



Access to data sets

*Erncip Thematic Group
on Video Surveillance for
Security of Critical
Infrastructure*

Lucio Marcenaro, PhD
University of Genoa, Italy

July 2016

The research leading to these results has received funding from the European Union as part of the European Reference Network for Critical Infrastructure Protection project.

Access to data sets

This publication is a technical report by the Joint Research Centre, the European Commission's in-house science service. It aims to provide evidence-based scientific support to the European policymaking process. The scientific output expressed does not imply a policy position of the European Commission. Neither the European Commission nor any person acting on behalf of the Commission is responsible for the use which might be made of this publication.

JRC Science Hub

<https://ec.europa.eu/jrc>

JRC103341

ISBN 978-92-79-62654-8

doi:10.2788/504779

© European Union, 2016

Reproduction is authorised provided the source is acknowledged.

All images © European Union 2016

Contents

Contents	2
Acknowledgements	4
Abstract	5
1. Introduction	7
1.1. Background	7
1.2. Purpose of the report	8
2. Complexity of video analytics	10
2.1. Sensors	10
2.2. Functionalities	11
2.3. Scenarios	11
2.4. Real-time requirements	12
3. Importance of data sets	13
3.1. Common data	13
3.2. Exhaustive conditions	13
3.3. Algorithms comparisons	14
3.4. Automatic performance evaluation	14
4. Critical issues of data sets	15
4.1. Complexity	15
4.2. Ground truth	15
4.3. Suitability	16
4.3.1. Balance	16
4.4. Video quality	17
4.5. Privacy issues	18
4.6. Open access	19
5. Data set construction checklist	20
6. Main features	21
7. Existing data sets	22
8. Conclusion	25
8.1. Future work	25
References	26
List of abbreviations and definitions	33
List of figures	38
List of tables	39
Appendix A. Data sets description	40
A.1. HDA person data set	40
A.2. WWW crowd data set	42
A.3. MOT benchmark	43

A.4. ChokePoint data set	44
A.5. VIRAT.....	46
A.6. Comprehensive cars (CompCars).....	48
A.7. INRIA person data set	49
A.8. TUGRAZ ICG long-term pedestrian data set	50
A.9. Crowd data set	51
A.10. PEdesTrian attribute (PETA) data set.....	52
A.11. CUHK crowd data set	53
A.12. GRAZ-02	54
A.13. Person Re-ID (PRID) 2011	55
A.14. MuHAVi.....	56
A.15. GRAZ-01	58
A.16. Mall data set	59
A.17. KTH action	60
A.18. Weizmann actions	61
A.19. UT-Interaction	62
A.20. i-LIDS.....	63
A.21. NIST digital video 1	64
A.22. Pedestrian walking path data set (Grand Central data set)	65
A.23. PETS 2007	66
A.24. PETS 2006	67
A.25. PETS 2009	68
A.26. PETS 2015	69
A.27. UCF aerial action data set.....	70
A.28. Mini-drone video data set (DronesProtect data set)	71

Acknowledgements

The author gratefully acknowledges the contributions, suggestions and reviews of the other members of the ERNCIP Thematic Group on Video Surveillance for Security of Critical Infrastructure, the ERNCIP Office and the University of Genova colleagues.

Abstract

The main objective of this report is to analyse existing data sets for video analytics (VA) and to determine how best to enable collection/common access to data sets in the EU for testing/evaluation of video surveillance software.

This report presents a critical analysis of video analytic data sets with specific attention to the protection of critical infrastructures. The introductory part of the report describes the importance of VA and the growth of the related market. In this scenario the importance of the usage of a common data set is highlighted. The main reason of the fundamental importance of data sets in video analysis is the intrinsic complexity of VA-related techniques: a common set of video sequences is seen as a powerful boost in the design, development and testing of VA algorithms.

This report describes different aspects that make VA so complex and demonstrates the importance of having common and widespread data sets. Data sets must also rely on the availability of standards related to several aspects of the VA for the protection of critical infrastructures: refer to (Ferryman, 2016) for an overview of standards in video surveillance including the need for standards, an overview of existing relevant standardisation efforts including gaps, and a roadmap for the development of future standards.

A detailed description and analysis of critical issues of VA data sets are provided, and a simple but effective 'data set construction checklist' are proposed.

In Appendix A, several existing data sets are summarised and commented in relation with the use cases highlighted in the report (van Rest, 2015a). Moreover, the impact of each data set in the scientific community is estimated by considering the total number of referencing papers and the most relevant research using the data set for computing the performances of a proposed technique.

With this report, we follow up on the recommendations regarding test data sets for VA use cases of (van Rest, 2015b) and (van Rest, 2015a). In particular:

- together with (van Rest, 2015b) and the *Video Analytics Adoption - Key considerations for the end user* (Doyle, 2016), this report helps build an argument for why data sets matter in the boardroom of critical infrastructure end users and industry;
- this report gives the requirements for creating high-quality and relevant data sets.

VA modules represent the core components of automatic video surveillance systems: these modules are able to process video sequences acquired from single or multiple video sensors, extract high-level information and automatically identify situations of interest or potentially dangerous for maintaining an appropriate level of safety for the considered environment.

One of the typical key requirements from critical infrastructures' operators for VA modules is that they must guarantee a sufficient level of performance 24 hours per day, 7 days per week. Unfortunately, because of the high variability of the visual information, even in a simple video surveillance installation, this feature is typically extremely tough to achieve.

Moreover, in real video surveillance systems, an extremely wide variety of heterogeneous sensors can be found with a significant number of functionalities in ever-changing scenarios.

A typical approach for solving these problems is to test each video analysis module against a wide variety of video sequences: for this reason, standard data sets play a fundamental role in the design and implementation of real market-ready video surveillance systems.

This report deals with VA data sets, considering main features of these data and highlighting pros and cons of all the considered sets. Features are identified by

considering their importance in solving the 24/7 requisites: from the results of the report it is clear that certain sequences are better suited for specific functionalities and not all the existing data sets can be used in all the real environments. The existence of some kind of ground truth for the considered data set represents a very important feature of it, as it may allow an objective and quantitative evaluation of the VA module.

1. Introduction

1.1. Background

Nowadays, the increased demand of security is a particularly relevant need in our society. Therefore, systems able to automatically interpret interactions, both among people and between people and the environment, represent an actual domain of research which still lack efficient solutions and open problems. Over the last few years, the video surveillance market has experienced an impressive growth thanks to the improved technological solutions (e.g. network high-resolution cameras) and the continuous cutting of hardware prices. The VA market is expected to grow from USD 1 537.9 million in 2015 to USD 3 971.2 million by 2020, at an estimated compound annual growth rate of 20.9 % during the forecast period from 2015 to 2020. Considering this strong projected market growth, VA algorithms only showed moderate growth during the last few years, mainly because of poor performances in real situations. Market analysts agree that VA will dramatically increment its growth rate as soon as a sufficiently robust killer application is proposed and developed by research groups worldwide (Markets and Markets, 2016).

More in detail, visual tracking represents a fundamental processing step for most VA applications, where the aim is to automatically understand the action performed by the objects present in the monitored scene (Dore, 2010). This problem has been widely investigated in the last decades, but a solution that is valid in general situations is still to be found. Typically, video analysis is the core component behind a scene-understanding framework that can be based on multilevel trackers that are able to enrich global trajectories information by including other features such as scale, pose and shape in the object description with the aim of accomplishing advanced scene-interpretation tasks.

In video surveillance projects, automatic and real-time event-detection solutions are required to guarantee an efficient and cost-effective use of the infrastructure. Many solutions to automatically detect a variety of events of interest have been proposed. However, not all solutions and technologies satisfy all the requirements of the surveillance scenario. For this reason, performance evaluation of existing event-detection solutions becomes an important step in the deployment of video surveillance projects. In literature, several practical approaches exist that aim at minimising the problem of the ground truth generation as well as the expertise required to evaluate and compare the results by introducing specific requirements for event-detection scenarios. This approach is believed to be applicable for an initial evaluation of candidate solutions to a specific surveillance scenario before more exhaustive tests in an integrated environment.

Object tracking, pattern recognition and, in general, image analysis and image understanding techniques today offer different approaches to the automatic detection of events of interest in specific surveillance applications. Not all detection strategies behave equally well: a performance evaluation study is necessary to select the most appropriate solution to a specific problem.

The evaluation is often done directly by the users of the surveillance infrastructure. In fact, system integration is needed before the evaluation is possible and users need to monitor the behaviour of the automatic event detector for days. For economic reasons, this complex, long and expensive process cannot be repeated for a large number of candidate solutions.

Although this approach guarantees that all the constraints of the surveillance scenarios are taken into account, it introduces a number of limitations:

- it is not suitable for comparing several event-detection solutions on the same detection task and video sources;
- it requires time to guarantee the statistical completeness of the data: some events may be very rare;

- it requires active feedback from the users who may not have the time to monitor performance continuously;
- it makes it difficult to learn from the observed limitations and to update the solution accordingly;
- it hinders external auditing based on independent quality standards (e.g. the role that I-LIDS had in the United Kingdom).

An alternative and more systematic approach to the above strategy is to focus the performance evaluation effort on the modules constituting the event-detection solution (e.g. the object-tracking performances, the background update strategy, etc.). These approaches are successfully used by the academic community and enable the researchers to test, compare and improve specific image analysis modules on reference data sets.

