11-2016

Summarization of egocentric videos: A comprehensive survey

Ana GARCIA DEL MOLINO

Cheston TAN

Joo-Hwee LIM

Ah-hwee TAN
Singapore Management University, ahtan@smu.edu.sg

Summarization of Egocentric Videos: A Comprehensive Survey

Ana Garcia del Molino, Cheston Tan, Joo-Hwee Lim, and Ah-Hwee Tan,

***Abstract*—The introduction of wearable video cameras (e.g. GoPro) in the consumer market has promoted video life-logging, motivating users to generate large amounts of video data. This increasing flow of first-person video has led to a growing need for automatic video-summarization adapted to the characteristics and applications of egocentric video. With this paper, we provide the first comprehensive survey of the techniques used specifically to summarize egocentric videos. We present a framework for first-person view summarization, and compare the segmentation methods and selection algorithms used by the related work in the literature. Next, we describe the existing egocentric video datasets suitable for summarization, and then the various evaluation methods. Finally, we analyze the challenges and opportunities in the field, and propose new lines of research.**

***Index Terms*—Egocentric Vision, First Person View, Survey, Video Summarization**

I. Introduction

A LOT of things have changed since Steve Mann introduced his wearable camera to the community in the 1990s [1]. Wearable devices, from smart wristbands to smart glasses, are not only developed by and for researchers anymore, as a consumer market has emerged and grown steadily in recent years. The affordability of devices such as the Narrative Clip and GoPro cameras allows mass-market consumers to continuously record for many hours, producing huge amounts of unconstrained data. However, the device wearer (the person recording the video) may never revisit much of those recorded visual memories, and the few important episodes could be hidden among many repetitive images or long uninteresting segments.

Thus, it is clear that if we want wearable video devices to be really attractive to the potential consumer, there is a need to identify and locate those meaningful and interesting segments and make browsing and retrieving fast and efficient, or even piece segments together into a coherent summary for a better story- telling experience. This issue has been addressed in different ways since Lifelogging (the practice of continuously capturing and recording images and videos of one's life) was first introduced. Whereas some researchers target the management of such large amounts of data by providing indexing and retrieval systems [2–9], others try to summarize the content of the videos or image sets, so that the user can appreciate the overall meaning and experience of the recorded memory in a much shorter time [10–30]. However, even though retrieval can be used to provide personalized summaries, its use as a tool for summarization is not well explored yet.

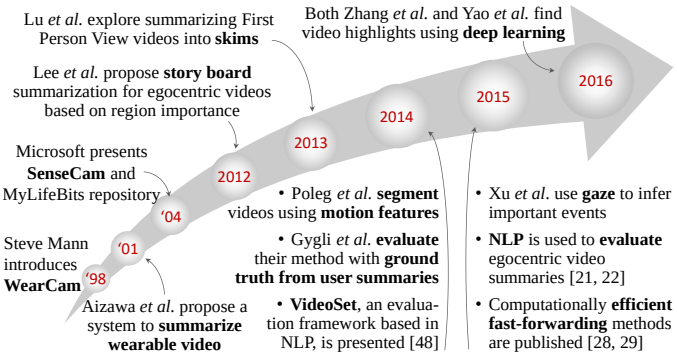

Fig. 1. Milestones in First Person View video summarization.

Two recent surveys [31, 32] study, respectively, the methods used to summarize sets of egocentric pictures; and the state-of-the-art for six different egocentric objectives and their sub-tasks, such as *Object Recognition and Tracking, Activity Recognition*, and *Interaction Detection*. Bolanos et al. [31] review the different approaches for storytelling through Lifelogging (with cameras typically taking 2 pictures per minute), whereas Betancourt et al. [32] review the First Person View (FPV) video summarization problem briefly in one subsection.

We expand on their work by providing an extensive analysis of FPV video summarization approaches. Even though the existing literature regarding non-FPV summarization is already documented ([33, 34]), the specific characteristics of egocentric videos make such Third Person View (TPV) techniques inapplicable to FPV videos, as described in section II-B.Fig. 1 presents schematically the milestones achieved in FPV video summarization, showing the increasing interest that the field has arisen lately.

In this paper we first introduce the need of video summarization techniques for the multiple egocentric contexts (section II-A), the characteristics of FPV, and how FPV summarization techniques differ from TPV (section II-B). We then present a general framework for FPV video summarization (as shown in Fig. 2), and review and organize the literature according to it. The presented framework is data-oriented, depending on the given input —images or video— and desired output —story boards, video skimming or fast-forwarding, as defined in section III. It consists of two clear steps: segmentation of the input data (section III-B), and selection of the relevant segments or key frames (section IV). We also analyze in-depth the datasets used for this task (section V), and the obtained

This work was partially supported by A*STAR JCO REVIVE project grant 1335h00098, the Singapore International Graduate Award (SINGA), and the Obra Social "La Caixa" and Casa Asia Fellowship.

A. G. del Molino, C. Tan and J. H. Lim are with the Institute for Infocomm Research (I2R) – Agency for Science, Technology and Research (A*STAR), Singapore. (e-mail: stugdma, cheston-tan, joohwee@i2r.a-star.edu.sg).

A. H. Tan (asahtan@ntu.edu.sg) is with the School of Computer Science and Engineering at the Nanyang Technological University of Singapore.

results and evaluation approaches (section VI). We finalize by giving some insight on the promising research directions and challenges.

II. Preliminaries: Egocentric Videos

To understand why TPV summarization approaches cannot be directly applied to FPV, we first need to define FPV and its differences from TPV (e.g. consumer videos from smartphones, professional recordings such as movies or documentaries, etc), as well as the motivation of these egocentric recordings. FPV (or egocentric) recordings comprise images and videos taken with (hands-free) wearable cameras, and approximate the wearer's visual experience[1]. Videos recorded with devices such as the Narrative Clip, Autographer, Looxcie, Google Glass, GoPro, Tobii, etc, are typical examples of FPV videos.

A. Summarizing egocentric video for its different applications

About 8 million wearable cameras were sold in 2014, and the number of shipments is expected to reach 30 million units by 2020 [35]. Being able to record what we see without compromising our mobility or the use of our hands clearly opens a wide range of opportunities, as outlined in Table I. However, most of them require specific summarization tools in order to extract the relevant data. Here we list some of these applications:

- **Law enforcement and security:**
  Almost 100,000 police officers in the United States, the United Kingdom, and parts of Asia are already recording their whole day with wearable cameras to assure their good practices while patrolling, and the number may increase considerably in the coming years [35]. Currently, such recordings are used as evidence of what happened on that specific incident. In a future, however, a summarization algorithm will be able to find behavior patterns and detect dangerous situations, to assess the police force beforehand, ignoring aesthetics or emotionally pleasing constraints.
- **Caregivers supervision and memory digitalization – from Lifelogging to daily activities:**
  "Are my parents safe living alone?" or "What are my new patient's routines?" are examples of doubts that could be solved by automatically analyzing wearable videos, getting an overview of the daily recordings given specific queries. Moreover, the daily recordings could be automatically summarized to keep only the most relevant events (classifying differently daily routines from unique and rare events). Since batteries and memory capacities nowadays do not permit for continuous video recording, life-logging devices taking pictures at a fixed interval can be used in cognitive therapy or as a means of memory preservation (even if in digital form) [36, 37]. On the other hand, if taking videos of our daily activities only sporadically, we could share with family and friends an extract of our life (e.g. an afternoon in the park) by selecting a representative variety of actions, interactions and emotions along the different events.
- **Special life events –celebrations and holidays:** The main usage of personal video cameras is preserving memorable events such as celebrations and holidays. However, recording the experience generally means not being able to fully be a part of it. Wearable cameras allow the cameraman to press the record button and forget about it while enjoying with the others. In a summary of such experiences (either as a short video or a photo album) we would expect to find happy faces, emotional moments, interactions with other people or animals, and beautiful scenery.
- **Extreme experience sharing –sports and adventure:** The cheaper wearable action cameras such as GoPro get, the more athletes and adventure-lovers record their full experiences to share the videos afterwards with friends or the general public. Summarizing such huge amount of video is a burdensome task that could be simplified by automatic systems. These systems should discriminate the thrilling or visually attractive shootings from shaky or dull ones.
- **Instructional video –navigation and remote assistance:** Both for navigation purposes or remote assistance, wearable devices with augmented capabilities are becoming of great help. They can help the user find his way (e.g. when moving around a new building or a never-explored area), relying on somebody else's directions. They can also help the user perform a specific task, by viewing somebody else's first person view experience (e.g. the manufacturer's technicians or a chef cooking a recipe). Those systems are trained recording all possible routes or steps, to find afterwards the subshots the user will need as guidance.

