The Evolution of First Person Vision Methods: A Survey

A. Betancourt\textsuperscript{1,2}, P. Morerio\textsuperscript{1}, C.S. Regazzoni\textsuperscript{1}, and M. Rauterberg\textsuperscript{2}

\textsuperscript{1}Department of Naval, Electric, Electronic and Telecommunications Engineering DITEN - University of Genova, Italy
\textsuperscript{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 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, an increasing demand of methods to process these videos, possibly in real-time, is expected. Current approaches present a 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 2014, highlighting, among others, most commonly used features, methods, challenges and opportunities within 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. Up to now, no consensus has yet been reached in literature with respect to naming this video perspective. First Person Vision (FPV) is arguably the most commonly used, but other names, like Egocentric Vision or Ego-vision has also recently grown in popularity. The idea of recording and analyzing videos from this perspective is not new in fact, several such devices have been developed for research purposes over the last 15 years [1, 2, 3, 4, 5]. Figure 1 shows the growth in the number of articles related to FPV video analysis between 1997 and 2014. Quite remarkable is the seminal work carried out by the Media lab (MIT) in the late 1990s and early 2000s [6, 7, 8, 9, 10, 11], and the multiple devices proposed by Steve Mann who, back in 1997 [12], described the field with these words:

\begin{quote}
Let’s 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).
\end{quote}

Recently, in the awakening of this technological trend, several companies have been showing interest in this kind of devices (mainly smart-glasses), and multiple patents have been presented. Figure 1 shows the devices patented in 2012 by Google and Microsoft. Together with its patent, Google also announced Project Glass, as a strategy to test its device among a exploratory group of people. The project was introduced by showing short previews of the Glasses’ FPV recording capabilities, and its ability to show relevant information to the user through the head-up display.

Remarkably, the impact of the Glass Project (wich the most significant attempt to commercialize wearable technology up to date) is to be ascribed not only to its hardware, but also to the appeal of its underlying operating system. The latter continues to bring a large group of skilled developers, thus in turn making a significant boost in the number of prospective applications for smart-glasses, a phenomenon that has happened with smartphones several years ago. On one hand, the
This paper summarizes the state of the art in FPV video analysis and its temporal evolution between 1997 and 2014, analyzing the challenges and opportunities of this video perspective. It reviews the main characteristics of previous studies using tables of references, and the main events and relevant works using timelines. As an example, Figure 3 presents some of the most important papers and commercial announcements in the general evolution of FPV. We direct interested readers to the must read papers presented in this timeline. In the following sections, more detailed timelines are presented according to the objective addressed in the summarized papers. The categories and conceptual groups presented in this survey reflects our schematic perception of the field coming from a detailed study of the existent literature. We are confident that the proposed categories are wide enough to conceptualize existent methods, however due to the growing speed of the field they could require future updates. As will be shown in the coming sections, the strategies used during the last 20 years are very heterogeneous. Therefore, rather than provide a comparative structure between existing methods and features, the objective of this paper is to highlight common points of interest and relevant future lines of research. The bibliography presented in this paper is mainly in FPV. However, some particular works in classic video analysis are also mentioned to support the analysis. The latter are cited using italic font as a visual cue.

To the best of our knowledge, the only paper summarizing the general ideas of the FPV is [18], which presents a wearable device and several possible applications. Other related reviews include the following: [19] reviews the activity recognition methods with multiple sensors; [20] analyzes the use of wearable cameras for medical applications; [3] presents some challenges of an active wearable device.

In the remainder of this paper, we summarize existent methods in FPV, according to a hierarchical structure we propose, highlighting the more relevant works and the temporal evolution of the field. Section 2 introduces general characteristics of FPV and the hierarchical structure, which is later used to summarize the current methods according to their final objective, the sub-tasks performed and the 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 (FPV) Video Analysis

During the late 1990s and early 2000s, the advances in FPV analysis were mainly performed using highly elaborated devices, typically proprietarily developed by different research groups. The list of devices proposed is wide, where each device was usually presented in conjunction with their potential applications and a large array of sensors which only envy from modern devices in their design, size and commercial capabilities. The column “Hardware” in Table 2 summarizes these devices. The remaining columns of this table are explained in section 2.1. Nowadays, current devices could be considered as the embodiment of the futuristic perspective of the already mentioned pioneering studies. Table 1 shows the currently available commercial projects and their embedded sensors. Such devices are grouped in three categories:

- **Smart-glasses**: Smart-glasses have multiple sensors, processing capabilities and a head-up display, making them ideal to develop real time methods and to improve the interaction between the user and its device. Besides, smart-glasses are nowadays seen as the starting point of an augmented reality system. However, they cannot be considered a mature product until major challenges, such as battery life, price and target market, are solved. The future of these devices is promising, but it is still not clear if they will be adopted by the users on a daily basis like smartphones, or whether they will become specialized task-oriented devices like industrial glasses, smart-helmets, sport devices, etc.

- **Action cameras**: commonly used by sportsmen and lifeloggers. However, the research community has been using them as a tool to develop methods and algorithms while anticipating the commercial availability of the smart-glasses during the coming years. Action cameras are becoming cheaper, and are starting to exhibit (albeit still somewhat limited) processing capabilities.

- **Eye trackers**: have been successfully applied to analyze consumer behaviors in commercial environments. Prototypes are available mainly for research purposes, where multiple applications have been proposed in conjunction with FPV. Despite the potential of these devices, their popularity is highly affected by the price of their components and the obtrusiveness of the eye tracker sensors, which is commonly carried out using an eye pointing camera.

FPV video analysis gives some methodological and practical advantages, but also inherently brings a set of challenges that need to be addressed [18]. On one hand, FPV solves some problems of the classical video analysis and offers extra information:

- **Videos of the main part of the scene**: Wearable devices allow the user to (even unknowingly) record the most relevant parts of the scene for the analysis, thus reducing the
Steve Mann presents some experiments with wearable devices.

Media Lab (MIT) illustrates the potential of FPV playing "Patrol" (a real space-time game).

Explores the advantages of active wearable cameras.

