Video content analysis on body-worn cameras for retrospective investigation

Henri Bouma [1], Jan Baan, Frank B. ter Haar, Pieter T. Eendebak, Richard J.M. den Hollander, Gertjan J. Burghouts, Remco Wijn, Sebastiaan P. van den Broek, Jeroen H.C. van Rest

TNO, Oude Waalsdorperweg 63, 2597 AK The Hague, The Netherlands

ABSTRACT

In the security domain, cameras are important to assess critical situations. Apart from fixed surveillance cameras we observe an increasing number of sensors on mobile platforms, such as drones, vehicles and persons. Mobile cameras allow rapid and local deployment, enabling many novel applications and effects, such as the reduction of violence between police and citizens. However, the increased use of bodycams also creates potential challenges. For example: how can end-users extract information from the abundance of video, how can the information be presented, and how can an officer retrieve information efficiently? Nevertheless, such video gives the opportunity to stimulate the professionals' memory, and support complete and accurate reporting. In this paper, we show how video content analysis (VCA) can address these challenges and seize these opportunities. To this end, we focus on methods for creating a complete summary of the video, which allows quick retrieval of relevant fragments. The content analysis for summarization consists of several components, such as stabilization, scene selection, motion estimation, localization, pedestrian tracking and action recognition in the video from a bodycam. The different components and visual representations of summaries are presented for retrospective investigation.

Keywords: Surveillance, CCTV, security, bodycam, video content analysis (VCA), action recognition.

1. INTRODUCTION

The security professional is increasingly supported by technology. The camera has become an important instrument to help assess the situation, either live or retrospectively. Besides fixed surveillance cameras we see an increasing number of sensors on mobile platforms, such as UAVs (drones), vehicles and the security professional himself. This fits in the more generic development of wearable technology – such as mobile phones and smart glasses. Also in the defense sector, mobile technology is increasingly being used to support the deployed soldier. Wearable cameras allow for rapid local deployment, enabling many novel applications and effects, only now being discovered. For example, recent studies showed that the use of body-worn cameras (bodycams) by the police leads to a reduction of violence and complaints by civilians [3]. Defense forces in counter-insurgency scenarios may discover similar effects when using bodycams while interacting with the local population.

The increased use of bodycams creates several challenges and opportunities. The growing amount of video footage is challenging in terms of practical usefulness and searchability. How can we extract relevant information from such video, how can the information be presented, and how can an officer retrieve information efficiently? Nevertheless, such video gives the opportunity to stimulate the professionals memory and support complete and accurate reporting.

In this paper, we explore the opportunities for the application of bodycams and challenges that may be associated with the video footage it produces. We focus on how video content analysis (VCA) can address these challenges so that opportunities may be seized, including the creation of video summaries that allows quick retrieval of relevant fragments. This also implies developing VCA components, such as stabilization, scene selection, motion estimation, localization, pedestrian tracking and action recognition in bodycam videos. Different components and visual representations of summaries are presented for retrospective investigation. Our main contribution is that it gives an overview of state-of-the-art VCA capabilities related to the use cases for body-worn cameras. Several VCA capabilities had to be modified to allow usage on these moving cameras. Examples of those novel modifications are the selection of sharp frames for stabilization, periodic analysis for gait recognition, and the cueing virtual pan-tilt-zoom (PTZ) for action recognition.

[1] henri.bouma@tno.nl; phone +31 888 66 4054; http://www.tno.nl

H. Bouma, J. Baan, F.B. ter Haar, P.T. Eendebak, R.J.M. den Hollander, G.J. Burghouts, R. Wijn, S.P. van den Broek, J.H.C. van Rest, "Video content analysis on body-worn cameras for retrospective investigation," Proc. SPIE, vol. 9652, 2015. http://dx.doi.org/10.1117/12.2194436

The outline of the paper is as follows. The background is described in Section 2, which elaborates about the potential benefits of bodycams and the relevance of VCA for bodycams in security applications. The VCA capabilities are shown in Section 3. Section 4 describes the visual representation summary. Finally, the conclusions are presented in Section 5.

2. BACKGROUND

The public discussion about the use of bodycams has mainly focused on civil rights[2] and the reduction of violence between police forces and citizens. Unnecessary or excessive use of violence between civilians and police remains a source of concern. The use of bodycams can reduce the use of force by the police against civilians and the number of complaints about police misconduct [3]. Such results lead to high expectations of the bodycam elsewhere [19]. But realization and proof of any (positive) effects also requires implementing proper working procedures, ICT support and administration of incidents [19].

Other use cases are more in line with the main purpose of police itself: maintaining law and order. It appears that citizens are more accepting the outcome of an interaction with the police if they know that there is video of the incident. This has been shown for regular surveillance cameras (CCTV) [49], but is likely also the case for footage from bodycams. So, if the use of bodycams increases the chance that there is an adequate quality of video footage available of an incident, then this increases acceptance of outcomes, and may reduce the costs of subsequent legal procedures.

Although these are good reasons for implementing bodycams on a large scale, there are also challenges associated with bodycams. Specifically, they produce large amounts of data that need to be processed, stored and viewed by end-users. As more data is collected, more advanced and user-friendly ways to retrieve desired episodes become needed.

One way to do this is to present information to end users to aid stimulation of their memory. Not all events are remembered well by the human memory, but specifically those that seem relevant at the time. In contrast, the bodycam can capture data without loss of information. Recorded footage can therefore support the human memory, even *beyond* what was directly visible in the footage. Testimonies of all witnesses – including police officers – may be improved by reviewing the (bodycam) video footage. However, the amount of recorded data can be large and finding relevant events to stimulate or support memory can be time consuming, which forms an opportunity for the application of VCA.

2.1 Stimulating the memory

Episodic memories are memories of autobiographic events (*who, what, where* and *when* knowledge). The quality of the recall of this memory can be improved if one is presented with cues about the people involved in an event, the content of the event, the location and the time of the event etc. It was found that reviewing self-recorded images (similar to bodycam footage; reflecting episodic memory) gave rise to higher recall in both types of memories than well-known control images (reflecting only semantic memory). Good memorable cues are recognizable, distinctive (unusual or prototypical) and personally significant [30].

There are several ways to present information to a user in a graphical user interface (e.g., [11][18][25]). Presented cues should be closely related to key elements of memory triggers: *who* (e.g., number of people in the scene based on face detection), *what* (bright/dark, special/routine, semantic concepts), *where* (location, near / far), *when* (day/night, time of day, calendar selection, but also before / after). Four important types of cues can be expressed: *visual cues* (especially photos of locations, persons, actions and objects [30]), *location* (global positioning system: GPS), *temporal* (date and time) and *social cues* (people involved). These cues can be presented in different panes, e.g., showing video frames, a map with tracks and a timeline with event information.

Graphical presentation of visual cues (*Snaps*), location cues (*Tracks*) and the combination (*SnapTracks*) have been compared on its effect on memorizing events [26]. *Tracks* appeared to stimulate the inference instead of actual remembering and *Snaps* led to more recall of details than *SnapTracks*. The higher recall may result from the design of the *Snaps* interface, which shows multiple images without user intervention. This contrasts with *SnapTracks* – where users are first presented with the map visualization, which they then use to navigate to the images – making access to images less direct. However, the participants preferred the combined version, so that they could rapidly navigate through a large amount of data and then zoom in on details of interest.

[2] http://www.civilrights.org/press/2015/body-camera-principles.html

The combination of photos with textual annotations was also studied [25]. The study showed that people accessed the annotations more than photos. People felt the textual annotations provided more fine-grained indices when scanning for relevant information. For the photos it seemed more difficult to find an appropriate visual index.

Several studies showed that social cues are very important for recall [30], but the cues have not been tested in these studies since the good methods to represent social cues were not available [18]. Furthermore, vision-based shot detection performed poorly in moving body-worn cameras [11]. These studies emphasize the human perception aspect of the human-machine interface and the human recall of episodic memory, while more recent developments in video content analysis (VCA) were not taken into account.

