An Overview of First Person Vision and Egocentric Video Analysis for Personal Mobile Wearable Devices

Alejandro Betancourt[1,2], Pietro Morerio[1], Carlo S. Regazzoni[1], and Matthias Rauterberg[2]

[1]Department of Naval, Electric, Electronic and Telecommunications Engineering DITEN - University of Genova, Italy

[2]Designed Intelligence Group, Department of Industrial Design, Eindhoven University of Technology, Eindhoven, Netherlands

Abstract—The emergence of new wearable technologies such as action cameras and smart glasses has increased the interest of the computer vision scientists in the First Person perspective. Nowadays, this field is attracting attention and investments of companies aiming to develop commercial devices with First Person Vision recording capabilities. Due to this interest, it is expected to have an increasing demand of methods to exploit these videos. The current methods to process these videos present particular combinations of different image features and quantitative methods to accomplish specific objectives like object detection, activity recognition, user machine interaction and so on. This paper summarizes the evolution of the state of the art in First Person Vision video analysis between 1997 and 2013, highlighting the most commonly used features, methods, challenges and opportunities of the field.

Index Terms—First Person Vision, Egocentric Vision, Wearable Devices, Smart Glasses, Computer Vision, Video Analytics, Human-machine Interaction.

1 Introduction

Portable head-mounted cameras able to record dynamic high quality first-person videos, have become a common item among sportsmen over the last five years. These devices represent the first commercial attempts to record experiences from a first-person perspective. This technological trend is a follow-up of the academic results obtained in the late 1990s, combined with the growing interest of the people to record their daily activities.

The idea of recording and analyzing videos from first person perspective is not new. To mention some examples: In 1998 Mann proposed the WearCam [1]. Later in 2000, Mayol et al. proposed a necklace device [2], and in 2005 Mayol et al. developed an active shoulder mounted camera [3]. In 2006, the Microsoft Research Center started to use the SenseCam for research purposes [4], while Pentland et al. [5] developed a wearable data collector system (InSense). Finally, it is important to highlight the work of Mann, who, since 1978, has been working on his own family of devices. A total of five generations have been proposed up to date, the last one being presented in [6].

> *"Lets imagine a new approach to computing in which the apparatus is always ready for use because it is worn like clothing. The computer screen, which also serves as a viewfinder, is visible at all times and performs multi-modal computing (text and images)".*

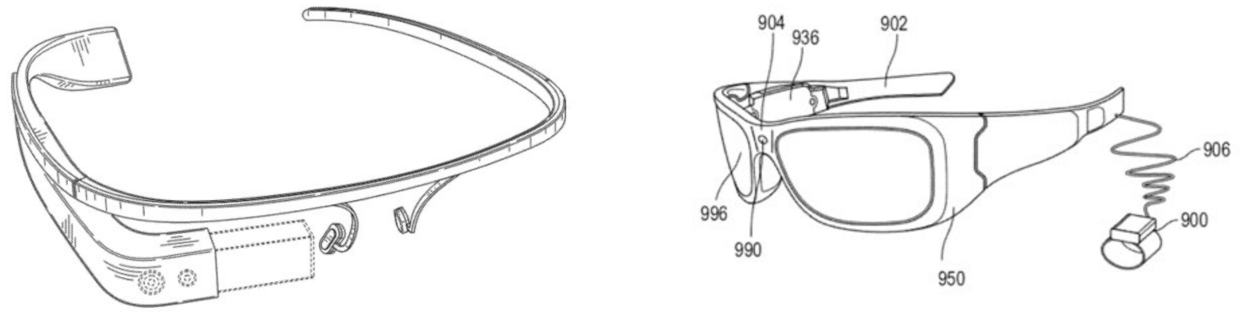
Up to date, no consensus has yet been reached in the literature with respect to naming this video perspective. *First Person Vision* (FPV) is arguably the most commonly used term, and will be used in the remainder of this paper. It is worth noting that the term *Egocentric Vision* has also recently grown in popularity, and can be used interchangeably with the FPV.

In the awakening of this technological trend, Google announced the Project Glass in 2012. The company started publishing short previews on the Internet demonstrating the Glasses FPV recording capabilities. This was coupled by the ability of the device to show relevant information to the user through the head-up display (Figure 1(a)). The main idea of the Project Glass is to use a wearable computer to reduce the time between intention and action [7].

By observing the trends in the number of apps dedicated to smartphones, it is feasible to predict a similar growth in the number of applications related to FPV. However, important challenges such as privacy issues should be solved before [8]. The range of potential applications is wide and involves different fields, e.g. military applications, enterprise applications, consumer behaviour [9], touristic purposes [10], massive surveillance [11], medical applications [12] to name a few.

This paper summarizes the state of the art in FPV video analysis, analyzing the challenges and opportunities cor-

alejandro.betancourt@ginevra.dibe.unige.it
pietro.morerio@ginevra.dibe.unige.it
carlo@dibe.unige.it
g.w.m.rauterberg@tue.nl



(a) Google glasses (U.S. Patent D659,741 - May 15, 2012)

(b) Microsof augmented reality glasses (U.S. Patent Application 20120293548 - Nov 22, 2012).

Fig. 1. Examples of the commercial smart patents. (a) Google patent of the smart glasses; (b) Microsoft patent of an augmented reality wearable device.

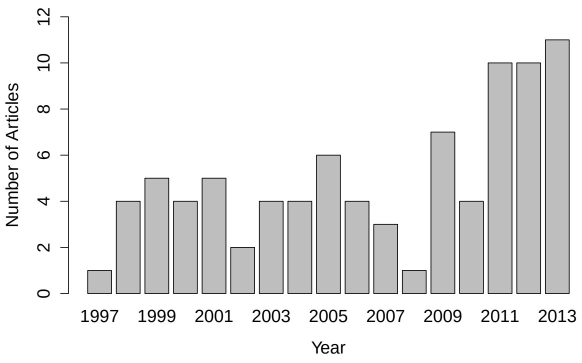


Fig. 2. Number of articles per year directly related to FPV video analysis.

responding to the production and usage of the relevant technological solutions. To the best of our knowledge, the only paper summarizing the general ideas of the FPV was [13], where a set of wearable devices were presented and the possible applications highlighted. Other related works included the following: [14] reviewed the activity recognition methods with multiple sensors; [15] analyzed the use of wearable cameras for medical applications; [3] analyzed some challenges of an active wearable device. Eventually, a remarkable work was carried out by the Media lab (MIT) in the late 1990s and early 2000s, foreseeing a clear research line in FPV [16, 17, 18, 19, 20, 21]. Figure 2 shows the growth in the number of articles related to FPV video analysis between 1997 and 2013.

In the remainder of this paper, we summarize existent methods in FPV according to a hierarchical structure. Section 2 introduces general characteristics of FPV and the hierarchical structure, which is later used to summarize the current methods according to the final objectives, subtasks performed and features used. In section 3 we briefly present the publicly-available FPV datasets. Finally, section 4 discusses some future challenges and research opportunities in this field.

2 First Person Vision (FPA) video analysis

During the late 1990s and early 2000s the advances in FPV analysis were mainly performed using highly elaborated devices, typically proprietary developed by the research groups themselves. The list of devices proposed during that time is wide and each device was usually presented in conjunction with their potential applications. However, the emergence of these devices showed a clear trend of embodiment with large arrays of sensors which only envy from modern devices in their design, size and commercial capabilities [6, 1, 2, 22, 3, 4, 23, 24]. Nowadays, current devices could be considered as the embodiment of the futuristic perspective of those studies [16, 17, 18, 19, 20, 21, 25].

Table 1 shows the currently-available commercial devices and their embedded sensors. These devices are presented in three groups: smart glasses, wearable cameras and eye trackers. Smart glasses have multiple sensors, processing capabilities and a head-up display. These groups of specifications make them ideal to develop new kind of methods and to improve the interaction between the user and its device. Smart glasses are nowadays seen as the starting point of an augmented reality system. On the other side, wearable cameras are commonly used by sportsmen and lifeloggers. However, the research community has been using them as a tool to develop methods and algorithms envisioning the commercial availability of smart glasses during the coming years. Finally, eye trackers have been successfully applied to analyze consumer behaviors in commercial environments.

