed of a linear transformation supervised by a gaussian bump created from the gaze fixation point, i.e. a 2D gaussian centred in the point the subject of the action was staring at. After that, both branches had a bidirectional LSTM and, following it, they were fused.

[202] aimed at demonstrating that a two-stream approach with an LSTM was suitable for classifying egocentric actions without any egocentric feature. Moreover, they also showed that resizing images to adjust the size of objects to those of Imagenet's images could potentially improve the results. [187] hypothesised how a CNN-RNN structure could focus on ROI to better discriminate actions and, for that, they analysed the shortcomings of the LSTM and proposed their alternative Long Short-Term Attention LSTA module. This new RNN introduced a built-in spatial attention and a revised output gating. They deployed their LSTA in a two-stream architecture and also proposed, for the cross-modality fusion of RGB and OF, a novel control of the bias parameter of one of the modalities using the other one. [118] aimed to learn spatio-temporal attention features using human gaze as supervision. For that, they proposed a two-stream network, in which each of the streams included the spatio-temporal attention module (STAM) they contributed. This module included a 3D inception module and a 3D convolutional layer to predict an attention map. This map was combined with the original feature of the stream to create more informative features. [228] advocated for the use of Inertial Measurement Unit IMU for the motion classification instead of the OF arguing that the latter's computation was rather demanding. Instead, they created a layered-like approach. The classification of the motion was performed first by an LSTM. Depending on the predicted label, samples were categorised into different motion groups (for example, "standing", "walking" and

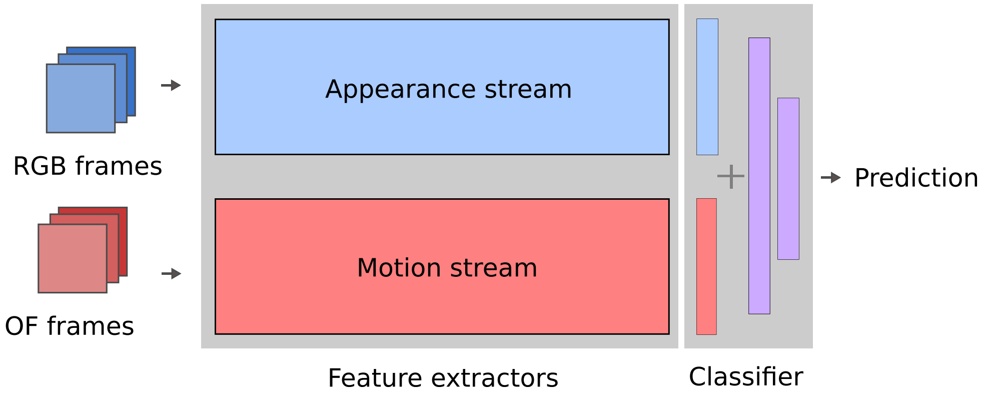

Fig. 6. Two-stream neural network. It is composed of a feature extractor based on convolutional networks and a classifier based on fully-connected layers.

so on). Within each group, various possible actions could be inferred, but the actions not associated to the motion of the group were discarded (e.g. actions in which it was impossible to be ”standing” are discarded if the sample is categorised as ”standing”). If the sample was contained within a group with only one action, then this action was predicted. In case there were various possibilities, the motion group was used as a prior for the other branch (the appearance branch), whose objective was to classify the sample among the possible actions of the group using visual features. To adapt the method for low and high frame-rate photo streams, two branches were used within the appearance stream. For a low frame-rate, a CNN was used and, for a high frame-rate, a CNN and an LSTM were employed. Similarly, [119] implemented a two-stream network in which one of the branches passed IMU data through an LSTM. The other branch employed a Recurrent Capsule Network (RecCapsNet) and a convlstm to extract spatio-temporal features. Then, both branches' features were fed to FC layers (separately), then combined by concatenation and, once again, the result was fed to a single FC layer. A softmax activation was finally used to provide an action probability distribution.

The application of the two-stream started becoming mainstream, as the architecture was being employed as a baseline. For example, [234] focused on hand-hygiene egocentric actions and proposed a method for first locating the action within an untrimmed video using low-cost hand mask and motion histogram features. In fact, once the action had been found, the classification was done using a two-stream network. [102] proposed a two-stream network in which one of the branches was composed of a self-attention based Graph Convolutional Network and the other one implemented a residual-connection enhanced bidirectional *Independently* RNN. [112] implemented a model that generated a Hierarchical Volumetric Representation (HVR) of the scene and employed a two-stream network. One branch took the visual input and processed it with an I3D network and the other one computed environment features. This allowed the model to sample possible action locations (learnt in a latent space) and to use those local 3D features for the action classification.

Multi-stream architectures. As two-stream architectures became popular, a natural extension of them arose including more branches and different input modalities. Each modality is assumed to be complementary to the rest and, thus, helpful to improve the classification of actions. Fig. 7 shows a general schema of a multi-stream architecture. [64], for the action anticipation task, used three complementary modalities of data: RGB (for appearance, using a Batch Normalised Inception), OF (for motion, using a TSN or TSN) and object features (confidence scores obtained from an object detector). They introduced their Modality ATTention (MATT) mechanism to fuse them, weighting each of them in an adaptive way to predict actions. The use of object detector information was again explored in the work of [207], who detected a shortcoming in the two-branched architecture (modelling appearance and motion): both failed to exploit local information as there was no position-aware information. In fact, just looking at the motion change or the collection of objects in the scene may not be enough for an annotator to understand the action, that is when position-aware features (referred to as privileged information) could help to drive the learning to action-relevant motion and objects. In addition, they contributed a Symbiotic Attention mechanism for Privileged information (SAP) that allowed for the communication of the three sources of information. A 3D CNN was used to process appearance and motion (outputting a single feature vector) while a Faster Region-based Convolutional Neural NetworkR-CNN was employed for the object features (extracted with RoIAlign). The motion and appearance features were individually fused with the detector's features and some learnt gate weights (from the opposite branch) were applied to them. One further attention step was applied using the opposite branch's features before obtaining the last feature vector for a branch. Both the verb and the noun were inferred separately and the predictions were combined and re-weighed by the training set's distribution to get the action prediction.

[195] leveraged depth information in their multi-stream deep neural network (MDNN), having two more branches fed with RGB and OF data. The contribution of this approach was that they aimed to preserve the distinctive characteristics of each stream and to explore the shareable information. That is, as features extracted from each stream were neither fully independent nor correlated, the fusion of these features lacked any meaning. Hence, they proposed a non-linear fusion strategy in which they mixed the shareable components and the distinctive components (both obtained with a non-linear mapping of the original features) with a weighted addition. In the loss function, apart from the categorical cross-entropy loss, they included two more terms: (i) a term to measure the correlation between the shareable terms (modelled with a Cauchy estimator) and (ii) a term to enforce the orthogonality constraint on both the shareable components and the distinctive ones. Moreover, they also included a hand module that was fed with the RGB frames. Within this module, a binary mask was generated to black out parts of the original RGB images that were later used for classification. In fact, the softmax output of this module was combined through a weighted fusion with the softmax of the original network.

In fact, multiple streams can arise in an intermediate step of the system, not only at the beginning, as in the case of the work of [74]. They presented a novel Mutual Context Network (MCN) that jointly learnt an action-dependent gaze prediction and a gaze-

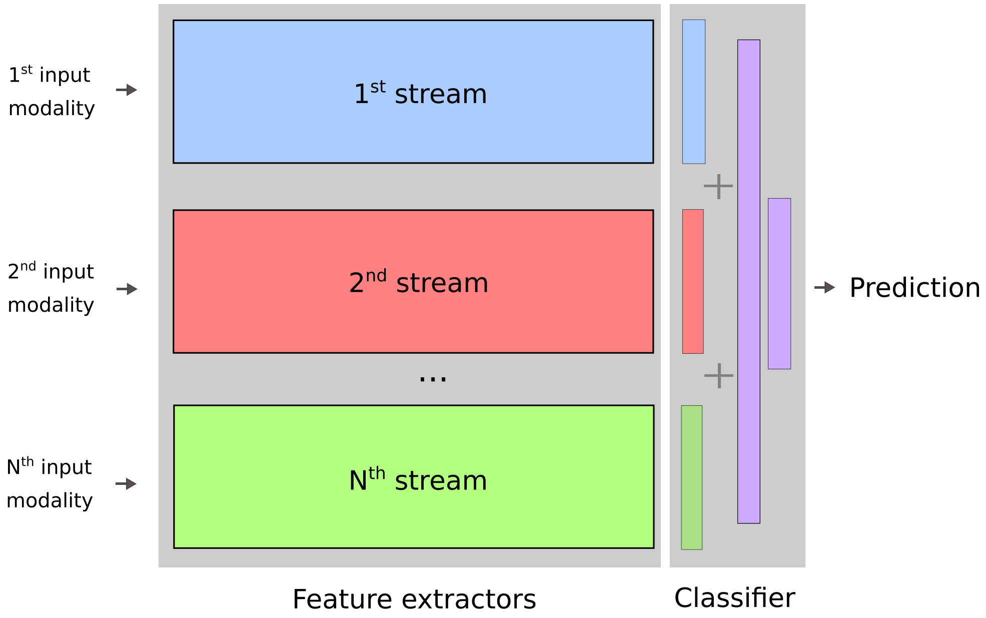

Fig. 7. Multi-stream neural network. Various data modalities are included as input, each one with its own feature extractor, and fused at some point (depending on the strategy, e.g. early, late or early + late). A classifier is used at the end for the prediction.

guided action prediction. RGB and OF frames were processed by an Inflated 3D Convolutional Neural Network I3D and fused by addition at the beginning to create the set of features F. Then, three parallel modules or branches could be found: (i) the saliency-based gaze prediction module that outputted a set of saliency map predictions (bottom-up attention); (ii) the action-based gaze prediction module that took action predictions, created kernels using them and then deconvolved F with these kernels to create another set of gaze prediction maps; and (iii) the gaze-guided action recognition module that took the generated combination of saliency and gaze maps, used them to pool the gaze and non-gaze regions of the input features, convolved them and obtained a probability distribution over the set of possible actions.

[83] proposed a branched method (for global and local characteristics) in which each branch had tree streams: for RGB, OF and warped OF, each one having as their backbone network a C3D [199]. The local branch, in contrast to the global one, was fed with the crops of the salient regions of the input frames, which were first aligned. To evaluate each stream's performance, they analysed early and late fusion strategies. To combine both branches, they came up with a cross-fusion strategy, in which pairs of different modalities of data (RGB and OF) and the same modality of data were mixed and a NN decided which features should have been attended. To generate the video-level prediction, they opted for the maximum-weighted-score voting, choosing the label with the highest weighted confidence score. [128] presented a multi-stream network that, apart from the RGB and OF data, included a mask image branch for the hand shape and position information. This was motivated by the fact that the recognition of actions may have suffered in different environments due to its poor generalisation ability. That is why they included hand information to solve this, outperforming the conventional methods and also confirming that their method had a higher robustness to different scenarios. [75] contributed a three-stream network that took RGB, OF and magnitude-orientations as input. An I3D network was used as the backbone network. For the high-level temporal modeling, a BiLSTM with attention was used to focus on the most important parts of videos.

Single-stream, multiple tasks. Some authors criticised the use of various streams and proposed a single stream approach that was trained for various tasks in order to improve the generalisation of the method. Fig. 8 illustrates a possible scheme of this type of network. [176] presented an initial single-stream Ego Convnet which consisted of two convolutional layers, each followed by a max pooling operation, a Rectified Linear Unit RELU non-linearity and local response normalization (LRN). At the head of the network, two FC layers were used. The inputs of the network were encoded egocentric cues: hands masks (as binary images, automatically segmented from the input images), camera motion (grayscale images representing the horizontal and vertical components separately, corrected using 2D homographies and RANSAC) and saliency maps (also as grayscale images). These features were taken from a set of L adjacent frames, creating a set of input features with a channel depth of $4 \times L$ (four features per frame). To handle the class imbalance, they proposed using the infogain multinomial logistic loss. [86] showed their multi-task learning approach to simultaneously learn verbs, objects, coordinates for hand locations and the gaze-based visual saliency. This allowed them to improve the generalisation ability of their model due to the network having to follow various objectives. In addition, the network was forced to exploit the secondary cues (hand locations and visual saliency) that, otherwise, may have been missed. They later proposed to extend this to multi-dataset multi-tasking in [89]. In each training batch, datasets were always mixed. A 3D CNN was used as backbone and task/dataset specific–heads were included on top.

[149] argued that appearance and motion information should have been jointly learned, as opposed to two-stream approaches with late fusion. They proposed their self-supervised first-person action recognition network (SparNet), a single-stream network that coupled both appearance and motion through a Motion Segmentation MS self-supervised task. This way, their training objective forced the network to learn the movement of objects. Their architecture was composed of a CNN-RNN structure (the RNN in this case was a convlstm), with a global average pooling and a FC classifier for action recognition. For the MS, the output of the CNN was fed to another convolutional layer and then to a FC layer. This output was compared with the ground truth using a per-pixel cross entropy loss. For the ground truth, Improved Dense Trajectories IDT [204] were computed. Then, each pixel was labelled as moving or static depending on whether movement had been detected for at least 10 frames in these features. The network was trained using both losses with equal relevance. The video

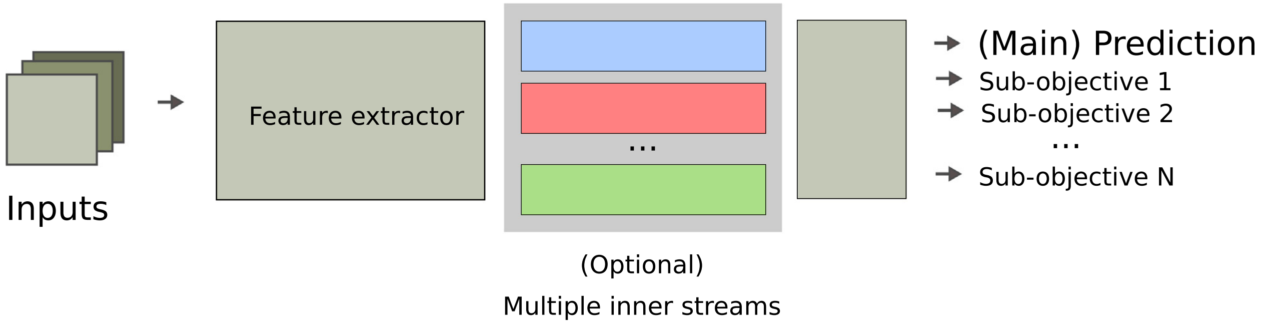

Fig. 8. General scheme of a single-stream neural network. Usually with a single input, optionally many inner branches, and possibly various sub-objectives (multi-tasking).

attention mechanism of [145] was used in the work of [146] in a single-stream architecture that branched at the end to predict the verb and object that composed an action. The backbone network used in the approach was the Temporal Shift Module (TSM) [107].