2.2 Other use cases of bodycams

Bodycams may also help in the preparation before a police deployment. For example, consider the scheduled visit of an important person (VIP) to a new, not yet secured location. Scouting the location before the visit with the use of a bodycam can help the police – and the VIP – to create a mental map of the environment, and to assess the behavior there. It may be desirable to reconstruct the scene and assess human behavior. Furthermore, it may be useful to analyze the path of the user of a bodycam and link the data to regular fixed CCTV footage of cameras with known locations to obtain good localization and further situation awareness.

A further use case could be the live detection of aggression, either applied by the wearer of the bodycam, or directed towards him. Currently available bodycams are already equipped with manual start, stop and alarm buttons (operated by the carrier or remotely). But this kind of functionality can also be automatically initiated by live processing of the video or other nearby signals. This may be an opportunity for tools that can live detect aggression in audio and video footage, such as people – including the wearer – swearing, shouting, running, fighting, falling and lying down. A first step may be the development of a tool that automatically generates summaries of riots where many officers used bodycams.

Yet another use case, one that typically triggers privacy concerns just on its own, is the use of bodycams to recognize or even identify people or vehicles that are registered as missing, wanted or that have a restraining order. On the one hand, this requires very high quality data in terms of resolution and contrast, which may be a complication. On the other hand, the sideways viewpoint from a bodycam is optimal for the capture of such data. This may be an opportunity for automated tools that capture and process number plates and (soft) biometrics, including face, voice, gait and clothing. A first step may be the development of an automated tool that scans historical bodycam footage for number plates of stolen vehicles or faces of missing persons. However, more public records could become a large burden on the capacity of a police force when citizens exercise a right to view data that is recorded of them, especially when the privacy of other people in the scene must be respected. This creates a demand for tools that implement privacy enhancing technologies, such as the automatic blurring of faces.

Other, perhaps more farfetched use cases include the automatic interpretation and translation of speech, the detection of lies and a social interaction support module, a system that assists the user during a social interaction, similar to what is currently in use at call centers.

2.3 Relevance of VCA for bodycams in security applications

VCA solutions are used to reduce the huge amount of data that CCTV cameras produce so that the operator can focus on the most relevant parts. For example, sterile-zone monitoring alerts operators for activities in industrial zones that should be abandoned at night. More advanced analysis is emerging to find suspects [7] and to detect suspicious behavior [9]. Besides fixed CCTV cameras we see an increasing number of sensors on mobile ‘platforms’, such as UAVs, vehicles, police officers and civilians. We expect that the rapid growing market of smart phones, smart glasses and bodycams will lead to an increasing need for VCA on mobile footage. However, the existing VCA technologies developed for CCTV cameras are not directly applicable to bodycams, because they often assume that the camera is (almost) static. The VCA techniques for bodycams should be able to cope with a moving camera viewpoint and a dynamic scene.

The VCA processing can be done *online* (i.e. ‘live’ or ‘real-time’; when the police officer is in action) or *offline* (i.e. ‘delayed’ or ‘retrospective’; at the end of the day or even later when there is a specific question from an investigation). There are two reasons why we believe that offline processing will be used before online processing. The first reason is that computation, communication and energy resources are scarce. Retrospective resources are less scarce than those needed for real-time processing. The second reasons is that, although *live* usage could benefit from an early warning system, the number of false alarms must be extremely low, while retrospective use could also benefit from a more

efficient interactive search or summary application. These applications can speed up the search time, even when they are not perfect. Therefore, we focus on VCA capabilities for retrospective summarization of videos.

Table 1 shows a list of use cases related to VCA in bodycams. For each of these use cases, we show for which security functions they are useful. In this table we see that three of the security functions potentially benefit the most from the application of VCA to bodycam footage: collect proof, improve testimony and efficient reporting. For these security functions many different use cases for the use of VCA can be defined. These three functions represent three different strategic drivers for innovation in security: obtaining results, reducing errors and reducing costs.

Table 1: Use case (and related VCA capability) versus security functions.

Relevant use case or example ***(related VCA capability between brackets)***	**Reduce violence between police / citizens**	**Effective surveillance**	**Collect proof**	**Improve testimony**	**Efficient reporting**	**Facilitate public record**	**Scout location**	**Officer status**	**Riot summary**	**Amber alert**
Live VCA useful	×	×						×	×	×
Delayed VCA useful	×		×	×	×	×	×	×	×	×
Allow fast forward playback for more efficient viewing of a long video (*stabilization*)	×		×	×	×	×	×		×	×
Summarized long video by presenting thumbnails for each different environment/scene (*scene detection*)			×	×	×	×	×		×	
Obtain position information for indoor situations where GPS fails (*ego-motion estimation and localization*)		×	×	×	×	×	×	×		
Allow alignment with GPS information, allow ego-action recognition (*estimate orientation change*)	×	×	×	×	×	×	×	×		
Allow detection of changes in the environment, e.g., left luggage (*spatial synchronization*)		×	×	×	×		×			
Actions of agent may indicate whether he needs assistance, e.g., the agent is running or falling (*ego-action recognition*)	×			×	×			×	×	
Recognize suspects or missing persons or objects (*face recognition, object recognition*)		×	×	×	×	×	×			×
Recognize the officer that is wearing the camera (*gait recogn.*)	×					×				
Finding suspects, victims or witnesses (*re-identification*)	×	×	×				×		×	×
Find stolen objects or owner of left luggage (*object detection*)		×	×	×	×		×			
Find a running suspect (*action recognition*)	×	×	×	×	×				×	

3. VCA CAPABILITIES

3.1 Framework overview

Just like in other application scenarios, VCA capabilities can be used together, which requires an encompassing framework. The framework for the analysis of video from a bodycam consists of the following VCA capabilities: stabilization, scene selection, motion estimation, localization, pedestrian tracking and action recognition (see Figure 1). Finally, the information of the different VCA components is summarized in a visual representation. The components and representation are described in the following subsections.

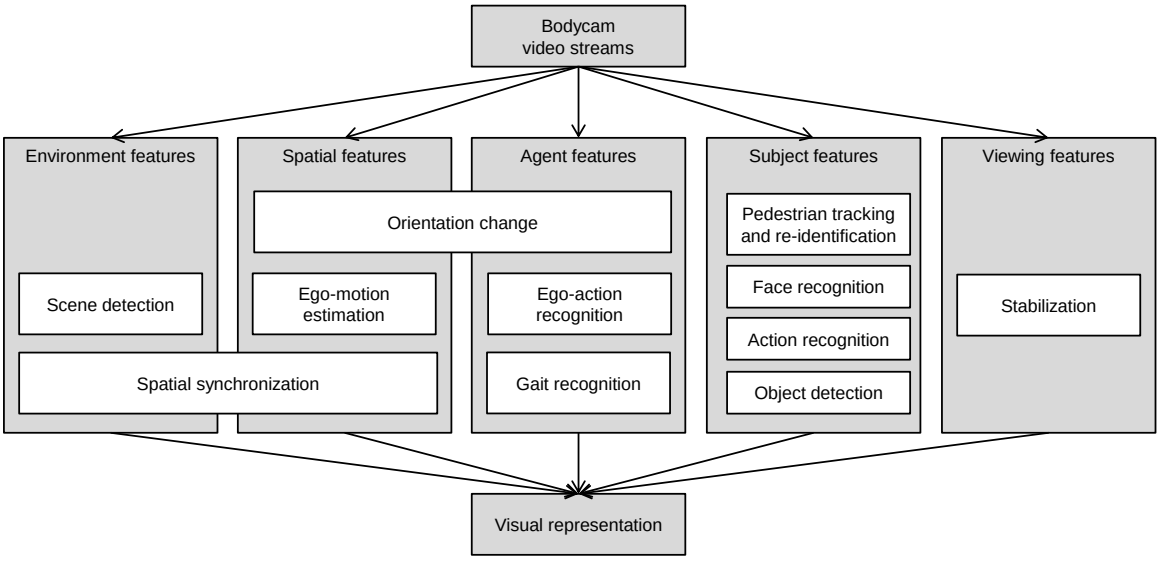

Figure 1: Overview of the framework.