[1]However, note that Lifelogging devices located at chest level differ in the recorded content from head-mounted devices, in the sense that they do not capture the sudden changes of head direction.

TABLE I
TYPE OF SUMMARIZATION AND OBJECTIVES FOR DIFFERENT WEARABLE VIDEO INTENTION

	Output			**Objectives**									
							Importance						
	Story Board	Skim	Fast Forward	Uniformity (in time)	Diversity/Uniqueness	Representativeness	Thrilling action	Image quality	Interactions	Semantic events	Emotion	Attention (wearer)
Law enforcement		✓	✓	✓	✓		✓		✓	✓		
Lifelogging	✓				✓	✓		✓	✓		✓	✓
Dailylife sharing		✓				✓			✓	✓	✓	✓
Celebrations	✓	✓			✓	✓		✓	✓	✓	✓	✓
Holidays	✓	✓			✓	✓		✓	✓	✓	✓	✓
Extreme sports		✓	✓	✓		✓	✓			✓		
Adventure		✓	✓		✓		✓	✓	(✓)			✓
Navigation		✓	✓		✓	✓				✓		
Remote Assistance		✓	✓	✓					✓	✓		✓

Examples of wearable applications and the most appropriate summarization criteria, both in terms of output type and objectives to maximize

B. Characteristics of First Person View Video

Based on the findings of Tan et al. [38], we observe and highlight the following principal discriminative characteristics of FPV in contrast to TPV:

- **Intention:** Egocentric videos are unconstrained in nature, lacking a proper structure for the purpose of the video. In general there is no specific intention in the recording, and so no focus on the relevant thing the wearer wanted to keep documented, if he ever wanted to record anything in particular. Unlike FPV videos, in TPV videos the camera man usually focuses on the item or experience to record, composing the scene around it. This makes it easier for computer vision techniques to find the interesting spots in the video, e.g. zooming towards a particular person or object. In contrast to TPV videos, FPV ones are manipulated by spontaneous human attention and hence can capture important cues that provide critical knowledge for video summarization. Attention can be inferred from head motion [19, 20, 39], which has also been used in the literature to characterize the performed activity [39, 40] or intention [41] —if I want you to look to a certain place, I may point there with my head.
- **Content:** Since Lifelogging is a hands-free action, there is no constraint on what to record and what to keep out of the camera field of view. The wearer may record everything while being free to fully enjoy that life experience. As a result, most of the logged data could very well be repetitive or irrelevant. Moreover, the video is a continuum of consecutive events, with smooth transitions from one to another, and without camera cuts to discriminate different sequences. On the other hand, TPV videos tend to record the experiences worth remembering, since the camera is turned on and off to avoid uninteresting scenes, and proper framing and focusing.
- **Quality:** Due to head or chest motion, wearable devices tend to result in videos with many blurry and shaky segments. This is unlike TPV videos, where the cameraman tries to stabilize the recording. Moreover, FPV videos are frequently unaligned, due to head tilt.

The highly unconstrained nature of FPV presented above makes traditional TPV summarization methods difficult to apply, since these are generally domain-specific (designed for sports, news, movies, TV dramas, music videos...). The analysis benefits from the rigid structure of those contexts, relying on speech excitement, applause, flash lights or "score" cuts, text captions in broadcast news and shows, background music, shot duration and silences, laughs for sitcoms, etc [33]. These cues are mostly absent in egocentric video [10, 31, 32], and so are not available for its analysis. Moreover, such long streams of data with very subtle boundaries (both temporal and spatial) add an additional challenge to FPV video segmentation, and the low quality of the recordings hampers accurate feature tracking. Therefore, applying TPV summarization techniques over FPV videos provide inaccurate results, even performing worse than uniform sampling in some cases [12].

Furthermore, as many of the reviewed works point out [21, 23, 42], the ideal summary is context dependent. As such, it is important to summarize each kind of video differently. Unlike for TPV, where the context is generally known beforehand and common throughout the video, FPV faces the problem of having to deal with a possibly unknown and diverse context. Therefore, algorithms need to predict the summarization objective, which may change during the recording.

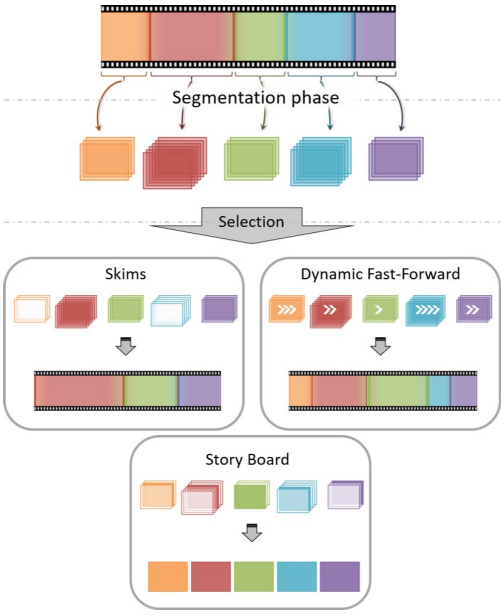

Fig. 2. A general framework for egocentric video summarization. The video is first segmented into scenes or subshots, from which: a) the most relevant subshots are selected for skim summaries; b) the subshot speed is determined for dynamic fast forwarding (traditional fast forwarding does not need prior segmentation); c) a key frame is extracted for story boards.

Yet, FPV offers a great advantage over TPV, which is the personal nature of FPV videos. The egocentric point-of-view allows for a privileged peek into interactions with objects, animals and other people. The video captures the wearer's ongoing activities and goals, thus following his or her gaze and attention patterns can allow the system to detect highlights [19, 20, 22, 39].

III. Approaches for Summarizing Egocentric Videos

New video summarization techniques have been explored lately for FPV videos. In general, these systems are bottom-up, relying mostly on low-level features and ego- motion characteristics [8, 11, 17, 18, 21, 28], but some works also apply supervised learning [9, 10, 12, 16, 19, 20, 23, 25] and exploit physiological data such as EEG signals [14, 15] and gaze [22] to select the relevant segments, as a top-down guidance. Table III on page 5 presents in a comprehensive and schematic way the features used in each analyzed paper, the cues used for the segmentation of events, and the objectives to be met when selecting subshots to represent the video.

When approaching the summarization problem, different possible outputs are considered from storyboards to video fast-forwarding. Table I outlines the preferred output (storyboards, skims or fast-forward, as defined below) for each kind of video intention, whereas Table II lists the reviewed summarization

papers along with their selected type of summary. Fig. 2 shows the typical framework for each approach, which begins by segmenting the video into different events or small subshots, in order to select the most adequate ones afterwards.