[188] proposed a CNN-ConvLSTM architecture that took RGB frames and the difference between consecutive frames as input to the network. Aiming at addressing the issue of the inherent ego-motion of egocentric videos, [115] introduced a single-stream architecture that started estimating the camera motion and then the temporal sequence was divided into various chunks using K-means clustering. For each chunk, the camera motion was compensated independently. Then, a slow-fast ResNet (only using its temporal branch) processed those chunks and a late fusion concatenation was used to gather them. Two softmax functions were used at the end to provide verb and object probability distributions. The action probability distribution was obtained with the outer product of the former two distributions and re-weighted using the training set distribution.

Although it was not specifically intended for an egocentric setting, [203] validated their approach using an egocentric dataset. They contributed a NN based on Capsule Networks in which they introduced Temporal Shift modules to allow the network to process temporal information without adding extra computation. An R*CNN [67] was used in the work of [34] to extract hands as primary regions and objects as secondary regions. This information was fed to a Hierarchical Long Short-Term Memory Network (HLSTM). In the first level, per-frame predictions were given within each shot of the video, being the last hidden step of the HLSTM the input for the shot-level LSTM. [150] presented their SparNet, a single-stream network that jointly learnt spatial and temporal features using an auxiliary pretext task. The objective of the latter was to estimate the motion associated to static images. This allowed the model to better focus on action-related features.

Combination of multiple features. Part of the literature advocates for the use of different features outside of the end-to-end deep learning systems. [182] employed ego-video and IMU data to tackle the action segmentation and recognition in cooking-related tasks. They explored in supervised and unsupervised settings the performance of Gaussian Mixture Model GMM, HMM and KNN. For high dimensional data, they observed that KNN worked better than the other two approaches. [171] combined gaze motion (using statistical features) and visual features (in a bag-of-features approach, more suitable for object and scene recognition). Both branches were independently processed and trained using an SVM. The results were combined to get an action prediction. [176] added two more streams to their single-stream approach with features learnt using deep networks: one that captured the appearance (being fed with RGB images) while the other one was in charge of the motion (taking a stack of OF images). Fisher Vectors were used to encode these features and an SVM was used for the classification. To fuse the three streams, weighted classifier scores were taken.

In a similar fashion to the two-stream network, [233] presented an architecture with two independent branches, being the motion features represented with a Fisher Vector encoding of various features, IDT among them, and being the appearance features encoded with CNN. They considered the use of early, early + late and late fusion approaches to train an SVM for action recognition. In the proposal of [79], the short-term motion was modelled with OF, while the long-term one was extracted using the POT representation proposed by [164] with different pooling operators. The appearance of a video was represented with its middle frame. With the concatenated features, an SVM was applied to classify actions. With a different objective compared to the previous authors, [155] contributed a novel Reinforcement Learning algorithm to save energy in wearable devices by trading off vision-based action recognition with a low power motion-based sensor. For the vision part, a CNN-RNN approach was used while an LSTM was employed to process the motion. The policy function approximator was implemented using an LSTM and trained using the Asynchronous Advantage Actor Critic (A3C) algorithm. [216] proposed a multi-task clustering using two methods: the earth movers distance multi-task clustering and the convex multi-task clustering. In addition, they had two test cases: home and office environments, in which different features were used as descriptors, as they argued that the feature selection was not important for their method. Therefore, for the home environment, they employed the object-centric features of the work of [148] while, for the office, both the eye motion and the head and body motion were considered together with OF images.

[134] employed image and accelerator features and aimed at predicting both the action and the energy expenditure (in terms of kilocalories). They concatenated both input features and used an LSTM to model their evolution in videos. [52] supplemented the inertial data from motion sensors (from a smart watch) with vision-based features and studied whether this addition could be helpful in settings in which sensor data alone was not enough to recognise actions. [239] introduced the gaze-informed ROI (GROI), the region where the gaze was fixated (and, supposedly, an area relevant to the task). Feature extraction was done in that area and then these were encoded and fed to a classifier. Specifically, the feature encoding consisted of (i) a normalisation step with RootSIFT (see the work of [10]); (ii) a dimensionality reduction with PCA; and (iii) a encoding step with Vectors of Locally Aggregated Descriptors (VLAD) [80] and Fisher Vectors [147]. Later, [238] proposed to enhance the local spatial and temporal feature extraction using saliency maps. [87] aimed to recognise hand-related actions based on the presence and position of detected ROI in the scene explicitly, without using visual features. For that, (i) they detected hands and tracked them across time and (ii) they also looked for objects that were relevant for actions. The problem

was modelled as the learning of the sequence of the detected spatial positions.

[226] had three different data streams: one extracting information from images (appearance stream), another one from IMU and GPS data (motion stream) and the last one adding external knowledge included by the user (external knowledge stream). In the appearance stream, a CNN processed images so that it extracted probability distributions from which several object probabilities could be obtained. The set of all the objects, as strings, were used to compute the basic belief assignment (BBA) for this stream. In the motion stream, several features such as the mean, standard deviation, correlation and so forth were extracted and an SVM was applied with those features as input. The third stream created a BBA using knowledge acquired from users: for some ranges of the time of the day, the possibility of actions occurring was provided by the user with some confidence values. Given an input timestamp, it was possible to acquire this time-based prior. To fuse all three streams, Dezert-Smarandache theory (DSmT) [51] based multi-source fusion was used. In a later approach, [227] constructed their framework that used the DSmT around three modalities: location, motion and vision data from a wearable hybrid sensor system. [35] argued that, in contrast to the classical ensemble of a CNN and a Random Deep Forest RDF, their late-ensemble technique could reduce the overfitting. The softmax output of the CNN, contextual metadata (day of week and time of day) and the global image information (histograms of colour) were given as features to the RDF, obtaining a better combination of both classifiers instead of the naive ensemble. This translated into a 5% accuracy improvement in their dataset of 19 activities.

The Visual Rhythm Texture Descriptor (VRTD), obtained as texture features over visual rhythms, was used in the work of [131]. To build the visual rhythms, they used various types of images: RGB, OF and motion boundaries. In a first instance, they built them using a whole video, and later from patches of videos. Aiming to obtain an efficient technique to classify actions, [94] applied a median filter, followed by the watershed segmentation algorithm. Features were extracted using the Histogram of Oriented Gradients, colours and the GiST descriptor. The combination of the features was passed through a genetic algorithm, reducing its size. The classification was performed using an SVM and a Random Forest.

Knowledge graphs. A visual example of a knowledge graph for egocentric actions can be seen in Fig. 9, although the structure has to be adapted depending on the proposed solution. For instance, [212] presented SEMBED (SEMantic emBEDding), an approach to embed egocentric object interactions within a semantic-visual graph (SVG). Their aim was to estimate the probability distribution over the potential semantic labels. The verb annotations of the interactions were unbounded (many verbs could describe the same interactions), thus, embracing ambiguity in order to capture the semantic relationships and the visual similarities of motion and appearance features. The SVG was built from the training set, in which (i) videos that were semantically linked (they had the exact same verb label) were also linked in the SVG (first type of edge), (ii) nodes that were visually similar, yet semantically different, were linked (second type of edge) and (iii) edge weights corresponded to the normalised visual similarity with neighbouring nodes. With the SVG created, they employed the Markov Walk (MW) of [57] and, taking the z nearest neighbours and t steps, they found the probability distribution over the possible labels. Similar to the work of [212,165] leveraged graphs for the representation. First, they extracted features from individual frames (with the whole frame and also dividing it into four bins) employing a Pyramidal Histogram of Oriented Gradients (PHOG) [25]. They also used the centre-surround model proposed by [166] to capture the ego-motion. With those features, they built a weighted Video Similarity Graph (VSG) in which nodes represented frames of a video and edges were measurements of the similarity between frames. To infer actions, they used a weakly supervised approach in which only 5% of the frames were required to be labelled and the remaining ones had their label predicted using a Random Walk execution.

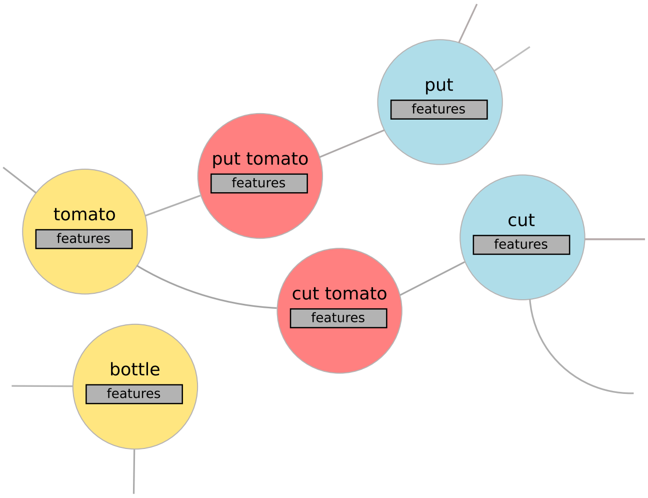

Fig. 9. Example of a knowledge graph for EAR. Each node may have one or various features associated, learnable or not. For visualisation purposes, colours are assigned to nodes depending on the information they contain: verbs, objects or actions.

Hand-based recognition. Unlike previous approaches in which only the shape and pose of hands or their interaction with objects were analysed, the following approaches exploit both their appearance and motion or leverage hand information combined with other features. Starting from the work of [236], in which they trained a Deconvolutional Neural Network DCNN to infer a pixel-to-pixel segmentation of hands from weakly and strongly supervised data, i.e. massive bounding box annotations and fully annotated segmentation masks, respectively. The network was trained using a novel Expectation–Maximisation (EM) like learning framework. For further improvement, the hand segmentation masks were paired with motion maps (OF) and object feature maps (the top-5 strongest object feature maps) with another DCNN to detect active object regions or the interactional foregrounds. Concerning the objective, both a softmax for object class probabilities and a bounding box regressor were used. They trained two such detectors, one for active objects and another one for passive objects, and histograms were extracted from these representations to model the appearance of actions. For the motion, IDT were extracted from the global image (global motion) and from the active object region (local motion). Combining all these features and applying an SVM classifier, they inferred the action. [66] focused on hand-related actions and introduced a dataset of 3D hand poses. They presented an extensive experimental evaluation of RGB-D and pose-based action recognition covering 18 baselines (state-of-the-art approaches). They concluded that the hand pose cue is of major help in the EAR field. [28] presented their work on desktop action recognition, i.e. actions performed while humans are sitting at a desk, and contributed a new dataset for the task. Specifically, they focused on actions involving the manipulation of objects. After extracting various features from hands (e.g. hand shape, position and motion, inner hand OF and so on), they analysed their discriminative potential to classify actions. The conclusion they extracted was that hand shape and motion were decisive to recognise desktop actions. .

Finally, there are some approaches used for challenges that are often based on ensemble models, i.e. models such as the ones presented by [185]. In their case, they employed an ensemble of CNN-LSTA [187] and also of Hierarchical Feature Aggregation (HF)-TSN [186].

2.4. Other approaches

The remaining of the literature opted for going in other research directions within the EAR field. For example, including other modalities of data or changing some standards of the field.

Sound modality. Sound has not been extensively researched in the literature of the EAR field, having few datasets with both RGB frames and sound. However, it is a promising solution for some actions in which the visual appearance and motion are not enough (see Fig. 10). For example, [9] suggested combining audio-visual features with multi-kernel learning MKL and multi-kernel boosting MKBoost. Specifically, MKL was used to learn the weights of different features, kernels and their parameters using the training set. Concerning the features, the authors thought about employing complementary features from different modalities: from videos, they extracted global features using the Grid OF-based Features (GOFF), Vision-based inertial features (VIF) (both from the work of [3]) and Log-Covariance (Log-C) features [70]; local features from videos leveraging Cuboids [99]; and, for audio features, they used Mel-frequency cepstral coefficients (MFCCs) [47]. For the classification part, apart from the MKL and the MKBOOST, they employed an SVM.

With the launch of the EPIC Kitchens dataset [43], we can expect more works exploring these areas. That is the case of [31], as they explored the usefulness of sound. Their system processed an audio spectrogram of the first four seconds of the video (fully covering more than the 80% of the videos with that length) using a VGG-11 network. For the spectrogram, they employed a short-time Fourier transform to filter human voices and focus on noises. In another work, [32] created a three-stream network, taking spatial, temporal and audio features and combining them using a late-fusion approach. Data was sparsely sampled from the video by dividing it into K uniformly distributed segments. Audio was again represented by its spectrogram. [90] proposed a new architecture for multi-modal temporal-binding (the combination of modalities within a range of temporal offsets) of RGB, OF and audio. In contrast to previous works, the modality fusion was performed before the temporal aggregation and per Temporal Binding Window (TBW), defined in the paper as ”a range of temporal offsets within which an individual is able to perceptually bind inputs across sensory modalities”. The width of the TBW was dependant on the length of the video (not to bias the network towards short or long actions) and K such windows were sampled from videos. For each window, a three-stream network (with three Batch Normalised Inceptions as backbone networks) processed each modality. All modalities were fused at a mid-level and a prediction was given for the window. For the video level prediction, the predictions of all windows were averaged. Baesd on their experiments, they demonstrated that audio is complementary to the appearance and motion representation of the RGB and OF inputs. [151] contributed a new loss function for multi-modal models (combining audio and visual information) that aligned better the contribution of both modalities.