3.2 Stabilization and frame selection

For body-worn cameras, stabilization and frame selection are important during live view or retrospective playback. The bodycams shake and move due to the gait of the police officer wearing it. Stabilization is required, especially for an efficient fast-forward playback. Stabilization is commonly used to compensate for small shaking movements of the camera such as those due to movement caused by wind or by vibrations from an engine. Recently, innovative visualizations were proposed to allow for extremely fast forward, such as the *BriefCam* video synopsis [40] (which is especially suitable for deserted static cameras) and hyper-lapse [27] (which is suitable for bodycams). Hyper-lapse was proposed as a way to compensate for large movements [27]. This approach is similar to the 'simultaneous localization and mapping' (SLAM) and three-dimensional (3D) reconstruction-based methods [53] and it constructs a smooth virtual camera path through the environment. Recently, a more efficient implementation was proposed for fast-forward playback [38]. Instead of uniform frame sampling (as a common baseline), they use a frame selection based on forward looking frames. However, a side effect of camera movement is the introduction of motion blur in the recorded images, especially in forward looking frames. Therefore, in this paper, we propose a novel frame-selection method, which we will call 'LuckyLapse', similar to lucky imaging [13][15]. The proposed method does not select the central forward looking frames – since the walking pattern causes much motion blur when the camera is in the center – but it selects the extrema points left and right, which contain hardly motion blur. Therefore, these points are used to obtain a sharper image. These images are valuable for the retrospective reporting and content analysis. Furthermore, it may be useful to create a virtual stereo camera and perform depth analysis. The relation between horizontal motion (measured with optic flow) and the image sharpness (measured with a local blur estimator [8]) are shown in Figure 2.

3.3 Scene selection and summarization

Scene selection is important to reduce the complete video to only the unique environments, which simplifies retrieval and may stimulate the memory of the professional. Selecting distinct scenes in a single continuous video (e.g., life-loggings) has some similarities to selecting different shots in a video composed of many fragments (e.g., movies, broadcast TV programs). In shot (or logical story unit (LSU) [47]) detection, the aim is to find transitions from one video fragment to the next. Various shot-detection methods have been proposed that search for a visual dissimilarity caused by the transition [46]. Yet, the difference for bodycams is that there are no discrete fragments. To solve this, an advanced method was presented to provide a video summary with only the most distinct scenes [54]. Sparse coding was used to select only the scenes that are hard to reconstruct given an online learned dictionary. The true novelty is that scenes are selected online, i.e., while the video is being generated. Another difference for bodycams is that a large portion of visual dissimilarities is caused by camera ego-motion. Recently, a new method coined *SenseCam* [11] was proposed to select the dissimilarities that relate to a change of scenes and to ignore other dissimilarities. Visual dissimilarity was expressed in terms of semantic characteristics.

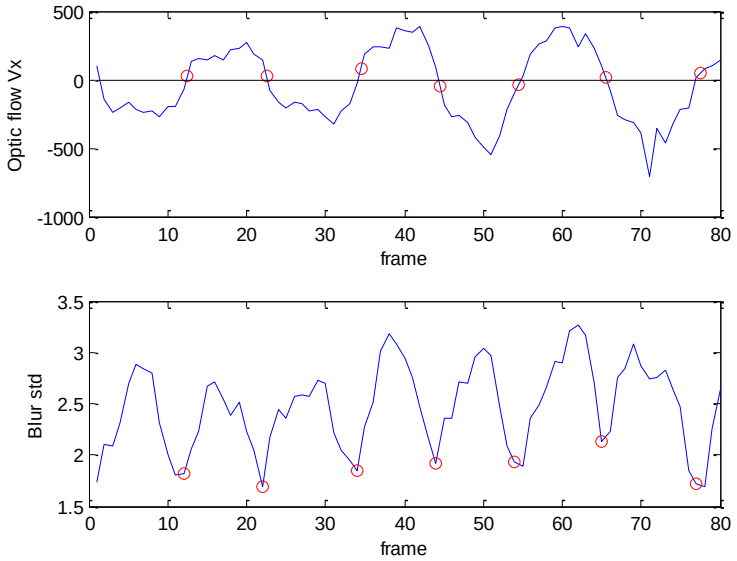

Figure 2: Horizontal optic flow (top) and the average local blur estimate (bottom) during forward motion (walk or run). The sharpest images (low blur, red circles) can be selected when the horizontal motion is minimal.

The advantage is that these suffer less from motion artefacts and that a user can relate to the data more easily. The semantic characteristics relate to the *who*, *what* and *where* of the visual content. Computer vision enables the search by determining – for instance – the number of people in the scene (who), whether the scene is bright or dark, or contains specific semantic concepts (what), and characteristics about the location such as buildings (where). An advanced fusion algorithm combines these characteristics into the scene selection.

3.4 Ego-motion estimation and localization

Localization of police officers may be valuable to assess the situation and to guide personnel in the field. GPS information provides absolute location information in outdoor settings without any drift, but the GPS signal reception is hampered by buildings and indoor situations. With 3D reconstruction techniques it is possible to compute simultaneously the ego-motion of the camera and the location of points on the observed surface in 3D space [53]. The video-based ego-motion estimation provides location in outdoor and indoor settings, but it is hampered by drift. This can result in a location estimate that is complementary to GPS information. GPS and ego-motion estimation can therefore best be combined, to compensate for the weakness of the other.

An alternative localization approach is the recognition of specific scene elements. In particular, the images in the bodycam can be matched to images in the environment, e.g. CCTV footage of fixed cameras. This matching can be performed with 'scale invariant feature transform' (SIFT) descriptors [32] (See Figure 3). The CCTV camera for which a close match is obtained reveals location of the bodycam. This localization approach requires the presence of unique and recognizable scene elements, and produces a smaller set of location estimates when compared to ego-motion estimation. However, it is less computationally demanding and does not suffer from drift.

Figure 3: Matching images from a bodycam (left) to images of a CCTV camera (right). This allows localization of the bodycam.

3.5 Estimation of orientation change

Estimates of the change in orientation are useful to analyze the activities of the police officer and to synchronize video data with GPS data. In this section we describe how to estimate the changes of the orientation of the user wearing the bodycams. To estimate the motion of a camera, SIFT keypoints are computed in the camera images. For pairs of consecutive frames the keypoints are matched [32] and further refined by calculating the fundamental matrix using a 'random sample consensus' (RANSAC) [14] approach. Under the assumption that the movement of the user is small between two frames, we can extract the rotation between the camera orientations between consecutive frames from the fundamental matrix [20]. Even though the motion estimation between two frames is ambiguous (we cannot determine the scale of the movement), the rotation is well defined. Since the bodycam is worn in upright position, we can extract the rotation of the user from the rotation of the camera. Although for each pair of frames the rotation is only a rough and noisy estimate, the integrated values over a sequence of frames are rather robust. In particular the roll and yaw components of the rotation show good correlations with activity of the wearer. The extracted changes in orientation can be used to derive information about the recording, e.g., whether the person moved, was walking or turning. In Figure 4, the roll between frames (with an interval of 0.2 seconds apart) is shown, and a clear difference is visible between the periodic movement while walking and the absence of movement when standing still. In Figure 5, the sideways turn (yaw), integrated over 5 seconds, is shown. The larger peaks correspond for example to a 180 degree clockwise turn (on the left, see also Figure 6) and an approximately 360 degree anti-clockwise turn (on the right).

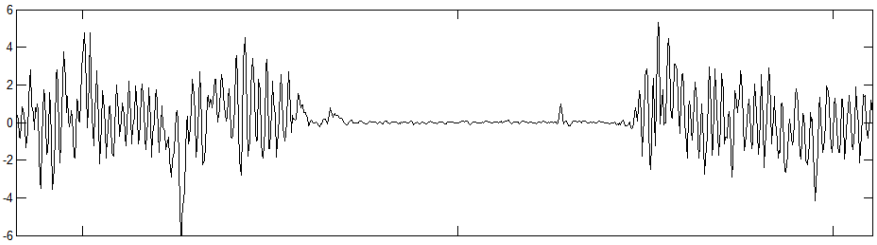

Figure 4: Roll changes in degrees, between images 0.2 seconds apart, during 2.3 minutes. The middle part is while standing still, the rest while walking.