- **Static Story Boards:** Though being a minority, some works summarize egocentric videos extracting the highlights in the form of a set of key frames to obtain a sort of photo album [10–13]. However, this approach is mainly used for Lifelogging data, where the input is a set of pictures from chest mounted cameras instead of video recordings [3, 5, 43–46].We will only extensively review summarization of FPV video content, since summarizing periodically taken pictures into story boards is well reviewed by Bolanos et al. [31].
- **Dynamic Video Skimming:** For personal recordings from head-mounted cameras (such as holiday or home videos), retaining the original video structure is generally preferred. The summary is done by selecting the most relevant segments of consecutive frames (subshots) to represent the full video [8, 9, 14–25].
- **Fast Forward:** Despite the shaky nature of egocentric videos and its inherent challenge for a faster video browsing,preserving the whole video content is important for adventure and extreme sports videos [26–29]. A variant of this is dynamic fast-forwarding, where the video is segmented into different sections and set a variable speed to each of them [26], as opposite to traditional fast-forwarding, where a constant speed is set along the video. When traditionally fast forwarding, however, there is no need for segmentation and selection, since all the video is kept. Such works are reviewed in section III-A.

TABLE II
SUMMARIZATION METHODS REVIEWED IN THIS SURVEY.

		Dataset	Task			Evaluation		
			Storyboard	Skim	Fast-Forward	Subjective	Annotation	Textual
Aizawa *et al.*, 2001	[14]	*not public*		✓			*none*	
Ng *et al.*, 2002	[15]	*not public*		✓			*none*	
Lee *et al.*, 2012	[10]	UT Ego	✓			✓		
Lu *et al.*, 2013	[16]	UT Ego, ADL		✓		✓		
Xiong *et al.*, 2014	[11]	UT Ego	✓				✓	
Gygli *et al.*, 2014	[17]	SumMe		✓			✓	
Zhao *et al.*, 2014	[18]	*not public*		✓			✓	
Okamoto *et al.*, 2014	[26]	*not public*			✓	✓		
Kopf *et al.*, 2014	[27]	Microsoft's			✓		*	
Lee *et al.*, 2015	[12]	UT Ego, ADL	✓			✓		
Varini *et al.*, 2015	[19]	*not public*		✓		✓		
Varini *et al.*, 2015	[20]	*not public*		✓		✓		
Gygli *et al.*, 2015	[21]	SumMe		✓			✓	✓
Xu *et al.*, 2015	[22]	GTEA-g.+, EgoSum+g.		✓			✓	✓
Lin *et al.*, 2015	[23]	*YouTube*		✓			✓	
Poleg *et al.*, 2015	[28]	Microsoft's, Disneyworld			✓		*	
Joshi *et al.*, 2015	[29]	*not public*			✓		*	
Bettadapura *et al.*, 2016	[13]	*not public*	✓			✓		
Zhang *et al.*, 2016	[24]	UT Ego	✓	✓		✓	✓	
Yao *et al.*, 2016	[25]	*YouTube*		✓	✓		✓	
Halperin *et al.*, 2016	[30]	Microsoft's, Disneyworld			✓		*	

The table includes the dataset used, task, and evaluation method for each paper. More information on the datasets can be found in Table IV.
* Fast-forwarding preserves the original video content. Computer Vision metrics (mean, standard deviation or difference of motion direction between consecutive frames) are evaluated against other methods (such as naive 10x uniform sampling or Instagram's hyperlapse).

A. Approaches for fast-forwarding egocentric video

As mentioned before, traditional fast-forwarding approaches do not discriminate events from one another. The features used for non-dynamic fast-forwarding are mainly motion related, to be used for stabilization. The first hyperlapse approach for FPV reconstructs the FPV video by changing the 3D virtual position of the camera [27]. The reconstruction is done using structure-from-motion algorithms, and the novel camera path is optimized to be close to the input one. Even if obtaining very good results, they come at a very high computational cost. To compensate this, [29] and [28] obtain timelapses —or hyperlapses— by selecting the optimal set of frames, avoiding non-aligned consecutive sampled frames. Joshi et al. [29] first estimate how well frames align with their temporal neighbors in order to minimize frame-to-frame motion (using RANSAC on sparse feature points), then, the camera path is smoothed considering only the selected frames. In [28, 30], frames are matched according to their viewing directions (estimating the epipolar point and direction of motion), promoting those with forward orientation. The discarded frames are used in [30] to widen the field of view, using the scene's available additional information.

B. Segmenting the input data

Whereas TPV segmentation approaches typically try to identify the shot boundaries comparing consecutive frames, or even using the frames' time-stamp [33, 34], this methodology cannot be directly applied over FPV, since egocentric videos consist of a single shot with extremely smooth transitions between consecutive events.

As can be observed in Table III, FPV segmentation is still mostly based in raw features, not considering human or perceptual cues. The second column section of this table provides an overview of the cues used for event or subshot segmentation, arranged by in-depth of the analysis: deterministic length or temporal proximity; image processing techniques; and finally attention analysis.

As well as segmenting the video deterministically, set to a specific number of frames or time [9, 18, 21, 23, 25, 26], we observe that the most frequently used features for egocentric video clustering or segmentation are **color** [10–12, 14–16] and **motion** cues such as optical flow and blurriness [14–17, 19, 20, 39, 40]. When combined, color can be used both to smooth the motion-based classification [14, 15] or to identify similar events separated in time [16]. As for the motion features, they are generally used to predict the

TABLE III
FEATURES, SEGMENTATION AND SELECTION METHODS USED IN THE REVIEWED PAPERS

		Image/video **features** used									**Segmentation** cues								**Selection** (Objectives)						
		Low-level (e.g. color)	SIFT, (D/H)oG, GIST	Saliency	Motion, blur	Deep learning	Object interaction	People recognition	Gaze	Sensors (metadata)	Uniform Length	Temporal proximity	Color similarity	GIST similarity	Ego-motion	Machine/deep learning	Attention/fixation	Location	Uniformity (in time)	Diversity/Uniqueness	Representativeness	Aesthetics	Importance (predicted)	Attention (wearer)	Query (user input)
Aizawa *et al.*, 2001	[14]	✓			✓					✓			✓		✓									✓	
Ng *et al.*, 2002	[15]	✓			✓					✓			✓		✓									✓	✓
Lee *et al.*, 2012	[10]	✓	✓				✓	✓				✓	✓										✓		
Lu *et al.*, 2013	[16]	✓	✓		✓		✓	✓					✓		✓					✓			✓		
Xiong *et al.*, 2014	[11]	✓	✓		✓					✓		✓	✓									✓			
Gygli *et al.*, 2014	[17]	✓		✓	✓		✓	✓							✓								✓		
Zhao *et al.*, 2014	[18]		✓		✓						✓									✓					
Okamoto *et al.*, 2014	[26]		✓		✓						✓												✓		
Poleg *et al.*, 2014	[39]				✓										✓				–						
Xiong *et al.*, 2015	[9]	✓	✓				✓	✓			✓								–						✓
Lee *et al.*, 2015	[12]	✓	✓				✓	✓				✓	✓						✓	✓			✓		
Varini *et al.*, 2015	[19]		✓		✓										✓		✓			✓				✓	✓
Varini *et al.*, 2015	[20]		✓		✓					✓					✓		✓			✓				✓	✓
Gygli *et al.*, 2015	[21]	✓				✓					✓								✓		✓		✓		
Xu *et al.*, 2015	[22]					✓			✓								✓			✓	✓			✓	
Lin *et al.*, 2015	[23]		✓								✓												✓		
Poleg *et al.*, 2015	[40]					✓										✓			–						
Bettadapura *et al.*, 2016	[13]	✓	✓		✓					✓				✓				✓				✓	✓		
Zhang *et al.*, 2016	[24]					✓										✓				✓					
Yao *et al.*, 2016	[25]					✓					✓												✓		
Usage count		**10**	**11**	1	**11**	5	5	5	1	5	6	3	**6**	1	**7**	2	3	1	2	**7**	2	2	**9**	5	4

Different cues are used to define, segment and summarize the video, mainly according to color and motion properties. When selecting the relevant segments, a mixture of different objectives is maximized, including importance (predicted, from the wearer attention, or a particular user's query) in most reviewed papers. – Selecting specific segments is not the objective of the paper

wearer's activity or attitude patterns and then segment the videos accordingly. Examples of these methodologies are the *Cumulative Displacement Curves* in [39], a Super Vector Machine–Hidden Markov Model pipeline in [19], and a 3D Deep Convolutional Neural Network in [40].