Microsoft Research releases the SenseCam.

Show the strong relationship between gaze and FPV.

GoPro Hero release.

Google releases the Glass Project to a limited number of people.

Brief summary of the devices proposed by Steve Mann.

Some of the more important works and commercial announcements in FPV.

TABLE 1

<table>
<thead>
<tr>
<th>Camera</th>
<th>Eye Tracking</th>
<th>Microphone</th>
<th>GPS</th>
<th>Accelerometer</th>
<th>Gyroscope</th>
<th>Magnetometer</th>
<th>Altitude</th>
<th>Light Sensor</th>
<th>Proximity Sensor</th>
<th>Body-Heat Detector</th>
<th>Temperature Sensor</th>
<th>Head-Up Display</th>
</tr>
</thead>
<tbody>
<tr>
<td>Google Glasses</td>
<td>✔</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Epson Moverio</td>
<td>✔</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Recon Jet</td>
<td>✔</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Vuzix M100</td>
<td>✔</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GlassUp</td>
<td>✔</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Meta</td>
<td>✔</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Optinvent Ora-s</td>
<td>✔</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SenseCam</td>
<td>✔</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lumus</td>
<td>✔</td>
<td>✔</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pivothead</td>
<td>✔</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GoPro</td>
<td>✔</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Looxcie camera</td>
<td>✔</td>
<td>✔</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Epiphany Eyewear</td>
<td>✔</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SMI Eye tracking Glasses</td>
<td>✔</td>
<td>✔</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Tobii</td>
<td>✔</td>
<td>✔</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

1 Other projects such as Orcam, 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, Samsung, Sony, Oakley could be working on their own versions of similar devices, however no information has been officially announced up to date. Microsoft recently announced the Hololens but not technical specifications have been officially presented.

3 This data is created on January 2015.

In [18] one multi-sensor device is presented for research purposes.

necessity for complex controlled multi-camera systems [23].

• 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, as discussed in section 3.

• 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 taking place, as in [24].

• Internal state inference: According to [25], eye and head movements are directly influenced by the person’s emotional state. As already done with smartphones [26], this fact can be exploited to infer the user’s emotional state, and provide services accordingly.

• 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, whereas hands tend to appear in the bottom left and bottom right part of the frames [27, 28].

On the other hand, FPV itself also presents multiple challenges, which particularly affect the choice of the features to be extracted by low level processing modules (feature selection is discussed in details in section 2.3):

• 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 [29]. Camera calibration is not possible and often scale, rotation and/or translation-invariant features are required in higher level modules.

- **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. Here shape descriptors may be preferred to color-based features [28].

- **Real time requirements**: One of the motivations for FPV video analysis is its potential of being used for real time activities. This implies the need for the real time processing capabilities [30].

- **Video processing**: Due to the embedded processing capabilities (for smart-glasses), it is important to define efficient computational strategies to optimize 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 turn the FPV video analysis into an applicable framework for daily use. However, a real time cloud processing strategy requires further development in video compressing methods and communication protocols between the device and the cloud processing units.

The rest of this chapter summarizes FPV video analysis methods according to a hierarchical structure, as shown in Figure 4, starting from the raw video sequence (bottom) to the desired objectives (top). Section 2.1 summarizes the existent approaches according to 6 general objectives (Level 1). Section 2.2 divides these objectives in 15 weakly dependent subtasks (Level 2). Section briefly introduces the most commonly used image features, presenting their advantages and disadvantages, and relating them with objectives. Finally, section 2.4 summarizes the quantitative and computational tools used to process data, moving from one level to the other. In our literature review, we found that existing methods are commonly presented as combinations of the aforementioned levels. However, no standard structure is presented, making it difficult for other researchers to replicate existing methods or improve the state of the art. We propose this hierarchical structure as an attempt to cope with this issue.

### 2.1 Objectives

Table 2 summarizes a total of 117 articles. The articles are divided in six objectives according to the main goal addressed in each of them. The left side of the table contains 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, and is plotted in Figure 1.

Note from the table that the most commonly 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 **Environment Mapping**. Another often studied 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 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 within a data-fusion framework. For each objective, several examples of data-fusion and multi-sensor approaches are mentioned.

#### 2.1.1 Object recognition and tracking

**Object recognition and tracking** is the most explored objective in FPV, and its results are commonly used as a starting point for more advanced tasks, such as activity recognition. Figure 5 summarizes some of the most important papers that focused on this objective.

In addition to the general opportunities and challenges of the FPV perspective, this objective introduces important aspects to be considered: i) Because of the uncontrolled characteristics of the videos, the number of objects, as well as their type, scale and point of view, are unknown [27, 77]. ii) Active objects, as well as user’s hands, are frequently occluded. iii) Because of the mobile nature of the wearable cameras, it is not easy to create background-foreground models. iv) The camera location makes it possible to build a priori information about the objects’ position [27, 28].

Hands are among the most common objects in the user’s field of view, and a proper detection, localization, and tracking could be a main input for other objectives. The authors in...
### TABLE 2
Summary of the articles reviewed in FPV video analysis according to the main objective

<table>
<thead>
<tr>
<th>Year</th>
<th>Objective Extra Categories</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Object Recognition and Tracking</td>
</tr>
<tr>
<td>1997</td>
<td></td>
</tr>
<tr>
<td>1998</td>
<td>[7, 6, 31]</td>
</tr>
<tr>
<td>1999</td>
<td>[8, 32]</td>
</tr>
<tr>
<td>2000</td>
<td>[33]</td>
</tr>
<tr>
<td>2001</td>
<td>[36]</td>
</tr>
<tr>
<td>2002</td>
<td></td>
</tr>
<tr>
<td>2003</td>
<td></td>
</tr>
<tr>
<td>2004</td>
<td></td>
</tr>
<tr>
<td>2005</td>
<td>[49, 50]</td>
</tr>
<tr>
<td>2006</td>
<td></td>
</tr>
<tr>
<td>2007</td>
<td>[56, 57]</td>
</tr>
<tr>
<td>2008</td>
<td>[60]</td>
</tr>
<tr>
<td>2009</td>
<td>[27, 50]</td>
</tr>
<tr>
<td>2010</td>
<td>[65]</td>
</tr>
<tr>
<td>2011</td>
<td>[68, 69]</td>
</tr>
<tr>
<td>2012</td>
<td>[76, 77, 78, 29, 79, 80]</td>
</tr>
<tr>
<td>2013</td>
<td>[85, 86, 30, 87, 88]</td>
</tr>
<tr>
<td>2014</td>
<td>* [96, 97, 98, 99, 100, 101]</td>
</tr>
</tbody>
</table>

