# Effects of Camera Position and Media Type on Lifelogging Images

**Katrin Wolf**[1]**, Yomna Abdelrahman**[2]**, David Schmid**[2]**, Tilman Dingler**[2]**, Albrecht Schmidt**[2]

[1]BTK - University of Art and Design
Berlin, Germany
k.wolf@btk-fh.de

[2]VIS, University of Stuttgart
Stuttgart, Germany
{firstname.lastname}@vis.uni-stuttgart.de

## ABSTRACT

With an increasing number of new camera devices entering the market, lifelogging has turned into a viable everyday practice. The promise of comprehensively capturing our life's happenings has caused adoption rates to grow, but approaches to do so greatly differ. In this paper we evaluate existing visual lifelogging capture approaches through a user study with two main capture dimensions: (1) comparing the body position where a lifelogging camera is worn: head versus chest (2) comparing the media captures: video versus stills. We equipped 30 participants with cameras on their heads and chests. That data was evaluated by subjective user ratings as well as by objective image processing analysis. Our findings indicate that (1) chest-worn devices are more stable and contain less motion blur through which feature detection by image processing algorithms works better than from head-worn cameras; 2) head-worn video cameras, however, seem to be the better choice for lifelogging as they capture more important autobiographical cues than chest-worn devices, e.g., faces that have been shown to be most relevant for recall.

## CCS Concepts

**•Human-centered computing → Empirical studies in ubiquitous and mobile computing;**

## Author Keywords

Lifelogging; wearable camera; ego-centric camera

## INTRODUCTION

To support human memory and archive information, personally taken camera images have always played an important role in people's lifes. The most recent evolution of individuals' life documenting technologies was the rise of wearable cameras to enable continuous lifelogging [24]. Previous research identified supporting recall and memory retrieval as one of the major motivations for the use of wearable lifelogging cameras [6, 11, 15, 21]. Devices like the SenseCam, the Narrative Clip, the GoPro camera or camera glasses allow continuous life capture. This development marks the switch from active to passive recording of visual (and sometimes audio) data: image recording is automatically done without the need to actively take out a camera, define its focus or release the shutter.

On the downside, capturing by default results in less control over the image focus or quality in comparison to traditionally explicit image capturing. Wolf et al. [26] mention the effects of body locations for lifelogging cameras on image quality: currently available lifelogging cameras have clips to attach the camera to clothes (e.g. the Narrative Clip), are included in glasses (e.g. Google Glass) or come with additional equipment to wear the camera on the head or in front of the chest (e.g. Gopro). The camera position thereby influences the quality of the pictures taken. For instance, when wearing the camera in front of the user's chest, hands often occlude the camera view, while rapid head movements result in blurry images of cameras that are attached to the head or embedded into glasses. Thus, lifelogging images often contain artifacts and noise, which is highly dependent on the body position the camera is attached.

In this paper, we aim to investigate the quality of images captured with lifelogging cameras. We collected lifelogging images during different everyday situations and evaluate their quality depending on the camera position as well as on both media types: static images and video. Existing tools like iPhoto support sorting, finding, and rediscovering personal images. Such tools apply image processing, which work better if the quality of the image is good, e.g. the image is sharp and showns the motive (face/object) without occlusion. Moreover, we judge images according their subjective qualitative attributes, like whether or not the content contains the cues that are relevant for recalling our memory.

Lifelogging image capturing produces large image data. Thus, automated image sorting algorithms will become an essential way to sort, navigate through, and view our personal images. Viewing images, of cause, has to meet our expectations of image quality that is influenced by traditional non-automated image capturing techniques. To support personal lifelogging image viewing, we need to better understand the characteristics of lifelogging images, including how different camera positions and media types influence the image quality. Hence, in this work we analyze how camera position (head vs. chest) and media type (video vs. still) influence the subjective quality of lifelog images and quantifiable image artifacts. We assess the media quality both qualitatively, in terms of user perception when recalling memory, and quantitatively by using image processing and computer vision algorithms.

*MUM '15, November 30-December 02, 2015, Linz, Austria*

© 2015 ACM. ISBN 978-1-4503-3605-5/15/11…$15.00

DOI: http://dx.doi.org/10.1145/2836041.2836065

This paper is motivated by the recent interest in body-worn cameras, and it contributes to the existing research body by evaluating existing lifelog capture approaches. Through both a user study and a computed image evaluation we show that video, if captured with a head-worn video camera, is the preferable choice for lifelogging regarding the image quality.

## RELATED WORK

Steve Mann proposed the use of wearable cameras as âĂŹvisual memory prostheticsâĂŹ [20], which are embedded in glass-like prototypes to enable ubiquitous real-life image capturing with the aim to create a personal photo/videographic memory prosthesis [21]. Gemmell *et al.* [11] developed a lifelogging platform called MyLifeBits to investigate the efforts of digitizing an entire lifetime and storing related documents. They considered all kinds of data, including a radio and TV capturing tool, GPS data, and to capture what analogue content people may see in their life. For data capturing they use the time-lapse camera SenseCam [12]. SenseCam is an on-body sensor-enhanced capture device that allows for passive picture taking including additional data, such as GPS. Hodges *et al.* [15] used a SenseCam in a 12-month clinical trial with a patient suffering from amnesia for reviewing experiences that had been forgotten. Chen and Jones presented within their iCLIPS project another on-body time-lapse camera for augmenting the human memory [6]. Despite video being able to capture every moment when recording, Chen and Jones argue to use photographs as watching video streams may cause a heavy information load.

Tasks like watching or scanning lifelog images are challenged by the information overload resulting from passive and long-term image capturing. Image processing and computer vision techniques enable us to browse through large ego-centric (and thus very noisy) image material. Gurrin *et al.* [14] use date, time, and GPS location for the organization of personal collections in order to enable users to efficiently search their photo archives. In other works supervised learning is used to summarize videos by identifying and recognizing activities [10, 9, 22]. Others use unsupervised approaches including scene discovery [16], story-driven summarization [19], and key frame selection [7]. Techniques include automatically detecting novelty in an image sequence [1], by appearance and geometric cues based on alignment of the captured frame sequences of the daily activities, combined with background deviation for identifying novel activity. Identifying activity classes with ego-centric vision included the segmentation of hands with active objects based on foreground extraction from the first-person view and appearance model to detect objects with weakly supervised technique [10, 18]. Fathi et.al linked objects, hands, and actions to understand activities [9]. Another approach for representing activities includes the usage of Markov models with atomic events considering object-object and object-wrist interaction with prior learning phase to identify activities [4]. Lee et.al proposed people/object-driven summaries based on regional importance cues for egocentric video browsing [13]. Doherty *et al.* [8] built automatic classifiers for visual lifelogs to infer personal lifetraits, such as people's characteristics and behavior. They were able to extract 22 distinct activities, such as meeting friends or having lunch.