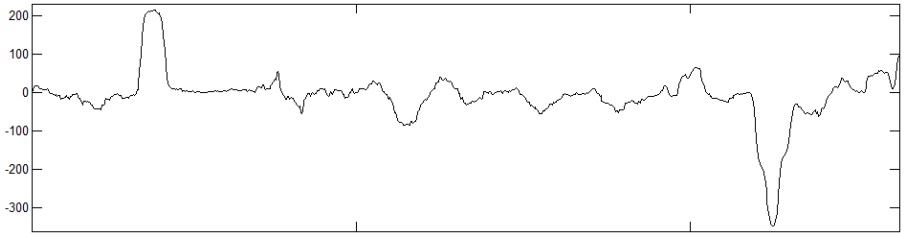

Figure 5: Sideways turn (yaw) in degrees, integrated over 5 second intervals, during a 2.5 minute period.

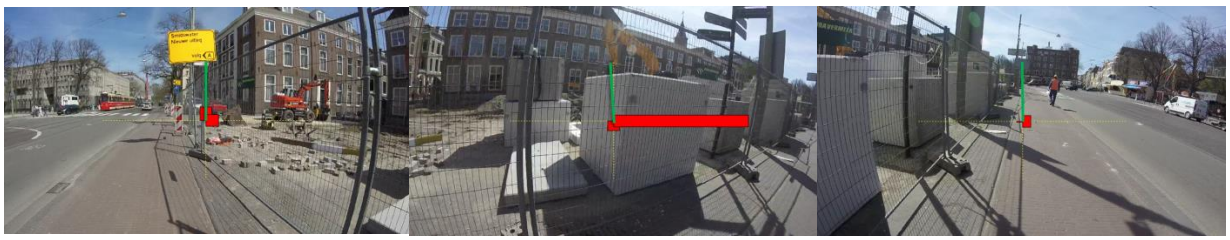

Figure 6: View during the 200-degree positive peak on the left in Figure 5. The thick red bars indicate horizontal and vertical orientation change rate, with the then (almost vertical) green line indicates roll.

3.6 Spatial synchronization

When analyzing a certain event using recorded video it can occur that the same location has been recorded multiple times. This can occur in recordings from different bodycams or in recordings from the same bodycam over time. In order to compare these recordings, it is useful to place them in the same frame of reference (e.g. spatial and temporal). In this section we describe how to synchronize two videos such that playing them simultaneously allows an operator to spot

differences easily. Suppose an officer wishes to focus on all video material at a certain location. It is easy to select all (segments of) videos that are relevant to this location, by using the GPS data associated with the cameras. If multiple videos are relevant, the streams can be synchronized in order to structure the playback of these videos.

The starting point of our method is a pair of videos (possibly from different times) that are recorded at the same location. Since the videos are recorded at the same location, it is likely that the individual frames in the videos have a large overlap and can be matched to each other. We do this by sampling the video streams at a low framerate (1-2 Hz) and calculating SIFT keypoints [32] for all sampled frames. For each pair of frames from video 1 and video 2 the keypoints are matched using the corresponding SIFT descriptors. The matching is then refined by estimation of the fundamental matrix [20]. The number of matches between the sampled frames is shown as a matrix in Figure 7. It is clear that there is a section in the middle of the videos with a good overlap (the number of matches is over 300). After calculation of the number of matches, the best order is determined for simultaneous playback of the two videos. We do this by creating a graph from the pairs of frames and determine a path with minimal costs. Each pair of frames corresponds to a node of the graph and nodes corresponding to pairs of frames that are close are connected with an edge. With each node a weight is associated that depends on the number of matches (more matches results in a lower weight or cost). We then search for the cheapest path in the graph that connects the first pair of frames to the last pair of frames. The path found by our method is shown in Figure 8 with a green line. Using the constructed path we can create a simultaneous playback of the videos. Each point of the path corresponds to a pair of frames that is used for playback. In order to playback the videos at normal speed (recall that we sampled frames for analysis at a low framerate) we smooth the path and use linear interpolation between the nodes of the path. A further improvement in the simultaneous playback would be to, not only align the frames, but also to align the pixels in the individual frames using the keypoints matches. Initial experiments indicate that this has to be done at a higher framerate in order to arrive at a good registration. Also parallax effects are strong, so it could be that alignment of all pixels using a simple homography (projective transformation) would not improve the experience of the operator. This is something to investigate in further research.

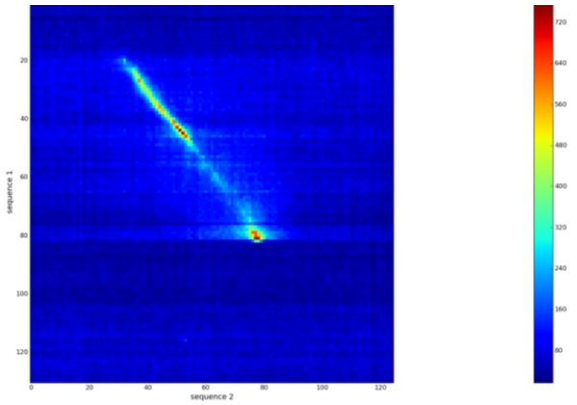

Figure 7: Number of matches between sequences (range 0 to 750 matches).

3.7 Ego-action recognition

There are two types of actions: the actions of pedestrians that are visible in the video and the actions of the person that is wearing the bodycam [5][37][39], which we will call ‘ego-actions’. Ego-action recognition is useful to analyze the actions of a police officer or another person wearing the bodycam. For example, when a police officer is running or falling, it may be relevant to send an alert to colleagues. From the orientation angles as signal in time, as described in Sec. 3.5, actions of the person wearing the camera can be recognized. As a simple example, from the roll and the yaw of the image, it was derived when the person was walking or not and when a turn was made. For walking, the amplitude of the roll signal (measured as the root-mean-square over a few seconds) is a good indication, but it would result in many false alarms. Determining whether the signal is periodic, turned out to be a robust indicator. The periodicity was checked by taking the autocorrelation of a 3 second interval, which should be periodic as well, showing clear peaks at constant intervals. An example is shown in Figure 9, which shows the autocorrelation of a 6 second interval of the signal in Figure 4. Figure 10 shows the resulting action recognition. Walking is defined as a periodic signal combined with high enough root-mean-square of roll. Turning indicates turns of over 45 degrees. The first two turns correspond to the 180 degree and 360 turns in Figure 5.