Task reformulation. Some researchers provided novel viewpoints, aiming at changing EAR conventions. For instance, [130] explored the inconsistencies in the annotation of the temporal boundaries of object interactions within and across annotators and datasets. They argued that this phenomenon is given mainly due to the limited understanding of the different phases of actions and proposed annotating these boundaries based on Rubicon Boundaries from the Rubicon Model of Action Phases presented by [68]. [213] claimed that there were three incorrect assumptions driving the recognition of object interaction as a standard one-vs-all classification problem: (i) classes were self-contained and had strict boundaries, yet when the number of classes increased, these tended to overlap; (ii) sequences could be split into segments containing only one object interaction, but multiple interactions could be given at the same time; and (iii) whether a verb could be used to label actions was a binary decision, whereas different annotators could label the same action in different ways. Therefore, in their work, they proposed to reformulate the recognition of object interactions as a multi-label classification, obtaining better results compared with the single-label approach.

To date, egocentric models have been pre-trained with exocentric data (as there are larger datasets for that). Recently, [104] criticised that approach, arguing that it created a major domain mismatch as exocentric models ignore egocentric vision properties. In fact, as stated in [172], pre-training only with exocentric data led to worse results compared to training with egocentric data. Other methods that map exocentric to egocentric data still required parallel datasets, which were difficult to gather. However, [104] proposed to use egocentric pre-defined tasks to steer the pre-training of an egocentric model with exocentric data. These tasks were (i) discerning between exocentric and egocentric videos, (ii) finding interactive objects and (iii) discovering hand-interaction regions. The ground-truth signals came from the output of other models, being these training signals (and their respective classifiers) similar to distillation models.

Privacy. Egocentric video data may be seen as privacy-sensitive due to the possible threats they represent. [198] analysed some of them: (i) recognising the wearer of the camera, (ii) detecting that two videos have been recorded by the same person with low error, (iii) extracting the gait signature of the wearer of the camera and (iv) matching that gait with the one of a third-person view to recognise the person. That is why the research on this topic of the EAR field is becoming crucial with the increase of available egocentric videos. [54] aimed at classifying actions while preserving the privacy of bystanders. In contrast to other approaches that selectively filtered regions of images, they proposed to blur images without suffering from a significant drop in the performance. In fact, they carried out a quantitative analysis with 640 users to assess the trade-off between privacy and performance, reaching the conclusion that degrading egocentric images led to a more positive perception of privacy, increasing the willingness of users to be captured. They employed CNN for the EAR and analysed the effect of different levels of blurring and the obtained performance. [183] argued that, within eyewear devices, the first-person camera in charge of mapping the user’s gaze to the visual scene can pose a security threat. To solve that, they proposed *PrivacEye*, a method to detect privacy-sensitive scenes and to automatically disable the eye tracker’s egocentric camera. To analyse the situation, they employed deep features combined with eye movement features. To activate the camera, eye movements alone were considered to measure the level of privacy of new situations. It is also possible that some people may want to check whether a person has been recorded or not and, for that, a query is required. In order to ease that type of query, [152] proposed a method to add a unique signature to videos of a given user using the patterns of the head motion. Unlike other methods, it was volatile and could only be used for a particular place and time.

Data sampling. Videos of actions are usually composed of frames that are relevant to actions and others that may be considered noise or that are simply identical to the adjacent ones. In fact, many methods uniformly sample a few frames of videos due to the fixed-length constraint of many DL architectures and the memory limitations, without taking into account the importance of those specific frames for the task. However, there are works such as that of [218] in which a plug-and-play module for any EAR solution was

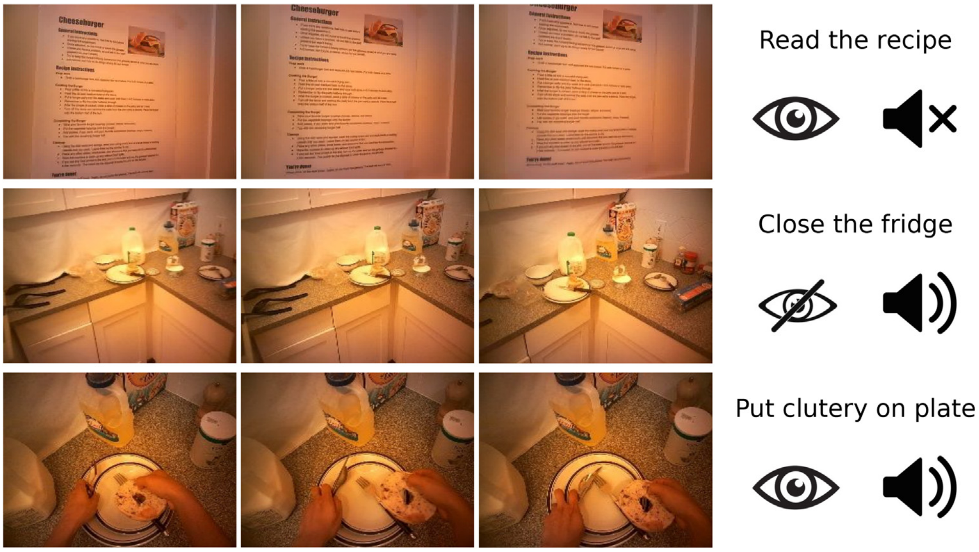

Fig. 10. Example of the importance of sound for the recognition of egocentric actions. Samples of actions that cannot be seen such as closing the fridge while staring at other part of the kitchen are impossible to predict without complementary information such as sound.

proposed. Given unprocessed noisy video clips, the module outputted a few informative frames. The basis of the module was the combination of the sampler and the evaluator modules. The first one was in charge of providing candidate frames while the second had to evaluate the selection. As there was no ground truth, the evaluation was done conditioned on the result of other EAR algorithms. That is, a selection of frames that obtained a good result in those algorithms was considered a good choice. This signal was used to train the sampler module in a student–teacher fashion.

3. Alternative Learning Paradigms for Egocentric Action Recognition

So far, all the works reviewed have followed a supervised setting approach, i.e. a labelled dataset is required for the training and evaluation stages. This means that each class requires several samples for the learning phase. As the number of executable actions in real-world scenarios is very high, it seems unrealistic to collect and annotate samples for any kind of action. Hence, alternative paradigms to supervised learning become very interesting to implement EAR systems in real-world applications.

These include the few-shot and zero-shot learning approaches (see Fig. 11 for an illustration), in which a few samples or no sample at all is required for the training and the system has to generalise to unseen samples and classes with the aid of prior knowledge or by exploiting the characteristics of the data. Furthermore, the unsupervised setting can be considered within this category, not requiring labels for the training stage. This section presents a review of egocentric vision-based works with the aforementioned characteristics.

Following the literature on EAR of Section 2, a popular strategy to classify actions, the two-stream approach, is inherently built to decompose the two main drivers or components of actions, the verb and the noun. This allows the two-stream model to acquire knowledge separately about verbs and objects and to leverage this information for the classification of actions. In fact, the idea of fusing verbs and objects to infer new combinations was already introduced in 2017 by [233]. They employed the Fisher Vector encoding of IDT and the Histogram of Oriented Gradients for the verb and visual CNN features for the noun. They argued that the specialised features they considered for each decomposed concept were the key to success in zero-shot tasks. In addition, they analysed various fusion methods among *early*, *late* and *early + late* stage fusion. [6] used a Myo armband sensor data and an MLP for verbs and a GoogleNet that took crops of video frames for nouns. The latter crops were extracted from the gaze region. Any new action composed from the combination of the learnt verbs and objects could be inferred using their system. Similarly, [141] used a DL two-stream approach, having both branches separated, predicting the verb and the object separately. The outer product of the probability distributions of verbs and objects was employed to generate an action probability distribution. Within this setting, they contributed a way to re-weight that distribution with external knowledge coming from text corpora. This way, action frequencies from real-world corpora could be included in the system, allowing for better predictions. The improvement came from (i) the removal of non-existing actions and (ii) the re-weighting of not common actions and frequent actions, making it a suitable approach for real-world applications.

[210] had as objective the creation of a representation suitable for cross-modal search, i.e. video-to-text or text-to-video queries, in which the query and the target were from different modalities. They proposed two Multi-Modal Embedding Networks (MMEN) that embedded video features and text features into the same space, one for verbs and the other one for nouns. Specifically, the same video features were sent to the two MMEN while, in the case of the text, the verb and the set of nouns were first extracted and then sent to their corresponding modules. The representations obtained from the MMENs were further encoded to get features for verbs and also for objects. The network leveraged intra- and cross-modality losses to preserve the neighbourhood structure within verb and noun spaces and to ensure that the representation of a query and a relevant item for that query from a different modality were closer than the representation of the query and a non-relevant item. With their approach, they aimed to create verb and noun spaces that were suitable for actions, e.g. for a given verb, independently of the objects accompanying it, the representation should have been able to capture its essence.

To join both zero-shot learning and few-shot learning, [169] proposed the *cross-modal few-shot generalisation setting*. First, they embedded videos using CNN features and deep metric learning so that they could create an embedding space with discriminative

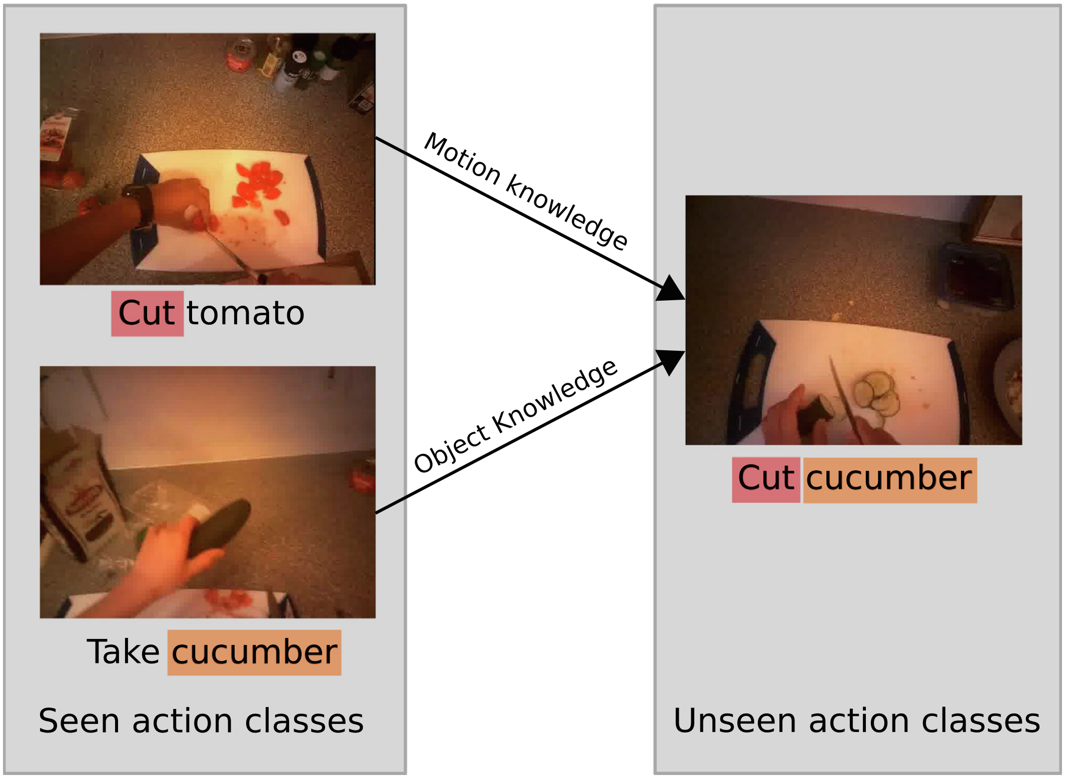

Fig. 11. ZSEAR approach using motion and object recognition to predict unseen action classes. Frames taken from the EGTEA Gaze + dataset.

features. Then, they aimed to align word-embedding features of class labels with video embedding features with a combined loss term. In fact, word embeddings were passed through another embedding to match the dimensionality of video features and to become class-agnostic. This enabled the cross-modality metric-learning between the video embedding and the twice-embedded class labels. They even proposed a new split of EPIC-Kitchens to evaluate their approach. [123] built a model, the Spatial–Temporal Interaction Network (STIN), that could reason about the relations between objects and the agent of the action (hands). Using an object detector and a tracker, they created object-graph representations that included hands and constituent object nodes. With this, spatio-temporal reasoning was performed using the BB, aiming at understanding how the relation between the subject and the objects change over time. This included the modelling of the transformation of the geometric relations of objects.

So far, zero-shot approaches had relied on data to train a latent space. However, going one step beyond, it is possible to have a zero-shot application with zero supervision as in the case of [2]. They proposed one of the first works on open-world EAR. They employed commonsense knowledge from ConceptNet [109,181] to solve the demand for training data quality and quantity. By means of relations between several concepts, they were able to infer the action with the highest probability. This methodology was grounded on the Pattern Theory formalism of [69,1].

Apart from the strategy of dividing the learning of the verb and the noun, other approaches could be employed for the zero-shot task. The approaches followed by [212,211] (discussed in Section 2), as the authors mentioned, could be employed for zero-shot EAR, although that was not their original purpose.

In their fully unsupervised system, [23] used the two-stream autoencoder approach in which both streams were aggregated by late fusion. Then, a cascade of LSTM autoencoders were used, each one with a different temporal resolution and offset. At the end, K-means clustering was applied for the learning of groups. The results showed that their method led to the discovery of semantically meaningful action groups. According to the authors, this unsupervised methodology may had its applications for privacy-sensitive data as, in contrast to supervised methods, there were no pre-defined labels. [45] had the aim of discovering task-relevant objects in an unsupervised way. To focus the attention on objects, they employed appearance, position and motion features and gaze fixations. Using clustered motion features, they were able to predict modes of interaction (movement patterns, as in the case of actions). [92] used sparse OF vectors as their motion features to encode the ego-motion in a fully unsupervised scenario. They found out that this feature itself was a strong descriptor for actions related to sports.