In particular, FPV video analysis gives some methodological and practical advantages but also inherently

TABLE 1
Commercial approaches to wearable devices with FPV video recording capabilities

	Camera	Eye Tracking	Microphone	GPS	Accelerometer	Gyroscope	Magnetometer	Altitude	Light Sensor	Proximity Sensor	Body-Heat Detector	Temperature Sensor	Head-Up Display
Google Glasses	✓		✓	✓	✓	✓	✓		✓	✓			✓
Recon Jet	✓		✓	✓	✓	✓	✓					✓	✓
Vuzix M100	✓		✓		✓	✓	✓		✓	✓			✓
GlassUp	✓				✓	✓	✓		✓				✓
Meta	✓		✓	✓	✓	✓							✓
Iptinvent Ora-s	✓		✓	✓	✓	✓	✓		✓				✓
SenseCam	✓				✓			✓	✓		✓	✓	
Lumus	✓				✓	✓	✓						✓
Pivothead	✓												
GoPro	✓								✓				
Looxcie camera	✓												
Epiphany Eyewear	✓												
SMI Eye tracking Glasses	✓	✓											
Tobii	✓	✓	✓										

[1] Other projects such as *Nissan*, *Telepathy*, *Olympus MEG4.0*, *Oculon* and *Atheer* have been officially announced by their producers but no technical specifications have been already presented.

[2] According to unofficial online sources, other companies like *Apple*, *Microsoft*, *Samsung*, *Sony*, *Oakley* could be working on their own versions of similar devices, however no information has been officially announced up to date.

[2] This data is created on April 2014.

[3] In [13] one multi-sensor device is presented for research purposes.

bring a set of challenges that need to be addressed [13]. On one hand, FPV solves some problems of classic video analysis and offers extra information:

- *Videos of the main part of the scene:* Wearable devices allow the user to (even unknowingly) record the parts of the scene most relevant for the analysis, thus reducing the necessity for complex controlled multi-camera systems [26].
- *Variability of the datasets:* Due to the increasing commercial interest of the technology companies, a large number of FPV videos is expected in the future, making it possible for the researchers to obtain large datasets that differ among themselves significantly.
- *Illumination and scene configuration:* Changes in the illumination and global scene characteristics could be used as an important feature to detect the scene in which the user is involved, e.g. detecting changes in the place where the activity is held [27].
- *Internal state inference:* According to [28], the eye movements are directly influenced by the person's emotional state. In our opinion, this relationship could be bidirectional, making possible to estimate the user's internal state.
- *Object positions:* Because users tend to see the objects while interacting with them, it is possible to take advantage of the prior knowledge of the hands' and objects' positions, e.g. active objects tend to be closer to the center and hands tend to appear in the bottom left and bottom right part of the frames [29]

On the other hand FPV itself also presents some challenges:

- *Non static cameras:* One of the main characteristics of FPV videos is that cameras are always in movement. This fact makes it difficult to differentiate between the background and the foreground [30].
- *Illumination conditions:* The locations of the videos are highly variable and uncontrollable (e.g. visiting a touristic place during a sunny day, driving a car at night, brewing coffee in the kitchen). This makes it necessary to deploy robust methods for dealing with the variability in illumination.
- *Real time requirements:* One of the motivations for FPV video analysis is its potential of being used in various daily activities. This implies the need for the real time processing capabilities [31].
- *Video processing:* Due to the embedded processing capabilities of smart glasses, it is important to define optimal computational strategies to optimize the battery life, processing power and communication limits among the processing units. At this point, cloud computing could be seen as the most promising candidate tool to make from FPV video analysis an applicable framework for daily use.

In addition to the general opportunities and challenges presented above, there are other important aspects related to the objects that appear in FPV videos: i) The number of objects to be detected is variable and un-

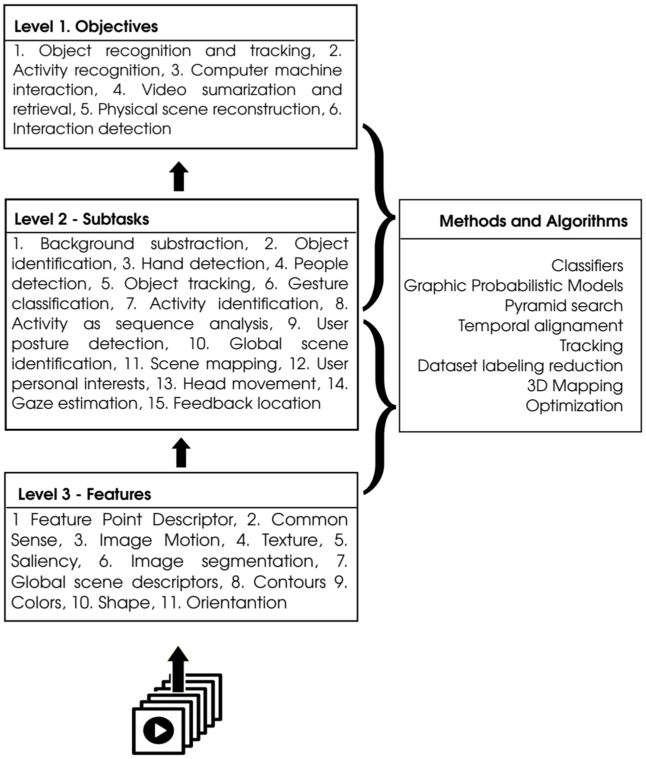

Level 1. Objectives
1. Object recognition and tracking, 2. Activity recognition, 3. Computer machine interaction, 4. Video sumarization and retrieval, 5. Physical scene reconstruction, 6. Interaction detection
Level 2 - Subtasks
1. Background substraction, 2. Object identification, 3. Hand detection, 4. People detection, 5. Object tracking, 6. Gesture classification, 7. Activity identification, 8. Activity as sequence analysis, 9. User posture detection, 10. Global scene identification, 11. Scene mapping, 12. User personal interests, 13. Head movement, 14. Gaze estimation, 15. Feedback location
Level 3 - Features
1 Feature Point Descriptor, 2. Common Sense, 3. Image Motion, 4. Texture, 5. Saliency, 6. Image segmentation, 7. Global scene descriptors, 8. Contours 9. Colors, 10. Shape, 11. Orientantion
Methods and Algorithms
Classifiers
Graphic Probabilistic Models
Pyramid search
Temporal alignament
Tracking
Dataset labeling reduction
3D Mapping
Optimization

Fig. 3. Hierarchical structure to explain the state of the art in FPV video analysis.

bounded (e.g. 1 hand and 1 object, 2 hands and 1 object, 2 hands and no other objects). ii) The type of objects is undefined, especially in uncontrolled environments [29]. iii) Objects appear at different scales (e.g. Looking at cars from a panoramic view is different that looking at cars while driving in a highway). iv) Objects could be occulted by other objects [32]. v) Objects appear in different states: ”A fridge looks different when its door is open“ [32].

The rest of this chapter summarizes FPV video analysis methods according to a hierarchical structure, as shown in Figure 3. In the top of the figure, general objectives and in the bottom the video sequence are shown. Section 2.1 summarizes the existent approaches according to 6 general objectives (Level 1). Section 2.2 divides those objectives in 15 weakly dependent subtasks (Level 2). Afterwards, section 2.3 briefly introduces the most used image features (Level 3). Finally, section 2.4 summarizes the quantitative and computational tools used to transform data from one level to the other. In the literature review, we found that existent methods are commonly presented as combinations of the aforementioned 4 parts. However, no standard structure is presented and it is difficult for other researchers to replicate existing methods or improve the state of the art. We propose this hierarchical structure as an effort to alleviate this issue.