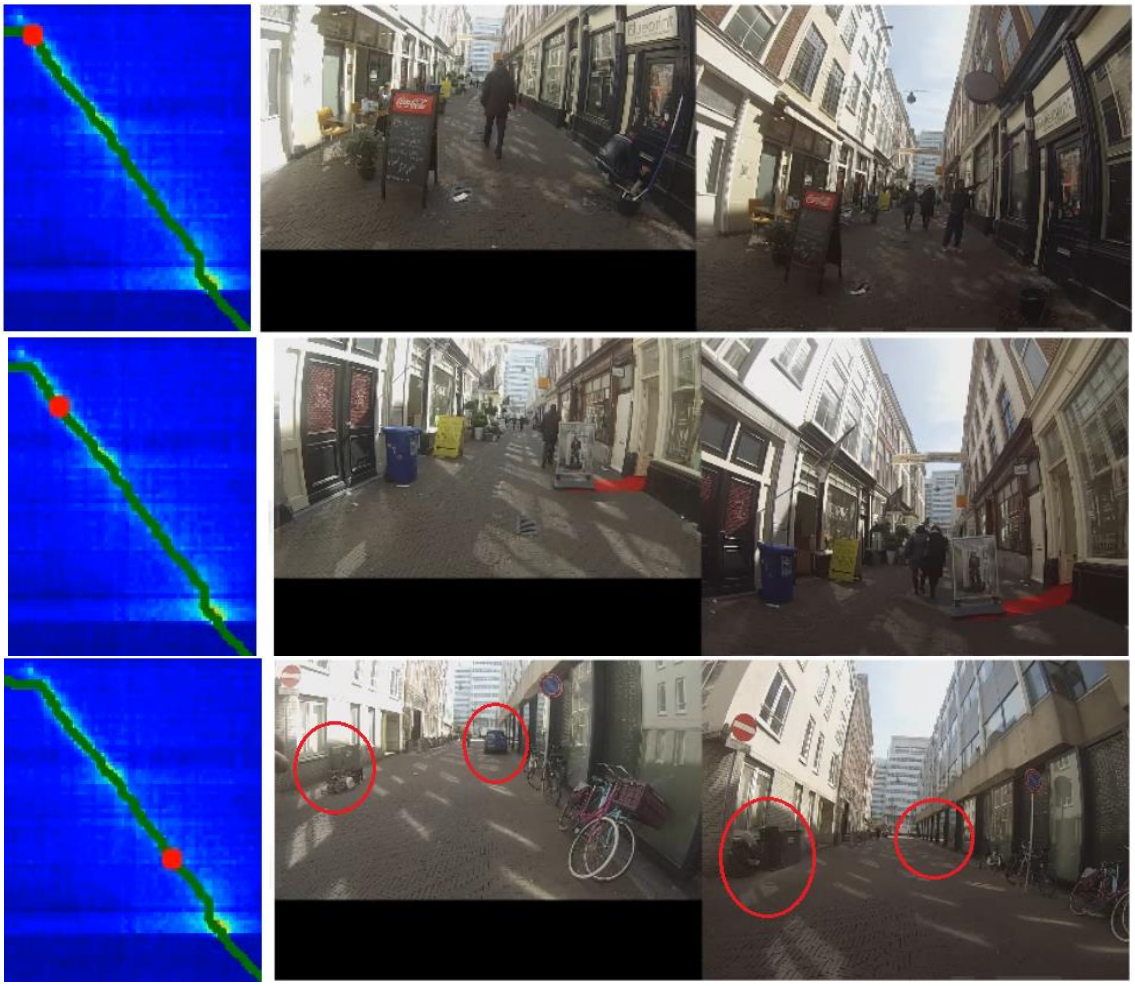

Figure 8: Three examples of synchronous playback of matching frames. This can help to find changes in the scene, such as the bicycle and the car.

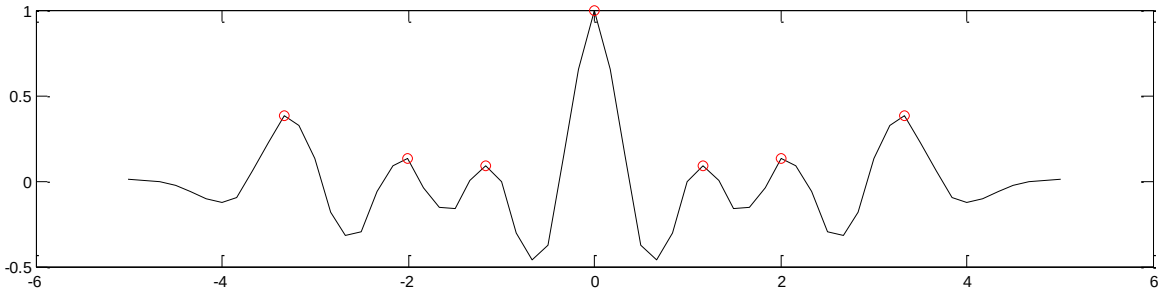

Figure 9: Autocorrelation of the roll angle in a 6 second interval during walking.

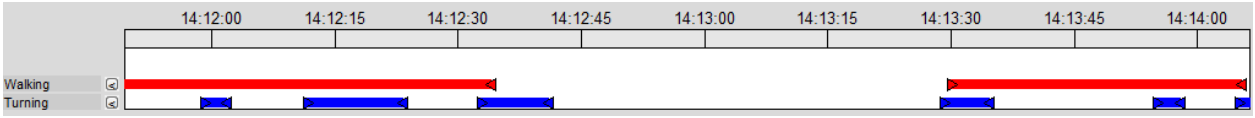

Figure 10: Ego-action recognition, showing intervals of walking and occasions of larger turns.

3.8 Gait recognition

Soft biometrics such as gait recognition, could be used to recognize the officer based on other bodycam footage which can be traced back to him. This could be useful if the information is lost that describes which bodycam was used by which officer. People have a periodic pattern in their walking pattern in the order of a second, but there are variations between persons. This periodicity is visible in the estimated orientation changes. As a simple test, 12 recordings of six persons with four cameras were used. Recordings were cut in parts of about 5 minutes, for which the orientation changes were estimated. The period of the signals while walking (as determined by the Ego-action recognition of Sec. 3.7), and the root-mean-square of the signals themselves were determined over a single interval as training. Of all other intervals, these values were compared to the training values, using the standard deviation as a weight, and choosing the closest match to label the recording. This results in the confusion matrix in Table 2. The average correct recognition is 70%.

Table 2: Confusion matrix for gait recognition.

True labels (P1 – P6)	**Class label estimates (P1 – P6)**					
	P1	**P2**	**P3**	**P4**	**P5**	**P6**
Person P1	**64%**	0	27%	9%	0	0
Person P2	0	**93%**	7%	0	0	0
Person P3	16%	21%	**47%**	16	0	0
Person P4	0	0	0	**100%**	0	0
Person P5	0	0	13%	7%	**40%**	40%
Person P6	0	0	16%	0	11%	**74%**

3.9 Face recognition

The recognition of faces is useful to detect missing people, people with restraining orders or wanted criminals. A similar usefulness holds for number-plate recognition. Initial experiments with face recognition on bodycam footage have been performed with a commercial off-the-shelf face-recognition application and a low quality observation chain [35].

3.10 Pedestrian tracking and re-identification

Pedestrian tracking and re-identification (i.e. ‘forensic search’ or ‘people recognition’) is relevant to retrieve suspects or witnesses quickly (Figure 11). A system for person re-identification typically consists of pedestrian detection [12], tracking [21] and matching [1][4][7][33][34][48][51][55]. In order to handle the severe motion of the bodycam, we used a tracking algorithm that uses appearance (template matching) and motion estimates (of camera and pedestrian). Re-identification is important as social cue to stimulate memory, but also to answer the typical surveillance questions, such as: “Where did a suspect go to?” or “Where did he come from?”. The graphical user-interface with a time axis, a camera axis and many thumbnails of tracks [7] can be helpful to present the social cue and find similar people (or faces) efficiently.

3.11 Object detection

Object recognition and localization is important to understand the content of video and allow flexible querying in a large number of cameras. Typically, the object detectors that perform well on public benchmarks are trained on large collections. The deep convolutional neural network (CNN) has been demonstrated to be an effective approach of which several implementations have been proposed, such as Caffe [23], Overfeat [45] and R-CNN [17]. Recently, a real-time general-purpose search engine was developed that allows users to pose natural language queries to retrieve corresponding images [43]. Top-down, this demonstrator interprets queries, which are presented as an intuitive graph to collect user feedback. Bottom-up, the system automatically recognizes and localizes concepts in images and it can incrementally learn novel concepts. A smart ranking combines both and allows effective retrieval of relevant images. Alternative methods can also be used to find specific items, e.g., based on change detection (difference with background), instance search with SIFT descriptors, or with color information. For example, the blue bag below can be retrieved efficiently without learning a convolutional neural network (Figure 12).