In the system proposed by [230], the approach to represent the video was fully unsupervised and dictionary-free. They pooled features from sub-intervals using a temporal feature pooling function. These were temporally aligned and a recursive Fast Fourier Transform was applied on a pyramidal temporal structure. Although this procedure did not require labels, they need them to train the action classifier. Similarly, [156] created a video descriptor without supervision. First, they employed the trajectory-pooled deep-convolutional descriptor (comprising both spatial and temporal information). Sub-actions were adaptatively localised in time, then the features were aggregated by temporal pooling and rank pooling was used to determine the temporal evolution of videos. The Hilbert-Huang transform was finally applied to obtain the final descriptor. Nevertheless, the training of the SVM classifier required labelled data.

4. Egocentric Video Datasets

Since 2009, several datasets for EAR have been proposed in the literature, being the main resource for researchers to evaluate their proposals. Table 2 summarises the most relevant datasets of the literature and their characteristics. Specifically, we show whether they contain BB annotations, their publication year, the number of action clips (instances used for training and evaluating machine learning models) and the number of action, verb, and object classes. In the case of Charades-Ego, the dataset is partially egocentric, having part of its content filled with third-person videos. The annotation of actions in all the presented datasets consists of a verb and a set of nouns, creating an action when combined. That may be one of the reasons why popular methods such as the two-stream network approach have adapted well to the egocentric vision, i.e. as the motion and the object features can be decomposed, there are labels to train two separated classifiers and/or to jointly train two branches.

There are other egocentric datasets that are not suitable for EAR due to their intrinsic purpose (the task, a focus on activities or interactions rather than on actions and so on) and/or due to the lack of labels. We only present datasets that are, to the best of our knowledge, publicly available.

The **University of Texas at Austin Egocentric (UT Ego)** dataset [101] is composed of 4 videos (10 in total, but only 4 public) with a length of 3-5 h and recorded in an uncontrolled setting. The videos capture a variety of activities such as eating, shopping, attending a lecture, driving and cooking.

The **JPL First-Person Interaction dataset (JPL-Interaction dataset)** [163] is an egocentric dataset composed of activities of interactions (e.g. shake hands, hug or punch) with the wearer of the camera.

The **NUSFPID - NUS First Person Interaction Dataset** [137] is composed of 8 interactions in both egocentric and exocentric perspectives.

The **Stereo Ego-Motion Dataset**[1] contains videos that show a person walking around objects or animals under no special circumstances. The first two objects, a car and a chair, show no motion whereas the cats and dogs of the next two cases have strong articulated motion.

The **LENA (Life-logging Ego-ceNtric Activities)** [178] includes 13 activities recorded with the Google Glass such as read, watch videos, walk straight and so forth.

The **EGO-GROUP and EGO–HPE datasets** [8,7] are aimed for ego-vision applications: social group detection and head pose estimation, respectively.

The **Egocentric Dataset of the University of Barcelona - Segmentation (EDUB-Seg)** [194,53] is a dataset acquired with Narrative Clip, taking a picture every 30 s, containing 18,735 frames from seven users. For the sake of variety, each user recorded their actions in different scenarios: attending a conference, on holiday, during the weekend and during the week. It contains annotations to segment events in time under the condition that those events can be inferred using visual features, i.e. there is enough visual information in that segment to infer the event.

The **Multimodal Egocentric Activity Dataset** [179] contains 20 activities, having each activity short clips of up to 15 s. For example, it includes writing sentences, organising files and running. Furthermore, images are accompanied by sensor signals.

The UTokyo collection of datasets, composed of **UTokyo Paired Ego-Video (PEV)** dataset [221], the **UTokyo Navigation** dataset [222] and the **UTokyo Ego-Surf** dataset [220,223], are a family of datasets developed by the University of Tokyo. The first one contains videos from dyadic (between two persons) conversations, capturing interactions. The second one has videos of people walking around a university campus to visit landmarks, but the videos per se are not available (due to privacy concerns), yet already extracted features can be obtained. The third one contains 8 groups of videos recorded synchronously during face-to-face conversations.

The **EgoFoodPlaces** dataset [168] involves 12 users in their daily food-related activities. The classes for this dataset are localisations where the activities are held.

The **Dataset of Multimodal Semantic Egocentric Videos (DoMSEV)** [173] is a 80-h dataset containing information about the scenes that were being recorded. This includes the type of scene (indoor, urban, crowded environment or nature), the activity performed (walking, running, standing, browsing, driving, biking, eating, cooking, eating, observing, in conversation, playing or shopping), if there was something special that caught the attention of the observer and also interactions with some objects.

The **EGOcentric–Cultural Heritage dataset (EGO-CH)** [157] is a dataset for cultural sites' visitors behaviour understanding. The dataset includes 60 videos, 26 environments and over 200 Point of InterestPOI. Moreover, it is annotated with temporal labels including the location of the visitor and the observed POI, a BB annotation around POI and the survey associated to each video filled by the visitor. The dataset is aimed at providing 4 tasks: room-based localisation, POI or object recognition, object retrieval and survey prediction.

The **EgoK360** dataset [22] is an egocentric 360° video analysis dataset. It contains several activities with actions within them, being quite challenging due to the distortion and the wide field of view.

5. Applications of Egocentric Video Analysis

The analysis of egocentric videos serve for several purposes. Although the datasets shown in Section 4 may provide some hints on the kind of applications that can be given, we review the applications found in the literature. Given that the field is still relatively new, many new applications may arise in the future.

Ambient Assisted Living. One of the current main challenges for the public administration is to promote active and healthy ageing for as long as possible. Achieving it would pose positive consequences for the society and the socio-sanitary services, such as reducing the costs from medicines and other treatments. The latter expenses are becoming more and more worrying with the ageing of society. For example, Spain dedicated the 9.8% of its GDP to elderly care in 2014[2]. Given that reports estimate that the world's older population is going to duplicate by 2050[3], the magnitude of the problem may become unmanageable. Due to this, public administrations are investing in research projects which may help alleviating or avoiding this problem in the future, creating an active and healthy older population. Although the research projects using computer vision approaches have mainly focused on the third-person vision [29], nowadays the use of wearable systems is more abundant [37,39]. [126] proposed a system to support clinicians for the care of dementia patients and [231] used smart glasses with a first-person system that could warn people with cognitive impairments of dangerous situations. But not only is it useful for supporting health professionals, aiding caregivers is also a potential application of first-person systems. [136] described a method leveraging a first-person camera to evaluate the tender dementia-care technique. They obtained the 3D facial distance, pose and eye-contact states between caregivers and receivers and performed statistical analysis to assess the caregiver's skills. These types of approaches can be grouped in the AAL paradigm, which promotes the use of modern ICT technolo-

[1] https://lmb.informatik.uni-freiburg.de/resources/datasets/StereoEgomotion/

[2] https://www.imserso.es/InterPresent2/groups/imserso/documents/binario/112017001_informe-2016-persona.pdf

[3] https://www.nih.gov/news-events/news-releases/worlds-older-population-grows-dramatically

Table 2
Summary of the most relevant egocentric action recognition datasets ordered by their publication year. *Only for 4 objects. **Manually computed, there is no official number.

Dataset	Year	Object BB?	Action clips	Action classes	Verb classes	Object classes
Intel Egocentric Vision [162]	2009	×	922	42	42	42
CMU [48]	2009	×	516	31	16	33
ADL [148]	2012	✓	436	32	24	42
GTEA Gaze [60]	2012	×	511	94	10	33
GTEA Gaze+ [60]	2012	×	3,371	44	9	29
BEOID [44]	2014	×	742	34	15	20
EGTEA Gaze+ [105]	2018	×	10,325	106	19	53
Charades-Ego [172]	2018	×	30,516	157	33	36
First-Person Hand Action (FPHA)[66]	2018	✓*	1,175	45	27	26
EPIC-Kitchens [43]	2018	✓	50,547	2,747	93	272
EPIC-Tent [78]	2019	×	921	11	6	9
EPIC-Kitchens-100 [41]	2020	✓	89,979	4,025	97	300
Meccano [158]	2021	✓	8,857	61	12	20
H20 [96]	2021	✓	184**	36	11	8

gies to assist the elderly in their ADL. The main objective of the AAL is to avoid the dependence of elderly people on other people in their daily living activities. In particular, EAR becomes a key enabler for AAL approaches.

Hand recognition. Hands are of special importance for humans, allowing us to interact with objects and environments. As a consequence, the daily life of a person with impaired or reduced hand functionality may be drastically affected and the recovery of hands should be a priority [17]. Even though health related issues may be grouped within the AAL field, this is a special case for egocentric videos. As seen throughout the document, hands play a key role in egocentric actions and, therefore, this use case is separated from the aforementioned. The recognition of hands includes their localisation in the space, their segmentation, their identification (left or right) and the pose estimation (fingertips, for example). From this information, it is possible to remotely assess the functioning of hands. Another application of the recognition of hands is to be able to understand children's visual attention [15], as it seems that parents' hands drive their attention.

The augmented reality (AR) and the virtual reality (VR) technologies, which are becoming more popular, require the egocentric recognition of hands for natural user interfaces that need to know the position and movements of the hands [17]. For example, [71] proposed an interface to move 3D objects using hands and, thus, they implemented a virtual hand interaction technique. In the work of [77], they aimed at simultaneously detecting click actions and estimating occluded fingertip positions. [197] introduced a solution to allow users to inspect 3D objects using their hands, requiring to estimate the 6D palm pose and the gesture performed. [76] focused on the rotation of 3D objects. By performing the "holding" gesture, virtual objects could be summoned into the palm, allowing another gesture to trigger their function. [26] argued that it was difficult to correctly detect hands in cluttered backgrounds with varying illuminations and, hence, they proposed a solution for indoor and outdoor environments.

Social Interaction Analysis. People's social behaviour can be analysed and classified using egocentric videos. [58], for example, aimed at detecting social interactions in a day-long activity. First, the context provided by faces was obtained and used to estimate the location that was being attended. Second, based on the patterns of people, roles were assigned to them. By analysing temporal patterns of roles and locations, they were able to detect and recognise social interactions. They also explored the inclusion of head movement as an extra feature. [163] focused on interactions with the wearer of the camera, including both friendly and aggressive interactions. [217] had as objective the extraction of interaction features (IF), features that are common between interactions. These are mainly composed of physical information of head, body languages and emotional expression. An HMM was used to model the sequence.

When considering a group, based on the concept of the F-formation [91,8] tracked through a video sequence a group of people, estimating their head pose and 3D location, to predict the affinity of a two people in the scene. Again following the F-formation concept, [5] aimed at detecting when a social interaction was given.

Pedestrian movement anticipation. Using an egocentric camera, it is possible to analyse the patterns of movements of the pedestrians in front of the wearer and anticipate their movements. This may even have applications for autonomous vehicles for pedestrian safety [184,36,108].

Nutritional behaviour analysis. The analysis of egocentric videos could be interesting when we are performing actions related to eating. This could lead to analyse our nutritional behaviours, diet and lifestyle as proposed by [82]. Moreover, as mentioned by [168], the food intake and its duration are of major relevance to protect against diseases. That is why they developed a model to detect the food intake events during the day. [24] aimed at both localising and recognising food simultaneously.

6. Conclusions

Throughout this survey four main distinct ways to categorise the EAR proposals have been introduced: those solutions based on objects or the appearance, the ones employing motion as their main driver, hybrid approaches that consider both the appearance, and the motion and other approaches (still not that abundant) that consider more modalities like the sound or contribute on other topics of the field. Moreover, alternative learning paradigms for the EAR and potential applications of this research field have been summarised.

Although the EAR field advances are still far from being completely transferable to real-world applications, many steps towards that goal have been taken. There are larger and larger datasets to train deeper and deeper models, allowing to obtain models with better performance and generalisation ability. The range of egocentric actions that are considered in the literature is also increasing with the evolution of datasets, considering rarer or more difficult events. But this advance does not only come from the data, new important modalities of data such as sound, crucial for actions that are recognised only by that feature or in which this may play an important role, are being included in the literature and the datasets.

6.1. Future work

Many ideas to tackle the EAR have been proposed throughout this document. Many of them are still taking their first steps while others have a larger trajectory. Nonetheless, their potential is shown when comparing different solutions using standard benchmarking datasets. Between them, the following research lines should be taken into account:

- The use of the sound seems promising (see Section 2.4) despite models using it can not compare directly with methods that do not employ it. However, apart from the simple comparison between models to achieve the best possible accuracy, solutions including sound have appeared to provide a solution for new action classes that did not have an easy way to be distinguished. For example, consider an action that is not seen by the camera but can be heard, such as a fridge closing while the camera wearer is turning back (performing the action while looking away from the fridge). By including sound it would now be possible to recognise this action. Clever ways to fuse sound information with RGB, OF and so on need to be proposed to push the real-world recognition of egocentric actions.
- The use of complementary information, apart from the sound, to the traditional RGB and OF setting. For example, the object-centric features extracted from RPN modules in hybrid approaches. This seems to lead to competitive results [206] while exploiting one of the most important features in the egocentric vision: objects. There are also works including hand information. It is possible that including hands just like objects are could lead to an improvement due to the inclusion of hands' shape, trajectory and so on, as some actions can only by distinguished by discerning those cues. As an example, imagine trying to distinguish turning on or off a burner. Visually, both actions look the same, there is only a variation in the motion of the hands. There should also be more research including left and right hand variations, as so far the field has focused on right-handed actions when only one hand is necessary.
- Creating attention mechanisms that are specific to the egocentric setting. There may be a suitable way to improve the results and the information captured by models without making networks bigger and deeper. In fact, the scaling of networks towards bigger and bigger versions is reaching hardware limitations and, thus, alternative ways to increase the performance are even more necessary.
- Multi-tasking approaches such as [121,86,89] have obtained the best results among many EAR solutions using the GTEA Gaze + and EGTEA Gaze + datasets. This type of approach may be a key enabler of the breaking of the performance barrier that can be achieved with single-objective methods. This includes, for example, aiming at learning egocentric features and/or verb, object and action labels at the same time, following the literature of the EAR field. If more than a single objective is considered, the results obtained by these works may suggest that a stronger generalisation is achieved.
- Alternative paradigms for learning egocentric actions in order to be able to apply an EAR system in the real-world should also be considered, including the zero-, one- and few-shot learning. These require none, one or few samples, respectively, related to the task and they usually extract the information required for the learning (if any) from prior knowledge or auxiliary datasets. They may also exploit characteristic of the data (hands or objects) or use unsupervised algorithms such as clustering, i.e. grouping data points by specific features. This allows to create models that may be able to generalise better when there is a scarcity of data for a given task, making them more suitable for real-world purposes.