In summary, cameras are increasingly used as wearable lifelogging devices on various body positions, and thus, the vision of augmenting the memory starts to become reality. The position of lifelogging cameras influence the image content and quality, and video may - compared to stills - produces an amount of data that makes it hard to watch all captured material. However, image processing algorithms are used to pre-categorize the images for easier and time-efficient lifelogging information retrieval. To our best knowledge no work has been carried out to comprehensively investigate the effect of the body position where the camera is mounted (head or chest) and of the captured media type (video and stills) on image quality, neither on quality perceived by users nor on quality evaluated through automated image processing.

## METHOD

For comparing the effect of lifelogging camera characteristics (camera position and media type) we designed an experiment where participants were wearing cameras in three different situations at the two most common body positions: head and chest. The video data was afterwards used to generate to different media types: video and stills.

The goal was to have typical lifelogging images (still and video) evaluated by the people that had worn the cameras as well as by computers through automated image analysis. For the human evaluation we measured image quality ratings, such as the value of the image to recall the captured situation as well as perspective, camera motion, and occlusion. For the computer evaluation we compared quantitative aspects of image processing and computer vision techniques between both, the head- and the chest-worn camera data. These aspects are further classified into 1. image quality such as video sharpness, 2. feature extraction including hand and face detection and 3. scene classification i.e. foreground and background segmentation.

### *Participants*

We recruited 30 participants (25 males, 5 females) comprising university employees and participants of a one week lasting research seminar with different academic backgrounds, such as computer science, psychology, and digital media design. Their age ranged between 23 and 50 years (M=30.7, SD=7.4).

### *Apparatus*

For image capturing we used camera glasses (OctaCam HDC-700) that record video with 30fps. During the study participants were wearing two camera glasses at the same time, one on the nose and the other one mounted at the chest using a chest-band, see figure 1. We used the same cameras on both body positions to ensure equal image quality of the two camera positions.

As video encoding often introduces additional compression artifacts, we aimed to reduce quality loss. Thus, for the video (of which also the stills were extracted) we chose a very high quality video codec, a high resolution, and a high bit rate (video codec: H264 - MPEG - 4 AVC (part 10) (avc1), resolution: 1280px x 720px, frame rate: 30fps, bit rate: 8 Mbps).

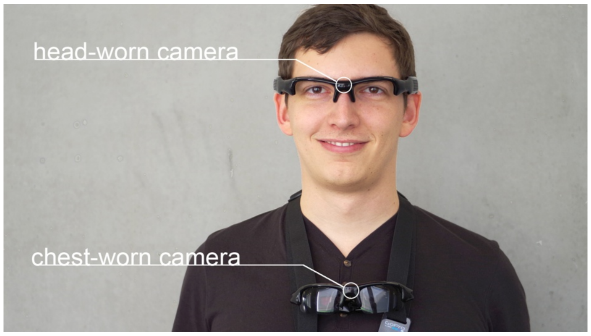

**Figure 1. Participant equipped with cameras on head and chest.**

For the video and video still evaluation, we developed a software tool that trimmed the two videos of each participant to an equal number of chunks of 30 seconds. To create the video stills, the first frame of each 30 second clip was exported as a snapshot.

That procedure served as preparation to present the two different media types (video, still) taken from two camera positions (head, chest) in smaller pieces to the participants one day after they had captured the material. To focus on the effects of the camera position on image quality no video clip contained audio. This step also helped to avoid privacy concerns of the participants.

During the image evaluation, the tool presented the 30sec video clips and the stills from both, the head- and the chest-worn camera in random order to the participants. The video clips were shown in twice the speed to reduce the overall session time. Underneath each video clip and still four Likert-Scale structured questions were presented. In addition, before finishing the session, an open questionnaire was presented to the participants.

### *Tasks*

The study was split in two tasks: (1) a video/still capture and (2) a video/still review task. During the first task, participants were asked to wear the two cameras in three different situations that were chosen to cover situations that challenge lifelogging capturing in various ways:

1. Attending a meeting
2. Walking over university campus
3. Having lunch or dinner with a group of colleagues

Thus, we had (1.) a situation where the chest most probably would be stable while the head may move a lot (2.) a mobile scenario (3.) a situation that contained manual activities with potential camera occlusions. We divided the participants into three groups of 10 each, where each participant was exposed to one of the three situations.

### *Measurements*

In summary, we recorded 60 videos, 30 videos with a head- and 30 with a chest-worn camera with 30fps and a resolution of 1280x720px. In total we collected 545 minutes (272.5 per camera), 18.2 minutes per participant. The evaluation tool recorded 7 item Likert-scale ratings answering the following statements that were presented after each still image and video clip:

- The perspective was perfect: Totally - Not at all
- Camera motion decreased the image quality: Totally - Not at all
- Occlusion decreased the image quality: Totally - Not at all
- The captured material shows information that helps me remembering the situation: Totally - Not at all

Furthermore, the tool recorded qualitative comments according to the following question presented at the end of the session and after all still images and video clips have been evaluated:

- What content shown in the images/video helped most to recall the situation?

In summary, we collected 1744 ratings, 872 per camera position for each media type (video, still). Further, 30 qualitative comments were collected (one per participant).

### *Procedure*

The experiment consisted of two sessions: during the first session and after filling the consent form, we equipped participants with the camera glasses for wearing them in one of the three situations described. For the second session, we invited participants to come back to our lab one day after the first session as then the events of the previous day were already committed to long-term memory [2]. In this session we asked participants to watch the captured material from the previous day, which had been prepared as video and as still by the evaluation tool. The image type (video or still) as well as the camera position (head, chest) was randomized. After watching each still image or 30-second video clip, participants answered the four Likert-scale structured questions regarding perceived image quality. When all material was presented and evaluated, participants were asked the open question on content that supports memory recall. Demographic questions were presented by the evaluation tool before the session started.