Such data set-based evaluation strategies should be considered of primary importance, not only in the academic world, but also in the actual set up of video surveillance solutions on the field. Law enforcement agencies and critical infrastructures' operators may be facing a waste of time, resources and ultimately money in deploying physical surveillance networks and VA solutions without a preliminary assessment of its modules and components. In fact, data sets can provide an offline verification phase of the performances of specific VA modules and functionalities, allowing not only to actually test them in a structured and exhaustive manner, but also to systematically evaluate and compare the level of accuracy they can offer. This can thus make end users aware of the potentialities and limitations of the modules deployed and enable them to opt for different solutions.

Particularly popular are the initiatives that provide common data sets for evaluating object-tracking or object-segmentation techniques (e.g. PETS). Since object tracking and segmentation provide the information that can be used to detect a broad class of events, this approach can be a priori extended to several surveillance scenarios with different detection requirements. However, these strategies are driven by academic research and are often too generic to be successfully applied to real surveillance contexts.

They require a manual ground truth data generation, which is an extremely long and subjective process. Moreover, the evaluation results require a high level of expertise to be analysed correctly. Finally, these approaches fail to provide a metric that enables an easy comparison of results, either against each other or against the ground truth.

In practice, these modular approaches do not fully answer the industrial needs for performance evaluation. The main limitation is that the detection of events of interest is not necessarily correlated to the performances of the constituent modules used to achieve the detection. In some cases, for instance, it is not necessary to have an extremely high-performing object tracking to count people or vehicles in a scene.

If we select the event detection only based on the quality of the tracking module, we may choose the wrong solution. Moreover, these methods do not take into account the impact of the implementation or the way the modules are combined: all of these choices are difficult to document and may make the difference between a good and a bad solution. Finally, these approaches cannot be used to compare methods based on completely different modules.

1.2. Purpose of the report

The main objective of this report is to analyse existing data sets for VA and to determine how best to enable collection/common access to data sets in the European Union for testing/evaluation of video surveillance software. This report deals with VA data sets, considering main features of these data and highlighting pros and cons of all the considered sets. Features are identified by considering their importance in solving the 24/7 requisites: from the results of the report it is clear that certain sequences are

better suited for specific functionalities and not all the existing data sets can be used in all the real environments.

The target audience for this report is primarily the security managers of critical infrastructure operators.

2. Complexity of video analytics

VA modules must be robust and able to cope with the high variability of real environments. The main challenge when producing VA modules is represented by heterogeneity (Narayanan, 2014).

Heterogeneity can either involve:

- employed sensors;
- functionalities required, i.e. of the tasks to be addressed by the deployed VA module;
- scenarios which can exhibit dramatic changes both in different locations and in time.

Due to this high variability of conditions, no VA module usually works as it is; in fact, accurate parameter tuning is often required in order to have it functioning properly in different situations (Greiffenhagen, 2000), as is sensors' calibration (Ramesh, 2005).

Application fields of VA are numerous; among others:

- indoor/outdoor surveillance;
- crowd analysis;
- traffic management;
- automotive safety;
- robotics;
- human-machine interfaces;
- video indexing;
- video content retrieval;
- health care;
- entertainment;
- domotics (home automation).

For a parallel analysis of the heterogeneous factors introducing complexity in VA as well as of their relationships and dependencies, the reader can refer to the ERNCIP document (van Rest, 2015b) 'Surveillance and video analytics — Factors influencing the performance' by Jeroen van Rest, MSc., TNO. The report proposes a morphological analysis of the surveillance domain in order to describe a surveillance system in its context. However, being too abstract to highlight differences between subcomponents, the morphological analysis is extended to cover the subdomain of VA (MAVA). In particular, Appendix C of (van Rest, 2015b) proposes a categorisation of the relevant factors for the MAVA. Analogies are highlighted when relevant.

2.1. Sensors

A surveillance sensory setup can be extremely heterogeneous. Not only can CCTV cameras be set out in a variety of arrangements, but new technologies can also be more appropriate for specific tasks (Foresti, 2003).

Possible sensors include:

- CCTV cameras;
- PTZ cameras;
- infra-red cameras
- thermal cameras;
- depth cameras.

VA modules must of course be designed ad hoc for the specific data stream they are intended to analyse.

Possible arrangements of such sensors may comprise:

- fixed camera;
- moving camera;
- multiple sensors arrangement (camera networks);
- aerial surveillance;
- egocentric vision.

The abovementioned concepts include the MAVA C.2. ('Camera').

2.2. Functionalities

VA modules can be designed to address a huge variety of tasks, ranging from low-level ones, such as motion detection, to higher inference, such as scene-situation assessment. The design of such functionalities is often addressed in an application-driven fashion. Such a strategy (i.e. restricting the application field) is indeed indicated in case heterogeneity is an issue.

Functionalities of VA modules include, among others (ordered by inference complexity):

- motion detection;
- tracking;
- crowd density/motion estimation;
- object detection;
- object counting (e.g. vehicles, pedestrians);
- object recognition/classification (e.g. face recognition);
- event detection (e.g. abandoned items detection);
- scene understanding (situation detection).

Each functionality of course requires dedicated algorithms, whose complexity increases as inference levels grow. Low levels are rather related to video processing in the strictest sense of the word, while higher levels may rely on more abstract concepts and mathematical tools.

Functionalities are included in the MAVA C.3. ('Video processing chain'), more in detail in C.3.2. The MAVA C.3. also considers that heterogeneity in the 'Video signal' (C.3.1.) to be a factor to be taken into account.

2.3. Scenarios

As already mentioned, heterogeneity of scenarios represents a big challenge in VA. Changing conditions are a matter of fact in the real world and VA modules must be able to cope with them. More in detail, variability may be characterised as both spatial and temporal.

Spatial variability issues include, for instance, indoor–outdoor variations, illumination changings between different locations and background shift. They may arise when moving sensors are employed or when an extensive network of cameras is employed (Ashani, 2009).

On the contrary, time variability refers to differences in scenarios that are caused by their own evolution in time: VA modules are often required to operate 24/7, possibly all year long. Illumination may change depending on weather conditions, the season, the time of day (night time is a particularly challenging scenario) (Hampapur, 2009).

Scenarios are also addressed in the MAVA C.1. ('Scene').

2.4. Real-time requirements

VA algorithms are often required to work in real time. Depending on the specific task a VA module is dedicated to, some events must be detected within few seconds or less. For instance, fire detection modules must trigger an alarm as fast as possible, since fire can spread surprisingly fast. Real-time requirements may also affect the quality of the video stream. Low-resolution frames can be employed in order to reduce processing time; however, this results in information loss due to quality degradation. Recently, methods for video stream acquisition, processing and analytics exploiting the Cloud were proposed in order to overcome issues related to limited availability of storage and compute resources (Abdullah, 2014) (Anjum, 2016).

3. Importance of data sets

Data sets are essential not only for evaluating functionalities of VA algorithms, but also for designing them. They usually consist of a set of videos that are shot with the purpose of testing specific algorithms and are often provided together with the so-called ground truth, i.e. the expected output of a VA module, be it a classification, a tracking or a detection module. Although nowadays there is an unrestrained proliferation of new data sets, most of them are very limited and lack substantial structure. They often come along with new methods and are proposed for the specific purpose of testing them.

VA is a continuously evolving research field and the whole VA community is finding novel application fields as new technologies become available (e.g. thermal cameras, time-of-flight sensors, etc.). The process of building a new data set is extremely complex and many different factors should be considered when looking for a new data set or when creating a new one for solving specific needs. When searching for a VA data set or when creating a new one, it is important to carefully consider and analyse the factors that influence the success and usability of the data set. This section highlights what the important aspects of data sets are and what features have to be considered in order to have a good data set: **common data** for the scientific community to work on and **exhaustive working conditions** with regards to the use case that is considered are fundamental. The next section underlines the **common criticalities** of a VA data set that must be taken into account when selecting or creating a data set.

3.1. Common data

Sharing data from testing methods is regarded as good practice within the scientific community. In fact, this allows having common data to test algorithms. Indeed, many data sets have become a baseline for the evaluation of methods addressing specific tasks. Common data sets not only compare results but also prevent wasting resources in collecting new data that are already available. Moreover, for critical infrastructure operators and their respective policymakers, using common test data sets allows for external auditing based on independent quality standards, thereby assuring stakeholders of the quality of this security measure.

3.2. Exhaustive conditions

Good data sets should cover (where possible) all the possible conditions under which a VA module could work. That is, it must be challenging with respect to the task(s) it has to evaluate. A general challenge is represented, for instance, by different illumination conditions of the videos composing the data set. More specific challenges are usually task related. Some examples are provided in Table 1.

Table 1. Examples of task-related data set challenges

Task	Challenge(s)
Motion detection	<ul style="list-style-type: none"> • Shining surfaces • Moving surfaces • Illumination changes
Tracking	<ul style="list-style-type: none"> • Multiple crossing trajectories causing association uncertainty • Target occlusions
Object detection	<ul style="list-style-type: none"> • Occlusions or partial occlusion

	<ul style="list-style-type: none"> • Size change • Deformable objects
Object counting	<ul style="list-style-type: none"> • Close objects • Occlusions or partial occlusion
Object recognition	<ul style="list-style-type: none"> • Intra-class similarity

3.3. Algorithms comparisons

Common data availability also allows a so-called fair comparison between algorithms. In fact, the evaluation of algorithms performance is a task that strongly depends on the testing data and, if the algorithm requires a training phase, may also depend on the training data.