* [107, 108, 109, 110, 111, 112, 113, 114, 16]  
** [115, 116, 117, 118, 119, 120, 121]

[122] develops a pixel by pixel classifier to locate human skin in videos.  
[49] proposes the first dataset for object recognition in FPV.  
[69] uses Multiple Instance Learning to reduce the labeling requirements in object recognition.

Fig. 5. Some of the more important works in object recognition and tracking.

[122] highlight the difference between hand-detection and hand-segmentation, particularly in the framework of wearable devices where the number of deployed computational resources directly influences the battery life of the devices. In general, due to the hardware availability and price, hand-detection and tracking is usually carried out using RGB videos. However, [111, 112] uses a chest-mounted RGB-D camera to improve the hand-detection and tracking performance in realistic scenarios. According to [49], hand detection could be divided into model-driven and data-driven methods.

Model-driven methods search for the best matching configuration of a computational hand model (2D or 3D) to recreate the image that is being shown in the video [123, 124, 125, 50]. These methods are able to infer detailed information of the hands, such as the posture, but in exchange large computational resources, highly controlled environments or extra sensors (e.g. Depth Cameras) could be required.

Data-driven methods use image features to detect and segment users’ hands. The most commonly used features for this purpose are the color histograms looking to exploit the particular chromaticism of human skin, especially in suitable color spaces like HSV and YCbCr [30, 13, 85, 86]. Color-based methods can be considered as the evolution of the pixel-by-pixel skin classifiers proposed in [122], in which color histograms are used to...
decide whether a pixel represents human skin. Despite their advantages, the color-based methods are far from being an optimal solution. Two of their more important restrictions are:
i) The computational cost, because in each frame they have to solve the $O(n^2)$ problem implied by the pixel-by-pixel classification.
ii) Their results highly influenced by significant changes in illumination, for example indoor and outdoor videos[28]. To reduce the computational cost, some authors suggest the use of superpixels[13, 30, 86], however, an exhaustive comparison of the computational times of both approaches is still pending, and computationally efficient superpixel methods applied to video (especially FPV video) are still at an early stage[126].

Regarding the noisy results, the authors in [85, 13] train a pool of models and automatically select the most appropriate depending on the current environmental conditions.

In addition to hands, there is an uncountable number of objects that could appear in front of the user, whose proper identification could lead to development of some of the most promising applications of FPV. An example is “The Virtual Augmented Memory (VAM)” proposed by [33], where the device is able to identify objects, and to subsequently relate them to previous information, experiences or common knowledge available online. An interesting extension of the VAM is presented in [127], where the user is spatially located using his video, and is shown relevant information about the place or a particular event. In the same line of research, recent approaches have been trying to fuse information from multiple wearable cameras to recognize when the users are being recorded by a third person without permission. This is accomplished in [110, 128] using the motion of the wearable camera as the identity signature, which is subsequently matched in the third person videos without disclosing private information such as the face or the identity of the user.

The augmented memory is not the only application of object recognition. The authors in [77] develop an activity recognition method which based only a list of the used objects in the recorded video. Despite the importance of these applications, the problem of recognition is far from being solved due to the large amount of objects to be identified as well as the multiple positions and scales from which they could be observed. It is here that machine learning starts playing a key role in the field, offering tools to reduce the required knowledge about the objects [69] or exploiting web services (such as Amazon Turk) and automatic mining for training purposes [129, 29, 130, 58].

Once the objects are detected, it is possible to track their movements. In the case of the hands, some authors use the coordinates of center as the reference point [30], while others go a step further and use dynamic models [46, 55]. Dynamic models are widely studied and are successfully used to track hands, external objects [59, 56, 59, 60, 57], or faces of other people [31].

### 2.1.2 Activity recognition

An intuitive step in the hierarchy of objectives is Activity Recognition, aimed at identifying what the user is doing in a particular video sequence. Figure 6 presents some of the most relevant papers on this topic. A common approach in activity recognition is to consider an activity as a sequence of events that can be modeled as Markov Chains or as Dynamic Bayesian Networks (DBNs) [6, 8, 34, 5, 63]. Despite the promising results of this approach, the main challenge to be solved is the scalability to multiple user and multiple strategies to solve a similar task.

Recently, two major methodological approaches for activity recognition are becoming popular: object based and motion based recognition. Object based methods aim to infer the activity using the objects appearing in video sequence [63, 71, 77], assuming of course that the activities can be described by the required group of objects (e.g. prepare a cup of coffee requires coffee, water and a spoon). This approach opens the door to highly scalable strategies based on web mining to know the objects usually required for different activities. However, after all, this approach depends on a proper Object Recognition step and on its own challenges (Section 2.1.1). Following an alternative path, during the last 3 years, some authors have been using the fact that different kind of activities create different body motions and as consequence different motion patterns in the video, for example: walking, running, jumping, skiing, reading, watching movies, among others [73, 80, 99]. It is remarkable the discriminative power of motion features for this kind of activities and the robustness to deal with the illumination and the color skin challenges.

**Activity recognition** is one of the fields that has drawn most benefits from the use of multiple sensors. This strategy started growing in popularity with the seminal work of Clarkson et al. [34, 32] where basic activities are identified using FPV video jointly with audio signals. An intuitive realization of the multi-sensor strategy allows to reduce the dependency between Activity Recognition and Object Recognition, by using Radio-Frequency Identification (RFID) tags in the objects [58, 132, 133, 134]. However, the use of RFIDs reduces the applicability to environments previously tagged. The list of multiple sensors does not end with audio and RFIDs, it also contains Inertial Measurement Units [62], multiple sensors of the “SenseCam” [70, 67], GPS [29], and eye-trackers [61, 78, 83, 74, 89].