### *Design*

Our study had a 3x2x2 mixed design with the between-groups variable task (attending a meeting (1), walking across university campus (2), having a meal with colleagues (3)) and the within-subject variables camera position (head, chest) and media type (video, still). Each participant group consisted of 10 people, and each group solved one of the three tasks. The dependent variables were perceived image quality (measured as property ratings and qualitative opinions) and objective image quality (measured in an image and feature evaluation through image processing and computer vision).

The image features and properties were chosen in respect to how they generally support lifelogging video sorting and finding. From autobiographical memory research we know, that persons, activities, and places are important information to recall past live events [23]. Feature recognition techniques, such as face and hand detection, are already used to analyze life-logging videos and to identify activities and events [9,

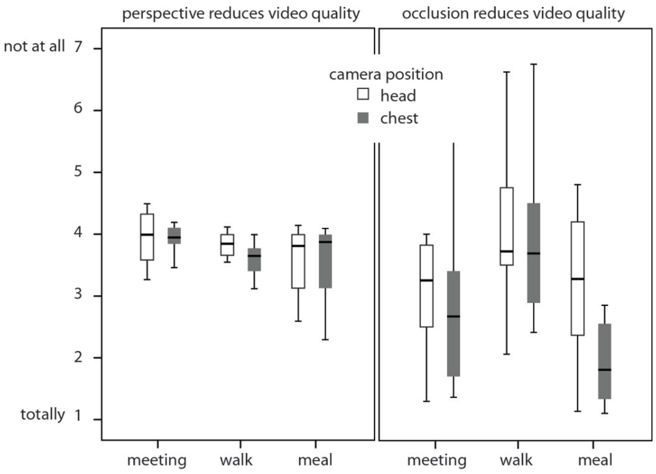

Figure 2. Boxplots for perceived video quality reduction through occlusion or perspective (min=1: quality was totally reduced through occlusion or perspective, max=7: quality was not at all influenced) per task.

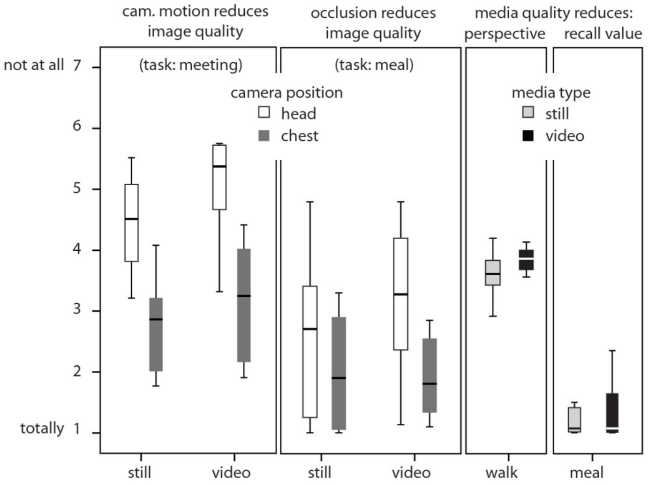

Figure 3. Boxplots for camera motion based media quality reduction during meetings (left), occlusion based media quality reduction during meals (center), and media quality based recall reduction during meals (min=1: totally reduced quality, max=7: quality was not effected).

13]. Aiming to explore which camera position works best for lifelogging, we used standard feature detection and image analysis techniques to compare their success for lifelogging material captured with head- versus chest-worn camera. To investigate whether video or stills provide better lifelogging data, we also compared the outcome of image processing algorithms using the video as well as the still data.

## PERCEIVED IMAGE QUALITIES

Using structured Likert-scale as well as open questions, we collected data to investigate the perceived image qualities (camera perspective, camera motion-based noise, and occlusion) and the type of content that supports recalling the captured situation.

### Results

For analyzing the quantitative results, we conducted Kruskal-Wallis Tests to indicate significant differences between the within-subjects variable *task*, and Post-hoc analysis with Mann-Whitney U tests were conducted with a Bonferroni correction applied, resulting in a significance level set at .017. Wilcoxon Signed-Rank Tests were used to analyze the between-subjects variables *camera position* and *media type*. For the qualitative results we used a bottom-up analysis and aggregated the comments into memory recall cue categories.

#### *Perspective*

For perspective ratings, we found significant differences between the three tasks regarding the video captured with the chest-worn camera (chest_video: $H(2)$=6.543, $p$=.038, see Figure 2, left), while perspective got no significantly different ratings in the other three conditions (head_still: $H(2)$=1.661, $p$=.436; chest_still: $H(2)$=3.814, $p$=.149; head_video: $H(2)$=1.667, $p$=.435). Post-hoc tests showed for the video captured with the chest-worn camera that the perspective while walking was significantly worse rated than during meeting situations ($U$=15.5, $p$=.009, see figure 2, left), but no significant differences regarding the perspective ratings were found between the other tasks (meeting vs. meal: $U$=30.5, $p$=.140; walk vs. meal: $U$=39.0, $p$=.405). Furthermore, video was rated better than still regarding the perspective when using the head-worn camera while walking ($Z$=-2.040, $p$=.041, see figure 3, right). The other comparisons of the within variables did not show significant effects on perspective ($p>$.05).

#### *Camera motion-based noise*

No significant difference was found in the camera motion based ratings between the three tasks, neither for stills captured with a head-worn camera (head_still: $H(2)$=1.454, $p$=.483) nor for stills captured with the chest-worn camera (chest_still: $H(2)$=0.314, $p$=.855).

For the meeting task, we found that wearing the camera on the chest reduces the motion-caused image quality significantly in comparison to wearing the camera on the head for both media types (still: $Z$=-2.293, $p$=.022; video: $Z$=-2.497, $p$=.013, see figure 3, left). Moreover, camera motion was perceived worse for stills compared to video when the camera was worn on the head during meetings ($Z$=-2.803, $p$=.005). Furthermore, the rating of camera motion for stills while walking show that the image quality was disturbed more using the chest- than using head-worn camera ($Z$=-2.091, $p$=.037). The other within-variable comparisons did not yield significant differences for the walking task ($p>$.05). During the meals, the camera motion reduced the perceived image as well as the video quality less when the camera was worn on the head (still: $Z$=-1.988, $p$=.047; video: $Z$=-2.499, $p$=.012).