6.2. Challenges

One of the major challenges that needs to be addressed with future works is how authors disseminate their models and results. It is already known that there is an issue with the reproducibility of Deep Learning results [55]. In fact, this also applies to the EAR community: there is a need for better description of models, datasets employed, the data splits created and so on. It is also specially important to establish appropriate metrics for the sake of comparison, as the accuracy is extensively used on its own. Due to the accuracy paradox and the unbalanced nature of EAR datasets, the accuracy is not a suitable metric and it does not allow to correctly compare different solutions. Moreover, how the results are provided is still not usually specified. That is, given the randomness associated to Deep Learning, providing a single result may be misleading and how this result has been computed should be specified. This problem is described by [55], whose authors propose to compare models using a budget (i.e. time to train, number of hyper-parameters and so forth).

Another aspect to improve is the collection of egocentric datasets we have. In fact, this is an important issue to address in order to push forward the research. In Section 4 the available datasets were analysed. Among them, the largest and most complete is the EPIC Kitchens dataset. In contrast to the exocentric vision, this community did not have a very large dataset to be used for pre-training or just to have a common dataset for benchmarking until the appearance of EPIC Kitchens, limiting the research and performance that could be obtained, having to pre-train EAR models with exocentric datasets. Nonetheless, even larger datasets need to be created (or the existing ones need to be extended), as it is known that video datasets are still small in comparison to static image datasets. In fact, in the egocentric community there is also a need for variety. The most used datasets, the GTEA family and the EPIC Kitchen dataset, target kitchen related actions. This limits the scope of actions and the possibility to apply to the real-world models that learnt from them. Moreover, this could also lead to a data bias, as models that used these datasets can be considered specialists in kitchen actions, neglecting other tasks.

CRediT authorship contribution statement

Adrián Núñez-Marcos: Conceptualization, Methodology, Investigation, Writing - original draft. **Gorka Azkune:** Conceptualization, Supervision, Writing - review & editing. **Ignacio Arganda-Carreras:** Conceptualization, Supervision, Writing - review & editing.

Declaration of Competing Interest

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

Acknowledgement

We gratefully acknowledge the support of the Basque Government's Department of Education for the predoctoral funding of the first author. This work has been supported by the Spanish Government under the FuturAAL-Context project (RTI2018-101045-B-C21) and by the Basque Government under the Deustek project (IT-1078–16-D).

References

[1] Sathyanarayanan Aakur, Fillipe de Souza, Sudeep Sarkar, Generating open world descriptions of video using common sense knowledge in a pattern theory framework, Quarterly of Applied Mathematics 77 (2) (2019) 323–356.

[2] Sathyanarayanan N Aakur, Sanjoy Kundu, and Nikhil Gunti. Knowledge guided learning: Towards open domain egocentric action recognition with zero supervision. arXiv preprint arXiv:2009.07470, 2020..

[3] Girmaw Abebe, Andrea Cavallaro, Xavier Parra, Robust multi-dimensional motion features for first-person vision activity recognition, Computer Vision and Image Understanding 149 (2016) 229–248.

[4] Nachwa Aboubakr, James L Crowley, and Rémi Ronfard. Recognizing manipulation actions from state-transformations. arXiv preprint arXiv:1906.05147, 2019..

[5] Maedeh Aghaei, Mariella Dimiccoli, Petia Radeva, With whom do i interact? detecting social interactions in egocentric photo-streams, in: 2016 23rd International Conference on Pattern Recognition (ICPR), IEEE, 2016, pp. 2959–2964.

[6] Mohammad Al-Naser, Hiroki Ohashi, Sheraz Ahmed, Katsuyuki Nakamura, Takayuki Akiyama, Takuto Sato, Phong Xuan Nguyen, and Andreas Dengel. Hierarchical model for zero-shot activity recognition using wearable sensors. In ICAART (2), pages 478–485, 2018..

[7] Stefano Alletto, Giuseppe Serra, Simone Calderara, Rita Cucchiara, Understanding social relationships in egocentric vision, Pattern Recognition 48 (12) (2015) 4082–4096.

[8] Stefano Alletto, Giuseppe Serra, Simone Calderara, Francesco Solera, Rita Cucchiara, From ego to nos-vision: Detecting social relationships in first-person views, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2014, pp. 580–585.

[9] Mehmet Ali Arabacı̇, Fatih Özkan, Elif Surer, Peter Jančovič, and Alptekin Temizel. Multi-modal egocentric activity recognition using audio-visual features. arXiv preprint arXiv:1807.00612, 2018..

[10] Relja Arandjelović, Andrew Zisserman, Three things everyone should know to improve object retrieval, in: 2012 IEEE Conference on Computer Vision and Pattern Recognition, IEEE, 2012, pp. 2911–2918.

[11] Maryam Asadi-Aghbolaghi, Albert Clapés, Marco Bellantonio, Hugo Jair Escalante, Víctor Ponce-López, Xavier Baró, Isabelle Guyon, Shohreh Kasaei, and Sergio Escalera. Deep learning for action and gesture recognition in image sequences: A survey. In Gesture Recognition, pages 539–578. Springer, 2017..

[12] Khalid E.L. Asnaoui, Aksasse Hamid, Aksasse Brahim, Ouanan Mohammed, A survey of activity recognition in egocentric lifelogging datasets, in: 2017 International Conference on Wireless Technologies, Embedded and Intelligent Systems (WITS), IEEE, 2017, pp. 1–8.

[13] Sikai Bai, Qi Wang, Xuelong Li, Mfi: Multi-range feature interchange for video action recognition, in: 2020 25th International Conference on Pattern Recognition (ICPR), IEEE, 2021, pp. 6664–6671.

[14] Sven Bambach. A survey on recent advances of computer vision algorithms for egocentric video. arXiv preprint arXiv:1501.02825, 2015..

[15] Sven Bambach, John Franchak, David Crandall, and Chen Yu. Detecting hands in children's egocentric views to understand embodied attention during social interaction. In Proceedings of the Annual Meeting of the Cognitive Science Society, volume 36, 2014..

[16] Sven Bambach, Stefan Lee, David J Crandall, Yu. Chen, Lending a hand: Detecting hands and recognizing activities in complex egocentric interactions, in: Proceedings of the IEEE International Conference on Computer Vision, 2015, pp. 1949–1957.

[17] Andrea Bandini, José Zariffa, Analysis of the hands in egocentric vision: A survey, IEEE Transactions on Pattern Analysis and Machine Intelligence (2020).

[18] Herbert Bay, Tinne Tuytelaars, Luc Van Gool, in: Surf: Speeded up robust features In European conference on computer vision, Springer, 2006, pp. 404–417.

[19] Ardhendu Behera, Matthew Chapman, Anthony G Cohn, and David C Hogg. Egocentric activity recognition using histograms of oriented pairwise relations. In 2014 International Conference on Computer Vision Theory and Applications (VISAPP), volume 2, pages 22–30. IEEE, 2014..

[20] Ardhendu Behera, David C Hogg, Anthony G Cohn, Egocentric activity monitoring and recovery, in: Asian Conference on Computer Vision, Springer, 2012, pp. 519–532.

[21] Alejandro Betancourt, Pietro Morerio, Carlo S Regazzoni, Matthias Rauterberg, The evolution of first person vision methods: A survey, IEEE Transactions on Circuits and Systems for Video Technology 25 (5) (2015) 744–760.

[22] Keshav Bhandari, Mario A DeLaGarza, Ziliang Zong, Hugo Latapie, Yan Yan, Egok360: A 360 egocentric kinetic human activity video dataset, in: 2020 IEEE International Conference on Image Processing (ICIP), IEEE, 2020, pp. 266–270.

[23] Bharat Lal Bhatnagar, Suriya Singh, Chetan Arora, CV Jawahar, and KCIS CVIT. Unsupervised learning of deep feature representation for clustering egocentric actions. In IJCAI, pages 1447–1453, 2017..

[24] Marc Bolaños, Petia Radeva, Simultaneous food localization and recognition, in: 2016 23rd International Conference on Pattern Recognition (ICPR), IEEE, 2016, pp. 3140–3145.

[25] Anna Bosch, Andrew Zisserman, Xavier Munoz, Representing shape with a spatial pyramid kernel, in: Proceedings of the 6th ACM International Conference on Image and Video Retrieval, 2007, pp. 401–408.

[26] Nadia Brancati, Giuseppe Caggianese, Maria Frucci, Luigi Gallo, Pietro Neroni, Robust fingertip detection in egocentric vision under varying illumination conditions, in: 2015 IEEE International Conference on Multimedia & Expo Workshops (ICMEW), IEEE, 2015, pp. 1–6.

[27] Andreas Bulling, Jamie A Ward, Hans Gellersen, Gerhard Troster, Eye movement analysis for activity recognition using electrooculography, IEEE Transactions on Pattern Analysis and Machine Intelligence 33 (4) (2010) 741–753.

[28] Minjie Cai, Lu. Feng, Yue Gao, Desktop action recognition from first-person point-of-view, IEEE Transactions on Cybernetics 49 (5) (2018) 1616–1628.

[29] Fabien Cardinaux, Deepayan Bhowmik, Charith Abhayaratne, Mark S Hawley, Video based technology for ambient assisted living: A review of the literature, Journal of Ambient Intelligence and Smart Environments 3 (3) (2011) 253–269.

[30] Joao Carreira, Andrew Zisserman, Quo vadis, action recognition? a new model and the kinetics dataset, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 6299–6308.

[31] Alejandro Cartas, Jordi Luque, Petia Radeva, Carlos Segura, and Mariella Dimiccoli. How much does audio matter to recognize egocentric object interactions? arXiv preprint arXiv:1906.00634, 2019..

[32] Alejandro Cartas, Jordi Luque, Petia Radeva, Carlos Segura, Mariella Dimiccoli, Seeing and hearing egocentric actions: How much can we learn?, in: Proceedings of the IEEE International Conference on Computer Vision Workshops, 2019

[33] Alejandro Cartas, Petia Radeva, and Mariella Dimiccoli. Contextually driven first-person action recognition from videos..

[34] Alejandro Cartas, Petia Radeva, Mariella Dimiccoli, Modeling long-term interactions to enhance action recognition, in: 2020 25th International Conference on Pattern Recognition (ICPR), IEEE, 2021, pp. 10351–10358.

[35] Daniel Castro, Steven Hickson, Vinay Bettadapura, Edison Thomaz, Gregory Abowd, Henrik Christensen, and Irfan Essa. Predicting daily activities from egocentric images using deep learning. In proceedings of the 2015 ACM International symposium on Wearable Computers, pages 75–82, 2015..

[36] Mohamed Chaabane, Ameni Trabelsi, Nathaniel Blanchard, Ross Beveridge, Looking ahead: Anticipating pedestrians crossing with future frames prediction, in: The IEEE Winter Conference on Applications of Computer Vision, 2020, pp. 2297–2306.

[37] Alexandros André Chaaraoui, Pa.u. Climent-Pérez, Francisco Flórez-Revuelta, A review on vision techniques applied to human behaviour analysis for ambient-assisted living, Expert Systems with Applications 39 (12) (2012) 10873–10888.

[38] François Chollet, Xception: Deep learning with depthwise separable convolutions, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 1251–1258.

[39] Pa.u. Climent-Pérez, Susanna Spinsante, Alex Mihailidis, Francisco Florez-Revuelta, A review on video-based active and assisted living technologies for automated lifelogging, Expert Systems with Applications 139 (2020) 112847.

[40] Darwin Ttito Concha, Helena De Almeida Maia, Helio Pedrini, Hemerson Tacon, André De Souza Brito, Hugo De Lima Chaves, and Marcelo Bernardes Vieira. Multi-stream convolutional neural networks for action recognition in video sequences based on adaptive visual rhythms. In 2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA), pages 473–480. IEEE, 2018..

[41] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Antonino Furnari, Jian Ma, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, and Michael Wray. Rescaling egocentric vision. CoRR, abs/2006.13256, 2020..

[42] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Sanja Fidler, Antonino Furnari, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, and Michael Wray. Scaling egocentric vision: The epic-kitchens dataset. In European Conference on Computer Vision (ECCV), 2018..

[43] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Sanja Fidler, Antonino Furnari, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, et al. Scaling egocentric vision: The epic-kitchens dataset. In Proceedings of the European Conference on Computer Vision (ECCV), pages 720–736, 2018..

[44] Dima Damen, Teesid Leelasawassuk, Osian Haines, Andrew Calway, Walterio W Mayol-Cuevas, You-do, i-learn: Discovering task relevant objects and their modes of interaction from multi-user egocentric video, BMVC 2 (2014) page 3.

[45] Dima Damen, Teesid Leelasawassuk, Walterio Mayol-Cuevas, You-do, i-learn: Egocentric unsupervised discovery of objects and their modes of interaction towards video-based guidance, Computer Vision and Image Understanding 149 (2016) 98–112.

[46] Pratyusha Das, Antonio Ortega, Symmetric sub-graph spatio-temporal graph convolution and its application in complex activity recognition, in: ICASSP 2021–2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2021, pp. 3215–3219.

[47] Steven Davis, Paul Mermelstein, Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences, IEEE Transactions on Acoustics, Speech, and Signal Processing 28 (4) (1980) 357–366.