2.1 Objectives

Table 2 summarizes a total of 85 articles. The articles are divided in six objectives according to the main task addressed in each of them. On the left side of the table the six objectives described in this section, and on the right side, extra groups related to hardware, software, related surveys and conceptual articles, are given. The category named Particular Subtasks is used for articles focused on one of the subtasks presented in section 2.2. The last column shows the positive trend in the number of articles per year. Last column is also plotted in Figure 2.

Note from the table that the most explored objective is *Object Recognition and Tracking*. We identify it as the base of more advanced objectives such as *Activity Recognition*, *Video Summarization and Retrieval* and *Physical Scene Reconstruction*. Another highly explored objective is *User-Machine Interaction* because of its potential in Augmented Reality. Finally, a recent research line denoted as Interaction detection allows the devices to infer the situations in which the user is involved. Along with this section, we present some details of how existent methods have addressed each of these 6 objectives. One important aspect is that some methods use multiple sensors in the framework of data-fusion systems. At the end of some objectives , several examples of data-fusion and multi-sensor approaches are mentioned.

2.1.1 *Object recognition and tracking*

Methods in this group address the problem of recognizing and tracking objects appearing in the user’s visual field. Those objects could be defined prior to the analysis, or could be arbitrary objects appearing in the recorded environment [84]. *Object recognition and tracking* is the most explored objective, and its results are commonly used as the inputs for more advanced inferences. One of the first applications of this objective is known as “the virtual augmented memory device”, a system able to differentiate objects and subsequently relate them to previous information and experiences [35, 38]. The virtual augmented memory used to be considered as a promising tool in the research community. However, due to its restrictions with respect to the number of objects, its real applicability has not been achieved yet.

The authors in [51] proposed a Bayesian method to recognize objects. This paper is commonly considered as a modern seminal paper in object recognition. Later, [29] published the first FPV challenging public dataset, in conjunction with a first benchmark of object identification. That baseline is known by other authors as the first benchmark for identification of handled objects. Afterwards, [66] showed that chaining background subtraction and object recognition is possible in order to achieve a better object identification. Later, [70] developed a

TABLE 2
Summary of the articles reviewed in FPV video analysis according to the main objective

	Objective						Extra Categories					
Year	Object Recognition and Tracking	Activity Recognition	User-Machine Interaction	Video Summarization and Retrieval	Physical Scene Reconstruction	Interaction Detection	Particular Subtasks	Related Software	Hardware	Conceptual Articles	Related Surveys	# Articles Reviewed
1997									[6]			1
1998	[17, 16, 33]	[16]							[1]			4
1999	[18, 34]	[18, 21]	[19]						[19]			5
2000	[35]	[36]	[37]						[2]			4
2001	[38]		[39, 40]	[41]			[42]					5
2002				[43, 44]				[44]				2
2003			[45]	[46, 47]						[11]		4
2004			[48, 49]	[22]				[22]	[50]			4
2005	[51, 52]	[51]	[53]	[54, 55]				[55]	[3]			6
2006		[23]	[56, 57]						[4, 23]			4
2007	[58, 59]				[58, 60, 59]							3
2008	[61]				[61]							1
2009	[29, 52]	[62, 63, 64]	[65]							[8]		7
2010	[66]	[12, 67]		[67, 68]								4
2011	[69, 70]	[71, 72]		[73, 71, 74]			[75, 76]		[24]		[14]	10
2012		[77, 32, 78, 30, 79, 80]	[81]			[26]	[82],		[13]		[13]	10
2013	[83, 84, 31, 85]	[86]	[10]	[27]		[87]			[88]	[7]	[15]	11

supervised method to reduce the video labeling requirements of the training stage. Recently, [32] highlighted the importance of taking into account the changes in the appearance of the objects when they are seen from different perspectives. In 2013, Wang et al. proposed an algorithm to recognize people in FPV videos [85].

A particular object to be detected and subsequently segmented is the user's hand. Hand detection and segmentation are widely studied fields. However, it is far from being considered a solved problem, particularly if changes in illumination and skin colors are taken into account. According to [51], hand detection could be divided into model and video based methods. In [51], a mixture of both approaches is presented. Model-based methods select the best matching configuration of hand model (2D or 3D) to recreate the image that is being shown in the video [89, 90, 91, 52]. These methods are able to infer detailed information of the hands such as approximate posture, but usually require large computational resources and are restricted to highly controlled environments

Video-based methods use image features such as color histograms to detect and segment the user's hands. These methods can be considered as the evolution of the pixel-by-pixel skin classifiers proposed in [92], in which color histograms are used to decide if a pixel can be classified as human skin or not. In [31], a mixture of color histograms and visual flow was proposed, while [10] used super-pixels and a random forest classifier. Recently, [84] analyzed the discriminative power of different features in pixel-by-pixel hand segmentation. Afterwards, the same authors proposed in [83] a model recommendation system which by using global features of the frames is able to decide the best hand-segmentator from a pool of models.

Once the objects are detected, it is possible to track their movements. In the case of the hands, some authors use as the reference point the coordinates of the hands [31], while others go a step further and use dynamic models [49, 57]. Dynamic models are widely studied and are successfully used to track hands, external objects [60, 58, 60, 61, 59], or faces of other people [33].

2.1.2 Activity recognition

An intuitive step following object recognition is to use the detected objects to infer the activity that the user is performing (e.g. making tea, cooking, watching tv, etc). In [64, 72], the relationship between activities and objects is explored and modeled using a Bayesian network. Later, [32] uses Support Vector Machines (SVM) to recognize the activities based only on the specific set of required objects. According to [32], ”it’s all about the objects being interacted with“. To support this argument the authors propose an experiment assuming the ground truth as perfect object detector.

Activity recognition could be considered as the field that has taken particular benefits of using multiple sensors. The use of multiple sensors started to grow in popularity with the work of Clarkson et al. Here, information from the FPV video and audio signals were fused to discriminate a set of basic activities [36, 34]. Later, [93] extended the methods presented in [94, 95, 96], by using Radio-Frequency Identification (RFID) to avoid explicit human labeling of the objects. The authors in [63] use Inertial Measurement Units. In [71, 68] the multiple sensors of the SenseCam are used to recognize activities in a large dataset recorded by multiple people. [1] Recently, [30] provided the theoretical formulation for adding GPS measurements to the task of activity recognition.

Eye trackers are other sensors that have been successfully fused with FPV videos for activity recognition. The power of this fusion is explained by the capability of eye trackers to differentiate between the users gaze and the camera view. In words of Yarbus [28], “eye movement reflects the human thought processes; so the observer’s thought may be followed to some extent from records of eye movement”. In [62], a mixture of eye-tracker, hand-tracker, and a FPV video is used to understand the sequence of actions performed by the user to accomplish an activity. Afterwards, [78] improved previous results in activity recognition by adding the user’s gaze to the model. Recently, different authors have worked on methods for inferring the gaze using FPV videos [76, 75, 86].

2.1.3 User-machine interaction

As mentioned before, smart glasses open the door to new ways of capturing interaction between the user and its device. As expected, this interaction has been addressed in two different ways, i.e. hardware and software. Due to the scope of this article, we are only interested in the software part, particularly, in the methods that are based on FPV video. In order to simplify the explanation of the current methods we divide this objective in two parts: i) when the user sends information to the device and ii) when the device gives a feedback to the user.

User-machine interaction using a camera is not a recent idea. However, the applicability of classical methods in FPV is limited, due to the uncontrolled characteristics of this video perspective [91]. First approaches in this field were motivated by the traditional way of user-computer interaction. In those studies, the main idea was to use the hands instead of a mouse to interact with the computer [37, 39].