Besides these, other approaches are explored for egocentric video segmentation, such as the use of **GIST** difference over a given window [13]; similarity between the R-CNN hashes extracted from the fixation region (using **gaze**) [22]; and setting the subshot boundaries where **semantic** labels change [8].

IV. SELECTING THE OPTIMAL SEGMENTS OR KEYFRAMES

Once the video is segmented, the next natural step is to select the most appropriate parts for the summary. This is done by maximizing a combination of objectives, as summarized in the last group of columns in Table III. These are grouped into video coherence (such as diversity of events or temporal uniformity), visual pleasantness through aesthetics, and inherent importance of the segment, either for the viewer (the user to watch the generated summary) or the wearer (the person recording the original video). Even if importance is the main target for most works, each objective has its shortcomings, and so the selection of objectives must be consistent with the type of input data and the purpose of the summary, as shown in Table I.

Features such as color, SIFT,DoG, HoG or HoF are frequently used to analyze video coherence [10, 12, 16, 18–21], as well as the use of deep learning [8, 21, 22, 24, 40]. Aesthetics is generally estimated using features such as color, SIFT, GIST and blurriness [11, 13]. Finally, importance may be estimated with supervised learning [23, 25, 26]; inferred from impersonal cues such as saliency [17], people and object interaction [10, 12, 16, 17, 19] and location [13]; or predicted from the wearer attention patterns, using sensors [14, 15, 22] or motion analysis [19, 20, 39].

A. Important to the viewer

The importance value of a frame or segment can be estimated for any user —universal predictors that do not consider the wearer's or specific viewer's interest. However, the absolute importance of each segment is context dependent, and cannot be equally estimated for, e.g., extreme sports and law enforcement video. As such, each reviewed system may predict importance differently. This importance score is frequently combined with video coherence objectives.

1) Considering only importance to obtain Story Boards: In the case of story boards, the objective is to select the best frames instead of full segments, and so the interest relies in each individual image. The interest predictor in [10] is trained with regions containing important **people and objects**, and

uses three kinds of features: egocentric, object-like and region properties —comprising a total of 14 features such as SIFT and DoG matches and region size. This approach, however, ignores the impact of the overall history in the generated story board, and works better with videos of daily activities.

Looking into the picture **aesthetics**, both [11] and [13] present a method to obtain a nice holiday or biographic photo album out of the captured videos, but the most representative or meaningful events can be left out of it. A predictor is used in [11] to select images that could have been intentionally taken. Said predictor is trained using images from the web and cues for composition or intention, such as picture alignment or accelerometer data. Without the need of a trained classifier, the authors of [13] look for picturesque images with good artistic properties: composition (considering the artistic *rule-of-thirds*), symmetry (on local SIFT features) and color vibrancy. Only videos from places of interest are analyzed (using GPS data), and the frame with minimal head tilt is selected out of the highest scored frames.

2) Diversity and uniformity in Story Boards: Uniqueness — or diversity— and uniformity parameters can be used alongside importance (as in [10]) to solve the story-line problem [12]. Uniqueness is defined as the absence of similar objects in consecutive selected frames and is computed as the color histogram difference. Uniformity, on the other hand, is related to the frame index. Albeit considering the overall narrative, the performance of this method is still subject to the context of the video to summarize, performing better in daily manipulation activities. Moreover, the selected frames can be of poor visual quality, even if informative enough.

In opposition to Story Boards, when selecting subshots to obtain Skims the whole set of frames needs to be considered, and the use of temporal features has been explored:

3) Considering only importance to convey Skims: User studies suggest that static images' interestingness is related to factors such as saliency, edges and colorfulness, object interactions, and the presence of landmarks, people or faces [47]. This assumption (and the belief that for a segment to be interesting it has to be inherently important) is used in [17] to rate the relevance of each segment as the sum of its frames' interestingness. Albeit introducing the *superframe*, a very interesting segmentation method based in ego-motion optimization, this summary lacks of narrative guidance, only relying on independent frame information.

Inspired by [42], the authors of [23] argue that each type of video (such as "skating", "gymnastics", "dog" or "surfing") must be summarized differently. Therefore, they train a different highlight detector for each context. Context detection models are also pre-trained using STIP features, since it may vary along the video. This methodology allows for almost *on-the-go* summarization, reducing the amount of data to keep, and solving the problem of memory storage. However, this method does not consider the story line to convey the summary. Yao *et al.* [25] also train a highlight detection classifier from human-generated summaries. They aim at obtaining the highlights of the video while considering the temporal dynamics. Their model fuses the highlight estimation of two different DCNNs: one trained on AlexNet CNN features (objects and animals) and an other on a 3D network output, containing the temporal information of the segment.

In [26], on the other hand, the objective is an adaptive fast-forwarding for pedestrian navigation instructions. A relevancy parameter is estimated through crosswalk detection and ego-motion cues to fast forward those scenes not containing crossings or changes of direction. Being specifically designed for navigation purposes, this method is absolutely context-driven.

4) Diversity and influence: Diversity is computed as a comparison of GIST and SIFT descriptors between consecutive segments in [16]. This is combined with the importance score estimated as in [10] and the influence of each segment to the general story to convey a story-driven summary. As in the case of [10, 12], this method strongly relies on supervised learning, both to predict the importance and influence of each segment. As such, it will perform better on contexts already seen in the training phase.

5) Representativeness and uniformity: To convey an approximation of the ideal summary, a submodular maximization of objectives can be learned using reference summaries [21]. The objectives chosen here are importance (using a classifier trained with deep features from the data provided by [10]), representativeness (defined as the most similar instances to the rest of the video) and uniformity (temporal coherence). As the authors point out, a good summary is not absolute, and depends both on the intention of the recording, the context and user preferences. Thus, the method can be improved by incorporating context-specific characteristics, or user likings knowledge.

6) Personalization: Both [8, 9] propose systems that can retrieve subshots from the stored videos given a video [8] or a story-based query [9]. Albeit these systems could provide a personalized summary by concatenating the retrieved shots, the authors did not explore such possibility in their presented works. In [19, 20], the summary is personalized by looking for scenes relevant to the cultural interest of the viewer or the wearer (more in subsection IV-B2), using DBpedia.

B. Important to the wearer

Importance can also be inferred from the wearer recording patterns (such as time spent at a certain place, or interacting with a certain item or person), or from physiological data recorded alongside the video. However, only a few works use the personal characteristics of the recording to provide a wearer-personalized summary. We present them here.

1) Physiological measurements: Some works use physiological measurements such as gaze and EEG signals to detect the wearer's interest, as is the case of [14, 15, 22]. Whereas Aizawa et al. [14, 15] use EEG signals on the α and β bandwidths to detect interest in the scene through brain activation and select segments with this sole objective, Xu et al. [22] use gaze to predict the attention given to each event. Their algorithm also encourages both **representativeness** and **diversity** by maximizing the entropy of the segments' descriptors. Each segment is defined with the Recursive Convolutional Neural

Network (R-CNN) hash computed around its frames' centroids. The attention score of each segment, on the other hand, is computed as the amount of its frames containing fixation, and is added to the equation to be maximized. Those summaries provide a better insight into the wearer's feelings while recording, capturing moments of higher attention, at the cost of having to use costly and uncomfortable sensor devices.

2) Estimation of attention: The authors of [39] also consider gaze information a very useful feature to infer important events, and so propose a gaze fixation estimator based on ego-motion features with 75% accuracy. In the same way, [19, 20] use their Hidden Markov Model (HMM) with motion observables and GPS information (added in [20]) to estimate the level of attention. This attention score is combined with **diversity** —from Bag of Words (BoW) distance between consecutive segments— and relevance to the user preferences. To measure the relevance, a semantic classifier is trained with images from the web for a given keyword query, using the BoW approach. This summary, therefore, can be user oriented, changing depending on the user preferences and query keywords, and is specifically designed for a cultural or touristy experience.