#### 2.1.3 User-machine interaction

As already mentioned, smart-glasses open the door to new ways for interaction between the user and his device. The device, being able to give feedback to the user, allows to close the interaction loop originated by the visual information captured and interpreted by the camera. Due to the scope of this paper, only approaches related to FPV video analysis are presented (we omit other sensors, such as audio and touch panels), categorizing them based on two approaches: i) the user sends information to the device, and ii) the device uses the information of the video to show the feedback to the user. Figure 7 shows some of the most important works concerning **User-machine interaction**.

In general, the interaction between the user and his device starts with intentional or unintentional command. An **intentional command** is a signal sent by the user using his hands through his camera. This kind of interaction is not a recent idea and several approaches have been proposed, particularly using static cameras [136, 137], which, as mentioned in section 2.1.1, can not be straightforwardly applied to FPV due to the mobile nature of wearable cameras. A traditional approach is

1. Wearable device developed by Microsoft Research in Cambridge with accelerometers, thermometer, infrared and light sensor
to emulate the mouse of computers with the hands [125, 35, 37], allowing the user to point and click at virtual objects created in the head-up display. Other approaches look for more intuitive and technology focused ways of interaction. For example, the authors in [13] develop a gesture recognition algorithm to be used in an interactive museum using 5 different gestures: “point out”, “like”, “dislike”, “OK” and “victory”. In [92], the head movements of the user are used to assist a robot in the task of finding a hidden object in a controlled environment. Under this perspective some authors combine static and wearable cameras [135, 7]. Quite remarkable are the results of Starner in 1998, being able to recognize American signal language with an efficiency of 98% with a static camera and head mounted camera. As is evident, hand-tracking methods can give important cues in this objective [138, 139, 140, 46], and make it possible to use features such as position, speed or acceleration of the users’ hands.

Unintentional commands are triggers activated by the device using information about the user without his conscious intervention, for example: i) the user is cooking by following a particular recipe (Activity Recognition), and the device could monitor the time of different steps without the user previously asking for it. ii) The user is looking at a particular item [Object Recognition] in a store [GPS or Scene Recognition] then the device could show price comparisons and reviews. Unintentional commands could be detected using the results of other FPV objectives, the measurements of its sensors, or behavioural routines learned from the user while previously using his device, among others. From our point of view, these kinds of commands could be the next step of user-machine interaction for smart-glasses, and a main enabler to reduce the required time to interact with the device [95].

Regarding the second part of the interaction loop, it is important to properly design the feedback system to know when, where, how, and which information should be shown to the user. In order to accomplish this, several issues must be considered in order to avoid misbehaviour of the system that could work against the user’s performance in addressing relevant tasks [42]. In this line, multiple studies develop methods to optimally locate virtual labels in the user’s visual field, without occluding the important parts of the scene [64, 51, 81].

2.1.4 Video summarization and retrieval

The main task of Video summarization and retrieval is to create tools to explore and visualize the most important parts of large FPV video sequences [24]. The objective and main issue is perfectly summarized in [39] with the following sentence: "We want to record our entire life by video. However, the problem is how to handle such a huge data". In general, existing methods define importance functions to select the more relevant subsequences or frames of the video, and later cut or accelerate the less important ones [119]. Recent studies define the importance function using the objects appearing in the video [29], their temporal relationships and causalities [24], or as a similarity function, in terms of its composition, between them and intentional pictures taken with a traditional cameras [115]. A remarkable result is achieved in [73, 99] using motion features to segment videos according to the activity performed by the user. This work is a good example of how to take advantage of the camera movements in FPV, usually considered as a challenge, to achieve good classification rates.

The use of multiple sensors is common within this objective, and remarkable fusions have been made using brain measurements in [39, 40], gyroscopes, accelerometers, GPS, weather information and skin temperature in [43, 44, 52], and online available pictures in [115]. An alternative approach to video
summarization is presented in [82] and [120], where multiple FPV videos of the same scene are unified using the collective attention of the wearable cameras as an importance function. In order to determine whether the two videos recorded from different cameras are pointing to the same scene, the authors in [141] use superpixels and motion features to propose a similarity measurement. Finally, it is significant to mention that “Video summarization and retrieval” has led to important improvements in the design of the databases and visualization methods to store and explore the recorded videos [41, 47]. In particular, this kind of developments can be considered an important tool for reducing computational requirements in the devices, as well as alleviate privacy issues related with the place where videos are stored.

2.1.5 Environment Mapping

Environment Mapping aims at the construction of a 2D or 3D virtual representation of the environment surrounding the user. In general, the of variables to be mapped can be divided in two categories: physical variables, such as walls and object locations, and intangible variables, such as attention points. Physical mapping is the more explored of the two groups. It started to grow in popularity with [59], which showed how, by using multiple sensors, Kalman Filters and monoSLAM, it is possible to elaborate a virtual map of the environment. Subsequently, this method was improved by adding object identification and location as a preliminary stage [56, 57]. Physical mapping is one of the more complex tasks in FPV, particularly when 3D maps are required due to the calibration restrictions. This problem can be partially alleviated by using a multi-camera approach to infer the depth [60, 18]. Research on intangible variables, can be considered an emerging field in FPV. Existent approaches define attention points and attraction fields, mapping them in rooms with multiple people interacting [120].

2.1.6 Interaction detection

The objectives described above are mainly focused on the user of the device as the only person that matters in the scene. However, they hardly take into account the general situation in which the user is involved. We label the group of methods aiming to recognize the types of interaction that the user is having with other people as Interaction Detection. One of the main purposes in this objective is social interaction detection, as proposed by [23]. 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 [93], which detected different behaviors of the people surrounding the user (e.g. hugging, punching, throwing objects, among others). Despite not being widely explored yet, this objective can be considered one of the most promising and innovative ones in FPV due to the mobility and personalization capabilities of the coming devices.