Other approaches go one step ahead, looking for more intuitive ways of interaction using the hands as the input. [17] used a desk-mounted and a head-mounted camera to recognize American signal language with an efficiency of 92% and 98%, respectively. Later, [97] used the hand position captured by an external camera to create an interactive environment in conjunction with a head-up display. Recently, [10] proposed a gesture recognition method to interact with the device. The proposed method recognizes five gestures: point out, like, dislike, OK, victory. As it is evident, hand-tracking methods can give important cues in this objective [98, 49, 99, 100, 49, 57], and make it possible to use variables like position, speed or acceleration of the user’s hands.

In addition, the device needs to present relevant information to the user through the head-up display. An optimal design of this feedback system is important, and important issues must be considered. The authors in [45] analyzed this kind of interaction and highlighted that poorly designed feedback system could work against the user’s performance in addressing relevant tasks. In [65, 53, 81] the authors presented a group of methods to optimally locate virtual labels in the user’s visual field, without occluding the important parts of the scene.

2.1.4 Video summarization and retrieval

The objectives and issues of video summarization and retrieval is perfectly captured in [41] with the following sentences: We want to record our entire life by video. However, the problem is how to handle such a huge data. Purpose of the video summarization and retrieval is to explore and visualize the most important parts of large FPV video sequences [27]. This objective is commonly achieved under two different strategies, the first one consisting of software and databases to store the information, and the second one consisting of methods to summarize and explore the video sequences. The first strategy is out of the scope of this survey. However, interested reader is referred to [44, 22] for relevant examples.

For the second strategy, [74] proposed a method for extracting the important frames of the video using visual flows. Later, [30] selected a short subsequence of the video using the importance of the objects as the decision criteria. Recently, [27] improved the importance criterion of [30] by using the temporal relation of the objects appearing in the video.

Video summarization has been also improved by the use of multiple sensors. The authors in [41] use brain measurements to summarize FPV videos. Afterwards, [43] improved the work of [41] by using brain measurements only in the training phase. The use of other sensors such

[1] Wearable device developed by Microsoft Research in Cambridge, UK.

as gyroscopes, accelerometers, GPS , weather information and skin temperature, have been studied as well [46, 47, 54].

2.1.5 Physical scene reconstruction

This objective aims at reconstructing a 2D or 3D map of a scene recorded by a wearable camera. Physical scene reconstruction in FPV started to grow in popularity with [60], which showed how, by using multiple sensors, Kalman Filters and monoSLAM, it is possible to elaborate a map of the user environment. This work was extended in [58, 59] by adding object identification as a preliminary stage. Furthermore, the same authors in [61] used multiple cameras to improve the results. Recently, [13] found that, by adding a second camera to the smart device, it can be possible to reconstruct the 3D map of the scene. The author highlighted the potential application of his approach for blind navigation.

2.1.6 Interaction detection

The objectives described above are only focused on the user of the device as the only person that matters in the scene. However, none of them take into account the general situation in which the user is involved. We label the group of methods to recognize the types of interaction that the user is having with other people as “Interaction Detection”. The first approach in this group is social interaction detection as proposed by [26]. In their paper, the authors inferred the gaze of the other people and used it to recognize human interactions as monologues, discussions or dialogues. Another approach in this field was proposed by [87] which detected different behaviors of the people surrounding the user (e.g. hugging, punching, throwing objects, among others).

2.2 Subtasks

As explained before, the objectives are highly dependent among them. Moreover, it is common to find that the output of one objective is subsequently used as the input for the other (e.g. Activity recognition usually depends upon object recognition). Because of this, researchers usually address small subtasks and latter merge them to accomplish the main objective. Based on the literature review, we identify a total of 16 subtasks. Table 3 shows the number of articles that use a subtask (columns) in order to address a particular objective (rows). It is important to highlight the many-to-many relationship among objectives and subtasks, which means that a subtask could be used to address different objectives, and one objective could require multiple subtasks. As examples: i) hand detection as a subtask could be the objective itself in object recognition, [31], but could also give important cues in activity recognition, [78]; moreover, it could be the main input in the user-machine interaction [10]. ii) The authors in [32] performed object recognition to subsequently identify the activity.

For the sake of simplicity, we omit separate explanation of each of the considered subtasks. However, we feel that the names are self-explanatory, with the possible exceptions of the following: i) *Activity as a Sequence* analyzes an activity as a set of ordered steps; ii) *Scene mapping* builds a 2D or 3D representation of the recorded scene; iii) *User Personal Interests* identifies the interesting parts for the user in the video sequence; iv) *Feedback location* identifies the optimal place in the head-up display to locate the virtual feedback without interfering with the user visual field.

It is important to highlight the importance of Hand detection from Table 3, which provides validation for its use as the base for other methods. Global scene identification, as well as object identification stand out as two important subtasks for activity recognition. Particularly, object recognition supports the idea of [32], which states that activity recognition is “all about objects”. Finally, the use of gaze estimation in multiple objectives confirms the ad-vantages of using eye-trackers in conjunction with FPV videos.

2.3 Video and image features

As mentioned before, FPV implies dynamic changes in the attributes and characteristics of the scene. Due to these changes, an appropriate selection of the features to use becomes important. The use of a particular feature could imply a better performance - or a failure - in dealing with the challenges of this video perspective [101] (e.g changes in illumination, skin color variations, rotation, translations, changes in the movement of the camera [66, 29, 13]).

As is well known, feature selection is not a trivial task, and usually implies an exhaustive search in the literature to identify which of them lead to better results. Even more, the literature is full of particular modifications to deal with specific challenges. Table 4 shows the most commonly used features in FPV used to address a particular subtask. The features are listed in the rows and the subtasks in the columns. In general, all the features are extracted from the video sequence. However, we also show a category named “common sense”, which groups the recent practices of data mining and cloud workforce to extract information using the web [102, 30, 103, 93]. A common practice is to mix or chain multiple features to improve the performance or to accomplish independent subtasks.

Note from Table 4 that *color histograms* are by far the most commonly used feature for almost all the subtasks, despite being highly criticized due to their dependence on illumination and skin changes. Other group of features frequently used for different subtasks is the *image*

TABLE 3
Number of times that a subtask is performed to accomplish a specific objective

	Background Subtraction	Object Identification	Hand Detection	People Detection	Object Tracking	Gesture Identification	Activity Identification	Activity as Sequence Analysis	User Posture Detection	Global Scene Identification	Scene Mapping	User Personal Interests	Head Movement	Gaze Estimation	Feedback Location
(O1) Object recognition and tracking	4	1	7	2	8										
(O2) Activity recognition	1	5	3				6	2	1	8			2	4	
(O3) User-machine interaction			6		3	2									3
(O4) Video summarization and retrieval		2	1	4			2	1		4		1	1	1	
(O5) Physical scene reconstruction		3			4						4				
(O6) Interaction detection				1			2						1	1	

TABLE 4
Number of times that each features is used in each subtask

		Background Subtraction	Object Identification	Hand Detection	People detection	Object Tracking	Gesture Identification	Activity Identification	Activity as Sequence Analysis	User Posture Detection	Global Scene Identification	Scene Mapping	User Personal Interests	Head Movement	Gaze estimation	Feedback Location
Feature Point Descriptor	SIFT	2	1								1					
	BRIEF			1												
	SURF										2					
	ORB			1												
	STIP	1	1													
Common Sense	Priori spatial information			3											3	
	Objects required							1								
	Web mining							1								
	Amazon turk		1													
Image Motion	Optical flow	3		2		2		3			1			1	4	
	Motion vectors					1					1		1	1		
	Temporal templates			1												
Texture	Wiccest							1								
	Edge histogram												1			1
	Wavlets				1											
Saliency	GBVS														2	
	MSSS															1
Img. segmentation	Super-pixels		2	3												
	Blobs			1	1											
Glob. scene	GIST										1				1	
Contour	OWT-UCM	2	2													
Color	Histograms	3	6	17	4	4	1	2			6		1		1	
Shapes	HOG		2	2												
Orientation	Gabor							1								

motion. The use of *Feature Point Descriptors* (FPD) is also worth noting. As expected, FPD are popular for object identification. However, their application in FPV for global scene identification is of particular interest. Finally, *common sense* stands out as a useful strategy to take advantages of the internet information in order to reduce the training labeling requirements.