In addition, as already mentioned in Section 2, due to high variability of conditions no VA module usually works as it is, but needs accurate parameter tuning in order to function properly in different situations. The more the performance depends on the tuning, the less robust the method is considered. Availability of common data also allows proving a method's robustness against this issue.

Data sets are sometimes provided with not only a ground truth, but also with a so-called baseline algorithm. This algorithm represents a starting point for a performance comparison.

3.4. Automatic performance evaluation

Another relevant component that can be provided with a data set is an automatic performance evaluation tool. Given the ground truth in some standard formats (xml is a very common format), comparison with the output of a VA module can be automatised. This implies existence and agreements over standardised evaluation schemes and evaluation scores, which is not always given for granted.

For instance, for a background–foreground segmentation VA module, the problem can be evaluated as a binary classification task. Standard ways of measuring accuracy include the wide class of F_β scores: $\beta = 1$ is the most common choice, but others can also be done. Again, with regard to classification, the n -fold cross validation scheme is usually employed, but the choice of n must be agreed depending on the problem addressed.

4. Critical issues of data sets

Critical issues that can be encountered in recording, designing or testing a data set are discussed in this section. Most of the issues in some way reflect one of the main challenges of the VA discussed above, namely heterogeneity of data.

4.1. Complexity

As discussed above, sequences in a data set should exhaustively introduce challenging conditions in order to test the limits of VA modules. However, a video could be too complex in some cases or not complex enough in others. Namely, it could provide challenging conditions for a kind of task but trivial ones for other functionalities.

In fact, the design of data sets should, in principle, follow a task-oriented approach. Specific sets of videos should be collected and accurately selected in order to properly evaluate each of the desired functionalities of the VA module under investigation.

As an example, a crowded scene like the one depicted in Figure 1 may be introducing just the right amount of complexity in a crowd-density estimation module, while it could be virtually impossible to manage with multiple tracking from a single camera. However, better results are achieved by exploiting multiple camera views (Krahnstoeber, 2009).



Figure 1. A sample frame from the PETS 2009 data set (Ferryman, 2009)

4.2. Ground truth

Many data set issues are related to the ground truth.

First, the ground truth is not always provided, thus impoverishing the video sequences of a consistent piece of data. Not being shared, the whole data set loses part of its generality and is not suitable for fair comparisons anymore, since any user could provide its own labelling of the data with consistent variability among them.

If not provided, a ground truth may be extremely time consuming to produce, especially for low-level tasks such as tracking and background/foreground segmentation. In the former case, the user must manually work out an object's position at each frame, while in the latter labels it must be assigned pixel by pixel.

Even when they are provided, ground truths are often presented in a manifold of formats (among others plain text, CSV and XML) and some additional software must usually be written in order to read and use it.

Further issues concern the methodology used to collect ground truths. Manual labelling is by far the most common methodology employed. Although not immune from bias, human judgement is the best tool available; however, users are sometimes supported by programmes that label ground truths. When this is the case, little control over the reliability of the provided data is left to the final user of the data set. To sum up, ground truths must be trusted, but information about its reliability is often missing or difficult to evaluate.

In many VA use cases, it is important to find the exact position, in world coordinates, of a specific object or event (e.g. left luggage, secured indoor area, intrusion, visitor threats, etc.). Image coordinates can be transformed into world coordinates if certain additional camera parameters (calibration) are available. As discussed in (van Rest, 2015a) calibration and auto-calibration methods constitute a fundamental aspect of many VA systems for critical infrastructure protection. The ground truth for a specific data set is enriched with calibration data, typically in the form of a set of parameters representing intrinsic and extrinsic camera parameters. Extrinsic parameters are related to the specific position where the video sensor is installed, while intrinsic parameters can be used for modelling camera distortions and lenses aberrations. These parameters must be estimated for each sensor in the system and are generally fixed for static cameras. In case of moving cameras such as PTZ or egocentric sensors, calibration data are typically time varying.

Moreover, auto-calibration from video data can be a capability of the VA system, especially for large-scale installations. Thus, data sets can also be used for testing the quality and precision in the estimation of calibration parameters.

4.3. Suitability

As already mentioned, a good practice in the design of data sets should be a task-oriented approach. The critical infrastructure operator should then find or acquire suitable sets of videos to properly evaluate functionalities of VA modules being tested. Unfortunately, it is typically impossible to find or create an exhaustive data set for all the considered functionalities. In fact, a collection of sequences will hardly ever be comprehensive enough to catch all the functionalities that may be required by a VA system. Moreover, the operator should consider that, even if almost complete, a data set might prove to be inadequate for testing future functionalities. The operator should then consider the possibility to use more than one single data set and to continuously revise the one being used for testing when a novel functionality is added to the surveillance system.

The length of a video must also be adequate. A long sequence allows capturing a wide range of environmental changes and constructing reliable models of the scene (Mittal, 2004).

4.3.1. Balance

Many VA modules rely on classification algorithms (e.g. object recognition is usually addressed as a multi-class classification problem; background/foreground segmentation often exploits a pixel-by-pixel binary classifier; detection can rely on a Haar cascade classifier). Such methods require both training and testing data. A data set comprising

only the testing of a video may thus be unsuitable for this class of modules. Instead, a well-designed data set will provide a balanced (e.g. in positive and negatives, or in multiple classes) training set, possibly together with an evaluation scheme, as discussed in the previous section (Betancourt, 2015).

In many cases, VA might not work out of the box: it is then important to select a data set together with the VA provider and work together to set up or even customise the algorithm (see recommendations (van Rest, 2015b) and (van Rest, 2015a)).

4.4. Video quality

Resolution is defined as the number of pixels (usually expressed in terms of width x height) that make up each picture frame. This is a one-dimensional definition used to describe an image, which is what an individual frame is made up of. FPS (frames per second) is used instead to represent the frame rate, which actually encodes the motion quality of a video stream.

Therefore, while resolution indicates the quality of single images displayed, FPS indicates the quality of the video motion. A high resolution combined with high FPS results in a high-quality video stream, however, requires higher bandwidth and higher storage requirements for video streams. Although the capacity of hard disks is now a minor issue, the size of sequences in a data set must be taken into consideration.

Also for the quality-related parameters, the design of a data set should follow a task-oriented design. The quality of the data set should be representative of the quality of the video that will be used in operational use, including resolution, frame rate, compression, quality of synchronisation and quality of calibration. It is always good to balance resolution and FPS based on the application requirements. Certain applications like face recognition, licence plate reading, etc. might require a higher resolution while not being demanding for what concerns FPS. On the contrary, others, like traffic monitoring and perimeter security, might be satisfied with a low resolution while requiring a higher frame rate. Some sequences could thus be extremely challenging with respect to some tasks, due to the characteristics of the video stream.

Video surveillance cameras used today usually support the following common resolutions:

- QVGA cameras — 320 x 240 pixels;
- VGA cameras — 640 x 480 pixels;
- megapixel cameras — 1280 x 1024 pixels;
- HDTV cameras — 1280 x 720 pixels, 1920 x 1080 pixels.

There might of course be other resolutions as well. This again raises the issue of the wide heterogeneity of data sets.

The commonly used frame rates might vary from 15 frames per second to 30 frames per second. In many cameras, one can actually select the desired frame rate based on the video image quality/available bandwidth/storage space, etc. A frame rate of at least 10 frames per second is usually recommended for the human eye to be able to comprehend the motion properly.

In certain applications, like face recognition, etc., a higher resolution might be more important than a higher frame rate because the images need to be clear enough for people to identify certain individual aspects in order to aid the investigation process. Some video surveillance applications allow video streams to be transmitted at different frame rates and resolutions. For example, a video can be transmitted to a monitor at a different frame rate/resolution, but the same can be recorded at a different frame rate/resolution as well.

Moreover, when considering VA data sets, video compression standards and qualities must be taken into account. Compression techniques must be used both when broadcasting and when storing video data. Video files can be extremely large if they have a long duration or were recorded in megapixel resolution. If the compression is too high, the image quality can be compromised. The following are compression standards that are mostly used for video surveillance.

- H.264: it is the more recent and most efficient video compression codec. It works by exploiting both spatial and temporal correlation in small groups of consecutive frames.
- Motion JPEG (MJPEG): it considers each frame of the video separately and compresses them as individual JPEG images.

More recent surveillance cameras can have multiple video streams and may be able to use multiple video compression codecs or different levels of compression. This allows one to configure different streams for mobile viewing, live viewing and long-term storage.

The actual bandwidth needed to transmit and store compressed videos depends on all the characteristics discussed earlier, including video format, frame rate and compression type. The following list summarises typical bandwidths for different video formats and frame rates.

- Megapixel camera — 1280 x 1024 pixels at 30 FPS with H.264 compression: a bandwidth of approx. 4 Mbps per camera should be expected.
- QVGA cameras — 320 x 240 pixels at 5 FPS with H.264 compression: a bandwidth of approx. 125 Kbps per camera should be expected.
- Megapixel camera — 1280 x 1024 pixels at 30 FPS with MJPEG compression: a bandwidth of approx. 12 Mbps per camera should be expected.

These bandwidths can vary depending on the quality of compression and on the actual content on the video: a higher level of motion in the scene means larger bandwidths.

In case of multicamera data sets, the correct synchronisation between video streams represents an important issue. Therefore, every video of the data set should be tagged with the acquisition time in such a way that VA algorithms can actually correlate and fuse information extracted from the multicamera system. Several synchronisation techniques exist in the state of the art, but the most used one is the well-known network time protocol. Designed to synchronise the clocks on network nodes of an IP network with a reliable time source, it can be efficiently used for guaranteeing the synchronisation in camera networks. The global positioning system or code division multiple access signals can be used as an accurate timing source.