2.2 Subtasks

As explained before, the proposed structure is based on objectives which are highly co-dependent. 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 on object recognition). For this reason, a common practice is to first address small subtasks, and later merge them to accomplish main objectives. Based on the literature review, we propose a total of 15 subtasks. Table 3 shows the number of articles analyzed in this survey 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. To mention some: i) hand detection, as a subtask, could be the objective itself in object recognition, [30], but could also give important cues in activity recognition [78]; moreover, it could be the main input in the user-machine interaction [13]. ii) The authors in [77] performed object recognition to subsequently infer the performed activity. As we reckon that their names are self-explanatory, we omit separate explanation of each of the subtasks, with the possible exceptions of the following: i) Activity as a Sequence analyzes an activity as a set of ordered steps; ii) 2D-3D Scene Mapping builds a 2D or 3D virtual representation of the scene recorded; iii) User Personal Interests identifies the parts in the video sequence potentially interesting for the user using physiological signals such as brainwaves [40]; iv) Feedback location identifies the optimal place in the head-up display to locate the virtual feedback without interfering with the user’s visual field.

As can be deduced from table 3, Hand detection plays an important role as the base for advanced objectives such as Object Recognition and User-Machine interaction. Global scene identification, as well as Object Identification, stand out as two important subtasks for activity recognition. More in detail, the tight bound between the Activity Recognition and the Object Recognition supports the idea of [77], which states that Activity Recognition is “all about objects”. Moreover, the use of gaze estimation in multiple objectives confirms the advantages of the recent trend of using eye-trackers in conjunction with FPV videos. Finally, it can be noted that Background Subtraction has lost some of its reputation if compared with fixed camera scenarios, due to the highly unstable nature of the backgrounds when observed from the First-person perspective.

2.3 Video and image features

As mentioned before, FPV implies highly dynamic changes in the attributes and characteristics of the scene. Due to these changes, an appropriate selection of the features becomes critical in order to alleviate the challenges and exploit the advantages presented in section 2. As is well known, feature selection is not a trivial task, and usually implies an exhaustive search in the literature and extensive testing to identify which method leads to optimal results.

The process of feature extraction is carried out at different levels, starting from the pixel level, with color channels of the image, and subsequently extracting more elaborated indicators at the frame level, such as saliency, texture, superpixels, gradients, etc. As expected, these features can be used to address some of the subtasks, such as object recognition or scene identification. However, they do not include any kind of dynamic information. To add dynamic information in the analysis, different approaches can be followed, for example analyzing the geometrical transformation between two frames to obtain image Motion features such as optical flow, or aggregating frame
large illumination changes, in [28] we highlight how Color-
get close to real time applications. On the other side, under
to reduce the computational complexity of the methods and
that the “dynamic objectives” like
Activities Recognition
places such as touristic hotspots [72, 15, 66]. Note from the table
but it is also remarkable their application to identify relevant
noting. As expected, they are popular for object identification,
remarkable results are for
Activity Recognition
Image Motion
used for several subtasks is
Sensor that is able to identify the user recording the video in
Convolutional Neural Network (CNN) to create a biometric
in [119], and recently as the input of a
Video Summarization
In our previous works we performed a comparison of the
2.4 Methods and algorithms
Once that features are 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 qual-
ity of the results, ultimately showing whether the advantages
of the FPV perspective are being exploited, or whether the FPV-
related challenges are impacting the objectives negatively. Table
5 shows the number of occurrences of each method (rows)
being used to accomplish a particular objective or a subtask
(columns).

The table highlights classifiers as the most popular tool in
FPV, which is commonly used to assign a category to an
array of characteristics (see [142] for a more detailed survey
on classifiers). The use of classifiers is wide and varies from
general applications, such as scene recognition [69], to more
specific, such as activity recognition given a set of objects
[78]. Particularly, we found that the most used are the Support
Vector Machines (SVM) due to their capability to deal
with non-separable non-linear multi-label problems using low
computational resources. On the other hand, SVMs require
large labeled training sets which restricts the range of potential
applications.
In our previous works we performed a comparison of the
performance of multiple features (HOG, GIST, Color) and

---

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

<table>
<thead>
<tr>
<th>Objective</th>
<th>Background Subtraction</th>
<th>Object Identification</th>
<th>Hand Detection and Segmentation</th>
<th>People Detection</th>
<th>Object Tracking</th>
<th>Gesture Identification</th>
<th>Activity as Sequence Analysis</th>
<th>Activity Recognition</th>
<th>User Posture Detection</th>
<th>Global Scene Identification</th>
<th>2D-3D Scene Mapping</th>
<th>User Personal Interests</th>
<th>Head Movement</th>
<th>Gaze Estimation</th>
<th>Feedback Location</th>
</tr>
</thead>
<tbody>
<tr>
<td>Object Recognition and Tracking</td>
<td>4</td>
<td>15</td>
<td>13</td>
<td>3</td>
<td>10</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Activity Recognition</td>
<td>3</td>
<td>8</td>
<td>3</td>
<td>1</td>
<td>13</td>
<td>2</td>
<td>1</td>
<td>8</td>
<td>1</td>
<td>6</td>
<td>6</td>
<td>3</td>
<td>1</td>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td>User-Machine Interaction</td>
<td>6</td>
<td></td>
<td>3</td>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1</td>
<td>3</td>
<td></td>
<td>1</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>Video Summarization</td>
<td>1</td>
<td>4</td>
<td>1</td>
<td>4</td>
<td>5</td>
<td>1</td>
<td>4</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>Environment Mapping</td>
<td>3</td>
<td></td>
<td>4</td>
<td>5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>Interaction Detection</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td></td>
<td></td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
</tr>
</tbody>
</table>
### TABLE 4
Number of times that each feature is used in to solve an objective or subtask