Table contains three columns without denoted features: this is because *Activity as a sequence* is usually chained with the output of short activity identification [21, 62, 73]. *Scene Mapping* initially detects the objects and subsequently maps them using a SLAM algorithm [59, 58, 61, 60]. Finally, *User Posture* is done in [23] using GPS and accelerometers.

2.4 Methods and algorithms

Once that features have been selected and estimated, the next step is to use them as inputs to reach the objective (outputs). At this point, quantitative methods start playing the main role, and as expected, an appropriate selection directly influences the quality of the results. Table 5 shows the number of occurrences of each method (rows) being used to accomplish a particular subtask (columns).

The table shows that the most commonly used methods are the *classifiers*, which are commonly used to assign a category to an array of characteristics (see [104] for a more detailed survey on classifiers). The use of classifiers is wide and varies from general applications, such as scene recognition [70], to more specific, such as activity recognition given a set of objects [78]. Particularly, we found that the most commonly used classifiers are the SVM and the k-means. Other group of tools that stands out are the *Probabilistic Graphical Models* (PGM), which are commonly used as a framework to combine multiple sensors or to chain results from different methods (e.g. to recognize the object and subsequently use it to infer the activity). Other popular methods are *trackers* such as Kalman filters and particle filters. These methods could also be represented as PGMs, but we present them as an extra category due to the specific subtasks that they address.

The use of machine learning methods introduces an important question: how to train the algorithms on realistic data without restricting their applicability? This question is widely studied in the field of Artificial Intelligence, and two different approaches are commonly followed, namely unsupervised and supervised learning [105]. Unsupervised learning requires less human interaction in training steps, but requires human interpretation of the results. Additionally, unsupervised methods have the advantage of being easily adaptable to changes in the video (e.g. new objects in the scene or uncontrolled environments [63]). Regarding the supervised methods, their results are easily interpretable but imply higher requirements in the training stage. Supervised methods use a set of inputs, previously labeled, to be trained and subsequently test their performance. Once the method is trained, it can be used on new instances without any additional human supervision. However, supervised methods are more dependent upon the training data, which could work against their performance when used on newly-introduced data [32, 27, 63, 26, 93, 30]. In order to reduce the training requirements and take advantage of the internet information, some authors create their datasets using services like Amazon Mechanical Turk [102, 30] or automatic web mining [103, 93].

Weakly supervised learning is another commonly used strategy, considered a middle point between supervised and unsupervised learning. This strategy is used to improve the supervised methods in two aspects: i) extend the capability of the classifier to deal with unexpected data; and ii) reduce the necessity for large training datasets. Following this, the authors of [67, 12] used Bag of Features (BoF) to monitor the activity of people with dementia. Later, [70, 72] used Multiple Instance Learning (MIL) to recognize objects using general categories. Afterwards, [73] used BoF and Vector of Locally Aggregated Descriptors (VLAD) to temporally align a sequence of videos.

3 Public Datasets

In order to support their results and create benchmarks in FPV video analysis, some authors have provided their datasets for public use to the academic community. The first publicly available FPV dataset is released by [51], and consists of a video containing 600 frames in a controlled office environment recorded from the left shoulder while the user interacts with five different objects. Later, [29] proposes a larger dataset with two people interacting with 42 object instances. The latter one is commonly considered as the first challenging FPV dataset because it guarantees the requirements identified by [19]: i) Scale and texture variations, ii) Frame resolution, iii) Motion blur and iv) Occlusion by hand.

Table 6 presents a list of the publicly-available datasets, along with their characteristics. Of particular interest are the changes in the camera location, which have evolved from shoulder-based to the head-based. These changes are clearly explained by the trend of the smart glasses and action cameras (see Table 1). Also noticeable are the changes in the objectives of the datasets, moving from low level such as object recognition, to more complex objectives like social interaction, gaze estimation and user-machine interaction. It should also be noted that less controlled environments have recently been proposed to improve the robustness of the methods in realistic situations. In order to highlight the robustness of their

TABLE 5
Mathematical and computational methods used in each subtask

	Background Subtraction	Object Identification	Hand Detection	People detection	Object Tracking	Gesture Identification	Activity Identification	Activity as Sequence Analysis	User Posture Detection	Global Scene Identification	Scene Mapping	User Personal Interests	Head Movement	Gaze estimation	Feedback Location
Classifiers	3	9	12	3		1	8			1		1		2	
Probabilistic Graphical Models	2	1				1	8	1		5				6	
Pyramid search		3	1												
Temporal alignment								1							
Tracking					7								1		
Dataset Labeling reduction		3								2					
SLAM methods											4				
Optimization															1

methods, several authors evaluated them on Youtube sequences recorded using goPro cameras [74].

Another aspect to highlight from the table is the availability of multiple sensors in some of the datasets. For instance, the Kitchen dataset [63] has four sensors, and the GTEA approach [78] has eye tracking measurements.

4 Challenges and future research

Wearable devices such as smart glasses will presumptively constitute a significant share of the technology market during the coming years, bringing new challenges and opportunities in video analytics. The interest in the academic world has been growing in order to satisfy the methodological requirements of this emerging technology. This survey provides a summary of the state of the art from the academic and commercial point of view, and summarizes the hierarchical structure of the existent methods. This paper shows the large number of developments in the field during the last 20 years, highlighting some of the up-coming lines of study.

From the commercial and regulatory point of view, important issues must be faced before the proper commercialization of this new technology can take place. Nowadays, the privacy of the recorded people is one of the most discussed, as these kinds of devices are commonly perceived as intruders [8]. Another important aspect is the legal regulations imposed on these devices, which commonly change depending on the country.

From the academic point of view, the research opportunities in FPV are wide. Under the light of this bibliographic review, we identify 4 interesting fields:

- Existent methods are proposed and executed in videos previously recorded. However, none of them seems to be able to work in a closed-loop fashion, by continuously learning from users experiences. We believe that a cognitive perspective could give important cues to this aspect and could aid the development of self-adaptive devices.
- The personalization capabilities of smart glasses open the door to new learning strategies. Incoming methods should be able to receive personalized training from the owner of the device. This kind of approach can help alleviating problems such as changes in the color skin.
- This survey points to the evident focus of current methods to address tasks accomplished by only one user and its wearable device. However, cooperative devices would be useful to increase the number of applications in areas such as scene reconstruction, military applications, cooperative games, sports and so on.
- Finally, regarding the processing units of the devices, important developments should be made in order to optimally compute FPV methods without draining the battery. Important cues to this problem could be found in cloud computing and high performance computing.

5 Acknowledgment

This work was supported in part by the Erasmus Mundus joint Doctorate in interactive and Cognitive Environments, which is funded by the EACEA, Agency of the European Commission under EMJD ICE.

TABLE 6
Current datasets and sensors availability

						Sensors					# Objects			Cam. Location		
		Year	Scene Recorded	Controlled Conditions	Objective	Video	IMUs	Body Media	eWatch	Eye Tracking	Activities	objects	people	Shoulder	Chest	Head
Hand Activity	[51]	2005	Desktop	✓	O1	✓						5	1	✓		
Intel	[29]	2009	Multiple locations		O1	✓						42	2	✓		
Kitchen.	[63]	2009	Kitchen recipes	✓	O2	✓	✓	✓	✓		3		43			✓
GTEA	[72]	2011	Kitchen recipes	✓	O2	✓					7		4			✓
VINST	[73]	2011	Going to the work		O2	✓							1		✓	
UEC Dataset	[74]	2011	Park		O2	✓					29		1			✓
ADL	[32]	2012	Daily activities		O2	✓					18		20		✓	
UTE	[30]	2012	Daily activities		O4	✓							4			✓
Disney	[26]	2012	Thematic park		O6	✓							8			✓
GTEA (Gaze, Gaze+)	[78]	2012	Kitchen recipes	✓	O2	✓				✓	7		10			✓
EDSH	[84]	2013	Multiple locations		O1	✓					-	-	-			✓
JPL FPV Int. D.S	[87]	2013	Office building		O6	✓					7		1			✓
EGO-HSGR	[10]	2013	Library		O3	✓					5		1			✓

REFERENCES

[1] S. Mann, ""WearCam" (wearable camera): personal imaging systems for long-term use in wearable tetherless computer-mediated reality and personal photo/videographic memory prosthesis," in *Wearable Computers*, (Pittsburgh), pp. 124–131, IEEE Comput. Soc, 1998.