[48] Fernando De la Torre, Jessica Hodgins, Adam Bargteil, Xavier Martin, Justin Macey, Alex Collado, and Pep Beltran. Guide to the carnegie mellon university multimodal activity (cmu-mmac) database. 2009..

[49] Ana Garcia Del Molino, Cheston Tan, Joo-Hwee Lim, Ah-Hwee Tan, Summarization of egocentric videos: A comprehensive survey, IEEE Transactions on Human-Machine Systems 47 (1) (2016) 65–76.
[50] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, Li Fei-Fei, Imagenet: A large-scale hierarchical image database, in: 2009 IEEE Conference on Computer Vision and Pattern Recognition, Ieee, 2009, pp. 248–255.
[51] Jean Dezert and Florentin Smarandache. Advances and applications of dsmt for information fusion. Am. Res. Press, Rehoboth, 1, 2004..
[52] Alexander Diete, Timo Sztyler, Lydia Weiland, Heiner Stuckenschmidt, Improving motion-based activity recognition with ego-centric vision, in: 2018 IEEE International Conference on Pervasive Computing and Communications Workshops (PerCom Workshops), IEEE, 2018, pp. 488–491.
[53] Semantic regularized clustering for egocentric photo streams segmentation, Mariella Dimiccoli, Marc Bolaños, Estefania Talavera, Maedeh Aghaei, Stavri G Nikolov, and Petia Radeva. Sr-clustering, Computer Vision and Image Understanding 155 (2017) 55–69.
[54] Mariella Dimiccoli, Juan Marín, Edison Thomaz, Mitigating bystander privacy concerns in egocentric activity recognition with deep learning and intentional image degradation, Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies 1 (4) (2018) 1–18.
[55] Jesse Dodge, Suchin Gururangan, Dallas Card, Roy Schwartz, and Noah A Smith. Show your work: Improved reporting of experimental results. arXiv preprint arXiv:1909.03004, 2019..
[56] Jeffrey Donahue, Lisa Anne Hendricks, Sergio Guadarrama, Marcus Rohrbach, Subhashini Venugopalan, Kate Saenko, and Trevor Darrell. Long-term recurrent convolutional networks for visual recognition and description. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 2625–2634, 2015..
[57] Chen Fang, Lorenzo Torresani, in: Measuring image distances via embedding in a semantic manifold In European Conference on Computer Vision, Springer, 2012, pp. 402–415.
[58] Alircza Fathi, Jessica K Hodgins, James M Rehg, Social interactions: A first-person perspective, in: 2012 IEEE Conference on Computer Vision and Pattern Recognition, IEEE, 2012, pp. 1226–1233.
[59] Alireza Fathi, Ali Farhadi, James M Rehg, Understanding egocentric activities, in: 2011 International Conference on Computer Vision, IEEE, 2011, pp. 407–414.
[60] Alireza Fathi, Yin Li, James M Rehg, Learning to recognize daily actions using gaze, in: European Conference on Computer Vision, Springer, 2012, pp. 314–327.
[61] Alireza Fathi, James M Rehg, Modeling actions through state changes, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2013, pp. 2579–2586.
[62] Amy Fire, Song-Chun Zhu, Learning perceptual causality from video, ACM Transactions on Intelligent Systems and Technology (TIST) 7 (2) (2015) 1–22.
[63] Martin A. Fischler, Robert C. Bolles, Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography, Communications of the ACM 24 (6) (1981) 381–395.
[64] Antonino Furnari, Giovanni Maria Farinella, What would you expect? anticipating egocentric actions with rolling-unrolling lstms and modality attention, in: Proceedings of the IEEE International Conference on Computer Vision, 2019, pp. 6252–6261.
[65] Harshala Gammulle, Simon Denman, Sridha Sridharan, Clinton Fookes, Two stream lstm: A deep fusion framework for human action recognition, in: 2017 IEEE Winter Conference on Applications of Computer Vision (WACV), IEEE, 2017, pp. 177–186.
[66] Guillermo Garcia-Hernando, Shanxin Yuan, Seungryul Baek, Tae-Kyun Kim, First-person hand action benchmark with rgb-d videos and 3d hand pose annotations, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 409–419.
[67] Georgia Gkioxari, Ross Girshick, Jitendra Malik, Contextual action recognition with r* cnn, in: Proceedings of the IEEE International Conference on Computer Vision, 2015, pp. 1080–1088.
[68] Peter M Gollwitzer, Action phases and mind-sets. Handbook of motivation and cognition, Foundations of social behavior 2 (1990) 53–92.
[69] Ulf Grenander, Elements of pattern theory, JHU Press (1996).
[70] Kai Guo, Prakash Ishwar, Janusz Konrad, Action recognition from video using feature covariance matrices, IEEE Transactions on Image Processing 22 (6) (2013) 2479–2494.
[71] Taejin Ha, Steven Feiner, Woontack Woo, Wearhand: Head-worn, rgb-d camera-based, bare-hand user interface with visually enhanced depth perception, in: 2014 IEEE International Symposium on Mixed and Augmented Reality (ISMAR), IEEE, 2014, pp. 219–228.
[72] Mary Hayhoe, Vision using routines: A functional account of vision, Visual Cognition 7 (1–3) (2000) 43–64.
[73] Sepp Hochreiter, Jürgen Schmidhuber, Long short-term memory, Neural computation 9 (8) (1997) 1735–1780.
[74] Yifei Huang, Zhenqiang Li, Minjie Cai, and Yoichi Sato. Mutual context network for jointly estimating egocentric gaze and actions. arXiv preprint arXiv:1901.01874, 2019..
[75] Javed Imran, Balasubramanian Raman, Three-stream spatio-temporal attention network for first-person action and interaction recognition, Journal of Ambient Intelligence and Humanized Computing (2021) 1–16.
[76] Youngkyoon Jang, Ikbeom Jeon, Tae-Kyun Kim, Woontack Woo, Metaphoric hand gestures for orientation-aware vr object manipulation with an egocentric viewpoint, IEEE Transactions on Human-Machine Systems 47 (1) (2016) 113–127.
[77] Youngkyoon Jang, Seung-Tak Noh, Hyung Jin Chang, Tae-Kyun Kim, and Woontack Woo. 3d finger cape: Clicking action and position estimation under self-occlusions in egocentric viewpoint. IEEE Transactions on Visualization and Computer Graphics, 21(4), 501–510, 2015..
[78] Youngkyoon Jang, Brian Sullivan, Casimir Ludwig, Iain Gilchrist, Dima Damen, and Walterio Mayol-Cuevas. Epic-tent: An egocentric video dataset for camping tent assembly. In Proceedings of the IEEE International Conference on Computer Vision Workshops, pages 0–0, 2019..
[79] Ali Javidani, Ahmad Mahmoudi-Aznaveh, A unified method for first and third person action recognition, in: Iranian Conference on Electrical Engineering (ICEE), IEEE, 2018, pp. 1629–1633.
[80] Herve Jegou, Florent Perronnin, Matthijs Douze, Jorge Sánchez, Patrick Perez, Cordelia Schmid, Aggregating local image descriptors into compact codes, IEEE Transactions on Pattern Analysis and Machine Intelligence 34 (9) (2011) 1704–1716.
[81] Shuiwang Ji, Xu. Wei, Ming Yang, Yu. Kai, 3d convolutional neural networks for human action recognition, IEEE transactions on pattern analysis and machine intelligence 35 (1) (2012) 221–231.
[82] Wenyan Jia, Yuecheng Li, Ruowei Qu, Thomas Baranowski, Lora E Burke, Hong Zhang, Yicheng Bai, Juliet M Mancino, Guizhi Xu, Zhi-Hong Mao, et al. Automatic food detection in egocentric images using artificial intelligence technology. Public health nutrition, 22(7):1168–1179, 2019..
[83] Haiyu Jiang, Yan Song, Jiang He, and Xiangbo Shu. Cross fusion for egocentric interactive action recognition. In International Conference on Multimedia Modeling, pages 714–726. Springer, 2020..
[84] Takeo Kanade, Martial Hebert, First-person vision, Proceedings of the IEEE 100 (8) (2012) 2442–2453.
[85] Hongwen Kang, Martial Hebert, Takeo Kanade, Discovering object instances from scenes of daily living, in: 2011 International Conference on Computer Vision, IEEE, 2011, pp. 762–769.
[86] Georgios Kapidis, Ronald Poppe, Elsbeth van Dam, Lucas Noldus, Remco Veltkamp, Multitask learning to improve egocentric action recognition, in: Proceedings of the IEEE International Conference on Computer Vision Workshops, 2019.
[87] Georgios Kapidis, Ronald Poppe, Elsbeth van Dam, Lucas PJJ Noldus, and Remco C Veltkamp. Egocentric hand track and object-based human action recognition. arXiv preprint arXiv:1905.00742, 2019..
[88] Georgios Kapidis, Ronald Poppe, Elsbeth van Dam, Lucas PJJ Noldus, Remco C Veltkamp, Object detection-based location and activity classification from egocentric videos: A systematic analysis, in: Smart Assisted Living, Springer, 2020, pp. 119–145.
[89] Georgios Kapidis, Ronald Poppe, Remco C Veltkamp, Multi-dataset, multitask learning of egocentric vision tasks, IEEE Transactions on Pattern Analysis and Machine Intelligence (2021).
[90] Evangelos Kazakos, Arsha Nagrani, Andrew Zisserman, Dima Damen, Epic-fusion: Audio-visual temporal binding for egocentric action recognition, in: Proceedings of the IEEE International Conference on Computer Vision, 2019, pp. 5492–5501.
[91] Adam Kendon. Studies in the behavior of social interaction, volume 6. Humanities Press International, 1977..
[92] Kris M Kitani, Takahiro Okabe, Yoichi Sato, and Akihiro Sugimoto. Fast unsupervised ego-action learning for first-person sports videos. In CVPR 2011, pages 3241–3248. IEEE, 2011..
[93] K.P. Sanal Kumar, Activity recognition in egocentric video using svm, knn and combined svmknn classifiers, IOP Conference Series: Materials Science and Engineering, volume 225, IOP Publishing, 2017, 012226.
[94] K.P. Sanal Kumar, R. Bhavani, Human activity recognition in egocentric video using hog, gist and color features, Multimedia Tools and Applications 79 (5) (2020) 3543–3559.
[95] Heeseung Kwon, Yeonho Kim, Jin S Lee, Minsu Cho, First person action recognition via two-stream convnet with long-term fusion pooling, Pattern Recognition Letters 112 (2018) 161–167.
[96] Taein Kwon, Bugra Tekin, Jan Stuhmer, Federica Bogo, and Marc Pollefeys. H2o: Two hands manipulating objects for first person interaction recognition. arXiv preprint arXiv:2104.11181, 2021..
[97] Michael Land, Neil Mennie, Jennifer Rusted, The roles of vision and eye movements in the control of activities of daily living, Perception 28 (11) (1999) 1311–1328.
[98] Michael Land, Benjamin Tatler, Looking and acting: vision and eye movements in natural behaviour, Oxford University Press, 2009.
[99] Ivan Laptev, Marcin Marszalek, Cordelia Schmid, Benjamin Rozenfeld, Learning realistic human actions from movies, in: 2008 IEEE Conference on Computer Vision and Pattern Recognition, IEEE, 2008, pp. 1–8.
[100] Kyungjun Lee, Abhinav Shrivastava, Hernisa Kacorri, Hand-priming in object localization for assistive egocentric vision, in: The IEEE Winter Conference on Applications of Computer Vision, 2020, pp. 3422–3432.
[101] Yong Jae Lee, Joydeep Ghosh, Kristen Grauman, Discovering important people and objects for egocentric video summarization, in: 2012 IEEE conference on computer vision and pattern recognition, IEEE, 2012, pp. 1346–1353.
[102] Chuankun Li, Shuai Li, Yanbo Gao, Xiang Zhang, and Wanqing Li. A two-stream neural network for pose-based hand gesture recognition. arXiv preprint arXiv:2101.08926, 2021..

[103] Xiangyu Li, Yonghong Hou, Pichao Wang, Zhimin Gao, Mingliang Xu, and Wanqing Li. Trear: Transformer-based rgb-d egocentric action recognition. IEEE Transactions on Cognitive and Developmental Systems, 2021..

[104] Yanghao Li, Tushar Nagarajan, Bo Xiong, Kristen Grauman, Ego-exo: Transferring visual representations from third-person to first-person videos, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 6943–6953.

[105] Yin Li, Miao Liu, James M Rehg, In the eye of beholder: Joint learning of gaze and actions in first person video, in: Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 619–635.

[106] Yin Li, Zhefan Ye, James M Rehg, Delving into egocentric actions, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 287–295.

[107] Ji Lin, Chuang Gan, Song Han, Tsm: Temporal shift module for efficient video understanding, in: Proceedings of the IEEE International Conference on Computer Vision, 2019, pp. 7083–7093.

[108] Bingbin Liu, Ehsan Adeli, Zhangjie Cao, Kuan-Hui Lee, Abhijeet Shenoi, Adrien Gaidon, and Juan Carlos Niebles. Spatiotemporal relationship reasoning for pedestrian intent prediction. IEEE Robotics and Automation Letters, 5(2), 3485–3492, 2020..

[109] Hugo Liu and Push Singh. Conceptnet-a practical commonsense reasoning tool-kit. BT technology journal, 22(4):211–226, 2004..

[110] Jianbo Liu, Yongcheng Liu, Ying Wang, Veronique Prinet, Shiming Xiang, and Chunhong Pan. Decoupled representation learning for skeleton-based gesture recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 5751–5760, 2020..

[111] Jianbo Liu, Ying Wang, Shiming Xiang, and Chunhong Pan. Han: An efficient hierarchical self-attention network for skeleton-based gesture recognition. arXiv preprint arXiv:2106.13391, 2021..

[112] Miao Liu, Lingni Ma, Kiran Somasundaram, Yin Li, Kristen Grauman, James M Rehg, and Chao Li. Egocentric activity recognition and localization on a 3d map. arXiv preprint arXiv:2105.09544, 2021..