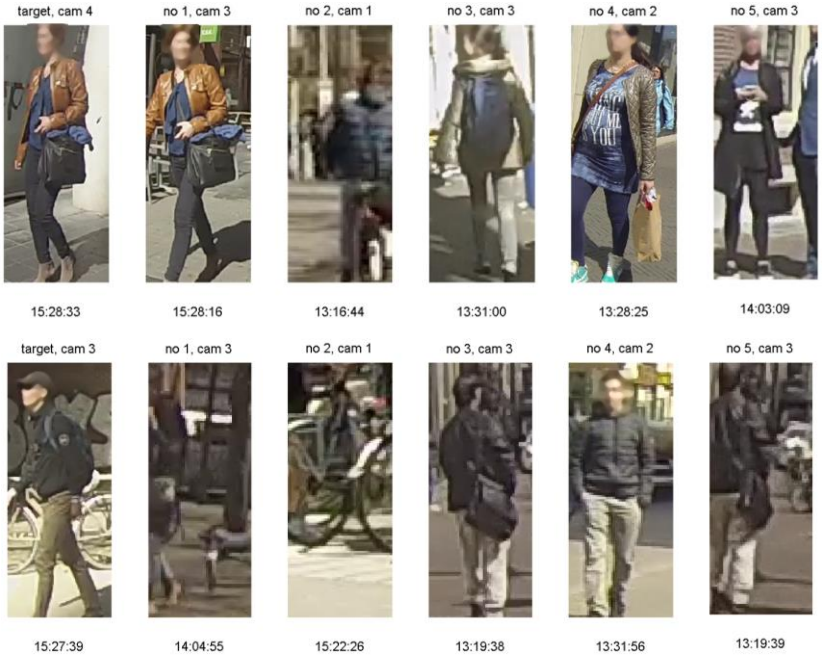

Figure 11: The user can select a query image of a person in one camera ('target') and find similar people in other cameras (candidate number 1 to 5). Faces have been anonymized manually for this publication.

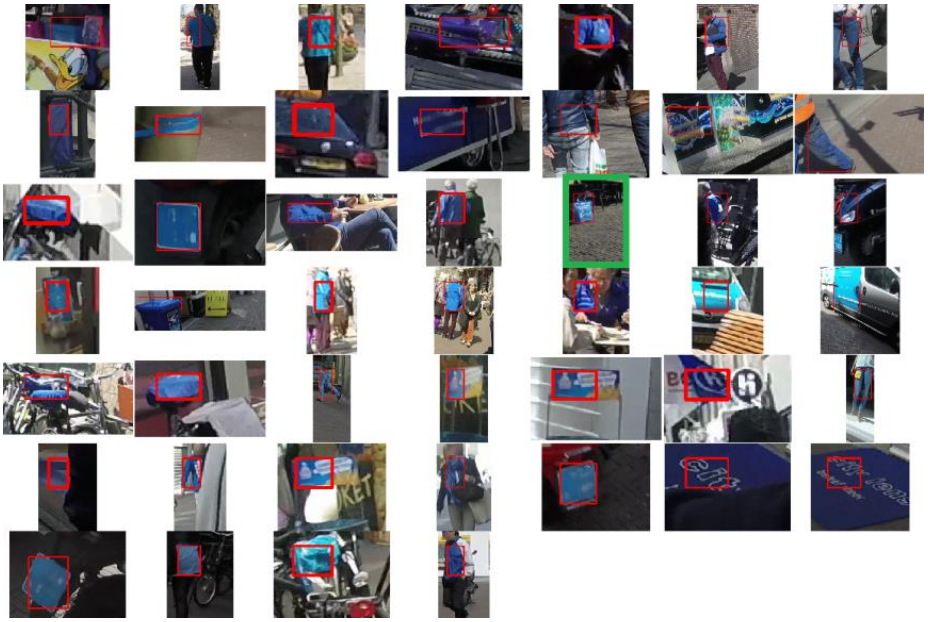

Figure 12: The retrieval of a blue bag in 30 minutes of video can be assisted by presenting objects with a similar color of blue to the user. Faces and number plates have been anonymized manually for this publication.

3.12 Action recognition

The purpose of action recognition is to describe what people – who are visible in the video – are doing. One of the existing pipelines consists of the computation of spatio-temporal interest points (STIP), k-means clustering and bag-of-words (BoW), and an support vector machine (SVM) classifier [Burghouts, AVSS]. Recently, we observe improved actions recognition with improved dense trajectories (IDTs), Fisher Vectors (FV) [50], interactions [42] and convolutional neural networks [22][24][44]. The IDTs contain the following features: histograms of oriented gradients (HOG), histograms of oriented flow (HOF), and motion boundary histograms (MBH). For each of these three features,

we compute principal components (PCA), cluster centers (k-means), Gaussian mixture models (GMM) and FV [6]. The resulting three vectors are concatenated and normalized, and subsequently classification is performed with a linear SVM.

For moving cameras, the IDTs have shown to give better performance than STIP features [50]. However, the IDT has two disadvantages. First, it provides only a limited description of the motion in a camera. E.g., forward motion of a bodycam results in a left-ward motion on the left-side of the image and a right-ward motion on the right side, which cannot be modelled very well. Second, the basic implementation, as commonly used, only provides a classification per frame, which is not suitable when multiple persons are present in the scene. To improve the performance of action recognition, we cued a 'virtual PTZ' based on the track of the person. The cueing resulted in footage that was stabilized for the motion of the camera and focuses on one person, which improves the action recognition.

The cued footage must be well tracked and stabilized. Motion is caused by three factors: motion of the camera, motion of the person and noise in the localization of the detector. Connecting detections of a pedestrian detector (e.g., [12]) has a low systematic motion error over a long time, but it introduces too much jitter. Template matching with a fast update which is based on the detection in the previous frame has a low stochastic error, but it will result in a drift over a long time. Therefore, our solution contains the following steps: tracking through pedestrian detections for low systematic error (but which suffers from jitter), frame-to-frame template matching for a low stochastic error which can compensate for rapid motions (but which suffers from drift), and a combination function that lets the systematic error slowly decay. The training material was obtained from HMDB [28], where we used clean examples of the classes *walk*, *run*, and *bike*. The tracking and cueing was also performed on the training data, which is independent from our bodycam data. Results on the bodycam data are shown in Figure 13.

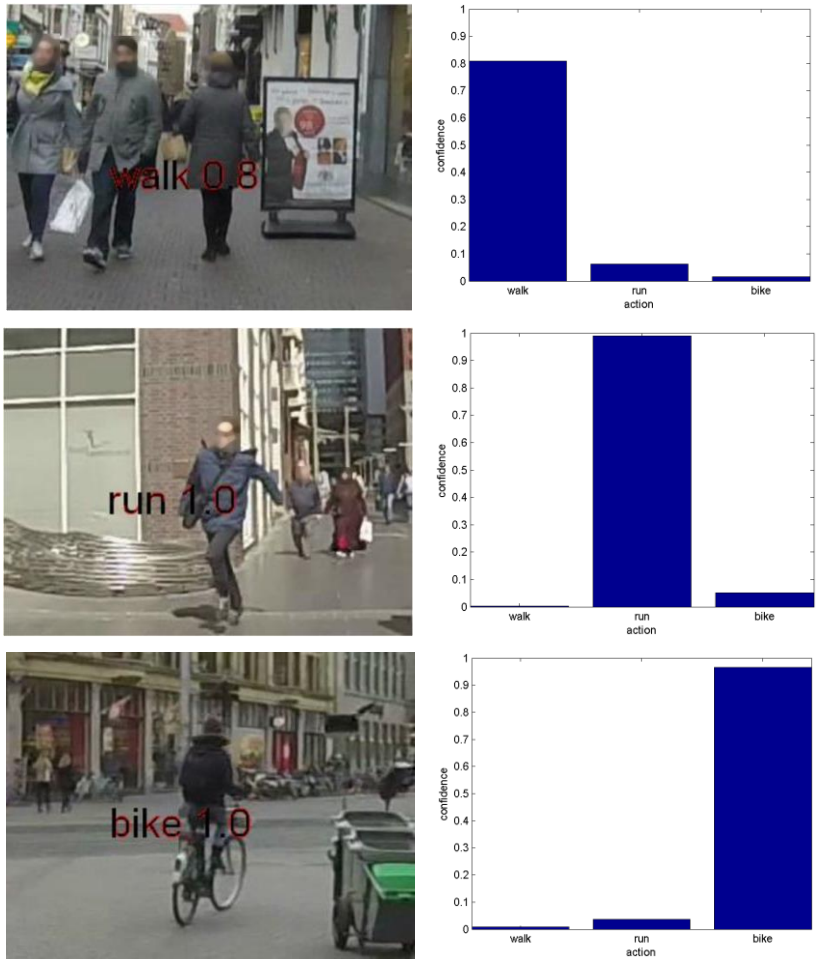

Figure 13: Example results of the actions classifiers on bodycam data. Based on the tracking and cueing results, actions such as *walk*, *run*, and *bike* can be recognized in short cued movie clips.