[2] W. Mayol, B. Tordoff, and D. Murray, "Wearable Visual Robots," in *Digest of Papers. Fourth International Symposium on Wearable Computers*, (Atlanta GA), pp. 95–102, IEEE Comput. Soc, 2000.

[3] W. Mayol, A. Davison, B. Tordoff, and D. Murray, "Applying active vision and slam to wearables," *Springer Tracts in Advanced Robotics*, vol. 15, no. 1, pp. 325–334, 2005.

[4] S. Hodges, L. Williams, E. Berry, S. Izadi, J. Srinivasan, A. Butler, G. Smyth, N. Kapur, and K. Wood, "Sensecam: a Retrospective Memory Aid," in *Proceedings of the 8th International Conference of Ubiquitous Computing (UbiComp 2006)*, pp. 177–193, Springer Verlag, 2006.

[5] M. Blum, A. Pentland, and G. Troster, "Insense: Interest-based life logging," *MultiMedia, IEEE*, vol. 13, no. 4, pp. 40–48, 2006.

[6] S. Mann, "Wearable computing: A first step toward personal imaging," *Computer*, vol. 30, no. 2, pp. 25–32, 1997.

[7] T. Starner, "Project Glass: An Extension of the Self," *Pervasive Computing, IEEE*, vol. 12, no. 2, p. 125, 2013.

[8] D. Nguyen, G. Marcu, G. Hayes, K. Truong, J. Scott, M. Langheinrich, and C. Roduner, "Encountering SenseCam: personal recording technologies in everyday life," in *Proceedings of UbiComp*, 2009.

[9] C. Thompson, W. Locander, and H. Pollio, "Putting Consumer Experience Back into Consumer Research: The Philosophy and Method of Existential-Phenomenology," *Journal of Consumer Research*, vol. 16, p. 133, Sept. 1989.

[10] G. Serra, M. Camurri, and L. Baraldi, "Hand Segmentation for Gesture Recognition in Ego-vision," in *Proceedings of the 3rd ACM international workshop on Interactive multimedia on mobile & portable devices*, (New York, NY, USA), pp. 31–36, ACM Press, 2013.

[11] S. Mann, J. Nolan, and B. Wellman, "Sousveillance: Inventing and Using Wearable Computing Devices for Data Collection in Surveillance Environments.," *Surveillance & Society*, vol. 1, no. 3, pp. 331–355, 2003.

[12] S. Karaman, J. Benois-Pineau, R. Megret, V. Dovgalecs, J. Dartigues, and Y. Gaestel, "Human Daily Activities Indexing in Videos from Wearable Cameras for Monitoring of Patients with Dementia Diseases," in *International Conference on Pattern Recognition*, pp. 4113–4116, Ieee, Aug. 2010.

[13] T. Kanade and M. Hebert, "First-person vision," *Proceedings of the IEEE*, vol. 100, pp. 2442–2453, Aug. 2012.

[14] D. Guan, W. Yuan, A. Jehad-Sarkar, T. Ma, and Y. Lee, "Review of Sensor-based Activity Recognition Systems," *IETE Technical Review*, vol. 28, no. 5, p. 418, 2011.

[15] A. Doherty, S. Hodges, and A. King, "Wearable Cameras in Health," *American Journal of Preventive Medicine*, vol. 44, no. 3, pp. 320–323, 2013.

[16] T. Starner, B. Schiele, and A. Pentland, "Visual contextual awareness in wearable computing," in *In Digest of Papers. 2 nd International Symposium on Wearable Computers*, pp. 50–57, IEEE Comput. Soc, 1998.

[17] T. Starner, J. Weaver, and A. Pentland, "Real-Time American Sign Language Recognition Using Desk and Wearable Computer Based Video," *Transactions on Pattern Analysis and Machine Intelligence*, vol. 20, no. 12, pp. 1371–1375, 1998.

[18] B. Schiele, T. Starner, and B. Rhodes, "Situation aware computing with wearable computers," in *Augmented Reality and Wearable Computers*, pp. 1–20, 1999.

[19] B. Schiele, N. Oliver, T. Jebara, and A. Pentland, "An Interactive Computer Vision System - Dypers: Dynamic Personal Enhanced Reality System," in *Internation Conference on Vision Systems*, (Las Palmas de Gran Canaria, Spain), ICVS, 1999.

[20] T. Starner, *Wearable Computing and Contextual Awareness*. PhD thesis, Massachusetts Institute of Technology, 1999.

[21] H. Aoki, B. Schiele, and A. Pentland, "Realtime personal positioning system for a wearable computer," in *Wearable Computers, 1999. Digest of Papers*, (San Francisco, CA, USA), pp. 37–43, IEEE Comput. Soc, 1999.

[22] J. Gemmell, L. Williams, and K. Wood, "Passive Capture and Ensuing Issues for a Personal Lifetime Store," in *First ACM Workshop on Continuous Archival and Retrieval of Personal Experiences (CARPE '04)*, (New York, NY), pp. 48–55, 2004.

[23] M. Blum, A. Pentland, and G. Tröster, "InSense : Life Logging," *MultiMedia*, vol. 13, no. 4, pp. 40–48, 2006.

[24] M. Devyver, A. Tsukada, and T. Kanade, "A Wearable Device for First Person Vision," in *3rd International Symposium on Quality of Life Technology*, 2011.

[25] T. Starner, S. Mann, B. Rhodes, J. Levine, J. Healey, D. Kirsch, R. Picard, and A. Pentland, "Augmented reality through wearable computing," tech. rep., 1997.

[26] A. Fathi, J. Hodgins, and J. Rehg, "Social Interactions: A First-Person Perspective," in *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference*, (Providence, RI), pp. 1226–1233, IEEE, June 2012.

[27] Z. Lu and K. Grauman, "Story-Driven Summarization for Egocentric Video," in *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference*, (Portland, OR, USA), pp. 2714–2721, IEEE, June 2013.

[28] A. Yarbus, B. Haigh, and L. Rigss, *Eye Movements and Vision*. New York, New York, USA: Plenum Press, 1967.

[29] M. Philipose, "Egocentric recognition of handled objects: Benchmark and analysis," in *2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, (Miami, FL), pp. 1–8, IEEE, June 2009.

[30] J. Ghosh and K. Grauman, "Discovering Important People and Objects for Egocentric Video Summarization," in *In Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference*, pp. 1346–1353, IEEE, June 2012.

[31] P. Morerio, L. Marcenaro, and C. Regazzoni, "Hand Detection in First Person Vision," in *Fusion*, (Genova), pp. 0–6, University of Genoa, 2013.

[32] H. Pirsiavash and D. Ramanan, "Detecting Activities of Daily Living in First-Person Camera Views," in *Computer Vision and Pattern . . .* , pp. 2847–2854, IEEE, June 2012.

[33] G. Bradski, "Real time face and object tracking as a component of a perceptual user interface," in *Applications of Computer Vision*, pp. 14–19, 1998.

[34] B. Clarkson and A. Pentland, "Unsupervised Clustering of Ambulatory Audio and Video," in *Acoustics, Speech, and Signal Processing*, (Phoenix, AZ), pp. 1520–6149, IEEE, 1999.