[113] Yang Liu, Ping Wei, Song-Chun Zhu, Jointly recognizing object fluents and tasks in egocentric videos, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 2924–2932.

[114] Yinan Liu, Wu. Qingbo, Liangzhi Tang, Hengcan Shi, Gaze-assisted multi-stream deep neural network for action recognition, IEEE Access 5 (2017) 19432–19441.

[115] Alejandro López-Cifuentes, Marcos Escudero-Viñolo, and Jesús Bescós. A prospective study on sequence-driven temporal sampling and ego-motion compensation for action recognition in the epic-kitchens dataset. arXiv preprint arXiv:2008.11588, 2020..

[116] David G Lowe, Distinctive image features from scale-invariant keypoints, International journal of computer vision 60 (2) (2004) 91–110.

[117] Lu. Minlong, Ze-Nian Li, Yueming Wang, Gang Pan, Deep attention network for egocentric action recognition, IEEE Transactions on Image Processing 28 (8) (2019) 3703–3713.

[118] Lu. Minlong, Danping Liao, Ze-Nian Li, Learning spatiotemporal attention for egocentric action recognition, in: Proceedings of the IEEE International Conference on Computer Vision Workshops, 2019.

[119] Yantao Lu and Senem Velipasalar. Human activity classification incorporating egocentric video and inertial measurement unit data. In 2018 IEEE Global Conference on Signal and Information Processing (GlobalSIP), pages 429–433. IEEE, 2018..

[120] Chih-Yao Ma, Asim Kadav, Iain Melvin, Zsolt Kira, Ghassan AlRegib, Hans Peter Graf, Attend and interact: Higher-order object interactions for video understanding, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 6790–6800.

[121] Minghuang Ma, Haoqi Fan, Kris M Kitani, Going deeper into first-person activity recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 1894–1903.

[122] Steve Mann. 'wearcam'(the wearable camera): personal imaging systems for long-term use in wearable tetherless computer-mediated reality and personal photo/videographic memory prosthesis. In Digest of Papers. Second International Symposium on Wearable Computers (Cat. No. 98EX215), pages 124–131. IEEE, 1998..

[123] Joanna Materzynska, Tete Xiao, Roei Herzig, Huijuan Xu, Xiaolong Wang, and Trevor Darrell. Something-else: Compositional action recognition with spatial-temporal interaction networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 1049–1059, 2020..

[124] Kenji Matsuo, Kentaro Yamada, Satoshi Ueno, Sei Naito, An attention-based activity recognition for egocentric video, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2014, pp. 551–556.

[125] Tomas McCandless and Kristen Grauman. Object-centric spatio-temporal pyramids for egocentric activity recognition. In BMVC, volume 2, page 3. Citeseer, 2013..

[126] Georgios Meditskos, Pierre-Marie Plans, Thanos G. Stavropoulos, Jenny Benois-Pineau, Vincent Buso, Ioannis Kompatsiaris, Multi-modal activity recognition from egocentric vision, semantic enrichment and lifelogging applications for the care of dementia, Journal of Visual Communication and Image Representation 51 (2018) 169–190.

[127] Xiao-Li Meng, Donald B Rubin, Maximum likelihood estimation via the ecm algorithm: A general framework, Biometrika 80 (2) (1993) 267–278.

[128] Shinya Michibata, Katsufumi Inoue, Michifumi Yoshioka, Atsushi Hashimoto, Cooking activity recognition in egocentric videos with a hand mask image branch in the multi-stream cnn, in: Proceedings of the 2020 Multimedia on Cooking and Eating Activities Workshop, 2020, pp. 1–6.

[129] Ajay K Mishra, Yiannis Aloimonos, Loong Fah Cheong, Ashraf Kassim, Active visual segmentation, IEEE Transactions on Pattern Analysis and Machine Intelligence 34 (4) (2011) 639–653.

[130] Davide Moltisanti, Michael Wray, Walterio Mayol-Cuevas, Dima Damen, Trespassing the boundaries: Labeling temporal bounds for object interactions in egocentric video, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 2886–2894.

[131] Thierry Pinheiro Moreira, David Menotti, Helio Pedrini, First-person action recognition through visual rhythm texture description, in: 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2017, pp. 2627–2631.

[132] Erik T Mueller, Commonsense reasoning: an event calculus based approach, Morgan Kaufmann, 2014.

[133] Tushar Nagarajan, Yanghao Li, Christoph Feichtenhofer, and Kristen Grauman. Ego-topo: Environment affordances from egocentric video. arXiv preprint arXiv:2001.04583, 2020..

[134] Katsuyuki Nakamura, Serena Yeung, Alexandre Alahi, Li Fei-Fei, Jointly learning energy expenditures and activities using egocentric multimodal signals, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 1868–1877.

[135] Tomoya Nakatani, Ryohei Kuga, Takuya Maekawa, Preliminary investigation of object-based activity recognition using egocentric video based on web knowledge, in: Proceedings of the 17th International Conference on Mobile and Ubiquitous Multimedia, 2018, pp. 375–381.

[136] Atsushi Nakazawa, Miwako Honda, First-person camera system to evaluate tender dementia-care skill, in: Proceedings of the IEEE International Conference on Computer Vision Workshops, 2019.

[137] Sanath Narayan, Mohan S Kankanhalli, Kalpathi R Ramakrishnan, Action and interaction recognition in first-person videos, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2014, pp. 512–518.

[138] Jean-Christophe Nebel, Francisco Florez-Revuelta, et al., Recognition of activities of daily living from egocentric videos using hands detected by a deep convolutional network, in: International Conference Image Analysis and Recognition, Springer, 2018, pp. 390–398.

[139] Thi-Hoa-Cuc Nguyen, Jean-Christophe Nebel, Francisco Florez-Revuelta, et al., Recognition of activities of daily living with egocentric vision: A review, Sensors 16 (1) (2016) 72.

[140] Xuan Son Nguyen, Luc Brun, Olivier Lézoray, Sébastien Bougleux, A neural network based on spd manifold learning for skeleton-based hand gesture recognition, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 12036–12045.

[141] Adrián Núñez-Marcos, Gorka Azkune, Eneko Agirre, Diego López-de Ipiña, and Ignacio Arganda-Carreras. Using external knowledge to improve zero-shot action recognition in egocentric videos. In International Conference on Image Analysis and Recognition, pages 174–185. Springer, 2020..

[142] Keisuke Ogaki, Kris M Kitani, Yusuke Sugano, Yoichi Sato, Coupling eye-motion and ego-motion features for first-person activity recognition, in: 2012 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, IEEE, 2012, pp. 1–7.

[143] Timo Ojala, Matti Pietikainen, David Harwood, Performance evaluation of texture measures with classification based on kullback discrimination of distributions, Proceedings of 12th international conference on pattern recognition, volume 1, IEEE, 1994, pp. 582–585.

[144] Timo Ojala, Matti Pietikäinen, David Harwood, A comparative study of texture measures with classification based on featured distributions, Pattern recognition 29 (1) (1996) 51–59.

[145] Juan-Manuel Perez-Rua, Brais Martinez, Xiatian Zhu, Antoine Toisoul, Victor Escorcia, and Tao Xiang. Knowing what, where and when to look: Efficient video action modeling with attention. arXiv preprint arXiv:2004.01278, 2020..

[146] Juan-Manuel Perez-Rua, Antoine Toisoul, Brais Martinez, Victor Escorcia, Li Zhang, Xiatian Zhu, and Tao Xiang. Egocentric action recognition by video attention and temporal context. arXiv preprint arXiv:2007.01883, 2020..

[147] Florent Perronnin, Christopher Dance, Fisher kernels on visual vocabularies for image categorization, in: 2007 IEEE Conference on Computer Vision and Pattern Recognition, IEEE, 2007, pp. 1–8.

[148] Hamed Pirsiavash, Deva Ramanan, Detecting activities of daily living in first-person camera views, in: 2012 IEEE Conference on Computer Vision and Pattern Recognition, IEEE, 2012, pp. 2847–2854.

[149] Mirco Planamente, Andrea Bottino, and Barbara Caputo. Joint encoding of appearance and motion features with self-supervision for first person action recognition. arXiv preprint arXiv:2002.03982, 2020..

[150] Mirco Planamente, Andrea Bottino, Barbara Caputo, Self-supervised joint encoding of motion and appearance for first person action recognition, in: 2020 25th International Conference on Pattern Recognition (ICPR), IEEE, 2021, pp. 8751–8758.

[151] Mirco Planamente, Chiara Plizzari, Emanuele Alberti, and Barbara Caputo. Cross-domain first person audio-visual action recognition through relative norm alignment. arXiv preprint arXiv:2106.01689, 2021..

[152] Yair Poleg, Chetan Arora, and Shmuel Peleg. Head motion signatures from egocentric videos. In Asian Conference on Computer Vision, pages 315–329. Springer, 2014..

[153] Yair Poleg, Chetan Arora, Shmuel Peleg, Temporal segmentation of egocentric videos, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2014, pp. 2537–2544.

[154] Yair Poleg, Ariel Ephrat, Shmuel Peleg, Chetan Arora, Compact cnn for indexing egocentric videos, in: 2016 IEEE Winter Conference on Applications of Computer Vision (WACV), IEEE, 2016, pp. 1–9.

[155] Rafael Possas, Sheila Pinto Caceres, Fabio Ramos, Egocentric activity recognition on a budget, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 5967–5976.

[156] Didik Purwanto, Yie-Tarng Chen, Wen-Hsien Fang, First-person action recognition with temporal pooling and hilbert–huang transform, IEEE Transactions on Multimedia 21 (12) (2019) 3122–3135.

[157] Francesco Ragusa, Antonino Furnari, Sebastiano Battiato, Giovanni Signorello, and Giovanni Maria Farinella. Ego-ch: Dataset and fundamental tasks for visitors behavioral understanding using egocentric vision. Pattern Recognition Letters, 131:150–157, 2020..

[158] Francesco Ragusa, Antonino Furnari, Salvatore Livatino, Giovanni Maria Farinella, The meccano dataset: Understanding human-object interactions from egocentric videos in an industrial-like domain, in: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2021, pp. 1569–1578.

[159] Joseph Redmon, Santosh Divvala, Ross Girshick, Ali Farhadi, You only look once: Unified, real-time object detection, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 779–788.

[160] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In Advances in Neural Information Processing Systems, pages 91–99, 2015..

[161] Xiaofeng Ren, Gu. Chunhui, Figure-ground segmentation improves handled object recognition in egocentric video, in: 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, IEEE, 2010, pp. 3137–3144.

[162] Xiaofeng Ren, Matthai Philipose, Egocentric recognition of handled objects: Benchmark and analysis, in: 2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, IEEE, 2009, pp. 1–8.

[163] Michael S Ryoo, Larry Matthies, First-person activity recognition: What are they doing to me?, in: Proceedings of the IEEE Conference on Computer vision and Pattern Recognition, 2013, pp 2730–2737.

[164] Michael S Ryoo, Brandon Rothrock, Larry Matthies, Pooled motion features for first-person videos, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 896–904.

[165] Abhimanyu Sahu, Rajit Bhattacharya, Pallabh Bhura, Ananda S Chowdhury, in: Action recognition from egocentric videos using random walks In Proceedings of 3rd International Conference on Computer Vision and Image Processing, Springer, 2020, pp. 389–402.

[166] Abhimanyu Sahu, Ananda S Chowdhury, Shot level egocentric video co-summarization, in: 2018 24th International Conference on Pattern Recognition (ICPR), IEEE, 2018, pp. 2887–2892.

[167] Abhimanyu Sahu, Ananda S Chowdhury, Together recognizing, localizing and summarizing actions in egocentric videos, IEEE Transactions on Image Processing 30 (2021) 4330–4340.

[168] Mostafa Kamal Sarker, Hatem A. Rashwan, Estefania Talavera, Syeda Furruka Banu, Petia Radeva, Domenec Puig, et al., Macnet: Multi-scale atrous convolution networks for food places classification in egocentric photo-streams, in: Proceedings of the European Conference on Computer Vision (ECCV), 2018.

[169] Tyler R Scott, Michael Shvartsman, and Karl Ridgeway. Unifying few-and zero-shot egocentric action recognition. arXiv preprint arXiv:2006.11393, 2020..

[170] Lei Shi, Yifan Zhang, Jian Cheng, Lu. Hanqing, Skeleton-based action recognition with multi-stream adaptive graph convolutional networks, IEEE Transactions on Image Processing 29 (2020) 9532–9545.

[171] Yuki Shiga, Takumi Toyama, Yuzuko Utsumi, Koichi Kise, Andreas Dengel, Daily activity recognition combining gaze motion and visual features, in: Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct Publication, 2014, pp. 1103–1111.

[172] Gunnar A Sigurdsson, Abhinav Gupta, Cordelia Schmid, Ali Farhadi, and Karteek Alahari. Charades-ego: A large-scale dataset of paired third and first person videos. arXiv preprint arXiv:1804.09626, 2018..

[173] Michel Silva, Washington Ramos, João Ferreira, Felipe Chamone, Mario Campos, and Erickson R. Nascimento. A weighted sparse sampling and smoothing frame transition approach for semantic fast-forward first-person videos. In 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pages 2383–2392, Salt Lake City, USA, Jun. 2018..

[174] Karen Simonyan and Andrew Zisserman. Two-stream convolutional networks for action recognition in videos. In Advances in Neural Information Processing Systems, pages 568–576, 2014..

[175] Suriya Singh, Chetan Arora, C.V. Jawahar, Generic action recognition from egocentric videos, in: 2015 Fifth National Conference on Computer Vision, Pattern Recognition, Image Processing and Graphics (NCVPRIPG), IEEE, 2015, pp. 1–4.

[176] Suriya Singh, Chetan Arora, C.V. Jawahar, First person action recognition using deep learned descriptors, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 2620–2628.

[177] Suriya Singh, Chetan Arora, and CV Jawahar. Trajectory aligned features for first person action recognition. Pattern Recognition, 62:45–55, 2017..