C. Importance independent

Not considering the interest of the segments, and targeting the problem of storage and the need for almost real-time summarization, [18] proposes a summarization method based on **uniqueness or diversity**. The summary is created in an on-line fashion while creating a dictionary of video sentences. Every new segment is analyzed by detecting spatio-temporal interest points and describing them with a concatenation of HoG and HoF features. If it is impossible to reconstruct it by using the learned dictionary, the segment is added to the summary and the dictionary is updated with the new features. In this way, all events are represented while avoiding repetitions, but not all events might be relevant to the overall story (e.g. a change in the background with no meaningful action). Moreover, only the first occurrence of each event is added to the summary, even if it is not the most representative or important of said event.

Using deep learning, the video frames can be encoded with a Long Short Term Memory (LSTM) network. To select diverse content, the pairwise analysis of the segments is used in [24] as input for a Determinantal Point Process (DPP), which will output the optimal summary.

V. Egocentric Datasets

There is a small but growing number of datasets available for egocentric video analysis, and most of them are included in the analysis by [31, 32, 38]. Among them, many are not suitable for FPV summarization, since for this purpose they must contain videos recorded by people with head mounted cameras (to see exactly what the wearer sees even with subtle sight movements), in totally unconstrained environments, and long enough as to compress a wide variety of sub- activities. Even if recorded with head mounted devices, this is the case of datasets such as the following: CMU-MMAC [51], recorded in a staged kitchen; GTEA [52] and GTEA- gaze [50], which contain very specific videos, sometimes staged and short; UEC [53], which is a compilation of short videos from YouTube and recordings of choreographed activities; or SumMe [17], in which even if the egocentric videos are annotated specifically for the summarization task, they are too short to be useful for longer egocentric summarization, being at around 2 minutes long. The most used publicly available datasets for the summarization task are described below, along with other recently published ones. All these datasets useful for FPV video summarization are outlined in Table IV, which includes for each one the amount of videos, wearers and typical length; the year of release, original task for which they were recorded and the available annotation; and the works in which they have been used.

These datasets are the following:

- **UT Egocentric** [10]: Originally recorded to summarize FPV based on the presence of important objects and people, this dataset contains long videos of many different daily activities. Unlike all the other reviewed datasets, this one was recorded at low- quality frame rate (15fps), and the video data includes objects annotation.
- **Activities of Daily Living** [49]: To record this dataset, users were asked to perform a set of pre-accorded activities at their homes in a continuous way, wearing a chest-mounted GoPro camera. Videos are densely annotated with objects, interactions and the actions performed.
- **GTEA-gaze+**[50]: Intended to improve de data collected for GTEA-gaze, this dataset contains videos of subjects preparing meals out of 7 different food recipes in a natural kitchen setting. The dataset is recorded with SMI eye-tracking glasses, and contains annotation of around 100 different actions. Summarization annotation was added later by [22].
- **Disneyworld**[41]: This dataset was originally recorded to evaluate social interactions during a full day at an amusement park. However, due to its highly unconstrained nature, the long duration of its videos, and the textual annotation provided by [48], it is very convenient to be used for the egocentric summarization task.
- **VideoSet** [48]: To overcome the non-standardized evaluation issues, this evaluation tool based on textual information was released in 2014. VideoSet provides summarization annotation for videos from Disneyworld and UT Egocentric, and tools to evaluate the generated summaries.
- **Huji EgoSet** [39]: This dataset was recorded to test motion segmentation on any kind of activity, location and illumination setting. It is in continuous development and, to this date, contains 37 videos of unconstrained daily activities (driving, chilling, walking, etc) taken with head mounted GoPro cameras. It also includes egocentric videos extracted from Youtube. The videos are annotated with motion and activity patterns.
- **Microsoft's sports dataset** [27]: Used for fast- forwarding objectives, it was recorded with a GoPro camera on a helmet and includes adventure activities such as mountain biking or climbing.

TABLE IV
FIRST PERSON VIEW VIDEO DATASETS USED FOR THE SUMMARIZATION TASK.

Dataset name	Year	Device	Short description	Num. subjects	Num. vids	Typ. length	Gaze metadata	Other sensors	Original task: Summary	Original task: Activity rec.	Original task: Object rec.	Original task: Other	Annotation	Used in
UT Ego [10]	2012	Looxcie	Unconstrained videos of natural daily activities	4	10	3-5h			✓				- People and objects: text and boundary * [48] adds text annotation	[11, 12, 16, 21, 48]
ADL [49]	2012	Go Pro (chest)	Predefined set of actions at home	20	20	30'				✓	✓		- 18 actions - 42 objects data	[12, 16]
GTEA-gaze+ [50]	2012	SMI eye-tracking	Cooking in a natural setting	10	30	10-15'	✓			✓	✓		- 100 actions * [22] adds annotation for summarization	[22]
DisneyWorld [41]	2012	GoPro	Experiences during a day at Disney-World	8	8	6-8h						✓	- Social interactions * [39] adds motion annotation for segmentation * [48] adds text annotation	[39, 40, 48]
VideoSet [48]	2014	-	Evaluation tool for the summarization task over UT Ego (10 videos) and DisneyWorld (3 videos)	7	13	3-8h			✓				- Activities: 1-sentence descriptions every 5 - Summaries in text form	Not released
Huji EgoSet [39]	2014	GoPro	Different activities in unconstrained settings	3	37	6-30'						✓	- 7 motion classes	[40]
Microsoft's [27]	2014	GoPro	Videos of different sportive activities	1	5	3-13'			✓				(none)	[28]
[19, 20]	2015	*head mounted*	Recorded by tourists walking, shopping, or visiting cultural spots	12	12	30'			✓				- 6 motion classes	Not released
EgoSum+gaze [22]	2015	SMI eye-tracking	Daily life in unconstrained setting	5	21	15'-1.5h	✓		✓				- Summarization: 5-15 sets of segments / video	Not released
[8]	2015	Google Glass	Unconstrained videos taken during office time and holidays, with activities from lunch to sports	1	50	15'-1h		✓				✓	(none)	Not released
[9]	2015	*egocentric camera*	Videos taken at Disneyland with multiple actors, attractions and events.	10	10	>5'						✓	- Locations and events every 10 frames	Not released
[13]	2016	Contour Cam	Unconstrained videos taken during a vacation trip between USA's coasts.	1	26.5h			✓	✓				(none)	Not released
[25]	2016	GoPro	Unconstrained sports videos (15 categories).	-	600	2-15'			✓				- Interestingness of the segments on a scale of 1 to 3.	Not released

- **EgoSum+gaze** [22]: Since no previous daily life video dataset with gaze information existed, this dataset was recorded to test and evaluate a gaze-driven summarization method. Recorded with SMI eye-tracking glasses and a Pupil eye-tracking device, it contains long unconstrained daily life videos and gaze information. The annotation was obtained both from the wearer and external experts, and includes 5$\sim$15 relevant events per video, where each of these events or blocks contain a variable number of subshots that are equally adequate to be selected as part of the summary.
- **Microsoft's video highlights** [25]: this dataset contains 100 hours of FPV sports videos, alongside 15 different categories. All videos are mined from YouTube searching for "category + GoPro" to ensure the egocentric point of view, and a shot analysis is performed to remove edited content. The videos are between 2 and 15 minutes long, and each five-second segment is annotated by 3 independent judges with its level of interestingness (1="boring", 3="highlight"). 12 annotators participated in this task.

To our judgment, the most suitable datasets to perform and evaluate egocentric video summarization are EgoSum+gaze, since it contains both annotation from the wearer and external users for life-logging videos, GTEA-gaze+ with the annotation provided by [22] for task-specific summarization, the videos supported by VideoSet, and Microsoft's summarization dataset. However, none of them are publicly available. Extensive work towards a unified benchmark for a wider range of tasks is needed.