4.5. Privacy issues

A critical but often neglected issue related to data sets is privacy. For instance, still pictures (frames) often clearly show faces or the registration plates of cars (Korshunov, 2014). Not only is the individual sometimes filmed without their permission, but their picture is also going to be distributed to third parties who will use it for their own purposes, including business-related purposes. Privacy consent forms and statements should come along with data sets where people are clearly recognisable. However, laws regulating privacy in videos may change from country to country, while data sets usually circulate worldwide.

4.6. Open access

Nowadays in the digital era, sharing multimedia content is a matter of a click. However, some data sets exist, but for some reason are closed, i.e. not publicly available. Open access to data is a delicate and controversial topic within the research community. Reasons why data sets may be closed are to be ascribed to some latent need to protect research from being 'stolen' more than to tangible copyright issues related to publishers. Contrarily, industries are reasonably protecting their data as part of a market based on competition.

A clear solution for gaining access to closed data sets may consist in contacting the owners to request the data, clearly stating the kind of use intended, the purpose and the people who will have access to it.

5. Data set construction checklist

This chapter proposes a simple checklist for when a data set has to be selected or created for a specific use case. Different aspects already mentioned in the previous sections of this document are considered and briefly summarised for the considered data set.

It is worth noticing that some of the following can be quite complex to achieve (e.g. ground truth, calibration, baseline algorithm) and expert researchers and scientists might be involved to efficiently address these points. Moreover, in order to allow end users to use the following checklist it should be transformed into a practical method that requires as little background knowledge as possible.

<input type="checkbox"/>	Enough sequences	The data set contains enough sequences, exhaustively covering all working conditions. List all the different conditions together with a brief description.
<input type="checkbox"/>	Readme	The data set contains a detailed description (in the form of a readme with a summary of the data set content).
<input type="checkbox"/>	Ground truth	The data set contains the ground truth. If yes, add the format that is used for storing the data.
<input type="checkbox"/>	Calibration	The data set contains camera calibration parameters: intrinsic and/or extrinsic for static cameras.
<input type="checkbox"/>	Availability	The data set should be public and stored onto a server with a robust and fast internet connection to allow multiple concurrent downloads.
<input type="checkbox"/>	Evaluation	Because of the presence of the ground truth, it is possible to automatically evaluate the performance of proposed techniques using that data set. The data set creators should release some sort of automatic performance evaluation tool, either offline or web-based. Describe if and how this automatic evaluation can be done.
<input type="checkbox"/>	Baseline algorithm code	The data set contains an open source code for the proposed baseline algorithm. Describe if this is available and how it can be used.
<input type="checkbox"/>	Quality and format	The data set quality (e.g. image size, frame rate, etc.) and format (e.g. video/audio codec) should be guaranteed.

6. Main features

Based on the considerations presented above, here are the following fundamental features for a data set to comply with.

- **Goal:** following the task-oriented approach sketched so far, specifications should be given about which VA functionalities the data sets are intended for.
- **Realism:** natural scenes showing people or vehicles performing normal actions in standard contexts, with uncontrolled and cluttered backgrounds.
- **Diversity:** multiple heterogeneous locations with a variety of camera viewpoints and resolutions are to be included; diverse backgrounds and illumination conditions; a wide range of human actions and interactions; various environmental conditions (weather).
- **Quantity:** a considerable amount of examples for each of the classes considered; classes of interactions, of actions and of objects to be recognised or located, etc.
- **Length:** it is important for sequences to cover a sufficiently long time interval, not only to provide a consistent amount of data but also to allow the evaluation of long-term environmental changes (e.g. background evolution during the day).
- **Video quality:** a wide range of resolutions and frame rates; 2-30 Hz frame rates and 10-200 pixels in person-height. As discussed above, different combinations of resolution and FPS can be more suited for certain tasks.
- **Type of sensor(s):** realism must also be attained in the type of sensors available, from standard cameras to thermal, depth, etc.
- **Ground truth:** its availability is a matter of importance; it should be provided in a readable format, possibly together with specifications about its collection.
- **Calibration data:** for fixed cameras, calibration parameters may be provided, if available.
- **Ease of use:** the data set should contain specifications and list the included sequences together with a short description of the purpose; file names should be adequate; formats should not require exotic video codecs; eventually an automatic testing tool could be provided.
- **Accessibility:** server storage of a data set should be reliable in order to make it available to users at any time.

Additional features (important but not fundamental) may include the following.

- **Multiple sensors:** possibly multiple synchronised views of the scene (if more cameras are available or if different kinds of sensory equipment are deployed).
- **Ground and aerial videos:** (synchronised) aerial views might be useful.
- **Cost:** is the data set freely available?

7. Existing data sets

In this chapter, a list of existing data sets for surveillance VA purposes is provided. The list has been compiled having availability and relevance in mind. This list may not be comprehensive since research community continuously adds many data sets. Interested readers may refer to online directories such as (Riemenschneider, 2016), (Fisher, 2016) or (Truyen, 2008) to see other lists of computer vision data sets. However, there are no guarantees that any of these lists are complete, or that they are even representative of end users' situations and needs. This is why the development and operation of a high quality online repository for relevant data sets is recommended (see also Section 4.1. of (van Rest, 2015a)).

Table 2 shows the list of data sets, each data set is cross-referenced with the Appendix A. Data sets description where detailed description of data set is provided. Total number of citations and some of the most significant works (in terms of number of received citations, publication year, relevance of the journal/conference) making use of the data set are also listed in Appendix A (for each data set).

Table 2 also shows goals and types of the objects in each data set. Furthermore, the potential use cases (van Rest, 2015a) for which the data set can be useful is and shown in the Table 2. These use cases are achieved by associating the data set goals and object types to the required video analytic functionalities in each use case (see Table 3 for association table).

It is worth noticing that the following list should not be intended as a recommendation or endorsement that a particular data set is of substantial quality or that it is representative of any use case.

Table 2. Existing data set

Data set	Use cases	VA goals	Objects
A.1. HDA person data set	left luggage, secured indoor area, intrusion, visitor threats	object tracking, person re-identification	people
A.2. WWW crowd data set	crowd control	crowd analysis	people
A.3. MOT benchmark	left luggage, secured indoor area, intrusion, visitor threats	object tracking	people
A.4. ChokePoint data set	left luggage, secured indoor area, intrusion, visitor threats	object tracking, person re-identification	people
A.5. VIRAT	secured indoor area	event detection, activity recognition	people, car
A.6. Comprehensive cars (CompCars)	bomb threat, cargo theft at highway	object detection, object classification	car
A.7. INRIA person data	left luggage, secured indoor area, bomb threat, intrusion,	object detection	people

set	visitor threats		
A.8. TUGRAZ ICG long-term pedestrian data set	left luggage, secured indoor area, bomb threat, intrusion, visitor threats	object detection, tracking	people
A.9. Crowd data set	crowd control	crowd analysis	people
A.10. PEdesTrian attribute (PETA) data set	left luggage	object recognition	people
A.11. CUHK crowd data set	crowd control	crowd analysis	people
A.12. GRAZ-02	left luggage, cargo theft at highway	object recognition	people, car, bike
A.13. Person Re-ID (PRID) 2011	left luggage, secured indoor area, intrusion, visitor threats	person re-identification	people
A.14. MuHAVi	secured indoor area, public order management	action recognition	people
A.15. GRAZ-01	left luggage, secured indoor area, bomb threat, cargo theft at highway, intrusion, visitor threats	object detection	people, car, bike
A.16. Mall data set	crowd control, left luggage, secured indoor area, intrusion, visitor threats	crowd analysis, trajectory analysis, tracking, object detection	people
A.17. KTH action	secured indoor area, public order management	action recognition	people
A.18. Weizmann actions	secured indoor area, public order management	action recognition	people
A.19. UT-Interaction	secured indoor area, public order management	interaction analysis	people
A.20. i-LIDS	secured indoor area	event detection	
A.21. NIST digital			

video 1			
A.22. Pedestrian walking path data set	crowd control, left luggage, secured indoor area, intrusion, visitor threats	crowd analysis, trajectory analysis, object detection	people
A.23. PETS 2007	left luggage, secured indoor area	event detection	people, luggage
A.24. PETS 2006	left luggage, secured indoor area	event detection	people, luggage
A.25. PETS 2009	crowd control, secured indoor area, intrusion, visitor threats	crowd analysis, object tracking, event detection	people
A.26. PETS 2015	secured indoor area, intrusion, visitor threats, public order management	object tracking, event detection, trajectory analysis, interaction analysis	people
A.27. UCF aerial action data set	secured indoor area, public order management	activity recognition	people, car
A.28. Mini-drone video data set	secured indoor area, public order management	activity recognition	people, car

Table 3. Use case to VA functionality association

Use case	Capability	VA functionalities	Object
left luggage	detection of left luggage	object detection, object recognition, object tracking, trajectory analysis, interactivity analysis	people, luggage
left luggage	determining owner	object detection, object tracking, person re-identification	people
left luggage	locating owner	object detection, object tracking, person re-identification	people
left luggage	following owner	object detection, object tracking, person re-identification	people
secured indoor area	detection of loitering	object detection, object tracking, person re-identification, trajectory analysis, activity recognition	people
secured indoor area	detection of tailgating	object detection, object tracking, person re-identification, trajectory analysis, interactivity analysis	people
secured indoor area	walking against the mandatory flow	object detection, object tracking, person re-identification, trajectory analysis, event detection	people
secured indoor area	detection of passing through a door	object detection, object tracking, person re-identification, trajectory analysis, event detection	door, people
secured indoor area	sterile zone detection		

secured indoor area	following intruder	object detection, object tracking, person re-identification, trajectory analysis	people
public order management	aggression detection against bodycam user	activity recognition, interactivity analysis	people, weapon
maintenance	auto-calibration of large VSS deployments		
crisis management	auto-calibration of heterogeneous VSS deployments		
crowd control		crowd analysis	people
bomb threat		object detection	people, bomb
cargo theft at highway		object detection, object recognition	car
intrusion		object detection, object tracking, person re-identification, trajectory analysis	people
visitor threats		object detection, object tracking, person re-identification, trajectory analysis	people

8. Conclusion

This report presented a critical analysis of VA data sets with specific attention towards protection of critical infrastructures. The introductory part of the report described the importance of VA and the growth of the related market. In this scenario, the importance of the usage of a common data set was highlighted. The main reason for the fundamental importance of data sets in video analysis is the intrinsic complexity of VA-related techniques: a common set of video sequences is seen as a powerful boost in the design, development and test of VA algorithms.