[178] Sibo Song, Vijay Chandrasekhar, Ngai-Man Cheung, Sanath Narayan, Liyuan Li, and Joo-Hwee Lim. Activity recognition in egocentric life-logging videos. In Asian Conference on Computer Vision, pages 445–458. Springer, 2014..

[179] Sibo Song, Ngai-Man Cheung, Vijay Chandrasekhar, Bappaditya Mandal, Jie Liri, Egocentric activity recognition with multimodal fisher vector, in: 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2016, pp. 2717–2721.

[180] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. arXiv preprint arXiv:1212.0402, 2012..

[181] Robert Speer, Catherine Havasi, Conceptnet 5: A large semantic network for relational knowledge, in: The People's Web Meets NLP, Springer, 2013, pp. 161–176.

[182] Ekaterina H Spriggs, Fernando De La Torre, Martial Hebert, Temporal segmentation and activity classification from first-person sensing, in: 2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, IEEE, 2009, pp. 17–24.

[183] Julian Steil, Marion Koelle, Wilko Heuten, Susanne Boll, Andreas Bulling, Privaceye: privacy-preserving head-mounted eye tracking using egocentric scene image and eye movement features, in: Proceedings of the 11th ACM Symposium on Eye Tracking Research & Applications, 2019, pp. 1–10.

[184] Oily Styles, Arun Ross, Victor Sanchez, Forecasting pedestrian trajectory with machine-annotated training data, in: 2019 IEEE Intelligent Vehicles Symposium (IV), IEEE, 2019, pp. 716–721.

[185] Swathikiran Sudhakaran, Sergio Escalera, and Oswald Lanz. Fbk-hupba submission to the epic-kitchens 2019 action recognition challenge. arXiv preprint arXiv:1906.08960, 2019..

[186] Swathikiran Sudhakaran, Sergio Escalera, and Oswald Lanz. Hierarchical feature aggregation networks for video action recognition. arXiv preprint arXiv:1905.12462, 2019..

[187] Swathikiran Sudhakaran, Sergio Escalera, Oswald Lanz, Lsta: Long short-term attention for egocentric action recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 9954–9963.

[188] Swathikiran Sudhakaran, Oswald Lanz, Convolutional long short-term memory networks for recognizing first person interactions, in: Proceedings of the IEEE International Conference on Computer Vision Workshops, 2017, pp. 2339–2346.

[189] Swathikiran Sudhakaran and Oswald Lanz. Attention is all we need: Nailing down object-centric attention for egocentric activity recognition. arXiv preprint arXiv:1807.11794, 2018..

[190] Li Sun, Ulrich Klank, Michael Beetz, Eyewatchme-3d hand and object tracking for inside out activity analysis, in: 2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, IEEE, 2009, pp. 9–16.

[191] Sudeep Sundaram, Walterio W Mayol, Cuevas, High level activity recognition using low resolution wearable vision, in: 2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, IEEE, 2009, pp. 25–32.

[192] Dipak Surie, Thomas Pederson, Fabien Lagriffoul, Lars-Erik Janlert, Daniel Sjölie, in: Activity recognition using an egocentric perspective of everyday objects In International Conference on Ubiquitous Intelligence and Computing, Springer, 2007, pp. 246–257.

[193] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 1–9, 2015..

[194] Estefania Talavera, Mariella Dimiccoli, Marc Bolanos, Maedeh Aghaei, Petia Radeva, R-clustering for egocentric video segmentation, in: Iberian Conference on Pattern Recognition and Image Analysis, Springer, 2015, pp. 327–336.

[195] Yansong Tang, Zian Wang, Lu. Jiwen, Jianjiang Feng, Jie Zhou, Multi-stream deep neural networks for rgb-d egocentric action recognition, IEEE Transactions on Circuits and Systems for Video Technology 29 (10) (2018) 3001–3015.

[196] Bugra Tekin, Federica Bogo, Marc Pollefeys, H+ o, Unified egocentric recognition of 3d hand-object poses and interactions, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 4511–4520.

[197] Daniel Thalmann, Hui Liang, Junsong Yuan, First-person palm pose tracking and gesture recognition in augmented reality, in: International Joint Conference on Computer Vision, Imaging and Computer Graphics, Springer, 2015, pp. 3–15.

[198] Daksh Thapar, Chetan Arora, and Aditya Nigam. Is sharing of egocentric video giving away your biometric signature? 2020..

[199] Du. Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, Manohar Paluri, Learning spatiotemporal features with 3d convolutional networks, in: Proceedings of the IEEE International Conference on Computer Vision, 2015, pp. 4489–4497.

[200] Amin Ullah, Jamil Ahmad, Khan Muhammad, Muhammad Sajjad, and Sung Wook Baik. Action recognition in video sequences using deep bi-directional lstm with cnn features. IEEE access, 6:1155–1166, 2017..

[201] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In Advances in Neural Information Processing Systems, pages 5998–6008, 2017..

[202] Sagar Verma, Pravin Nagar, Divam Gupta, Chetan Arora, Making third person techniques recognize first-person actions in egocentric videos, in: 2018 25th IEEE International Conference on Image Processing (ICIP), IEEE, 2018, pp. 2301–2305.

[203] Théo Voillemin, Hazem Wannous, Jean-Philippe Vandeborre, 2d deep video capsule network with temporal shift for action recognition, in: 2020 25th International Conference on Pattern Recognition (ICPR), IEEE, 2021, pp. 3513–3519.

[204] Heng Wang, Cordelia Schmid, Action recognition with improved trajectories, in: Proceedings of the IEEE International Conference on Computer Vision, 2013, pp. 3551–3558.

[205] Wei Wang, Vincent W Zheng, Han Yu, and Chunyan Miao. A survey of zero-shot learning: Settings, methods, and applications. ACM Transactions on Intelligent Systems and Technology (TIST), 10(2):1–37, 2019..

[206] Xiaohan Wang, Yu Wu, Linchao Zhu, and Yi Yang. Baidu-uts submission to the epic-kitchens action recognition challenge 2019. arXiv preprint arXiv:1906.09383, 2019..

[207] Xiaohan Wang, Yu Wu, Linchao Zhu, and Yi Yang. Symbiotic attention with privileged information for egocentric action recognition. arXiv preprint arXiv:2002.03137, 2020..

[208] Yaqing Wang, Quanming Yao, James T Kwok, Lionel M Ni, Generalizing from a few examples: A survey on few-shot learning, ACM Computing Surveys (CSUR) 53 (3) (2020) 1–34.

[209] Michael Wray and Dima Damen. Learning visual actions using multiple verb-only labels. arXiv preprint arXiv:1907.11117, 2019..

[210] Michael Wray, Diane Larlus, Gabriela Csurka, Dima Damen, Fine-grained action retrieval through multiple parts-of-speech embeddings, in: Proceedings of the IEEE International Conference on Computer Vision, 2019, pp. 450–459.

[211] Michael Wray, Davide Moltisanti, Dima Damen, Towards an unequivocal representation of actions, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2018, pp. 1127–1131.

[212] Michael Wray, Davide Moltisanti, Walterio Mayol-Cuevas, Dima Damen, in: Sembed: Semantic embedding of egocentric action videos In European Conference on Computer Vision, Springer, 2016, pp. 532–545.

[213] Michael Wray, Davide Moltisanti, Walterio Mayol-Cuevas, and Dima Damen. Improving classification by improving labelling: Introducing probabilistic multi-label object interaction recognition. arXiv preprint arXiv:1703.08338, 2017..

[214] SHI Xingjian, Zhourong Chen, Hao Wang, Dit-Yan Yeung, Wai-Kin Wong, and Wang-chun Woo. Convolutional lstm network: A machine learning approach for precipitation nowcasting. In Advances in Neural Information Processing Systems, pages 802–810, 2015..

[215] Yan Yan, Elisa Ricci, Gaowen Liu, Nicu Sebe, Recognizing daily activities from first-person videos with multi-task clustering, in: Asian Conference on Computer Vision, Springer, 2014, pp. 522–537.

[216] Yan Yan, Elisa Ricci, Gaowen Liu, Nicu Sebe, Egocentric daily activity recognition via multitask clustering, IEEE Transactions on Image Processing 24 (10) (2015) 2984–2995.

[217] Jen-An Yang, Chia-Han Lee, V. Shao-Wen Yang, Srinivasa Somayazulu, Yen-Kuang Chen, Shao-Yi Chien, Wearable social camera: Egocentric video summarization for social interaction, in: 2016 IEEE International Conference on Multimedia & Expo Workshops (ICMEW), IEEE, 2016, pp. 1–6.

[218] Lijin Yang. Egocentric action recognition from noisy videos. 2020..

[219] Siyuan Yang, Jun Liu, Lu. Shijian, Meng Hwa Er, Alex C Kot, Collaborative learning of gesture recognition and 3d hand pose estimation with multi-order feature analysis, in: European Conference on Computer Vision, Springer, 2020, pp. 769–786.

[220] Ryo Yonetani, Kris M Kitani, Yoichi Sato, Ego-surfing first-person videos, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 5445–5454.

[221] Ryo Yonetani, Kris M Kitani, Yoichi Sato, Recognizing micro-actions and reactions from paired egocentric videos, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 2629–2638.

[222] Ryo Yonetani, Kris M Kitani, Yoichi Sato, Visual motif discovery via first-person vision, in: European Conference on Computer Vision, Springer, 2016, pp. 187–203.

[223] Ryo Yonetani, Kris M Kitani, and Yoichi Sato. Ego-surfing: Person localization in first-person videos using ego-motion signatures. IEEE transactions on pattern analysis and machine intelligence, 40(11):2749–2761, 2017..

[224] Chen Yu and Dana H Ballard. Learning to recognize human action sequences. In Proceedings 2nd International Conference on Development and Learning. ICDL 2002, pages 28–33. IEEE, 2002..

[225] Yu. Chen, Dana H Ballard, Understanding human behaviors based on eye-head-hand coordination, in: International Workshop on Biologically Motivated Computer Vision, Springer, 2002, pp. 611–619.

[226] Yu. Haibin, Wenyan Jia, Zhen Li, Feixiang Gong, Ding Yuan, Hong Zhang, Mingui Sun, A multisource fusion framework driven by user-defined knowledge for egocentric activity recognition, EURASIP Journal on Advances in Signal Processing 2019 (1) (2019) 14.

[227] Yu. Haibin, Wenyan Jia, Li Zhang, Mian Pan, Yuanyuan Liu, and Mingui Sun. A hierarchical parallel fusion framework for egocentric adl recognition based on discernment frame partitioning and belief coarsening. Journal of Ambient Intelligence and Humanized, Computing (2020) 1–23.

[228] Yu. Haibin, Guoxiong Pan, Mian Pan, Chong Li, Wenyan Jia, Li Zhang, Mingui Sun, A hierarchical deep fusion framework for egocentric activity recognition using a wearable hybrid sensor system, Sensors 19 (3) (2019) 546.

[229] Yuan Yuan, Yang Zhao, Qi Wang, Action recognition using spatial-optical data organization and sequential learning framework, Neurocomputing 315 (2018) 221–233.

[230] Hasan FM Zaki, Faisal Shafait, and Ajmal Mian. Modeling sub-event dynamics in first-person action recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 7253–7262.

[231] Kai Zhan, Steven Faux, Fabio Ramos, Multi-scale conditional random fields for first-person activity recognition, in: 2014 IEEE international conference on pervasive computing and communications (PerCom), IEEE, 2014, pp. 51–59.

[232] Hong-Bo Zhang, Yi-Xiang Zhang, Bineng Zhong, Qing Lei, Lijie Yang, Du. Ji-Xiang, Duan-Sheng Chen, A comprehensive survey of vision-based human action recognition methods, Sensors 19 (5) (2019) 1005.

[233] Yun C Zhang, Yin Li, James M Rehg, First-person action decomposition and zero-shot learning, in: 2017 IEEE erence on Applications of Computer Vision (WACV), IEEE, 2017, pp. 121–129.

[234] Chengzhang Zhong, Amy R Reibman, Hansel Mina Cordoba, Amanda J Deering, Hand-hygiene activity recognition in egocentric video, in: 2019 IEEE 21st International Workshop on Multimedia Signal Processing (MMSP), IEEE, 2019, pp. 1–6.

[235] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, Antonio Torralba, Learning deep features for discriminative localization, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 2921–2929.

[236] Yang Zhou, Bingbing Ni, Richang Hong, Xiaokang Yang, Qi Tian, Cascaded interactional targeting network for egocentric video analysis, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 1904–1913.

[237] Yi Zhu, Zhenzhong Lan, Shawn Newsam, Alexander Hauptmann, Hidden two-stream convolutional networks for action recognition, in: Asian conference on computer vision, Springer, 2018, pp. 363–378.

[238] Zheming Zuo, Bo Wei, Fei Chao, Qu. Yanpeng, Yonghong Peng, Longzhi Yang, Enhanced gradient-based local feature descriptors by saliency map for egocentric action recognition, Applied System Innovation 2 (1) (2019) 7.

[239] Zheming Zuo, Longzhi Yang, Yonghong Peng, Fei Chao, Qu. Yanpeng, Gaze-informed egocentric action recognition for memory aid systems, IEEE Access 6 (2018) 12894–12904.

Adrián Núñez-Marcos is a PhD student in the University of Deusto. He is a BsC in Computer Science from the University of Basque Country (UPV/EHU), where he also obtained the MsC degree in Computational Engineering and Intelligent Systems. His research interests include computer vision and deep learning.

Gorka Azkune is an assistant professor in the University of Basque Country (UPV/EHU). He has published over 20 international peer-reviewed articles in journals and international conferences. He is a member of the IXA NLP group. His research interests include machine learning and multimodal deep learning. He received a PhD in Computer Science from the University of Deusto.

Ignacio Arganda-Carreras is an Ikerbasque Research Associate at the University of the Basque Country (UPV/EHU), in San Sebastian, Spain. His research interests include computer vision and bioimage analysis. He received a Ph.D. in computer science and electrical engineering from the Universidad Autonoma de Madrid, Spain.