VI. EVALUATION

Evaluation of video summarization is still a great challenge. An objective best-summary ground truth does not exist, as each person may like different summaries for different reasons, and this preference may also change over time. Moreover, the summary is generally task-dependent, and it should be evaluated differently according to the intention. When evaluating FPV video summaries, the personal nature of the original recording makes it even more complicated: who should be the judge of the summary, who should annotate the key items? The wearer,

the viewer, or both? Their understanding of a good summary may be completely different.

Unlike for image memorability [54], for egocentric summarization there is no empirical evidence showing that inter-subject consistency is actually relatively high. To the contrary, Gygli et al. [17] evaluated the human performance when summarizing their SumMe dataset, computing the f-measures of each human summary against all the other participants', achieving measures between 0.1 and 0.5 –mean of 0.25– on their egocentric videos. Moreover, the problem with objective ground-truth based evaluation is that it may not properly reflect what users truly want from a summary. Perhaps because of this, methodologies are evaluated as a relative comparison with other techniques, either in user preference or better precision-recall evaluation, not considering the standalone performance.

From Table II, we see that the trend has been that each new paper proposing a new method also comes with its own dataset and evaluation approach. In the following subsections we present the tests conducted for each summarization technique and the results obtained. However, comparing them results very difficult, since the only way would be to test on the same videos, with the same evaluation method.

Video summaries are generally either evaluated in a subjective way conducting user studies, objectively by measuring precision and recall over the presence of key objects, people or events, or using Natural Language Processing (NLP) techniques with textual annotation of the whole video parts. However, these key items and video parts must be previously annotated according to subjective criteria.

To the best of our knowledge, until the publication of [48] in 2014 and [22] in 2015 there was no standardized evaluation benchmark for long egocentric videos. To standardize the evaluation process, [48] provides text annotations for a large collection of egocentric videos from public datasets, releasing the VideoSet tool, and [22] provides summaries annotation by experts for extensive egocentric videos in their dataset EgoSum+gaze and also in GTEA-gaze+. More information on said datasets can be found in section V and Table IV.

A. Evaluation through human judgment of the generated summary

Even though [12, 16, 19, 20, 26] evaluate the accuracy of their segmentation or importance classifiers based on their ground truths, the summary evaluation is done through user studies. Generally, the judges are first presented with a speed-up or browsable version of the original video. Then, they may asses the quality of each summary against the others in blind tests [13, 16, 26]; rate each summary individually [12]; or score whether the summaries show the relevant events according to the user's preferences [19, 20].

Lee et al. [12] additionally obtain objective measurements —section VI-B1. From their user study, however, they observe that uniform sampling is the preferred choice for videos of low content complexity, and that, in general, it is also rated better than the TPV baselines tested.

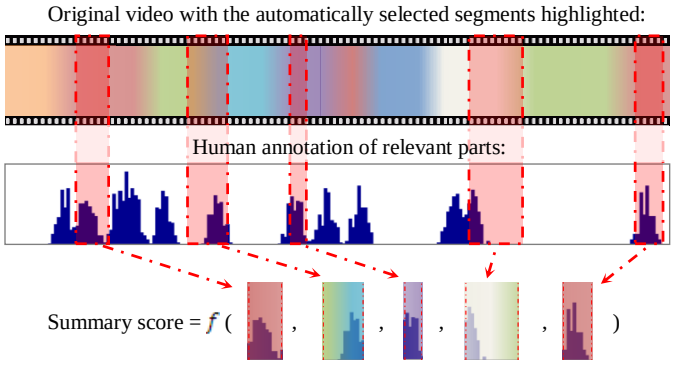

Fig. 3. Evaluation using human annotation of the important parts of the original video.
The summary score is computed based on the overlap between human annotation and the outputs automatically selected by the algorithm.

B. Evaluation using annotation

In order to obtain quantitative evaluations instead of user feedback, some researchers annotate the original videos with subjective cues. These annotations include relevant people, objects or events [10, 12], and aesthetics preferences [11]. However, the most common method is the annotation of important parts (several subjects analyze the video streams and select the relevant frames or segments which better summarize the video), obtaining a sort of summary histogram when putting them together [17, 18, 22–25]. In this way all benchmarks can be equally compared to the human selection. An illustrated example for this technique is shown in Fig. 3.

1) Important objects and people: In [10, 12] the object annotation is used to compute the recall rate —amount of important objects detected as a function of summary length. Results prove [10] to be better than uniform sampling and event-based adaptive uniform sampling (selecting evenly sampled frames from each predicted event) for long summaries, even though for 10-15 key frames all methods perform similarly. On the other hand, the method in [12] achieves overall better recall rates than two TPV video summarization state-of-the-art baselines.

2) Selection of visually pleasant pictures: To evaluate the precision and recall of the algorithm in [11], the authors selected 10,000 frames from their UTEgo dataset and distributed each of them to 5 workers from Amazon Mechanical Turk. Then, the workers were asked to rate the pictures in between 4 levels from intentional to accidental, as if they had a defective camera and had to sort the taken pictures. Using this annotation, they compare their method to four different baselines: saliency, blurriness, people likelihood and discriminative SVM, achieving better results in 3 of the 4 tested videos.

3) Subshot preference histograms: Gygli et al. [17] evaluate their interestingness-based method using the f-measure of the overlap between the obtained summary and the annotated human summaries. They find that only in 1 of the 4 egocentric videos tested their method is not the best or second best approach, compared to uniform sampling, color clustering and an estimation of the visual attention.

Xu et al. [22] compare their method to uniform sampling, k-means, and the same two baselines now using their gaze-based subshot segmentation. They conclude that their method is better in both tested datasets, and prove that using gaze for segmentation significantly outperforms not using it. To be used as benchmark, [22] made their EgoSum+gaze dataset publicly available (details can be found in section V).

LiveLight [18] is evaluated on a YouTube dataset, and obtains an average accuracy of 72%, whereas uniform sampling scores 39%, k-means clustering 52%, and a state-of-the-art method for TPV video summarization a 64%. It is also used as a baseline by the authors of [23]. LiveLight's average precision is better when the context is randomly chosen. However, if predicted beforehand, [23] outperforms all the baselines (also including a HoG/HoG Fisher Vector based SVM, and a consumer video state-of-the-art summarization method [55]) proving the importance of knowing the video context.

The LSTM-based summarization method presented in [24] is evaluated using the f-measure, precision and recall of the selected segments against the annotation, with both TPV and FPV videos. The authors observe that it performs better the more annotation is provided, being outperformed by the use of multi-layer neural networks (MLPs) (using neighboring frames as features) when the training data is scarce. Finally, [25] is proven better than the baselines in 11 of the 15 evaluated categories in terms of average precision of the highlight detection and normalized discounted cumulative gain. The authors note that the use of motion features boosts the performance. However, since all tested videos are sports-related, it is unclear whether their method can be applied to non-sport domains.

C. Evaluation using natural language

Since measuring text content similarity has long been explored by the Natural Language Processing (NLP) community and great progress has been achieved so far, the authors of [48] consider it is best to evaluate semantic summaries through text, even if that summary is visual. They propose an evaluation system in which the video summary can be compared to a nearly ideal one in an automated way, without direct human involvement.

However, for this method to work, all video segments must firstly be annotated in textual form, and a summary written. The video summary to evaluate is then converted to text mapping the annotated sentences of each selected segment. Finally, the distance from this converted summary to the human ground truth text summary can be computed using NLP measures.

In order to provide a standard evaluation benchmark for the summarization task, the VideoSet tool [48] provides sentence annotations for each 5-second long subshot in the supported datasets, and the written summary of each video. It uses the ROUGE package [56] (NPL evaluation techniques designed for text summarization) to measure distances.

Some works already use NLP measures to evaluate their summaries, as is the case of [22], to support their f-measure evaluation, and [21]. Gygli et al. [21] compare their method to a randomly generated one, uniform sampling, the method presented in [10] and video Maximal Marginal Relevance, a method adapted from text-summarization which rewards diversity. For both short and long summaries, the proposed one is found to be slightly better than uniform sampling. According to their evaluation, though, the technique of [10] produces worse results than uniform.