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>FPD</td>
<td>SIFT</td>
<td>9</td>
<td>5</td>
<td>1</td>
<td>4</td>
<td>3</td>
<td>2</td>
<td>14</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>GFTT</td>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>BRIEF</td>
<td>2</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1</td>
<td></td>
<td></td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>FAST</td>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>SURF</td>
<td>2</td>
<td>6</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td></td>
<td>1</td>
<td>2</td>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Diff. of Gaussians</td>
<td>1</td>
<td>1</td>
<td></td>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1</td>
<td></td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>ORB</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1</td>
<td></td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Texture</td>
<td>Wiccest</td>
<td>1</td>
<td>1</td>
<td></td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1</td>
<td></td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Laplacian Transform</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1</td>
<td></td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Edge Histogram</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1</td>
<td></td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Wavlets</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1</td>
<td></td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Other</td>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1</td>
<td></td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Saliency</td>
<td>GBVS</td>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>4</td>
<td></td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Other</td>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1</td>
<td></td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>MSSS</td>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1</td>
<td></td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Motion</td>
<td>Optical Flow</td>
<td>5</td>
<td>1</td>
<td>2</td>
<td>5</td>
<td></td>
<td></td>
<td>1</td>
<td>2</td>
<td></td>
<td>6</td>
<td>1</td>
<td>4</td>
<td>5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Motion Vectors</td>
<td>1</td>
<td>1</td>
<td>3</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Temporal Templates</td>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1</td>
<td></td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Glob. Scene</td>
<td>CRFH</td>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1</td>
<td></td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>GIST</td>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1</td>
<td></td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Img. Segment.</td>
<td>Superpixels</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>3</td>
<td></td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Blobs</td>
<td>2</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1</td>
<td></td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Contour</td>
<td>OWT-UCM</td>
<td>1</td>
<td>1</td>
<td>3</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>2</td>
<td></td>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Color</td>
<td>Histograms</td>
<td>21</td>
<td>20</td>
<td>11</td>
<td>10</td>
<td>3</td>
<td>8</td>
<td>20</td>
<td>4</td>
<td>4</td>
<td>1</td>
<td>5</td>
<td>7</td>
<td>1</td>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Shapes</td>
<td>HOG</td>
<td>6</td>
<td>4</td>
<td>3</td>
<td>1</td>
<td>2</td>
<td>5</td>
<td>1</td>
<td>1</td>
<td>3</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Orientation</td>
<td>Gabor</td>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1</td>
<td></td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

classifiers (SVM, Random Forest, Random Threes) to solve the hand-detection problem [28]. Our conclusion was that HOG-SVM was the best performing combination, achieving a classification rate of 90% and 93% of true positives and true negatives respectively. Another group of methods commonly used are clustering algorithms due to its simplicity, computational cost, and small requirements in the training datasets. Despite their advantages, clustering algorithms could require post-processing analysis of the results in order to endow them with human interpretation. Another promising group of tools are the Probabilistic Graphical Models (PGMs), which can be interpreted as a framework to combine multiple sensors and chain results from different methods in a unique probabilistic hierarchical structure (e.g. to recognize the object and subsequently use it to infer the activity). Dynamic Bayesian Networks (DBNs) are a particular type of PGMs which include time in their structure, in turn making them suitable for application in video analysis [143]. As an example, DBNs are frequently used to represent activities as sequences of events [6, 8, 34, 5, 63]. It is common to find that particular methods, such as Dirichlet Process Mixture Models (DPMM), are presented in their PGM notation, however given the promising recent results achieved in Activity Recognition and Video Segmentation, we decided to group them separately.

As stated in section 2.3, there is a large number of features that can be extracted for FPV applications. A common practice
TABLE 5
Mathematical and computational methods used in objective or each subtask

<table>
<thead>
<tr>
<th>Objective</th>
<th>Subtasks</th>
</tr>
</thead>
<tbody>
<tr>
<td>Object Recognition and Tracking</td>
<td>Object Recognition</td>
</tr>
<tr>
<td>Activity Recognition</td>
<td>User Machine interaction</td>
</tr>
<tr>
<td>Environment Detection</td>
<td>Interaction Detection</td>
</tr>
<tr>
<td>Object Identification</td>
<td>Hand Detection and segmentation</td>
</tr>
<tr>
<td>Object Tracking</td>
<td>Gesture Classification</td>
</tr>
<tr>
<td>Gesture Classification</td>
<td>Activity Identification</td>
</tr>
<tr>
<td>Activity Identification</td>
<td>User Posture Detection</td>
</tr>
<tr>
<td>User Posture Detection</td>
<td>Global Scene Identification</td>
</tr>
<tr>
<td>Global Scene Identification</td>
<td>Feedback Location</td>
</tr>
</tbody>
</table>

| 3D Mapping | Classifiers | 3 | 1 | 4 |
| Clustering | 21 | 29 | 3 | 9 | 1 | 2 |
| Comon sense | 4 | 8 | 3 | 5 | 2 |
| DPMM | 1 | 1 | 1 | 1 |
| Feature Encoding | 4 | 6 | 3 | 3 |
| Optimization | 1 | 1 | 1 | 3 |
| PGM | 6 | 17 | 3 | 3 |
| Pyramid Search | 1 | 1 | 1 | 1 |
| Regressors | 1 | 1 | 1 | 1 |
| Temporal Alignment | 4 | 3 | 5 | 1 | 2 |

**PGM** Probabilistic Graphical Models.
**DPMM** Dirichlet Process Mixture Models.

is to mix or chain multiple features before using them as input of a particular algorithm (table 5). This practice usually results in extremely large vectors of features that can lead to computationally expensive algorithms. In this context, the role of Feature Encoding methods, such as Bag-of-Words, is crucial to control the size of the inputs. We highlight the importance that some authors are giving to this tool, which, despite not being an automatic strategy like Linear Discriminant Analysis (LDA) and Principal Components Analysis (PCA), can nevertheless help to include human intuition in the analysis. As an example, the authors in [97] use BoW in Activity Recognition taking into account the presence, level of attention, and the role of the objects in the video.