This report described different aspects that make VA so complex and demonstrated the importance of having common and widespread data sets. Data sets must also rely on the availability of standards related to several aspects of the VA for critical infrastructures protection: refer to (Ferryman, 2016) for an overview of standards in video surveillance, including the need for standards, for an overview of existing relevant standardisation efforts, including gaps, and for a roadmap for the development of future standards.

A detailed description and analysis of critical issues of VA data sets were provided, and a simple but effective 'data set construction checklist' was proposed.

In the last part of the report, several existing data sets were summarised and commented in relation with the use cases highlighted in the report 'Surveillance use cases — Focus on ERNCIP video analytics', Thematic Group on Video Analytics and Surveillance, 2015. Moreover, the impact of each data set in the scientific community was estimated by considering the total number of referencing papers and the most relevant research using the data set for computing the performances of a proposed technique.

With this report, we follow up on the recommendations regarding test data sets for VA use cases of (van Rest, 2015b) and (van Rest, 2015a). In particular:

- together with (van Rest, 2015b) and the *Video Analytics Adoption - Key considerations for the end user* (Doyle, 2016), this report helps build an argument for why data sets matter in the boardroom of critical infrastructure end users and industry;
- this report gives the requirements for creating high-quality relevant data sets.

8.1. Future work

VA data sets play an important role for critical infrastructure protection, and this report can be considered as a first effort for enabling end users to find, select and create useful data sets for designing, testing and improving an adopted solution. However, many issues require further investigation. The proposed 'data set construction checklist' is currently very abstract and the study and development of a practical methodology would be very useful. A template procurement framework (to be used by critical infrastructures end users when procuring VA solutions) should be developed in such a way that it focuses on the importance of data sets in all the design, development and acceptance tests for a VA system.

Significant work should be done for designing, developing and maintaining a high-quality online repository for relevant data sets (see also Section 4.1. of (van Rest, 2015a))

As already noticed in the report (see Section 4.3.1.), algorithms will not work out of the box after the selection of a VA provider by using a data set; a period of co-development is recommended (see (van Rest, 2015b), (van Rest, 2015a)). For this reason, critical infrastructure operators may need to involve scientists to assist them in providing and using data sets. This report might be enriched by proposing a collaboration framework and guidelines between operators/end users and VA researchers and developers.

References

- Ferryman, J., 'Video Surveillance Standardization - Activities, Process and Roadmap', 2016.
- van Rest, J., 'Surveillance Use Cases: Focus on Video Analytics - ERNCIP Thematic Group Video Analytics and Surveillance', *JRC Publications Repository*. 2015a. <http://publications.jrc.ec.europa.eu/repository/handle/JRC100401>. Accessed Nov 2016.
- van Rest, J., 'Surveillance and Video Analytics: Factors influencing the performance - ERNCIP Thematic Group Video Analytics and Surveillance', *JRC Publications Repository*. 2015b. <http://publications.jrc.ec.europa.eu/repository/handle/JRC100399>. Accessed Nov 2016.
- Doyle, S., 'Video Analytics Adoption - Key considerations for the end user', *ERNCIP Project*. Aug 4, 2016. <https://erncip-project.jrc.ec.europa.eu/documents/video-analytics-adoption-key-considerations-end-user>. Accessed Nov 2016.
- Markets and Markets, 'Video Analytics Market by Type (Hardware, Video Analytics Software, and Services), Applications (Intrusion Management, Crowd Management, Situation Indication, License Plate Recognition, Pattern Recognition) - Global Forecast to 2020', TC 3523, 2016.
- Dore, A., Soto, M., and Regazzoni, C.S., 'Bayesian Tracking for Video Analytics', *Signal Processing Magazine, IEEE*, pp. 46-55, 2010.
- Narayanan, V., Datta, S., Cauwenberghs, G., Chiarulli, D., Levitan, S., and Wong, P., 'Video analytics using beyond CMOS devices', *Design, Automation and Test in Europe Conference and Exhibition (DATE)*, 2014, 2014.
- Greiffenhagen, M., Ramesh, V., Comaniciu, D., and Niemann, H., 'Statistical modeling and performance characterization of a real-time dual camera surveillance system', *Computer Vision and Pattern Recognition, 2000. Proceedings. IEEE Conference on*, Hilton Head, SC, USA, 2000.
- Ramesh, V., 'Real-time vision at Siemens Corporate Research', *Advanced Video and Signal Based Surveillance, 2005. AVSS 2005. IEEE Conference on*, 2005.
- Foresti, G.L., Regazzoni, C.S., and Varshney, P.K.
- Ashani, Z., 'Architectural Considerations for Video Content Analysis in Urban Surveillance', *Advanced Video and Signal Based Surveillance, 2009. AVSS '09. Sixth IEEE International Conference on*, 2009.
- Hampapur, A., Bobbitt, R., Brown, L., Desimone, M., Feris, R., Kjeldsen, R., Lu, M., Mercier, C., Milite, C., Russo, S., Shu, C.F., and Zhai, Y., 'Video Analytics in Urban Environments', *Advanced Video and Signal Based Surveillance, 2009. AVSS '09. Sixth IEEE International Conference on*, 2009.
- Abdullah, T., Anjum, A., Tariq, M.F., Baltaci, Y., and Antonopoulos, N., 'Traffic Monitoring Using Video Analytics in Clouds', *Utility and Cloud Computing (UCC), 2014 IEEE/ACM 7th International Conference on*, 2014.
- Anjum, A., Abdullah, T., Tariq, M.F., Baltaci, Y., and Antonopoulos, N., 'Video Stream Analysis in Clouds: An Object Detection and Classification Framework for High Performance Video Analytics', *Cloud Computing, IEEE Transactions on*, pp. 1, 2016.
- Krahnstoeber, N., Tu, P., Yu, T., Patwardhan, K., Hamilton, D., Yu, B., Greco, C., and Doretto, G., 'Intelligent Video for Protecting Crowded Sports Venues', *Advanced Video and Signal Based Surveillance, 2009. AVSS '09. Sixth IEEE International Conference on*, 2009.
- Mittal, A., Paragios, N., 'Motion-based background subtraction using adaptive kernel density estimation', *Computer Vision and Pattern Recognition, 2004. CVPR 2004. Proceedings of the 2004 IEEE Computer Society Conference on*, 2004.