[35] J. Farringdon and V. Oni, "Visual augmented memory (VAM)," in *2012 16th International Symposium on wearable computers*, no. 1997, (Atlanta GA), pp. 167–168, 2000.

[36] B. Clarkson, K. Mase, and A. Pentland, "Recognizing User Context Via Wearable Wensors," in *Digest of Papers. Fourth International Symposium on Wearable Computers*, (Atlanta, GA, USA), pp. 69–75, IEEE Comput. Soc, 2000.

[37] T. Kurata, T. Okuma, M. Kourogi, and K. Sakaue, "The hand-mouse: A human interface suitable for augmented reality environments enabled by visual wearables," in *Symposium on Mixed Reality*, pp. 4–5, 2000.

[38] T. Kurata, T. Okuma, and M. Kourogi, "VizWear: Toward human-centered interaction through wearable vision and visualization," *Lecture Notes in Computer Science*, vol. 2195, no. 1, pp. 40–47, 2001.

[39] T. Kurata and T. Okuma, "The Hand Mouse: Gmm Hand-color Classification and Mean Shift Tracking," in *Recognition, Analysis, and Tracking of Faces and Gestures in Real-Time Systems*, (Vancuver, Canada), pp. 119 – 124, IEEE, 2001.

[40] Y. Kojima and Y. Yasumuro, "Hand manipulation of virtual objects in wearable augmented reality," in *Virtual Systems and Multimedia*, pp. 463 – 469, 2001.

[41] K. Aizawa, K. Ishijima, and M. Shiina, "Summarizing wearable video," in *International Conference on Image Processing*, vol. 2, pp. 398–401, Ieee, 2001.

[42] M. Land and M. Hayhoe, "In What Ways Do Eye Movements Contribute to Everyday Activities?," *Vision research*, vol. 41, pp. 3559–65, Jan. 2001.

[43] Y. Sawahata and K. Aizawa, "Summarization of wearable videos using support vector machine," in *Proceedings. IEEE International Conference on Multimedia and Expo*, pp. 325–328, Ieee, 2002.

[44] J. Gemmell, R. Lueder, and G. Bell, "The MyLifeBits lifetime store," in *Proceedings of the 2003 ACM SIGMm*, (New York, New York, USA), pp. 0–5, ACM Press, 2002.

[45] R. DeVaul, A. Pentland, and V. Corey, "The memory glasses: subliminal vs. overt memory support with imperfect information," in *Seventh IEEE International Symposium on Wearable Computers*, pp. 146–153, Ieee, 2003.

[46] Y. Sawahata and K. Aizawa, "Wearable imaging system for summarizing personal experiences," in *Multimedia and Expo, 2003*, pp. I–45, Ieee, 2003.

[47] T. Hori and K. Aizawa, "Context-based video retrieval system for the life-log applications," in *Proceedings of the 5th ACM SIGMM international workshop on Multimedia information retrieval*, (New York, New York, USA), p. 31, ACM Press, 2003.

[48] R. Bane and T. Hollerer, "Interactive Tools for Virtual X-Ray Vision in Mobile Augmented Reality," in *Third IEEE and ACM International Symposium on Mixed and Augmented Reality*, no. Ismar, pp. 231–239, Ieee, 2004.

[49] M. Kolsch and M. Turk, "Fast 2d hand tracking with flocks of features and multi-cue integration," in *2004 Conference on Computer Vision and Pattern Recognition Workshop*, pp. 158–158, IEEE Comput. Soc, 2004.

[50] S. Mann, "Continuous lifelong capture of personal experience with Eye-Tap," in *Proceedings of the the 1st ACM workshop on Continuous archival and retrieval of personal experiences - CARPE'04*, (New York, New York, USA), pp. 1–21, ACM Press, 2004.

[51] W. Mayol and D. Murray, "Wearable Hand Activity Recognition for Event Summarization," in *Wearable Computers, 2005. Proceedings. Ninth IEEE International Symposium*, pp. 1–8, IEEE, 2005.

[52] L. Sun, U. Klank, and M. Beetz, "EYEWATCHME3D Hand and object tracking for inside out activity analysis," in *Computer Vision and Pattern . . .* , pp. 9–16, 2009.

[53] R. Tenmoku, M. Kanbara, and N. Yokoya, "Annotating user-viewed objects for wearable AR systems," in *Fourth IEEE and ACM International Symposium on Mixed and Augmented Reality*, pp. 192–193, Ieee, 2005.

[54] D. Tancharoen, T. Yamasaki, and K. Aizawa, "Practical experience recording and indexing of Life Log video," in *Proceedings of the 2nd ACM workshop on Continuous archival and retrieval of personal experiences*, (New York, New York, USA), p. 61, ACM Press, 2005.

[55] K. Aizawa, "Digitizing Personal Experiences: Capture and Retrieval of Life Log," in *11th International Multimedia Modelling Conference*, pp. 10–15, Ieee, 2005.

[56] R. Bane and M. Turk, "Multimodal Interaction with a Wearable Augmented Reality System," *Computer Graphics and Applications*, vol. 26, no. 3, pp. 62–71, 2006.

[57] M. Kölsch, R. Bane, T. Höllerer, and M. Turk, "Touching the visualized invisible: Wearable ar with a multimodal interface," in *IEEE Computer Graphics and Applications*, pp. 1–56, 2006.

[58] R. Castle, D. Gawley, G. Klein, and D. Murray, "Video-rate recognition and localization for wearable cameras," in *Procedings of the British Machine Vision Conference 2007*, (Warwick), pp. 112.1–112.10, British Machine Vision Association, 2007.

[59] R. Castle, D. Gawley, G. Klein, and D. Murray, "Towards simultaneous recognition, localization and mapping for hand-held and wearable cameras," in *Proceedings 2007 IEEE International Conference on Robotics and Automation*, pp. 4102–4107, Ieee, Apr. 2007.

[60] A. Davison, I. Reid, N. Molton, and O. Stasse, "MonoSLAM: Real-time Single Camera SLAM.," *IEEE transactions on pattern analysis and machine intelligence*, vol. 29, pp. 1052–67, June 2007.

[61] R. Castle, G. Klein, and D. Murray, "Video-rate localization in multiple maps for wearable augmented reality," in *2008 12th IEEE International Symposium on Wearable Computers*, pp. 15–22, Ieee, 2008.

[62] W. Yi and D. Ballard, "Recognizing behavior in hand-eye coordination patterns," *International Journal of Humanoid Robotics*, vol. 6, no. 3, pp. 337–359, 2009.

[63] E. Spriggs, F. De La Torre, and M. Hebert, "Temporal segmentation and activity classification from first-person sensing," in *2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, pp. 17–24, IEEE, June 2009.

[64] S. Sundaram and W. Cuevas, "High level activity recognition using low resolution wearable vision," *2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, pp. 25–32, June 2009.

[65] K. Makita, M. Kanbara, and N. Yokoya, "View management of annotations for wearable augmented reality," in *Multimedia and Expo*, pp. 982–985, Ieee, June 2009.

[66] X. Ren and C. Gu, "Figure-ground Segmentation Improves Handled Object Recognition in Egocentric Video," in *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 3137–3144, IEEE, June 2010.

[67] V. Dovgalecs, R. Megret, H. Wannous, and Y. Berthoumieu, "Semi-supervised learning for location recognition from wearable video," in *2010 International Workshop on Content Based Multimedia Indexing (CBMI)*, pp. 1–6, Ieee, June 2010.

[68] D. Byrne, A. Doherty, and C. Snoek, "Everyday concept detection in visual lifelogs: validation, relationships and trends," *Multimedia Tools and Applications*, vol. 49, no. 1, pp. 119–144, 2010.

[69] M. Hebert and T. Kanade, "Discovering Object Instances from Scenes of Daily Living," in *2011 International Conference on Computer Vision*, pp. 762–769, Ieee, Nov. 2011.