4. VISUAL REPRESENTATION DEMONSTRATOR

We implemented a graphical user interface (GUI) for the interaction with the VCA capabilities (see the demonstrator mockup in Figure 14). This GUI contains several panes. One of them is used to present spatial (GPS) information on a map, which is suitable to answer 'where' questions or to inspect footage at a certain location (e.g., the location of an incident). Another pane is used to show time segments, for the 'when' questions (e.g., when is the police officer walking) or to inspect events at a certain moment (e.g. the time of the incident). Another pane is used to present social information – such as snippets of persons – which is suitable to answer 'who' questions (e.g., to find a suspect).

In some cases, the interpretation of long videos can be enhanced by the efficient presentation of image information. In addition to the methods for faster playback (Sec. 3.2), a montage of thumbnails can help to represent the video without automatic content analysis. An example is shown in Figure 15. The left part of the figure shows an overview of 30 minutes of video. Within a glance, a user can see that the wearer of the bodycam was biking, visiting the Dutch house of parliament (*Binnenhof*), biking again, walking in a shopping area, waiting at the '*Xenos'*-shop, biking again, visiting the *Binnenhof* again, and finally biking. With a few clicks, the user can 'zoom in' on the time interval that represents the moment after leaving the *Xenos* shop. A view on the scene can help to find relevant moments and stimulate the memory. Table 3 shows a list of visual representations. For each of these representation, we show in which security related cases they are useful, and how they help to stimulate the memory and support an efficient search strategy.

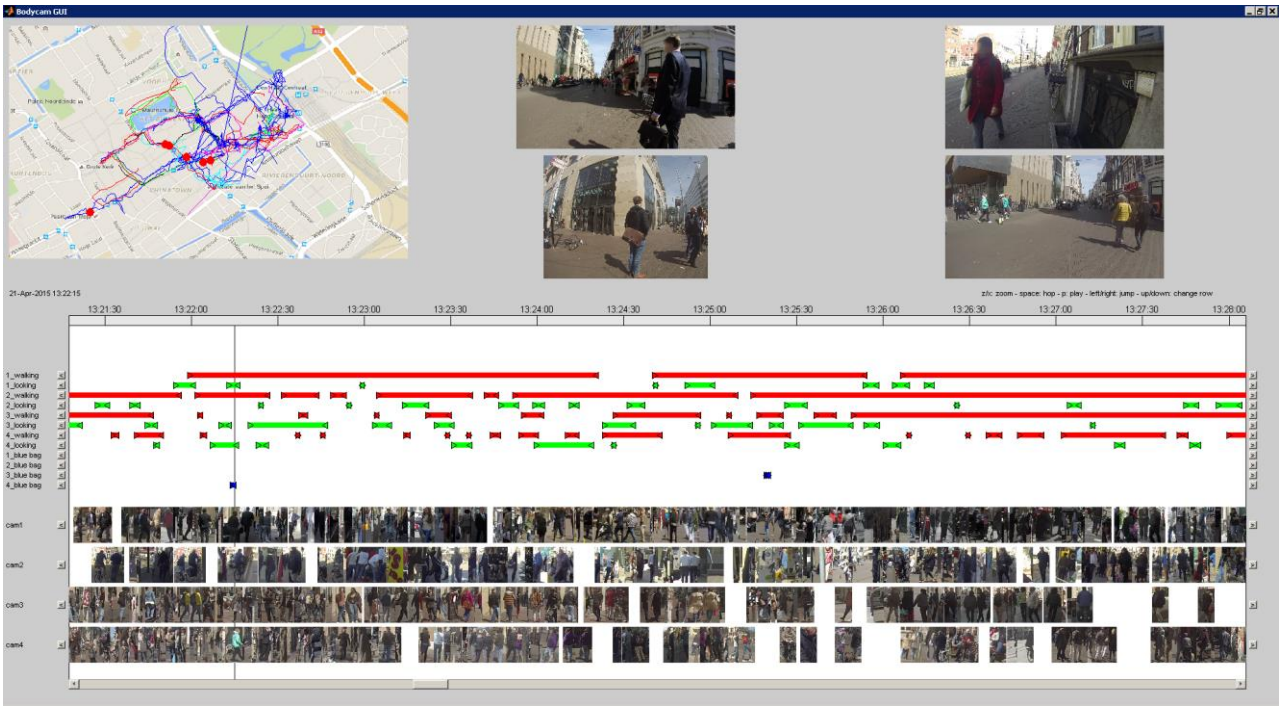

Figure 14: Graphical user interface with a map to present GPS position information ('where'), time segments to present temporal information ('when') of orientation and objects and person snippets to present social information ('who').

Figure 15: A montage of thumbnails of video with a small horizontal time-bar at the bottom. The red box at the time bar indicates the time interval that was selected. (Left) The thumbnails show an overview of 30 minutes of video. (Right) The thumbnails are retrieved at the small time interval of only a few minutes after leaving the yellow 'Xenos'-shop.

Table 3: Visual representation and the relation with relevant use-cases and cues that support the professional.

VCA capability / Visual representation	**Relevant use case / example**
Visual representation with textual summary	A textual summary can easily and flexibly be searched for elements that are of interest.
Visual representation with time segments	Finding moments where a police officer is running. 'When'-cue can stimulate the memory.
Visual representation with snippets	Finding suspects, victims or witnesses. The 'who'-cue can stimulate the memory.
Visual representation with GPS and map	To focus on video material that was recorded at the location of an incident. Location ('where') cues are important to select a region of interest.
Visual representation with montage of thumbnails	To present different scenes in the video efficiently. A view on the scene is a cue that can stimulate the memory.

5. CONCLUSIONS

The main contribution of this paper is that it gives an overview of state-of-the-art VCA capabilities related to the use cases for body-worn cameras in the security domain. VCA capabilities on bodycam footage can help to realize and improve security functions. For some of these functions VCA can improve in a relatively small number of ways (e.g. Amber Alert), for others there are many improvements possible (e.g. collect proof, improve testimony and efficient reporting). A complete summary of the video can be made automatically. A visual representation allows quick retrieval of relevant fragments. The content analysis for summarization consists of several components, such as stabilization, scene selection, motion estimation, localization, pedestrian tracking and action recognition in the video from a bodycam. Several VCA capabilities had to be modified to allow usage on these moving cameras. Examples of those novel modifications are the selection of high-quality frames for stabilization, gait recognition by periodic analysis, and the cueing virtual PTZ for action recognition.

ACKNOWLEDGEMENT

The work for this paper was conducted in the project 'Passive sensors' in the Netherlands top sector 'High Tech Systems and Materials' and the EU FP7 project TACTICS. We thank the Dutch national police for discussions and inspiration.