- Betancourt, A., Morerio, P., Barakova, E., Marcenaro, L., Rauterberg, M., and Regazzoni, CS., 'A Dynamic Approach and a New Dataset for Hand-Detection in First Person Vision', *International Conference on Computer Analysis of Images and Patterns*, 2015.
- Korshunov, P., Ebrahimi, T., 'UHD video dataset for evaluation of privacy', *Quality of Multimedia Experience (QoMEX), 2014 Sixth International Workshop on*, 2014.
- Riemenschneider, H., 'Yet Another Computer Vision Index To Datasets', YACVID. 2016. <http://riemenschneider.hayko.at/vision/dataset/>. Accessed Nov 2016.
- Fisher, R., 'CVOnline: Image Databases'. 2016. <http://homepages.inf.ed.ac.uk/rbf/CVonline/Imagedbase.htm>. Accessed Nov 2016.
- Truyen, R., 'Cantata'. 2008. <http://www.multitel.be/cantata/>. Accessed Nov 2016.
- Namibar, A., Taiana, M., Figueira, D., Nascimento, J.C., and Bernardino, A., 'A multi-camera video dataset for research on high-definition surveillance', *International Journal of Machine Intelligence and Sensory Signal Processing*, 2014.
- Shao, J., Kang, K., Loy, C.C., and Wang, X., 'Deeply learned attributes for crowded scene understanding', *In Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2015.
- Shao, J., Loy, CC., Kang, K., and Wang, X., 'Slicing Convolutional Neural Network for Crowd Video Understanding', *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016.
- Leal-Taixé, L., Milan, A., Reid, I., Roth, S., and Schindler, K., 'MOTChallenge 2015: Towards a Benchmark for Multi-Target Tracking', *arXiv:1504.01942*, 2015.
- Bewley, A., Ott, L., Ramos, F., and Upcroft, B., 'Alextrac: Affinity learning by exploring temporal reinforcement within association chains.', *International Conference on Robotics and Automation*, 2016.
- Xiang, Y., Alahi, A., and Savarese, S., 'Learning to track: Online multi-object tracking by decision making', *Proceedings of the IEEE International Conference on Computer Vision*, 2015.
- Manen, S., Timofte, R., Dai, D., and Van Gool, L., 'Leveraging single for multi-target tracking using a novel trajectory overlap affinity measure', *IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2016.
- Wong, Y., Chen, S., Mau, S., Sanderson, C., and Lovell, BC., 'Patch-based Probabilistic Image Quality Assessment for Face Selection and Improved Video-based Face Recognition', *IEEE Biometrics Workshop, Computer Vision and Pattern Recognition (CVPR) Workshops*, 2011.
- An, L., Kafai, M., and Bhanu, B., 'Dynamic bayesian network for unconstrained face recognition in surveillance camera networks', *Emerging and Selected Topics in Circuits and Systems, IEEE Journal on*, 2013.
- Kim, HI., Lee, SH., and Ro, YM., 'Face image assessment learned with objective and relative face image qualities for improved face recognition', *Image Processing (ICIP), 2015 IEEE International Conference on* (pp. 4027-4031). IEEE., 2015.
- Gou, G., Huang, D., and Wang, Y., 'Video face recognition via combination of real-time local features and temporal spatial cues', *Computer Vision, IET*, 8(4), pp.347-357., 2014.
- Oh, S., Hoogs, A., Perera, A., Cuntoor, N., Chen, CC., and Lee, J.T., 'A large-scale benchmark dataset for event recognition in surveillance video', *Computer Vision and Pattern Recognition (CVPR), 2011, IEEE Conference on*, 2011.
- Vondrick, C., Patterson, D., and Ramanan, D., 'Efficiently scaling up crowdsourced video annotation', *International Journal of Computer Vision*, No 101(1), pp. 184-204, 2013.

- Burgos-Artizzu, X.P., Dollár, P., Lin, D., Anderson, D.J., and Perona, P., 'Social behavior recognition in continuous video', *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, 2012.
- Vondrick, C., Ramanan, D., 'Video annotation and tracking with active learning', *Advances in Neural Information Processing Systems*, pp. 28-36, 2011.
- Yang, L., Luo, P., Loy, C.C., and Tang, X., 'A Large-Scale Car Dataset for Fine-Grained Categorization and Verification', *Computer Vision and Pattern Recognition*, 2015.
- Liu, H., Tian, Y., Yang, Y., Pang, L., and Huang, T., 'Deep Relative Distance Learning: Tell the Difference Between Similar Vehicles', *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016.
- Sochor, J., Herout, A., and Havel, J., 'BoxCars: 3D Boxes as CNN Input for Improved Fine-Grained Vehicle Recognition', *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016.
- Dalal, N., Triggs, B., 'Histograms of oriented gradients for human detection', *Computer Vision and Pattern Recognition*, 2005.
- Felzenszwalb, P.F., Girshick, R.B., McAllester, D., and Ramanan, D., 'Object detection with discriminatively trained part-based models', *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, No 32(9), pp. 1627-1645, 2010.
- Wang, J., Yang, J., Yu, K., Lv, F., Huang, T., and Gong, Y., 'Locality-constrained linear coding for image classification', *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, 2010.
- Felzenszwalb, P., McAllester, D., and Ramanan, D., 'A discriminatively trained, multiscale, deformable part model', *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, 2008.
- Roth, P.M., Sternig, S., Grabner, H., and Bischof, H., 'Classifier Grids for Robust Adaptive Object Detection', *CVPR*, 2009.
- Sternig, S., Roth, P.M., and Bischof, H., 'On-line inverse multiple instance boosting for classifier grids', *Pattern recognition letters*, No 33(7), pp. 890-897, 2012.
- Roth, P.M., SV, W.P., Lancelle, M., Birchbauer, J., Brändle, N., Havemann, S., and Bischof, H., 'Next-generation 3D visualization for visual surveillance', *Advanced Video and Signal-Based Surveillance (AVSS), 2011 8th IEEE International Conference on*, 2011.
- Htike, K.K., Hogg, D.C., 'Efficient non-iterative domain adaptation of pedestrian detectors to video scenes', *Proceedings-22nd International Conference on Pattern Recognition IEEE*, 2014.
- Lim, M.K., Kok, V.J., Loy, C.C., and Chan, C.S., 'Crowd Saliency Detection via Global Similarity Structure', *International Conference on Pattern Recognition*, 2014.
- Deng, Y., Luo, P., Loy, C.C., and Tang, X., 'Pedestrian attribute recognition at far distance', *Proceedings of ACM Multimedia (ACM MM)*, 2014.
- Tian, Y., Luo, P., Wang, X., and Tang, X., 'Pedestrian detection aided by deep learning semantic tasks', *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015.
- Shi, Z., Hospedales, T.M., and Xiang, T., 'Transferring a semantic representation for person re-identification and search', *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015.
- Ye, M., Liang, C., Wang, Z., Leng, Q., Chen, J., and Liu, J., 'Specific person retrieval via incomplete text description', *Proceedings of the 5th ACM on International Conference on Multimedia Retrieval*, 2015.

Shao, J., Loy, C.C., and Wang, X., 'Scene-Independent Group Profiling in Crowd', *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014.

Yi, S., Wang, X., Lu, C., and Jia, J., 'LO regularized stationary time estimation for crowd group analysis', *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014.

Wang, X., Wang, M., and Li, W., 'Scene-specific pedestrian detection for static video surveillance', *IEEE transactions on pattern analysis and machine intelligence*, No 36(2), pp. 361-374, 2014.

Shao, J.DNaZQ., 'An adaptive clustering approach for group detection in the crowd', *International Conference on Systems, Signals and Image Processing (IWSSIP)*, 2015.

Opelt, A., Pinz, A., Fussenegger, M., and Auer, P., 'Generic object recognition with boosting', *Pattern Analysis and Machine Intelligence, IEEE Transaction on*, pp. 416-431, 2006.

Ramirez, I., Sprechmann, P., and Sapiro, G., 'Classification and clustering via dictionary learning with structured incoherence and shared features', *In Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, 2010.

Opelt, A., Pinz, A., and Zisserman, A., 'Incremental learning of object detectors using a visual shape alphabet', *Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2006.

Leordeanu, M., Hebert, M., and Sukthankar, R., 'Beyond local appearance: Category recognition from pairwise interactions of simple features', *IEEE Conference on Computer Vision and Pattern Recognition*, 2007.

Hirzer, M., Beleznaï, C., Roth, P.M., and Bischof, H., 'Person Re-Identification by Descriptive and Discriminative Classification', *Scandinavian Conference on Image Analysis*, 2011.

Hirzer, M., Roth, PM., Köstinger, M., and Bischof, H., 'Relaxed pairwise learned metric for person re-identification', *European Conference on Computer Vision*, 2012.

Hirzer, M., Roth, PM., and Bischof, H., 'Person re-identification by efficient impostor-based metric learning', *Advanced Video and Signal-Based Surveillance (AVSS), 2012 IEEE Ninth International Conference on*, 2012.

Ma, AJ., Yuen, PC., and Li, J., 'Domain transfer support vector ranking for person re-identification without target camera label information', *Proceedings of the IEEE International Conference on Computer Vision*, 2013.

Singh, S., Velastin, SA., and Ragheb, H., 'MuHAVi: A Multicamera Human activity Video Dataset for the Evaluation of activity Recognition Methods', *2nd Workshop on Activity monitoring by multi-camera surveillance systems*, 2010.

Chaaroui, AA., Climent-Pérez, P., and Flórez-Revuelta, F., 'Silhouette-based human action recognition using sequences of key poses', *Pattern Recognition Letters*, No 34(15), pp. 1799-1807, 2013.

Cheema, S., Eweiwi, A., Thureau, C., and Bauckhage, C., 'Action recognition by learning discriminative key poses', *Computer Vision Workshops (ICCV Workshops), 2011 IEEE International Conference on*, 2011.

Eweiwi, A., Cheema, S., Thureau, C., and Bauckhage, C., 'temporal key poses for human action recognition', *Computer Vision Workshops (ICCV Workshops), 2011 IEEE International Conference on*, 2011.

Opelt, A., Fussenegger, M., Pinz, A., and Auer, P., 'Weak Hypotheses and Boosting for Generic Object Detection and Recognition', *ECCV*, 2004.

Lazebnik, S., Schmid, C., and Ponce, J., 'Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories', *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, 2006.

Boiman, O., Shechtman, E., and Irani, M., 'In defense of nearest-neighbor based image classification', *Computer Vision and Pattern Recognition. CVPR. IEEE Conference on*, 2008.

Chen, K., Loy, C.C., Gong, S., and Xiang, T., 'Feature Mining for Localised Crowd Countin', *British Machine Vision Conference*, 2012.

Loy, CC., Chen, K., Gong, S., and Xiang, T., 'Crowd Counting and Profiling: Methodology and Evaluation', *Modeling, Simulation and Visual Analysis of Crowds*, 2013.

Change Loy, C., Gong, S., and Xiang, T., 'From semi-supervised to transfer counting of crowds', *IEEE International Conference on Computer Vision*, 2013.

Zhang, Z., Wang, M., and Geng, X., 'Crowd counting in public video surveillance by label distribution learning', *Neurocomputing*, 2015.