#### *Occlusion*

Regarding the image quality reduction through occlusion ratings, we found significant differences between the tasks for the data captured with the chest-worn camera (chest_still: $H(2)$=6.289, $p$=.043; chest_video: $H(2)$=12.126, $p$=.002), while occlusion was not significantly different rated if the camera was worn on the head (head_still: $H(2)$=0.886, $p$=.642; head_video: $H(2)$=1.957, $p$=.376). Although, the omnibus

**Table 1. Beneficial memory cues represented in video and still content to recall the captured tasks (meeting, walk, meal).**

| Meeting | Walking | Having a meal |
|---|---|---|
| Persons (11) | Persons (12) | Persons (9) |
| - persons (6) | - persons (4) | - faces (2) |
| - faces (2) | - faces (3) | - gestures (2) |
| - emotions (1) | - passing people (2) | - passing people (2) |
| - persons speaking (1) | - persons speaking (2) | - persons (2) |
| - gestures (1) | - myself (own hands) (1) | - persons speaking (1) |
| Place (3) | Place (6) | Place (5) |
| - environmental overview (3) | - landmarks, e.g. restaurant name, buildings (6) | - environmental details, e.g. furniture (2) |
| | | - overview over the scene (2) |
| Objects (11) | Objects (2) | - menu with restaurant name (1) |
| - objects (4) | - objects (1) | |
| - PC / phone content (4) | - objects in my hand (1) | Objects (3) |
| - objects handed by persons (3) | | - food (3) |
| | Actions (2) | |
| Actions (1) | - specific situations (1) | Actions (1) |
| - situation change (1) | - weather (1) | - specific situations, e.g. people entering the scene (1) |
| | | Time (5) |
| | | - states refering to time, half eaten meal, half-full glass of juice (3) |
| | | - time on the phone (1) |
| | | - recording time of the video (1) |

test for the still data captured with chest was significant for the tasks, the post-hoc test could not confirm significant differences in occlusion ratings for the stills captured using a chest camera (meeting vs. walk: $U$=42.0, $p$=.041; meeting vs. meal: $U$=21.0, $p$=.028; walk vs. meal: $U$=23.0, $p$=.041; using a Bonferroni correction significance level set of .017). In contrast to the between-group test, within-factor test indicated significant differences, namely the video captured with the chest caused occlusion-based quality reduction during meals more than during walks ($U$=5.0, $p$=.001, see figure 2, right), while no difference was found between the other task pairs (meeting vs. walk: $U$=28.4, $p$=.096; meeting vs. meal: $U$=25.0, $p$=.059). During the meals, the chest camera was affected by occlusion significantly more for still and video than the head camera (still: $Z$=-2.310 $p$=.021, video: $Z$=-2.599 $p$=.009, as shown in figure 3, center).

### *Recall of still/video*

No significant difference was found regarding the memory recall value between the three tasks. Neither for stills extracted from video and recorded with a head-worn camera(head_still: $H(2)$=1.518, $p$=.468) nor for the camera worn on the chest (chest_still: $H(2)$=2.381, $p$=.304). Moreover, we found neither significant differences regarding the recall value ratings for the video recorded with a camera worn on the head (head_video: $H(2)$=1.053, $p$=.591) nor on the chest (chest_video: $H(2)$=4.494, $p$=.106). During the meals, the still images were rated to reduce the situation recall value of the lifelogging data when wearing a chest camera ($Z$=-2.100 $p$=.036, figure 3, right), while neither the media type was influencing the recall value of data captured with a head-worn camera nor did the camera position comparing same media types ($p>.05$).

### *Content to support memory recall*

We collected qualitative comments on still and video content that supports memory recall. That data was analyzed by grouping the participants' comments per task (meeting, walk, meal) into semantic categories that refer to memory cues, which are known from cognitive psychology to retrieve information of autobiographical memory: objects, actions, place, time, and (attitudes towards) persons, [23], see table 1.

Content-wise, capturing persons was rated to be most appreciated for video lifelogging, which includes faces, persons one talks to, passing people, the gestures of persons, and their own hands. Hands were also named in order to recall activities, for instance when they were holding objects. Objects in general were mentioned with a strong dependency on the situation, e.g. food during the meals and phones during meetings. Finally, places were named to help to remember the situation, where restaurant names as well as specific furniture or buildings are examples to recall a place.

## Summary

The perceived image quality of both, head- and chest-worn camera, depends highly on the task. For instance, the chest-worn camera stills are rated extraordinarily bad while having a meal because of occlusion, while the chest-camera perspective was rated worst while walking. Moreover, video led to better recall results than still images using a head-worn camera during meals, while the movements of the head-worn camera caused less image quality reduction than motions of the chest-worn camera. Thus, the quality perception is over all image properties mainly rated better for the head-worn camera captured lifelogging data regarding perspective, occlusion, camera motion, and situation recall value.

From the qualitative comments about naming content that helps to recall a situation, we found that (1) the categories for autobiographical memory cues can also be used to classify the lifelogging memory recall cues that we aggregated out of our participants' comments. (2) Most comments relate to

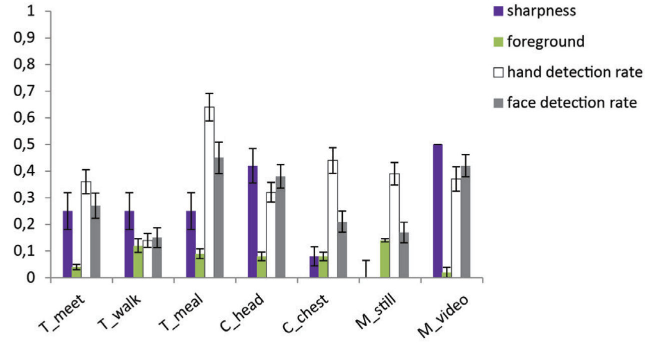

**Figure 4. Mean & SE of sharpness, foreground change, and hand as well as face detection rate per task (T), camera position (C), and media type (M). The values represent percentage with min=0, max=1.**

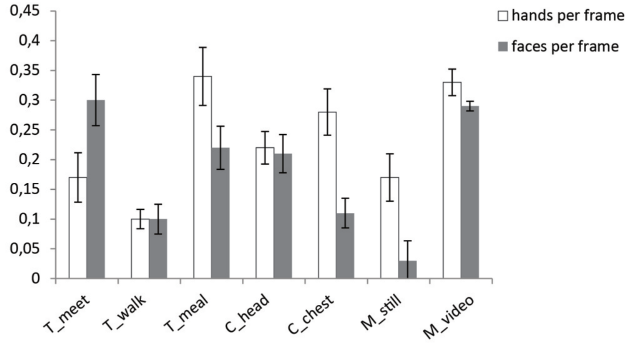

**Figure 5. Mean & SE of detected hand as well as detected face per frame for task (T), camera position (C), and media type (M). The values represent absolute numbers.**

the category *persons*, including faces and hands, followed by *objects* and *places*.