REFERENCES

[1] An, L., Kafai, M., Yang, S., Bhanu, B., "Reference-based person re-identification," IEEE AVSS, (2013).
[2] Arev, I., Park, H., Sheikh, Y., et al., "Automatic editing of footage from multiple social cameras," ACM Trans. Graphics 33(4), (2014).
[3] Ariel, B., Farrar, W., Sutherland, A., "The effect of police body-worn cameras on use of force and citizens' complaints against the police: a randomized controlled trial," J. Quant. Criminol., (2014).
[4] Bedagkar-Gala, A., Shah, S., "A survey of approaches and trends in person re-identification," Image and Vision Computing 32, 270-286 (2014).
[5] Behera, A., Chapman, M., Cohn, A., Hogg, D., "Egocentric activity recognition using histogram of oriented pairwise relations," Int. Conf. Comp. Vision Theory and Appl., (2014).
[6] Broek, S., Bouma, H., Hollander, R., et al., "Ship recognition for improved persistent tracking with descriptor localization and compact representations," Proc. SPIE 9249, (2014).
[7] Bouma, H., Baan, J., Landsmeer, S., Kruszynski, C., Antwerpen, G., Dijk, J., "Real-time tracking and fast retrieval of persons in multiple surveillance cameras of a shopping mall," Proc. SPIE 8756, (2013).
[8] Bouma, H., Dijk, J., Eekeren, A. van, "Precise local blur estimation based on the first-order derivative," Proc. SPIE 8399, (2012).
[9] Burghouts, G., Schutte, K., Hove, R. ten, et al., "Instantaneous threat detection based on a semantic representation of activities, zones and trajectories," Signal Image and Video Processing SIVP, (2014).
[10] Burghouts, G., Eendebak, P., Bouma, H., Hove, J. ten, "Improved action recognition by combining multiple 2D views in the bag-of-words model," IEEE AVSS, 250-255 (2013).
[11] Doherty, A., Pauly, K., Caprani, N., et al., "Experience of aiding autobiographical memory using the SenseCam," Human-Computer Interaction 27, 151-174 (2012).
[12] Dollar, P., Appel, R., Belongie, S., Perona, P., "Fast feature pyramids for object detection," IEEE Trans. PAMI 36(8), 1532-1545 (2014).
[13] Eekeren, A., Schutte, K., et al., "Turbulence compensation: an overview," Proc. SPIE 8355, (2012).
[14] Fischler, M., Bolles, R., "Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography," Comm. of ACMM 24(6), 381-395 (1981).
[15] Fried, D., "Probability of getting a lucky short-exposure image through turbulence," JOSA 68(12), (1978).
[16] Gee, A., Escamilla-Ambrosio, P., Webb, M., et al., "Augmented crime scenes: virtual annotation of physical environments for forensic investigation," Proc. ACM Multimedia in Forensics Sec. Intell., 105-110 (2010).
[17] Girshick, R., Donahue, J., Darrell, T., Malik, J., "Rich feature hierarchies for accurate object detection and semantic segmentation," IEEE CVPR, 580-587 (2014).
[18] Gouveia, R., Karapanos, E., "Footprint tracker: supporting diary studies with lifelogging," Proc. SigCHI Conf. Human Factors in Computing Systems, 2921-2930 (2013).
[19] Ham, T., Kuppens, J., Ferwerda, H., "Mobiel cameratoezicht op scherp; Effecten op geweld tegen de politie en het politieproces in beeld," Report Bureau Beke, (2011).
[20] Hartley, R., Zisserman, A., "Multiple view geometry in computer vision," Cambridge University Press, (2004).
[21] Hu, N., Bouma, H., Worring, M., "Tracking individuals in surveillance video of a high-density crowd," Proc. SPIE 8399, (2012).
[22] Ji, S., Xu, W., Yang, M., Yu, K., "3D convolutional neural networks for human action recognition," IEEE PAMI 35(1), 221-231 (2013).
[23] Jia, Y., Shelhamer, E., Donahue J., et al., "Caffe: convolutional architecture for fast feature embedding," ACM Multimedia, 675-678 (2014).
[24] Karpathy, A., Toderici, G., Shetty, S., et al. "Large-scale video classification with convolutional neural networks," IEEE CVPR, 1725-1732 (2014).
[25] Kalnikaite, V., Whittaker, S., "Beyond being there? Evaluating augmented digital records," Int. J. Human-Computer Studies 68, 627-640 (2010).
[26] Kalnikaite, V., Sellen, A., Whittaker, S., Kirk, D., "Now let me see where I was: Understanding how lifelogs mediate memory," Proc. SigCHI Conf. Human Factors Computing Systems, 2045-2054 (2010).
[27] Kopf, J., Cohen, M., Szeliski, R., "First-person hyper-lapse videos," ACM Trans. Graphics 33(4), (2014).
[28] Kuehne, H., Jhuang, H., Garrote, E., Poggio, T., Serre, T., "HMDB: a large video database for human motion recognition," IEEE ICCV, 2556-2563 (2011).
[29] Lu, Z., Grauman, K., "Story-driven summarization for egocentric video," IEEE CVPR, 2714-2721 (2013).

[30] Lee, M., Dey, A., "Providing good memory cues for people with episodic memory impairment," Proc. ACM SigAccess Conf. Computers and Accessibility, 131-138 (2007).
[31] Lee, Y., Ghosh, J., Grauman, K., "Discovering important people and objects for egocentric video summarization," IEEE CVPR, 1346-1353 (2012).
[32] Lowe, D., "Distinctive image features from scale-invariant keypoints," Int. J. Computer Vision 60(2), (2004).
[33] Ma, B., Su, Y., Jurie, F., "Covariance descriptor based on bio-inspired features for person re-identification and face verification," Image and Vision Computing 32, 379-390 (2014).
[34] Marck, J.W., Bouma, H., Baan, J., Oliveira Filho, J. de, Brink, M. van den, "Finding suspects in multiple cameras for improved railway protection," Proc. SPIE, (2014).
[35] Mu, M., Spreeuwers, L., Veldhuis, R., "Face identification in videos from mobile cameras," Proc. NCCV, (2014).
[36] Nixon, M., Tan, T., Chellappa, "Human identification based on gait," Springer Science and Business Media New York USA, (2010).
[37] Pirsiavash, H., Ramanan, D., "Detecting activities of daily living in first-person camera views," IEEE CVPR, 2847-2851 (2012).
[38] Poleg, Y., Halperin, T., Arora, C., Peleg, S., "Ego-sampling: Fast-forward and stereo for egocentric videos," preprint on arXiv, (2014).
[39] Poleg, Y., Arora, C., Peleg, S., "Temporal segmentation of egocentric videos," IEEE CVPR, 2537-2544 (2014).
[40] Pritch, Y., Rav-Acha, A., Peleg, S., "Nonchronically video synopsis and indexing," IEEE TPAMI 30(11), 1971-1984 (2008).
[41] Rest, J. van, Grootjen, F., Grootjen, M., et al., "Requirements for multimedia metadata schemes in surveillance applications for security," Multimedia Tools and Applications MTAP 70(1), 573-598 (2014).
[42] Ryoo, M., Matthies, L., "First-person activity recognition: what are they doing to me?," CVPR, (2013).
[43] Schutte, K., Bouma, H., Schavemaker, J., et al., "Interactive detection of incrementally learned concepts in images with ranking and semantic query interpretation," Content-Based Multimedia Indexing CBMI, (2015).
[44] Simonyan, K., Zisserman, "Two-stream convolutional networks for action recognition in videos," Adv. Neural Inf. Proc. Syst., 568-576 (2014).
[45] Sermanet, P., Eigen, D., Zhang X., et al., "Overfeat: integrated recognition, localization and detection using convolutional networks," ICLR, (2014).
[46] Smeaton, A., Over, P., Doherty, A., "Video shot boundary detection: seven years of TRECVid activity," CVIU 114(4), 411-418 (2010).
[47] Snoek, C., Worring, M., "Multi-modal video indexing: a review of state-of-the-art," Multimedia Tools and Applications 25, 5-35 (2005).
[48] Souded, M., "People detection, tracking and re-identification through a video camera network," PhD thesis Univ. Nice Sophia-Antipolis, (2013).
[49] Vigne, N. la, Lowry, S., Markman, J., Dwyer, A., "Evaluating the use of public surveillance cameras for crime control and prevention," Report Urban Institute Washington USA, (2011).
[50] Wang, H., Schmid, C., "LEAR-INRIA submission for the THUMOS workshop," THUMOS Challenge: ICCV Workshop on Action Recognition with Large Number of Classes, (2013).
[51] Wang, T., Gong, S., Zhu, X., Wang, S., "Person re-identification by video ranking," ECCV, 688-703 (2014).
[52] Whittaker, S., Kalnikaite, V., Petrelli, D., et al., "Socio-technical lifelogging: deriving design principles for a future proof digital past," Human-computer interaction 27, 37-62 (2012).
[53] Wieringa, F., Bouma, H., Eendebak, P., et al., "Improved depth perception with three-dimensional auxiliary display and computer generated three-dimensional panoramic overviews," J. Medical Imaging 1(1), (2014).
[54] Zhao, B., Xing, E., "Quasi real-time summarization for consumer videos," IEEE CVPR, 2513-2520 (2014).
[55] Zheng, W., Gong, S., Xiang, T., "Reidentification by relative distance comparison," IEEE Trans. Pattern Analysis and Machine Intelligence 35(2), 653-668 (2013).