Xu, J., Wu, Q., Zhang, J., Silk, B., Ngo, GT., and Tang, Z., 'Efficient People Counting with Limited Manual Interferences', *Digital Image Computing: Techniques and Applications (DICTA), IEEE International Conference on*, 2014.

Schuldt, C., Laptev, I., and Caputo, B., 'Recognizing Human Actions: A Local SVM Approach', *ICPR*, 2014.

Laptev, I., Marszalek, M., Schmid, C., and Rozenfeld, B., 'Learning realistic human actions from movies', *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, 2008.

Wang, H., Kläser, A., Schmid, C., and Liu, CL., 'Action recognition by dense trajectories', *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, 2011.

Jhuang, H., Serre, T., Wolf, L., and Poggio, T., 'A biologically inspired system for action recognition', *IEEE 11th International Conference on Computer Vision*, 2007.

Zhang, Z., Tao, D., 'Slow feature analysis for human action recognition', *IEEE Transactions on Pattern Analysis and Machine Intelligence*, No 34(3), pp. 436-450, 2012.

Blank, M., Gorelick, L., Shechtman, E., Irani, M., and Basri, R., 'Activities as space-time shapes', *ICCV*, 2005.

Gorelick, L.BM,SE,IMaBR., 'Actions as space-time shapes', *IEEE transactions on pattern analysis and machine intelligence*, No 29(12), pp. 2247-2253, 2007.

Ferrari, V., Marin-Jimenez, M., and Zisserman, A., 'Progressive search space reduction for human pose estimation', *Computer Vision and Pattern Recognition, CVPR*, 2008.

Brendel, W., Todorovic, S., 'Learning spatiotemporal graphs of human activities', *International Conference on Computer Vision IEEE*, 2011.

Ryoo, MS., Aggarwal, JK., 'Spatio-temporal relationship match: Video structure comparison for recognition of complex human activities', *ICCV*, 2009.

Ryoo, MS., 'Human activity prediction: Early recognition of ongoing activities from streaming videos', *International Conference on Computer Vision IEEE.*, 2011.

Bayona, Á., SanMiguel, JC., and Martínez, JM., 'Stationary foreground detection using background subtraction and temporal difference in video surveillance', *IEEE International Conference on Image Processing*, 2010.

Yi, S., Li, H., and Wang, X., ' Understanding pedestrian behaviors from stationary crowd groups', *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015.

Yi, S., Li, H., and ang, X., 'Pedestrian Travel Time Estimation in Crowded Scenes', *Proceedings of the IEEE International Conference on Computer Vision*, 2015.

Yi, S., Li, H., and Wang, X., 'Pedestrian Travel Time Estimation in Crowded Scenes', *Proceedings of the IEEE International Conference on Computer Vision*, Pedestrian Travel Time Estimation in Crowded Scenes.

Arsié, D., Schuller, B., and Rigoll, G., 'Multiple camera person tracking in multiple layers combining 2d and 3d information', *orkshop on Multi-camera and Multi-modal Sensor Fusion Algorithms and Applications-M2SFA2*, 2008.

Leach, MJ.SEP,RNM., 'Contextual anomaly detection in crowded surveillance scenes', *Pattern Recognition Letters*, No 44, pp. 71-79, 2014.

Hattori, H., Naresh Boddeti, V., Kitani, KM., and Kanade, T., 'Learning scene-specific pedestrian detectors without real data', *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015.

Godec, M., Sternig, S., Roth, PM., and Bischof, H., 'Context-driven clustering by multi-class classification in an active learning framework', *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition-Workshops*, 2010.

Shahrokni, A., Ellis, A., and Ferryman, J., 'Overall evaluation of the PETS2009 results', *IEEE International Workshop on Performance Evaluation of Tracking and Surveillance Miami*, 2009.

Ferryman, J., Shahrokni, A., 'An overview of the PETS 2009 challenge', *IEEE International Workshop on Performance Evaluation of Tracking and Surveillance*, 2009.

Yang, Y., Shu, G., and Shah, M., 'Semi-supervised learning of feature hierarchies for object detection in a video', *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2013.

Shu, G., Dehghan, A., and Shah, M., 'mproving an object detector and extracting regions using superpixels', *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2013.

Wang, J., Fu, W., Liu, J., and Lu, H., 'Spatiotemporal group context for pedestrian counting', *IEEE Transactions on Circuits and Systems for Video Technology*, No 24(9), pp. 1620-1630, 2014.

Milan, A., Leal-Taixé, L., Schindler, K., and Reid, I., 'Joint tracking and segmentation of multiple targets', *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015.

Li, L., Nawaz., T., and Ferryman, J., 'PETS2015: Dataset and Challenge', *IEEE International Conference on Advanced Video and Signal-based Surveillance (AVSS)*, 2015.

Bastani, V., Campo, D., Marcenaro, L., and Regazzoni, C., 'Online pedestrian group walking event detection using spectral analysis of motion similarity graph', *Advanced Video and Signal Based Surveillance (AVSS), 12th IEEE International Conference on*, 2015.

Nawaz, T., Boyle, J., Li, L., and Ferryman, J., 'Tracking performance evaluation on PETS 2015 Challenge datasets', *Advanced Video and Signal Based Surveillance (AVSS), 12th IEEE International Conference on*, 2015.

Bonetto, M., Korshunov, P., Ramponi, G., and Ebrahimi, T., 'Privacy in Mini-drone Based Video Surveillance', *Workshop on De-identification for privacy protection in multimedia*, 2015.

Ruchaud, N., Dugelay, JL., 'Privacy Protection Filter Using StegoScrambling in Video Surveillance', *MediaEval 2015 Workshop*, Wurzen, Germany, 2015.

Ferryman, J., Crowley, J.L., 'Eleventh IEEE International Workshop on Performance Evaluation of Tracking and Surveillance', 2009.

List of abbreviations and definitions

Term	Definition
24/7 requisites	A surveillance system must be designed and implemented to work 24 hours per day and 7 days per week (i.e. always without interruption).
Automotive safety	Refers to an automatic system that can be used in the vehicles to improve driver security (i.e. sleep detection).
Background shift	A typical situation with non-static cameras where it could be extremely difficult to build a reference (background) image because of the continuous movements of the sensor. The background is generally used for estimating objects of interest in the scene.
Baseline algorithm	The 'standard' technique that can be used for implementing a certain VA surveillance. It usually represents the reference technique to be improved with novel approaches and algorithms.
Camera calibration	Camera calibration estimates the parameters of a lens and the image sensor of an image or video camera. These parameters can be used to correct lens distortion, measure the size of an object in world units or determine the location of the camera in the scene. These tasks are used in applications, such as machine vision, to detect and measure objects. They are also used in robotics for navigation systems and 3-D scene reconstruction.
Crowd analysis	The capability to analyse and understand the evolution of crowds, detecting anomalous and potentially dangerous situations.
Data fusion	'A process dealing with the association, correlation and combination of data and information from single and multiple sources to achieve refined position and identity estimates and complete and timely assessments of situations and threats, as well as their significance. The process is characterised by continuous refinements of its estimates and assessments, and the evaluation of the need for additional sources, or modification of the process itself, to achieve improved results.'
Data set	A collection of data. In the scenarios considered by this report, a data set contains data acquired from video surveillance sensors. These data are typically constituted by video and possibly audio, but might be enriched with additional information (metadata) about sensors (type, model, manufacturer, position, calibration, etc.), timestamps, alarms or events.
Domotics	A set of hardware and software components specifically developed and installed for home automation.
Egocentric vision	The set of video images acquired from a camera that is mounted directly onto the user (i.e. on the head, body, glasses, etc.).

Term	Definition
End user	The person (or the group of persons) for whom a hardware or software product is designed from the developers, installers and servicers of the product. In this report, the end user is typically the critical infrastructure operator.
Event detection	The ability to find an event of interest in the monitored environment through automatic signal-processing algorithms.
Functionality	A specific feature of an automatic surveillance system (i.e. object tracking, people counting, traffic analysis, etc.).
Ground truth	Refers to the information collected by direct observation of the monitored scene and is generally considered as the reality to be used to compare results from automatic scene-understanding techniques.
Heterogeneity	Refers to the different types of features that can be extracted from the guarded environment and to the different types of sensor that can be used for monitoring purposes.
Human–machine interface	The interface between the operator and the automatic system. This is a software application that presents information to an operator or user about the state of a process, and accepts and implements the operator's control instructions.
Image coordinates	A bi-dimensional coordinate system that is integral with the image and defines the position of a pixel.
Image resolution	Refers to the number of pixels in an image. Resolution is sometimes identified by the width and height of the image as well as by the total number of pixels in the image.
Motion detection	An automatic technique that is able to detect moving parts in a sequence of images. It also represents one of the more simple functionalities of VA.
Pattern recognition	A branch of machine learning that focuses on the recognition of patterns and regularities in data. Pattern-recognition systems are in many cases trained from labelled 'training' data (supervised learning), but when no labelled data are available, other algorithms can be used to discover previously unknown patterns (unsupervised learning).
Performance evaluation	The process of measuring the quality of the results obtained from a signal-processing technique. The adopted metrics have to capture the overall performances of the considered algorithm, including the probabilities of false alarm and misdetections, but also the robustness and the computational complexity.
Pixel	A word invented from 'picture element', it is the basic unit of programmable colour on a computer display or in a computer image.