The quality of both media types, video and still, was perceived much higher when the lifelogging camera was worn on the head. Finally, from considering the subjective ratings of the lifelogging images, participants rated the memory recall value of videos higher than of stills.

## OBJECTIVE IMAGE QUALITY

As known from autobiographical memory theories and as shown in the previous section for lifelogging data, faces and actions are highly important cues to recall personal memory. Going through lifelogging data, such as watching personal photos or videos, is incredibly time consuming, and thus, image processing is an adequate approach to categorize such data to quickly find specific photographical stills or videos.

In general when applying image processing and computer vision we need a certain quality of sharpness. Otherwise useful feature detection algorithms, such as face and hand detection algorithms as well as for those extracting foreground and background, would fail. Moreover, foreground extraction can improve face and hand detection, which we found are relevant cues to recall situations out of lifelogging data. Thus, in the following section, we use standard image processing algorithms to investigate whether lifelogging images captured with head- or with chest-worn cameras as well as video or still material gain better results in image analyses and feature detection. Therefore, we conducted quantitative analyses of image property and feature extractions using our captured data. We utilized common image processing and computer vision techniques with the aid of openCV library. The property of interest includes image sharpness, while the features cover foreground changes as well as face and hand detection. Since we lack a reference image for evaluating the image property (sharpness), the evaluation for the values was done relative to the camera position.

### Results

The descriptive statistics (means and standard deviation) for the analyzed features and property are shown in table 2. The average values and standard errors of the sharpness property as well as of the features foreground change, face detection rate and hand detection rate are presented in figure 4. The average values of the absolute amount of detected faces and hands per frame including the standard error are shown in figure 5. We tested with mixed ANOVAs the influence of *task* as between-subjects factor and *camera position* and *media type* as within-subjects factors on image property and features. In case of a significant effect of *task*, we used Bonferroni corrected post-hoc tests to indicated significant differences between the three tasks.

#### *Image Sharpness*

One of the common ways of estimating image sharpness is using edge detection to define sharp intensity changes in an image. We applied Laplace filters to estimate the sharpness of the image as it computes the second spatial derivative of an image which results in zero values for the smooth i.e. blurred part of the image and peek values for edges. Picking the maximum value of the image reflects how sharp the image is which could be used for comparison purposes. To test and validate this approach we blurred the same image and compared the maximum values, which complied with our assumption (i.e. the sharper the image the higher the maximum value of the image data is). We computed the average frame sharpness for the captured videos from both cameras, and assigned binary values for each camera position, one for the position with higher average sharpness and zero for the one with less average sharpness.

A mixed ANOVA showed that pictures taken with a head-worn camera were significantly sharper than pictures taken with a chest-worn camera ($F_{1,27} = 24.324$, $p<.001$), but there were no sharpness differences between the three different *tasks* ($F_{2,27} = 0.895$, $p$=.413). There was a statistically significant interaction between the position of the camera and *media type* on sharpness, ($F_{1,27} = 14.738$, $p<.001$).

#### *Foreground Extraction and Change Detection*

The foreground is extracted by modeling and subtracting the background of an image. To accurately extract the foreground, a background model is computed using an accumulated weighted model for each pixel [27]. It is chosen to allow a dynamic adaptation of the background. The dynamic update is dependent on the learning rate parameter ($\alpha$) that controls

**Table 2. Mean values and SD per image property (sharpness) and image features (foreground change, hands detected per frame, hand detection rate, faces detected per frame, face detection rate) for each task (T), camera position (C), and media type (M).**

| | **Sharpness** | **Foreground** | **Hands pFrame** | **Hand dRate** | **Faces pFrame** | **Face dRate** |
|---|---|---|---|---|---|---|
| | (min=0/max=1) | (min=0/max=1) | (abs. values) | (min=0/max=1) | (abs. values) | (min=0/max=1) |
| Task_meet | 0.25 (0.44) | 0.04 (0.06) | 0.17 (0.27) | 0.36 (0.29) | 0.30 (0.26) | 0.27 (0.30) |
| Task_walk | 0.25 (0.44) | 0.12 (0.16) | 0.10 (0.10) | 0.14 (0.17) | 0.10 (0.16) | 0.15 (0.23) |
| Task_meal | 0.25 (0.44) | 0.09 (0.12) | 0.34 (0.31) | 0.64 (0.33) | 0.22 (0.23) | 0.45 (0.37) |
| CamPos_head | 0.42 (0.50) | 0.08 (0.13) | 0.22 (0.21) | 0.32 (0.29) | 0.21 (0.25) | 0.38 (0.34) |
| CamPos_chest | 0.08 (0.28) | 0.08 (0.12) | 0.28 (0.30) | 0.44 (0.37) | 0.11 (0.19) | 0.21 (0.30) |
| Media_still | 0.00 (0.00) | 0.14 (0.15) | 0.17 (0.17) | 0.39 (0.35) | 0.03 (0.06) | 0.17 (0.32) |
| Media_video | 0.50 (0.50) | 0.02 (0.04) | 0.33 (0.31) | 0.37 (0.32) | 0.29 (0.26) | 0.42 (0.30) |

how fast the background model is updated. The $\alpha$ value lies between zero and one and can be adjusted to either maintain a static or dynamic background model. The higher $\alpha$, the more sensitive the background model becomes to changes in the image sequence. Since the cameras are in a rapidly changing environment, we systematically tested the background modeling algorithm with different values. The $\alpha$ value of 0.1 showed best results in terms of capability of rapidly updating the background to adapt to the changing background yet extracting reasonable foreground. The extracted foreground is compared to previously extracted foregrounds to define the total difference resulting in values between 0 and 1 where 0 is identical and 1 is a totally different foreground.

The foreground change did neither significantly differ between the *tasks* ($F_{2,27} = 2.589$, $p$=.094) nor between the camera positions ($F_{1,27} = 1.345$, $p$=.256), but the *media type* influenced the foreground change significantly ($F_{1,27} = 29.124$, $p<.001$).

### Face detection

Using common face detection algorithm based on the openCV cascade classifier namely "haarcascade_frontalface_alt". We computed both the detection rate (as the percentage of true positives of all detected faces) and the amount of correctly detected faces per frame to reflect the influence of different tasks, of different camera positions and different media types on the ability to detect faces. Examples for correctly as well as for incorrectly detected faces per camera position including all tasks are presented in figure 6.