VII. Conclusion

In this survey, we have described the main differences between FPV and TPV and the summarization techniques associated. We have also illustrated the wide range of applications of wearable video cameras, and the need for specific summarization techniques for each type of FPV video.

Since summarization of FPV videos cannot be properly tackled with traditional TPV summarization techniques [10, 12, 21], this topic has attracted a lot of interest recently, particularly in the past two years. As such, we have analyzed all relevant techniques up to date, pointing out their main strengths and shortages, and giving special emphasis to the datasets used and evaluation performed.

However, the present situation, with each work using its own dataset and evaluation methodology, makes comparing summarization techniques a very hard task. Therefore, we deem it necessary to establish a benchmark to equally evaluate all summarization techniques on a common dataset.

As with all new research challenges, there are areas that can be improved or explored in more depth. We view the following areas to be more important: personalization of summaries; creation of extended datasets specific for the summarization task with complete annotations; looking into the aesthetics or visually pleasant moments; and use of multimodal cues such as accelerometer, audio, etc.

A. Personalization of summaries

From our point of view, the main challenge yet to be solved is the personalization of the summary, either from the perspective of the wearer or a third party viewing the summary. The summary quality perception is subjective, and so a good summary should strive to that specific user's preferences. To the best of our knowledge, only [14, 15] (using EEG) and [22] (using gaze) provide a personalized summary from the wearer's side, [8, 9] retrieve subshots that can be combined to summarize the video according to the user's query, and [19, 20] mix attention estimation with prior content preferences.

On one hand, we think the summary could be personalized for the person viewing the summary by unifying and formalizing retrieval and summarization (i.e. summary from query). The importance of each subshot should be related to the interest of the viewer, and not just an estimation over externally annotated data. On the other hand, the summary could be personalized to the wearer's life and feelings in different ways, for example:

- Use of the wearer's emotion and attention or interest —from physiological data, as suggested in [57], and/or speech;
- Making use of previously stored memories, and their relation to the video to be summarized;
- Mining for patterns in past memories to detect similar or relevant events in the newly uploaded video.

B. More specific and complete datasets

Our review shows there is still a gap for standardized egocentric datasets. A benchmark with variety of contexts and settings needs to be created. It should provide extensive videos of life events of unconstrained nature, with extended metadata —such as physiological measurements, gyroscope or GPS— and exhaustive annotation for summarization, both from the wearer perspective, an outsider's, and the different possible summary intentions. Nonetheless, user studies will still be needed to evaluate viewer-driven summarizations, along with the use of extensive annotation for precision-recall measurements on all possible queries, but a standard protocol should be proposed.

C. Aesthetics and enjoyable moments

Since the objective of the summarization is obtaining a meaningful and visually pleasant video or story board, we deem it necessary to detect and select enjoyable scenes, to cut in the right transition moments, and to stabilize the output videos. Firstly, the output video should contain the more appealing events, such as an emotional moment or great laugh (as explored in [58]), maybe even zooming to emphasize the moment. Secondly, transitions can be chosen to tell a more compelling story and avoid unfocused subshots (similar to the *superframe* approach [17]). Moreover, the generated summaries still keep the inherent shakiness of egocentric videos, making them dizzying to watch. Improvements on stabilization algorithms such as [59, 60] could be explored to solve this problem.

Some recent works explore the use of multiple cameras to convey a better final video cut [61, 62], selecting the best quality frames among those with common focus of attention. These types of approaches can also be of use to select events for which all camera wearers were paying attention.

D. Use of other multimodal cues

Finally, we have also realized that speech has never been considered before for FPV video summarization, even if it is widely used in TPV techniques. It could be because of the low audio quality of some devices, or because audio is often not recorded. In any case, the additional use of speech and other multimodal cues (e.g. sensors) is very likely to improve summarization when available.

REFERENCES

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

[2] K. Aizawa, D. Tancharoen, S. Kawasaki, and T. Yamasaki, "Efficient retrieval of life log based on context and content," in *Proceedings of the the 1st ACM workshop on Continuous Archival and Retrieval of Personal Experiences*, pp. 22–31, ACM, 2004.

[3] J. Gemmell, L. Williams, K. Wood, R. Lueder, and G. Bell, "Passive capture and ensuing issues for a personal lifetime store," in *Proceedings of the the 1st ACM workshop on Continuous archival and retrieval of personal experiences*, pp. 48–55, ACM, 2004.

[4] 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*, pp. 61–66, ACM, 2005.

[5] A. R. Doherty, C. Conaire, M. Blighe, A. F. Smeaton, and N. E. O'Connor, "Combining image descriptors to effectively retrieve events from visual lifelogs," in *Proceedings of the 1st ACM international conference on Multimedia information retrieval*, pp. 10–17, ACM, 2008.

[6] V. Chandrasekhar, W. Min, X. Li, C. Tan, B. Mandal, L. Li, and J. H. Lim, "Efficient retrieval from large-scale egocentric visual data using a sparse graph representation," in *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pp. 527–534, 2014.

[7] V. Chandrasekhar, C. Tan, W. Min, L. Liyuan, L. Xiaoli, and L. J. Hwee, "Incremental graph clustering for efficient retrieval from streaming egocentric video data," in *2014 22nd International Conference on Pattern Recognition (ICPR)*, pp. 2631–2636, IEEE, 2014.

[8] A. G. del Molino, B. Mandal, L. Li, and J. H. Lim, "Organizing and retrieving episodic memories from first person view," in *International Conference on Multimedia and Expo Workshops*, pp. 1–6, IEEE, 2015.

[9] B. Xiong, G. Kim, and L. Sigal, "Storyline representation of egocentric videos with an applications to story-based search," in *IEEE International Conference on Computer Vision*, pp. 4525–4533, 2015.

[10] Y. J. Lee, J. Ghosh, and K. Grauman, "Discovering important people and objects for egocentric video summarization.," in *Computer Vision and Pattern Recognition*, vol. 2, p. 6, 2012.

[11] B. Xiong and K. Grauman, "Detecting snap points in egocentric video with a web photo prior," *Computer Vision–ECCV*, pp. 282–298, 2014.

[12] Y. J. Lee and K. Grauman, "Predicting important objects for egocentric video summarization," *International Journal of Computer Vision*, pp. 1–18, 2015.

[13] V. Bettadapura, D. Castro, and I. Essa, "Discovering picturesque highlights from egocentric vacation videos," *arXiv preprint arXiv:1601.04406*, 2016.

[14] K. Aizawa, K. Ishijima, and M. Shiina, "Summarizing wearable video," in *International Conference on Image Processing*, vol. 3, pp. 398–401, IEEE, 2001.

[15] H. W. Ng, Y. Sawahata, and K. Aizawa, "Summarization of wearable videos using support vector machine," in *International Conference on Multimedia and Expo*, vol. 1, pp. 325–328, IEEE, 2002.

[16] Z. Lu and K. Grauman, "Story-driven summarization for egocentric video," in *Computer Vision and Pattern Recognition*, pp. 2714–2721, IEEE, 2013.

[17] M. Gygli, H. Grabner, H. Riemenschneider, and L. Van Gool, "Creating summaries from user videos," in *Computer Vision–ECCV*, pp. 505–520, Springer, 2014.

[18] B. Zhao and E. Xing, "Quasi real-time summarization for consumer videos," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2513–2520, 2014.

[19] P. Varini, G. Serra, and R. Cucchiara, "Egocentric video summarization of cultural tour based on user preferences," in *Proceedings of the 23rd Annual ACM Conference on Multimedia Conference*, pp. 931–934, 2015.

[20] P. Varini, G. Serra, and R. Cucchiara, "Personalized egocentric video summarization for cultural experience," in *Proceedings of the 5th ACM on International Conference on Multimedia Retrieval*, pp. 539–542, 2015.

[21] M. Gygli, H. Grabner, and L. Van Gool, "Video summarization by learning submodular mixtures of objectives," in *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3090–3098, 2015.