The use of machine learning methods (e.g. classifiers, clustering, regressors) introduces an important question to the analysis: 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 [144]. 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 [62]). The most commonly used unsupervised method in FPV are the clustering algorithms, such as k-means. In fact, the best performing superpixels are the result of an unsupervised clustering procedure applied over a raw image[145]. In [126] we proposed an optimization of the SLIC superpixels, and latter in [146] we introduced a new superpixel method based on Neural Networks. The proposed algorithm is a self-growing map that adapts its topology to the frame structure taking advantage of the dynamic information available in the previous frames.

Regarding the supervised methods, their results are easily interpretable but commonly imply higher requirements in the training stage. As an example, at the beginning of this section we highlighted some of the applications of SVMs. Supervised methods use a set of inputs, previously labeled, to parametrize the models. Once the method is trained, it can be used on new instances without any additional human supervision. In general, supervised methods are more dependent on the training data, fact which could work against their performance when used on newly-introduced cases [77, 24, 62, 23, 58, 29, 147]. In order to reduce the training requirements, and take advantage of the useful information available on Internet, some authors create their datasets using services like Amazon Mechanical Turk [129, 29], automatic web mining [130, 58], or image repositories [115]. We named this practice in table 5 as Common Sense.

**Weakly supervised learning** is another commonly used strategy, considered as a middle point between supervised and unsupervised learning. This strategy is used to improve the supervised methods in two aspects: i) extending the capability of the method to deal with unexpected data; and ii) reducing the necessity for large training datasets. Following this trend, the
authors of [66, 15] used Bag of Features (BoF) to monitor the activity of people with dementia. Later, [69, 71] used Multiple Instance Learning (MIL) to recognize objects using general categories. Afterwards, [72] used BoF and Vector of Locally Aggregated Descriptors (VLAD) to temporally align a sequence of videos. Eventually, let us mention Deep learning, a relatively recent approach which combines supervised and unsupervised learning techniques in a unified framework, where low level significant features are learned in an unsupervised fashion [148].

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 [49]. It consists of a video containing 600 frames recorded in a controlled office environment using a camera on the left shoulder, while the user interacts with five different objects. Later, [27] proposed 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 guaranteed the requirements identified by [9]: i) Scale and texture variations, ii) Frame resolution, iii) Motion blur, and iv) Occlusion by hand.

Implicitly, previous sections explain some of the main characteristics of FPV videos. In [149], these characteristics are compared for several FPV and Third Person Vision (TPV) datasets and their classification capabilities are evaluated. The authors reach a classification accuracy of 80.9% using blur, illumination changes, and optical flow as input features. In their study they also found a considerable difference in the classification rate explained by the camera position. The authors concluded that the more stable the camera, the less blur and motion and then the less discriminative power of these features. We highlight this difference as an important finding because it opens the door to an interesting discussion concerning which kind of videos, based on quantitative measurements, should be considered as FPV. Extra evidence about the role of the non-wearable cameras, such as hand-held devices when they are used to record from a first person perspective, is still pending. Our intuition points that, despite having some of the challenging characteristics of wearable cameras like mobile backgrounds and unstable motion patterns, hand-held videos would drastically differ in terms of features compared in [149].

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, such as social interaction 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 methods, several authors evaluated them on Youtube sequences recorded using goPro cameras [73].

Another aspect to highlight from the table is the availability of multiple sensors in some of the datasets. For instance, the Kitchen dataset [62] includes four sensors, the GTEA approach [78] includes eye tracking measurements, and the Egocentric Intel/Creative [111] was recorded with a RGBD camera.

4 Conclusion and future research

Wearable devices such as smart-glasses will presumably 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 main achievements and 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 ones, as these kinds of devices are commonly perceived as intruders [17]. Other important aspects are the legal regulations depending on the country, and the intention of the user to avoid recording private places or activities[113]. Another hot topic is the real applicability of smart-glasses as a massive consumption device or as a task-oriented tool to be worn only in particular scenarios. In this field, the technological companies are designing their strategies in order to reach out to specific markets. As an illustration, recent turn of events has seen Google move out of the glass project (originally intended to end with a massively commercialized product), in order to target the enterprise market. Microsoft, on the other hand, recently announced its task-oriented holographic device “HoloLens” embodied with a larger array of sensors.

From the academic point of view, the research opportunities in FPV are still wide. Under the light of this bibliographic review and our personal experience, we identify 4 main hot topics:

- Existing methods are proposed and executed in previously recorded videos. However, none of them seems to be able to work in a closed-loop fashion, by continuously learning from users’ experiences and adapt to the highly variable and uncontrollable surrounding environment. From our previous studies [150, 151], we believe that a cognitive perspective could give important cues to this aspect and could aid the development of the 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. We have found out, for instance, that this kind of approach can help alleviate problems, such as changes in the color skin models from different users [30] in a hand detection application. Indeed, color features, as stressed in 4, has proven to be extremely suitable to be exploited in this field.

- This survey focuses on methods for addressing tasks accomplished mainly by one user coupled with a single wearable device. However, cooperative devices would be useful to increase the number of applications in areas such as environment mapping, military applications, cooperative games, sports, etc.
Finally, regarding the real time requirements, important developments should be made in order to optimally compute FPV methods without draining the battery. This must be accomplished both from the hardware and the software side. On the one hand, progress still needs to be made on the processing units of the devices. On the other, lighter, faster and better optimized methods are yet to be designed and tested. Our personal experience lead us to explore fast machine learning methods [28] for hand detection, in the trend highlighted by table 5, and to discard standard features such as optic flow [30] because of computational restrictions. Promising methods in standard computer vision research, such as superpixel methods, were built from scratch in [146] in order to make them faster and better suited for video analysis [126]. Eventually, important cues to the problem of computational power optimization may also be found in cloud computing and high performance computing.

**References**


Hand: A Probabilistic Approach to Hand Disambiguation in Egocentric
Video,” in Computer Vision and Pattern Recognition, (Columbus, Ohio),

[105] S. Han, M. Filippou, A. Krishnamurthy, and D. Wetherall, “GlimpseData:
Towards Continuous Vision-Based Personal Analytics,” Journal of Experi-
mental Psychology: Human Perception and Performance, vol. 40, no. 2, p. 457,
2014.