[70] A. Fathi, X. Ren, and J. Rehg, "Learning to recognize objects in egocentric activities," in *Computer Vision and Pattern Recognition (CVPR)*, (Providence, RI), pp. 3281–3288, IEEE, June 2011.

[71] A. Doherty and N. Caprani, "Passively recognising human activities through lifelogging," *Computers in Human Behavior*, vol. 27, no. 5, pp. 1948–1958, 2011.

[72] A. Fathi, A. Farhadi, and J. Rehg, "Understanding Egocentric Activities," in *Proceedings of the 2011 IEEE International Conference on Computer Vision (ICCV)*, pp. 407–414, IEEE, Nov. 2011.

[73] O. Aghazadeh, J. Sullivan, and S. Carlsson, "Novelty detection from an ego-centric perspective," in *IEEE Computer Vision and Pattern Recognition (CVPR) 2011*, (Colorado, United States), pp. 3297–3304, Ieee, June 2011.

[74] K. Kitani and T. Okabe, "Fast Unsupervised Ego-action Learning for First-person Sports Videos," in *Computer Vision and Pattern Recognition (CVPR)*, (Providence, RI), pp. 3241–3248, IEEE, June 2011.

[75] K. Yamada, Y. Sugano, and T. Okabe, "Can saliency map models predict human egocentric visual attention?," in *Computer VisionACCV*, pp. 1–10, 2011.

[76] K. Yamada, Y. Sugano, and T. Okabe, "Attention prediction in egocentric video using motion and visual saliency," in *Pacific-Rim Symposium on Image and Video Technology*, 2011.

[77] A. Borji, D. Sihite, and L. Itti, "Probabilistic Learning of Task-Specific Visual Attention," in *2012 IEEE Conference on Computer Vision and Pattern Recognition*, (Providence, RI), pp. 470–477, Ieee, June 2012.

[78] A. Fathi, Y. Li, and J. Rehg, "Learning to Recognize Daily Actions Using Gaze," in *12th European Conference on Computer Vision*, (Florence, Itaty), pp. 314–327, Georgia Institute of Technology, 2012.

[79] K. Kitani, "Ego-Action Analysis for First-Person Sports Videos," *Pervasive Computing*, vol. 11, no. 2, pp. 92–95, 2012.

[80] K. Ogaki, K. Kitani, Y. Sugano, and Y. 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*, pp. 1–7, Ieee, June 2012.

[81] R. Grasset, T. Langlotz, and D. Kalkofen, "Image-driven view management for augmented reality browsers.," in *ISMAR*, pp. 177–186, Ieee, Nov. 2012.

[82] K. Yamada, Y. Sugano, T. Okabe, Y. Sato, A. Sugimoto, and K. Hiraki, "Attention Prediction in Egocentric Video Using Motion and Visual Saliency," *Proceedings of the 5th Pacific Rim Conference on Advances in Image and Video Technology (PSIVT'11)*, pp. 277–288, 2011.

[83] C. Li and K. Kitani, "Model Recommendation with Virtual Probes for Egocentric Hand Detection," in *ICCV 2013*, (Sydney), IEEE Computer Society, 2013.

[84] C. Li and K. Kitani, "Pixel-Level Hand Detection in Ego-centric Videos," in *2013 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3570–3577, Ieee, June 2013.

[85] H. Wang and X. Bao, "InSight: recognizing humans without face recognition," in *Proceedings of the 14th Workshop on Mobile Computing Systems and Applications*, (New York, NY, USA), pp. 2–7, 2013.

[86] Y. Li, A. Fathi, and J. Rehg, "Learning to Predict Gaze in Egocentric Video," in *International Conference on Computer Vision*, pp. 1–8, Ieee, 2013.

[87] M. Ryoo and L. Matthies, "First-Person Activity Recognition: What Are They Doing to Me?," in *IEEE Conference on Computer Vision and Pattern Recognition*, (Portland, OR, US), pp. 2730–2737, IEEE Comput. Soc, 2013.

[88] S. Mann, M. Ali, R. Lo, and H. Wu, "FreeGlass for developers,haccessibility, and AR Glass+ LifeGlogging Research in a (Sur/Sous) Veillance Society," in *Information Society (i-Society)*, pp. 51–56, 2013.

[89] M. Schlattmann, F. Kahlesz, R. Sarlette, and R. Klein, "Markerless 4 Gestures 6 Dof Real-time Visual Tracking of the Human Hand with Automatic Initialization," *Computer Graphics Forum*, vol. 26, pp. 467–476, Sept. 2007.

[90] M. Schlattman and R. Klein, "Simultaneous 4 Gestures 6 Dof Real-time Two-hand Tracking Without Any Markers," in *Proceedings of the 2007 ACM symposium on Virtual reality software and technology*, (New York, NY, USA), pp. 39–42, ACM Press, 2007.

[91] J. Rehg and T. Kanade, "DigitEyes: Vision-Based Hand Tracking for Human-Computer Interaction," in *In Proceedings of the workshop on Motion of Non-Rigid and Articulated Bodies*, pp. 16–22, IEEE Comput. Soc, 1994.

[92] M. Jones and J. Rehg, "Statistical Color Models with Application to Skin Detection 2 Histogram Color Models," in *Computer Vision and Pattern Recognition*, (Fort Collins, CO), pp. 1–23, IEEE Computer Society, 1999.

[93] J. Wu and A. Osuntogun, "A Scalable Approach to Activity Recognition Based on Object Use," in *Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference*, (Rio de Janeiro), pp. 1–8, IEEE, 2007.

[94] N. Krahnstoever, J. Rittscher, P. Tu, K. Chean, and T. Tomlinson, "Activity Recognition using Visual Tracking and RFID," in *2005 Seventh IEEE Workshops on Applications of Computer Vision (WACV/MOTION'05)*, vol. 1, (Breckenridge, CO), pp. 494–500, IEEE, Jan. 2005.

[95] D. Patterson, D. Fox, H. Kautz, and M. Philipose, "Fine-Grained Activity Recognition by Aggregating Abstract Object Usage," in *Ninth IEEE International Symposium on Wearable Computers (ISWC'05)*, pp. 44–51, IEEE, 2005.

[96] M. Philipose and K. Fishkin, "Inferring Activities from Interactions with Objects," *Pervasive Computing, IEEE*, vol. 3, no. 4, pp. 50–57, 2004.

[97] M. Kolsch, M. Turk, and T. Hollerer, "Vision-based Interfaces for Mobility," in *Mobile and Ubiquitous Systems: Networking and Services*, pp. 86 – 94, Ieee, 2004.

[98] M. Morshidi and T. Tjahjadi, "Gravity optimised particle filter for hand tracking," *Pattern Recognition*, vol. 47, no. 1, pp. 194–207, 2014.

[99] V. Spruyt, A. Ledda, and W. Philips, "Real-time, long-term hand tracking with unsupervised initialization," in *Proceedings of the IEEE International Conference on Image Processing*, (Melbourne, Australia), IEEE Comput. Soc, 2013.

[100] C. Shan, T. Tan, and Y. Wei, "Real-time hand tracking using a mean shift embedded particle filter," *Pattern Recognition*, vol. 40, pp. 1958–1970, July 2007.

[101] O. Popoola, "Video-Based Abnormal Human Behavior RecognitionA Review," *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 42, pp. 865–878, Nov. 2012.

[102] M. Spain and P. Perona, "Measuring and Predicting Object Importance," *International Journal of Computer Vision*, vol. 91, pp. 59–76, Aug. 2010.

[103] T. Berg and D. Forsyth, "Animals on the Web," *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Volume 2 (CVPR'06)*, vol. 2, no. 1, pp. 1463–1470, 2006.

[104] D. Lu and Q. Weng, "A survey of image classification methods and techniques for improving classification performance," *International Journal of Remote Sensing*, vol. 28, pp. 823–870, Mar. 2007.

[105] F. Camastra and A. Vinciarelli, *Machine Learning for Audio, Image and Video Analysis: Theory and Applications*. Springer, 2007.