The face detection rate was significantly higher for lifelogging data captured with the head-worn camera versus one worn at the chest ($F_{1,27} = 16.579$, $p<.001$). Moreover, the video *media type* led to detect significantly more faces correctly than the stills ($F_{1,27} = 36.366$, $p<.001$). The detection rate for faces differed significantly between the *tasks* ($F_{2,27} = 4.490$, $p$=.021): the face detection rate was significantly lower during the walking task versus the meal situation ($p$=.020), while meeting and walking ($p$=.968) as well as meeting and meal was not different ($p$=.188).

The amount of correctly detected faces did not differ significantly between the *tasks* ($F_{2,27} = 2.810$, $p$=.078). The actual amount of correctly detected faces per frame was significantly higher for material captured with a head-worn camera versus one worn at the chest ($F_{1,27} = 16.677$, $p<.001$). Moreover, the video *media type* leads to detect significantly more faces per frame than the stills ($F_{1,27} = 59.178$, $p<.001$).

### Hand detection

Based on the same evaluation approach used in detecting faces, a hand cascade classifier has been used to detect hands. We were only interested in detecting the user wearing the camera hands, thus further filtering was applied on the detected hands to define the detection rate of the hands. This filtering is realized by computing orientation of the hands, where the convex hull of the detected hands are computed and then finger positions are extracted based on the local maximum distance from the hands center. Based on the finger positions and hand center we estimate the orientation of the hands. For instance, conditions included the hand center should not exceed the finger tips. Both the recognition rate and the number of hands per frame were computed for the two camera positions. Examples for correctly as well as for incorrectly detected hands of the participants while capturing egocentric video per camera position and for all tasks are presented in figure 6.

The hand detection rate (true positives) out of the entire amount of detected hands (see below) was significantly higher for material captured with the chest-worn camera versus one worn at the head ($F_{1,27} = 6.757$, $p$=.015), while the media type had no significant influence on the hand detection rate ($F_{1,27} = 0.282$, $p$=.600). The percentage of correctly detected hands differed significantly between the *tasks* ($F_{2,27} = 20.707$, $p<.001$) and a Bonferroni corrected post-hoc test yielded that the percentage of true positives in the detected hands was significantly different for each *task*. Walking results in worst hand detection, and most hands were correctly detected during a *meal* (meeting vs. walk: $p$=.030, meeting vs. meal: $p$=.003, and walk vs. meal: $p<.001$).

The actual amount of correctly detected hands per frame was not significantly different for material captured with a chest-worn camera versus one worn at the head ($F_{1,27} = 3.966$, $p$=.057). Moreover, the video *media type* leads to detect significantly more hands per frame than the stills ($F_{1,27} = 20.099$, $p<.001$). The number of detected hands differed significantly between the *tasks* ($F_{2,27} = 7.236$, $p$=.003), while significantly less hands were detected during a walk compared with both, meetings ($p$=.018) and meals situation ($p$=.004). The amount

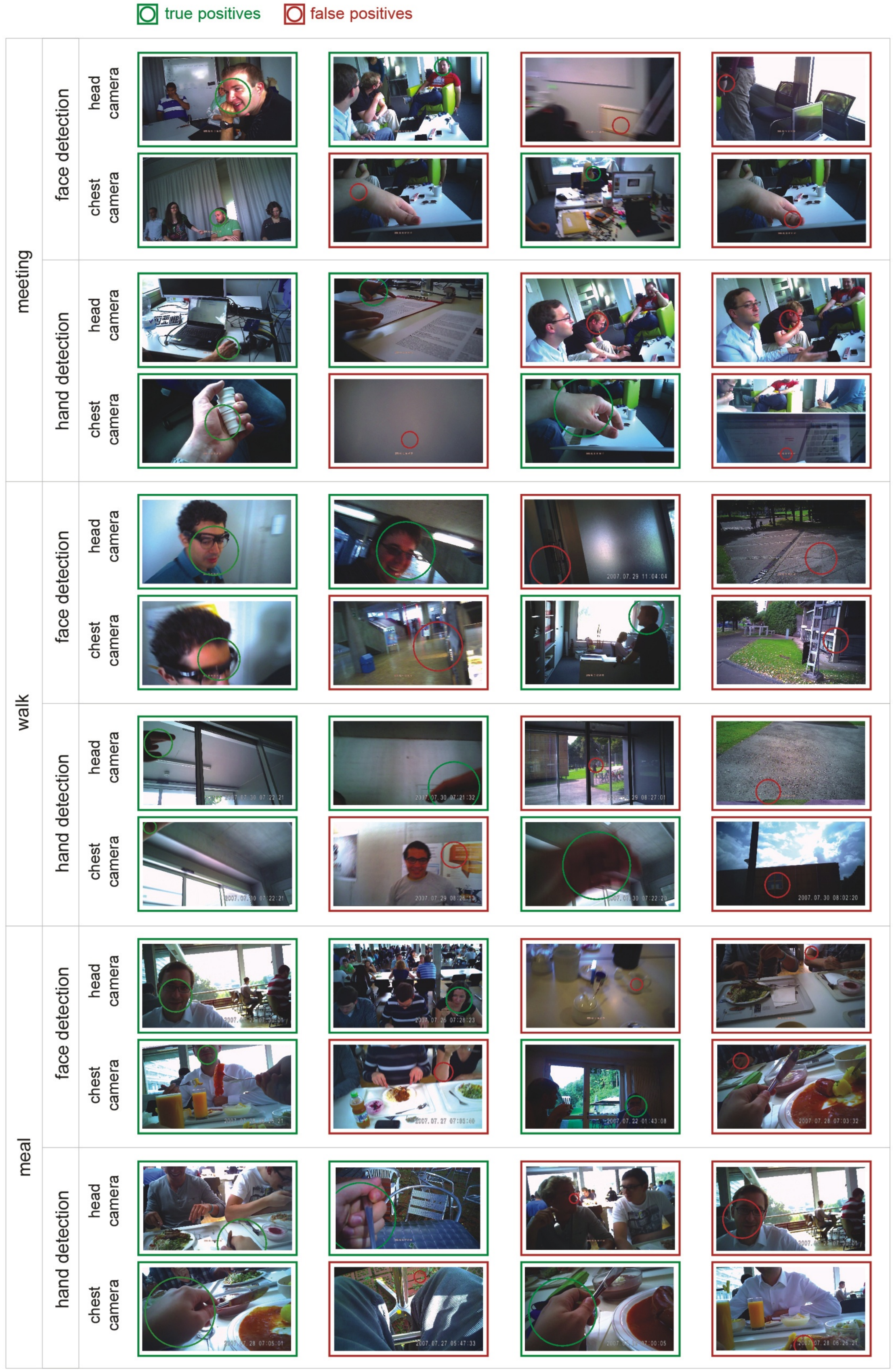

**Figure 6. Examples of correctly (true positives: marked with green frame) and incorrectly (true negative: marked with red frame) detected faces and hands per task (meeting, walk, meal), arranged in pairs: one is recorded with a head- & one with a chest-worn camera, both in the same moment.**

of detected hands did not differ between a meeting versus while having a meal ($p$=1.000) as shown with a Bonferroni corrected post-hoc test.