[107] J. Wang and C. Yu, “Finger-fist Detection in First-person View Based on
Monocular Vision Using Haar-like Features,” in Chinese Control Conference,

with Novel Set-of-Sets Representations,” in Computer Vision and Pattern

[109] S. Feng, R. Caire, B. Cortazar, M. Turan, A. Wong, and A. Ozcan, “Im-
munochromatographic Diagnostic Test Analysis Using Google Glass.,”

[110] Y. Poley, C. Arora, and S. Peleg, “Head Motion Signatures from Egocentric
Videos,” in Asian Conference on Computer Vision, (Singapore), Springer,
2014.

Images,” in ECCV Workshop on Consumer Depth Camera for Computer Vision,
2014.

in Four Lines of Code,” in Computer Vision and Pattern Recognition, 2015.

[113] R. Tempieman, M. Korayem, D. Crandall, and K. Apgu, “PlaceAvoider:
Steering first-person cameras away from sensitive spaces,” in Network and

Salience Models for Active Object Recognition in Egocentric Videos,” in
International Workshop on Perception Inspired Video Processing, (New York,

[115] B. Xiong and K. Grauman, “Detecting Snap Points in Egocentric Video
with a Web Photo Prior,” in International Conference on Computer Vision,
2014.

[116] W. Min, X. Li, C. Tan, B. Mandal, L. Li, and J. H. Lim, “Efficient Retrieval
from Large-Scale Egocentric Visual Data Using a Sparse Graph Representa-
tion,” in Computer Vision and Pattern Recognition Workshops, pp. 541–548,
Ieee, 2014.

[117] M. Bolanos, M. Carrozza, and P. Radeva, “Video Segmentation of Life-
Logging Videos,” in Articulated Motion and Deformable Objects, pp. 1–9, 2014.

Mixtures for Egocentric Video Segmentation,” in Computer Vision and

for Generating Walking Route Guidance,” in Computer Vision and Pattern
Recognition using Visual Tracking and RFID, in IEEE Workshops on
Applications of Computer Vision (WACV/MOTION’05), vol. 1, (Breckenridge,

[120] D. Patterson, D. Fox, H. Kautz, and M. Philipose, “Fine-Grained Activity
Recognition by Aggregating Abstract Object Usage,” in Ninth IEEE Inter-
national Symposium on Wearable Computers (ISWC’05), pp. 44–51, IEEE,
2005.

[121] M. Philipose and K. Fishkin, “Inferring Activities from Interactions with Objects,”

in Mobile and Ubiquitous Systems: Networking and Services, pp. 86 – 94, Ieee,
2004.

[123] G. Riva, F. Vatalaro, F. Davide, and M. Alcainz, eds., Ambient Intelligence –
The Evolution of Technology, Communication and Cognition Towards the Future

computer interaction applications of self-growing models,” Applied Soft
Information System Security.

[125] M. Morshidi and T. Tjahjadi, “Gravity Optimised Particle Filter for Hand

with Unsupervised Initialization,” in International Conference on Image

[127] C. Shan, T. Tan, and Y. Wei, “Real-time Hand Tracking Using a Mean Shift
2007.


[129] D. Lu and Q. Weng, “A Survey of Image Classification Methods and
Techniques for Improving Classification Performance,” International Journal

Switched Dynamic Bayesian Networks for Autonomous Cognitive Crowd
Monitoring,” in Augmented Vision and Reality, Augmented Vision and

[131] F. Camasra and A. Vinciarelli, Machine Learning for Audio, Image and Video

[132] A. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Susstrunk,
“Slic superpixels compared to state-of-the-art superpixel methods,” Pattern
Analysis and Machine Intelligence, IEEE Transactions on, vol. 34, no. 11,

Mixtures for Egocentric Video Segmentation,” in Computer Vision and

Detection 2 Histogram Color Models,” in Computer Vision and Pattern

tures 6 Dof Real-time Video Tracking of the Human Hand with Automatic

Two-hand Tracking Without Any Markers,” in Symposium on Virtual Reality

Human-Computer Interaction,” in Workshop on Motion of Non-Rigid and

superpixel clustering for real-time egocentric-vision applications,” IEEE
**Alejandro Betancourt** Alejandro Betancourt is PhD candidate of the Interactive and Cognitive Environment program between the Universita degli Studi di Genova and the Eindhoven University of Technology. Alejandro is a Mathematical Engineer and Master in Applied Mathematics from EAFIT University (Medellin, Colombia). Since 2011 Alejandro has been involved in research about Artificial Intelligence, Machine Learning and Cognitive Systems.

**Pietro Morerio** Pietro Morerio received his B.SC in Physics from the Faculty of Science, University of Milan (Italy) in 2007. In 2010 he received from the same university his M. Sc. in Theoretical Physics (summa cum laude). He was Research Fellow at the University of Genoa (Italy) from 2011 to 2012, working in Video Analysis for Interactive Cognitive Environments. Currently, he is pursuing a PhD degree in Computational Intelligence at the same institution.

**Matthias Rauterberg** Matthias Rauterberg is professor at the department of Industrial Design and the head of the Designed Intelligence group at Eindhoven University of Technology (The Netherlands). Matthias received the B.S. in Psychology (1978) at the University of Marburg (Germany), the B.S. in Philosophy (1981) and Computer Science (1983), the M.S. in Psychology (1981, summa cum laude) and Computer Science (1986, summa cum laude) at the University of Hamburg (Germany), and the Ph.D. in Computer Science/Mathematics (1995, awarded) at the University of Zurich (Switzerland).

**Carlo Regazzoni** Carlo S. Regazzoni received the Laurea degree in Electronic Engineering and the Ph.D. in Telecommunications and Signal Processing from the University of Genoa (UniGE), in 1987 and 1992, respectively. Since 2005 Carlo is Full Professor of Telecommunications Systems. Dr. Regazzoni is involved in research on Signal and Video processing and Data Fusion in Cognitive Telecommunication Systems since 1988.