[22] J. Xu, L. Mukherjee, Y. Li, J. Warner, J. M. Rehg, and V. Singh, "Gaze-enabled egocentric video summarization via constrained submodular maximization," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2235–2244, 2015.

[23] Y.-L. Lin, V. Morariu, and W. Hsu, "Summarizing while recording: Context-based highlight detection for egocentric videos," in *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pp. 51–59, 2015.

[24] K. Zhang, W.-L. Chao, F. Sha, and K. Grauman, "Video summarization with long short-term memory," *arXiv preprint arXiv:1605.08110*, 2016.

[25] T. Yao, T. Mei, and Y. Rui, "Highlight detection with pairwise deep ranking for first-person video summarization," in *Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[26] M. Okamoto and K. Yanai, "Summarization of egocentric moving videos for generating walking route guidance," *Image and Video Technology*, pp. 431–442, 2014.

[27] J. Kopf, M. F. Cohen, and R. Szeliski, "First-person hyper-lapse videos," *ACM Transactions on Graphics (TOG)*, vol. 33, no. 4, p. 78, 2014.

[28] Y. Poleg, T. Halperin, C. Arora, and S. Peleg, "Egosampling: Fast-forward and stereo for egocentric videos," in *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4768–4776, 2015.

[29] N. Joshi, W. Kienzle, M. Toelle, M. Uyttendaele, and M. F. Cohen, "Real-time hyperlapse creation via optimal frame selection," *ACM Transactions*

on Graphics (TOG), vol. 34, no. 4, p. 63, 2015.

[30] T. Halperin, Y. Poleg, C. Arora, and S. Peleg, "Egosampling: Wide view hyperlapse from single and multiple egocentric videos," *arXiv preprint arXiv:1604.07741*, 2016.

[31] M. Bolanos, M. Dimiccoli, and P. Radeva, "Towards storytelling from visual lifelogging: An overview," *arXiv preprint arXiv:1507.06120*, 2015.

[32] A. Betancourt, P. Morerio, C. S. Regazzoni, and M. Rauterberg, "The evolution of first person vision methods: A survey," *Circuits and Systems for Video Technology, IEEE Transactions on*, vol. 25, no. 5, pp. 744–760, 2015.

[33] A. G. Money and H. Agius, "Video summarisation: A conceptual framework and survey of the state of the art," *Journal of Visual Communication and Image Representation*, vol. 19, no. 2, pp. 121–143, 2008.

[34] W. Hu, N. Xie, L. Li, X. Zeng, and S. Maybank, "A survey on visual content-based video indexing and retrieval," *Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on*, vol. 41, no. 6, pp. 797–819, 2011.

[35] Tractica, "Wearable cameras: Global market analysis and forecasts," 2Q 2015.

[36] A. R. Doherty, S. E. Hodges, A. C. King, A. F. Smeaton, E. Berry, C. J. Moulin, S. Lindley, P. Kelly, and C. Foster, "Wearable cameras in health," *American journal of preventive medicine*, vol. 44, no. 3, pp. 320–323, 2013.

[37] A. R. Doherty, K. Pauly-Takacs, N. Caprani, C. Gurrin, C. J. Moulin, N. E. O'Connor, and A. F. Smeaton, "Experiences of aiding autobiographical memory using the sensecam," *Human–Computer Interaction*, vol. 27, no. 1-2, pp. 151–174, 2012.

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

[39] Y. Poleg, C. Arora, and S. Peleg, "Temporal segmentation of egocentric videos," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2537–2544, 2014.

[40] Y. Poleg, A. Ephrat, S. Peleg, and C. Arora, "Compact cnn for indexing egocentric videos," *arXiv preprint arXiv:1504.07469*, 2015.

[41] A. Fathi, J. K. Hodgins, and J. M. Rehg, "Social interactions: A first-person perspective," in *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pp. 1226–1233, IEEE, 2012.

[42] D. Potapov, M. Douze, Z. Harchaoui, and C. Schmid, "Category-specific video summarization," *Computer Vision–ECCV*, pp. 540–555, 2014.

[43] H. Lee, A. F. Smeaton, N. E. OConnor, G. Jones, M. Blighe, D. Byrne, A. Doherty, and C. Gurrin, "Constructing a sensecam visual diary as a media process," *Multimedia Systems*, vol. 14, pp. 341–349, 2008.

[44] A. R. Doherty, D. Byrne, A. F. Smeaton, G. J. Jones, and M. Hughes, "Investigating keyframe selection methods in the novel domain of passively captured visual lifelogs," in *Proceedings of the 2008 international conference on Content-based image and video retrieval*, pp. 259–268, ACM, 2008.

[45] R. Mestre, M. Bolaos, E. Talavera, P. Radeva, and X. Gir-i Nieto, "Visual summary of egocentric photostreams by representative keyframes," *arXiv preprint arXiv:1505.01130*, 2015.

[46] E. Talavera, M. Dimiccoli, M. Bolaos, M. Aghaei, and P. Radeva, "R-clustering for egocentric video segmentation," *Pattern Recognition and Image Analysis*, pp. 327–336, 2015.

[47] M. Gygli, H. Grabner, H. Riemenschneider, F. Nater, and L. Gool, "The interestingness of images," in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 1633–1640, 2013.

[48] S. Yeung, A. Fathi, and L. Fei-Fei, "Videoset: Video summary evaluation through text," *arXiv preprint arXiv:1406.5824*, 2014.

[49] H. Pirsiavash and D. Ramanan, "Detecting activities of daily living in first-person camera views," in *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pp. 2847–2854, IEEE, 2012.

[50] A. Fathi, Y. Li, and J. M. Rehg, "Learning to recognize daily actions using gaze," *Computer Vision–ECCV*, pp. 314–327, 2012.

[51] F. De la Torre, J. Hodgins, A. Bargteil, X. Martin, J. Macey, A. Collado, and P. Beltran, "Guide to the carnegie mellon university multimodal activity (cmu-mmac) database," *Robotics Institute*, p. 135, 2008.

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

[53] K. M. Kitani, T. Okabe, Y. Sato, and A. Sugimoto, "Fast unsupervised ego-action learning for first-person sports videos," in *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pp. 3241–3248, IEEE, 2011.

[54] P. Isola, J. Xiao, A. Torralba, and A. Oliva, "What makes an image memorable?," in *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pp. 145–152, IEEE, 2011.

[55] M. Sun, A. Farhadi, and S. Seitz, "Ranking domain-specific highlights by analyzing edited videos," *Computer Vision–ECCV*, pp. 787–802, 2014.

[56] C.-Y. Lin, "Rouge: A package for automatic evaluation of summaries," in *Text summarization branches out: Proceedings of the ACL-04 workshop*, vol. 8, 2004.

[57] G. Healy, C. Gurrin, and A. F. Smeaton, "Lifelogging and EEG: utilising neural signals for sorting lifelog image data," in *Quantified Self Europe Conference*, 2014.

[58] R. Cai, L. Lu, H.-J. Zhang, and L.-H. Cai, "Highlight sound effects detection in audio stream," in *International Conference on Multimedia and Expo, 2003. Proceedings of*, vol. 3, pp. III–37, IEEE, 2003.

[59] A. Goldstein and R. Fattal, "Video stabilization using epipolar geometry," *ACM Transactions on Graphics (TOG)*, vol. 31, no. 5, p. 126, 2012.

[60] A. Karpenko, D. Jacobs, J. Baek, and M. Levoy, "Digital video stabilization and rolling shutter correction using gyroscopes," *CSTR*, vol. 1, p. 2, 2011.

[61] Y. Hoshen, G. Ben-Artzi, and S. Peleg, "Wisdom of the crowd in egocentric video curation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pp. 573–579, 2014.

[62] I. Arev, H. S. Park, Y. Sheikh, J. Hodgins, and A. Shamir, "Automatic editing of footage from multiple social cameras," *ACM Transactions on Graphics (TOG)*, vol. 33, no. 4, p. 81, 2014.