### Summary

During our video and still analyses, we gained camera position based and media type dependent insights into image characteristics of lifelogging camera data. Hence, we documented our experience applying image processing on still as well as on video data that was driven by the aim to analyze effects on lifelogging image quality. In the following paragraphs, we provide a discussion of the limitations and benefits of wearing a lifelogging camera on the chest or on the head, and we discuss what media type may result in better feature analysis applying basic image processing algorithms. Finally, we suggest more advanced image processing approaches that in future works may improve the analysis' results by being more appropriate for lifelogging image processing than the basic algorithms we had applied.

#### *Head- versus chest-worn cameras*

In general, sharpness is indirectly related to face, hand, and activity detection as it can reduce the images quality to an extent that impairs image processing algorithms. We found that images from the head-worn cameras on average depicted higher sharpness than the chest-worn. That may be unexpected as the head most probably moves more than the chest, and motions reduce sharpness in image capturing. However, the chest-camera is often blurred with hand motions, e.g. while eating or handling with objects, which is a possible explanations for the sharpness lack of chest-worn cameras. Hence, the image processing analyses confirm what we have found when asking participants to rate the sharpness as they rated the head-worn camera images to be sharper than whose of the chest-worn device.

The camera positions also influenced the face detection rate as the head position camera captured more faces compared with the chest. An explaination of that result is that the head perspective often focuses on faces while the chest view might not include faces due to either occlusion by focussing lower positioned other objects (table or hands, see figure 6, row 3, 4, collumn 3, 4). Furthermore, different tasks influenced the face detection rate, which was due to the static position and the typical sitting situation highest during the meal and meeting, while it is lowest when walking. However, the different camera positions did not influence the hands per frame, more hands were detected correctly with the chest-worn camera.

#### *Video versus still*

The videos showed higher values for detected faces per frame than stills, which does not seem to be straight forward as the still is representing the same content but with less frames. Two reasons may cause the better face detection for video compared to stills: Firstly, the fewer faces per frame may be a results of more random content selection if only images taken twice a minute are considered to represent a scene. Secondly, the better images quality from the videos in terms of sharpness and foreground stability reflects a higher number of faces per frame and higher detection rate, where more face were detected correctly from video in comparison with the still.

#### *Suggestions for algorithm improvement*

In our approach we utilized basic computer vision algorithms for image features and properties extraction, however deploying more advanced techniques might lead to better results. For example, Kopf *et al.* [17] developed Hyperlapse, a method for converting video captured with a wearable camera with too many movements into a video that appears as if it would have been taken with a smoothly moving camera through reconstructing the 3D input camera path and then optimizing a novel camera path and cropping the output video accordingly.

However in this work we aimed to evaluate how well feature detection on lifelogging images works already when using basic algorithms, more advanced approaches are suggested for future work, including:

- Applying intensive image preprocessing before the analyses phase by applying:
  - Noise reduction filters (e.g. median filters).
  - Image stabilization and rectification
- Adding further constraints on the face detection like skin color and eye detection to reduce the false positive detection rate.

Furthermore, advanced capturing devices would significantly affect the results. For instance, wide angle cameras could be used to extend the filed of view enriching the amount of information in each frame, as well as, covering the blind spots experienced by each camera position. Also, the usage of additional sensors to detect camera motion may be useful to reduce motion blur.

## LIMITATIONS

We are aware that video stills are not a true representation of picture based lifelogging images, such as time lapse, as the shutter time of the stills differs from those of most photo cameras. In pre-study tests we made intensive use of the lifelogging photo camera Narrative Clip, and we experienced that the image quality of this state-of-the-art device is worse than images of traditional cameras and rather comparable with video stills. We accepted that the shutter time of the stills differs from those of most photo cameras in order to keep other image characteristics equal for both media types, such as camera position and perspective, resolution, and compression algorithm. Consequently, with our setup, the only main difference of our video stills - compared to common photo camera images - is the shutter time, which may add blur to the images. This can affect some of our analysis, e.g. face detection and perceived image quality. However, we had decided to not use two different camera types as we were particularly interested in comparing media that showed the same content captured at the same time, with exactly the same camera position, and with an equal perspective.

## DISCUSSION & CONCLUSION

Different image capturing devices have been proposed for lifelogging. They mainly differ regarding the wearing position and the media they capture. Our work shows that the choice of recorded media type as well as the camera position impact the value of the material captured. We analyzed the quality of video and stills for two camera positions and when the device was attached to the head or chest, through both objective ratings as well as image processing. We moreover gained some qualitative feedback about content that was important in such video to recall the situation.

Our analysis shows that video captured by head-worn cameras gives better memory hints about the situation recorded: the captured features were perceived more useful by participants, which is in line with previous research suggesting people, places, and time to be appropriate autobiographical memory cues (e.g., [3], [5], [25]).

The quality of both media types, video and still, was subjectively preferred when the camera was head-worn. This position was further favored for video recordings over images/stills, especially with regard to support memory recall. Since faces and the resulting ability to recognize people are a key memory cue, the perception of footage containing faces was favored by participants. Because manual review of lifelogging footage is highly time consuming, the ability to automatically extract faces is vital. Head-worn cameras produce images of higher objective quality than chest-worn devices. Moreover, more relevant content is captures, primary faces. Consequently, automated video analyses can better detect the relevant information, for example, using face detection algorithms. Hence, the head-worn position facilitates both recall and automated video indexing. However, due to limited battery life there is no wearable camera currently on the market that allows for 24h video recording, although this issue may be solved in the near future.

Thus, our conclusions are: (1) Faces are the most important content in people-centric situations in lifelogging video. (2) Head-worn video cameras are preferred for capturing lifelogging video as they provide video footage that contain faces of the people we interact with in life. Further, they produce material showing recall relevant content. (3) Finally, video better supports memory recall as it contains continuous life information while still/images miss some information and show, for example, less faces per frame than video.

However, there is a trade-off between information richness (video) and time required to watch the material (still). Hence, a more holistic view on lifelogging video representation is needed which is worth investigating in future works.

In summary, this work contributes to the domain of lifelogging image capture by showing that video provides more beneficial information than stills, while head-worn cameras capture more subjectively valuable content, e.g. faces. Hence, we conclude that a head-worn video camera works best as lifelogging camera device. Our study provides a comprehensive overview of lifelog camera images, thereby providing pointers as to when to select certain camera devices, positions, and media types.

## ACKNOWLEDGEMENTS

The authors thank the participants of our study and acknowledge the financial support of the Future and Emerging Technologies (FET) programme within the 7th Framework Programme for Research of the European Commission, under FET grant number: 612933 (RECALL).

## REFERENCES

1. Omid Aghazadeh, Josephine Sullivan, and Stefan Carlsson. 2011. Novelty detection from an ego-centric perspective. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*. IEEE, 3297–3304.
2. Richard C Atkinson and Richard M Shiffrin. 1968. Human memory: A proposed system and its control processes. *Psychology of learning and motivation* 2 (1968), 89–195.
3. Lawrence W Barsalou. 1988. The content and organization of autobiographical memories. *Remembering reconsidered: Ecological and traditional approaches to the study of memory* (1988), 193–243.
4. Ardhendu Behera, David C Hogg, and Anthony G Cohn. 2013. Egocentric activity monitoring and recovery. In *Computer Vision–ACCV 2012*. Springer, 519–532.
5. Christopher DB Burt. 1992. Retrieval characteristics of autobiographical memories: Event and date information. *Applied Cognitive Psychology* 6, 5 (1992), 389–404.
6. Yi Chen and Gareth JF Jones. 2010. Augmenting human memory using personal lifelogs. In *Proceedings of the 1st Augmented Human International Conference*. ACM, 24.
7. Aiden R Doherty, Daragh Byrne, Alan F Smeaton, Gareth JF Jones, and Mark Hughes. 2008. 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*. ACM, 259–268.
8. Aiden R Doherty, Niamh Caprani, Ciarán Ó Conaire, Vaiva Kalnikaite, Cathal Gurrin, Alan F Smeaton, and Noel E OâĂŹConnor. 2011. Passively recognising human activities through lifelogging. *Computers in Human Behavior* 27, 5 (2011), 1948–1958.
9. Alireza Fathi, Ali Farhadi, and James M Rehg. 2011a. Understanding egocentric activities. In *Computer Vision (ICCV), 2011 IEEE International Conference on*. IEEE, 407–414.
10. Alireza Fathi, Xiaofeng Ren, and James M Rehg. 2011b. Learning to recognize objects in egocentric activities. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference On*. IEEE, 3281–3288.
11. Jim Gemmell, Gordon Bell, and Roger Lueder. 2006. MyLifeBits: a personal database for everything. *Commun. ACM* 49, 1 (2006), 88–95.
12. Jim Gemmell, Lyndsay Williams, Ken Wood, Roger Lueder, and Gordon Bell. 2004. 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*. ACM, 48–55.

13. Joydeep Ghosh, Yong Jae Lee, and Kristen Grauman. 2012. Discovering important people and objects for egocentric video summarization. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 1346–1353.

14. Cathal Gurrin, Gareth JF Jones, Hyowon Lee, Neil O'Hare, Alan F Smeaton, and Noel Murphy. 2005. Mobile access to personal digital photograph archives. In *Proceedings of the 7th international conference on Human computer interaction with mobile devices & services*. ACM, 311–314.

15. Steve Hodges, Lyndsay Williams, Emma Berry, Shahram Izadi, James Srinivasan, Alex Butler, Gavin Smyth, Narinder Kapur, and Ken Wood. 2006. SenseCam: A retrospective memory aid. In *UbiComp 2006: Ubiquitous Computing*. Springer, 177–193.

16. Nebojsa Jojic, Alessandro Perina, and Vittorio Murino. 2010. Structural epitome: A way to summarize oneâĂŹs visual experience. In *Advances in neural information processing systems*. 1027–1035.

17. Johannes Kopf, Michael F Cohen, and Richard Szeliski. 2014. First-person hyper-lapse videos. *ACM Transactions on Graphics (TOG)* 33, 4 (2014), 78.

18. Cheng Li and Kris M Kitani. 2013. Pixel-level hand detection in ego-centric videos. In *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*. IEEE, 3570–3577.

19. Zheng Lu and Kristen Grauman. 2013. Story-driven summarization for egocentric video. In *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*. IEEE, 2714–2721.

20. Steve Mann. 1997. Wearable computing: A first step toward personal imaging. *Computer* 30, 2 (1997), 25–32.

21. Steve Mann. 1998. 'WearCam'(The wearable camera): personal imaging systems for long-term use in wearable tetherless computer-mediated reality and personal photo/videographic memory prosthesis. In *Wearable Computers, 1998. Digest of Papers. Second International Symposium on*. IEEE, 124–131.

22. Hamed Pirsiavash and Deva Ramanan. 2012. Detecting activities of daily living in first-person camera views. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*. IEEE, 2847–2854.

23. John A Robinson. 1976. Sampling autobiographical memory. *Cognitive Psychology* 8, 4 (1976), 578–595.

24. Abigail J Sellen and Steve Whittaker. 2010. Beyond total capture: a constructive critique of lifelogging. *Commun. ACM* 53, 5 (2010), 70–77.

25. Willem A Wagenaar. 1986. My memory: A study of autobiographical memory over six years. *Cognitive psychology* 18, 2 (1986), 225–252.

26. Katrin Wolf, Albrecht Schmidt, Agon Bexheti, and Marc Langheinrich. 2014. Lifelogging: You're Wearing a Camera? *IEEE Pervasive Computing* 13, 3 (2014), 8–12.

27. Zoran Zivkovic and Ferdinand van der Heijden. 2006. Efficient adaptive density estimation per image pixel for the task of background subtraction. *Pattern recognition letters* 27, 7 (2006), 773–780.