Term	Definition
Real time	The automatic system is able to acquire and process information from the guarded environment as the scene is naturally evolving over time. Results about scene understanding are available as soon as a certain event happens.
Re-identification	The ability to automatically recognise a certain object of interest in the field of view of a camera as the same object that is (or was) visible within the video stream acquired from a different video sensor.
Robustness	Refers to the fact that the automatic system must be able to work with acceptable performances (see performance evaluation) as well as with the increasing complexity of the scene (i.e. more people, environmental conditions, etc.).
Scene understanding	The ability to find high-level information about the monitored environment and the meaning of the (possibly coordinated) behaviours of the objects of interest.
Standard	An established norm or requirement with regard to technical systems. It is usually a formal document that establishes uniform engineering or technical criteria, methods, processes and practices. In contrast, a custom, convention, company product, corporate standard, etc. that becomes generally accepted and dominant is often called a de facto standard.
Surveillance	The monitoring of the activities of objects of interest (usually people, but also vehicles, etc.) for the purpose of influencing, managing, directing or protecting them. This report mainly considers video surveillance, i.e. surveillance by means of visual information acquired from cameras.
Tracking	The process of correlating the information extracted from processed signals about objects of interest over time. Multilevel tracking refers to the possibility to track subregions of the considered object.
Traffic management	A set of VA functionalities related to traffic, such as vehicle counting and classification, speed estimation, wrong-way detection, etc.
Video analytics	Video content analysis (also video content analysis) is the capability of automatically extracting 'high-level' contextual information from sequences of images.
Video-content retrieval	The ability to automatically use a previously created video index to allow better (faster and more efficient) access to the video itself for the human operator to be able to easily retrieve a specific part of the video.
Video indexing	The ability to automatically find interesting and useful (according to a certain definition) clues within a video.

Term	Definition
World coordinates	A tri-dimensional coordinate system that is independent of the sensors' locations.

Abbreviation	Full text
CCTV	closed circuit television
CI	critical infrastructure
CSV	comma separated values
ERNCIP	European Reference Network for Critical Infrastructure Protection
EU	European Union
FPS	frames per second
GB	gigabyte
HDTV	High-definition television
Hz	Hertz
I-LIDS	image library for intelligent detection systems
IR	infra-red
Kbps	kilo bytes per second
MAVA	morphological analysis on the subdomain of VA
MB	megabyte
Mbps	megabytes per second
NIST	National Institute of Standards and Technology
PETS	performance evaluation of tracking and surveillance
PTZ	pan-tilt zoom
QVGA	quarter video graphics array
VA	video analytics
VGA	video graphics array
VSS	video surveillance system
XML	eXtensible markup language

List of figures

Figure 1. A sample frame from the PETS 2009 data set [107]	15
--	----

List of tables

Table 1. Examples of task-related data set challenges	13
Table 2. Existing data set	22
Table 3. Use case to VA functionality association.....	24

Appendix A. Data sets description

A.1. HDA person data set

References	Size	Link	Citing works (4)
(Nambiar, 2014)		http://vislab.isr.ist.utl.pt/hda-dataset/	

The HDA data set is a multicamera high-resolution image sequence data set for research on high-definition surveillance. 18 cameras (including VGA, HD and full HD resolution) were recorded simultaneously during 30 minutes in a typical indoor office scenario at a busy hour (lunch time) involving more than 80 persons. In the current release (v1.1.), 13 cameras have been fully labelled.

The venue spans three floors of the Institute for Systems and Robotics (ISR-Lisbon) facilities. The following pictures show the placement of the cameras. The 18 recorded cameras are identified with a small red circle. The 13 cameras with a coloured view field have been fully labelled in the current release (v1.1.).



Each frame is labelled with the bounding boxes tightly adjusted to the visible body of the persons, the unique identification of each person, and flag bits indicating whether people are occluded or in a crowd.

- The bounding box is drawn so that it completely and tightly encloses the person.
- If the person is occluded by something (except by image boundaries), the bounding box is drawn by estimating the whole body extent.
- People partially outside the image boundaries have their bounding boxes cropped to image limits. Partially occluded people and people partially outside the image boundaries are marked as 'occluded'.
- A unique ID is associated to each person, e.g. 'person01'. In case of any doubt about a person's identity, the special ID 'personUnk' is used.
- Groups of people that are impossible to label individually are labelled collectively as 'crowd'. People in front of a 'crowd' area are labelled normally.

The following figures show examples of labelled frames: (a) an unoccluded person; (b) two occluded people; (c) a crowd with three people in front.



A.2. WWW crowd data set

References	Size	Link	Citing works (19)
(Shao, 2015)	40 GB	http://www.ee.cuhk.edu.hk/~jshao/WWWCrowdDataset.html	(Shao, 2016)

WWW crowd data set provides 10 000 videos with over 8 million frames from 8 257 diverse scenes, therefore offering a superiorly comprehensive data set for the area of crowd understanding. The abundant sources of these videos also enrich the diversity and completeness.



A.3. MOT benchmark

References	Size	Link	Citing works (38)
(Leal-Taixé, 2015)	2 GB	https://motchallenge.net	(Bewley, 2016) (Xiang, 2015) (Manen, 2016)

MOT offers a framework for the fair evaluation of multiple people tracking algorithms. This framework provides:

- a large collection of data sets, some already in use and some new challenging sequences;
- detections for all the sequences;
- a common evaluation tool providing several measures, from recall to precision to running time;
- an easy way to compare the performance of state-of-the-art tracking methods;
- several challenges with subsets of data for specific tasks such as 3D tracking, surveillance and sports analysis (updates coming soon).

The maintainers rely on the spirit of crowdsourcing and encourage researchers to submit their sequences to their benchmark in order for the quality of multiple object tracking systems can keep increasing and tackling more challenging scenarios.



A.4. ChokePoint data set

References	Size	Link	Citing works (92)
(Wong, 2011)	9 GB	http://arma.sourceforge.net/chokepoint/	(An, 2013) (Kim, 2015) (Gou, 2014)

ChokePoint is designed for experiments in person identification/verification under real-world surveillance conditions using existing technologies. An array of three cameras was placed above several portals (natural choke points in terms of pedestrian traffic) to capture subjects walking through each portal in a natural way (see [example](#)). While a person is walking through a portal, a sequence of face images (i.e. a face set) can be captured. Faces in such sets will have variations in terms of illumination conditions, pose and sharpness, as well as misalignment due to automatic face localisation/detection. Due to the three camera configurations, one of the cameras is likely to capture a face set where a subset of the faces is near frontal.

The data set consists of 25 subjects (19 male and 6 female) in Portal 1, and 29 subjects (23 male and 6 female) in Portal 2. The recording of Portals 1 and 2 are 1 month apart. The data set has a frame rate of 30 FPS and the image resolution is 800 x 600 pixels. In total, the data set consists of 48 video sequences and 64 204 face images. In all sequences, only one subject is presented in the image at a time. The first 100 frames of each sequence are for background modelling where no foreground objects were presented.

Each sequence was named according to the recording conditions (e.g. P2E_S1_C3) where P, S, and C stand for portal, sequence and camera, respectively. E and L indicate subjects either entering or leaving the portal. The numbers indicate the respective portal, sequence and camera label. For example, P2L_S1_C3 indicates that the recording was done in Portal 2, with people leaving the portal, and captured by camera 3 in the first recorded sequence.

To pose a more challenging real-world surveillance problem, two sequences (P2E_S5 and P2L_S5) were recorded with a crowded scenario. In addition to the aforementioned variations, the sequences were presented with continuous occlusion. This phenomenon presents challenges in identity tracking and face verification.

This data set can be applied, but not limited, to the following research areas:

- person re-identification;
- image set matching;
- face quality measurement;
- face clustering;
- 3D face reconstruction;
- pedestrian/face tracking;
- background estimation and subtraction.



A.5. VIRAT

References	Size	Link	Citing works (214)
(Oh, 2011)	40 GB	http://www.viratdata.org/	(Vondrick, 2013) (Burgos-Artizazu, 2012) (Vondrick, 2011)

The data set is designed to be realistic, natural and challenging for video surveillance domains in terms of its resolution, background clutter, diversity in scenes and human activity/event categories than existing action recognition data sets.

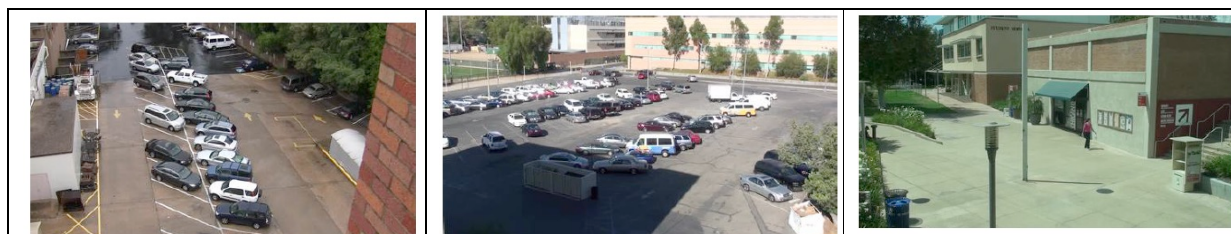
Compared to existing data sets, the data set has the following distinguishing characteristics.

- **Realism and natural scenes:** data were collected in natural scenes showing people performing normal actions in standard contexts, with uncontrolled and cluttered backgrounds. There are frequent incidental movers and background activities. Actions performed by directed actors were minimised; most were actions performed by the general population.
- **Diversity:** data were collected at multiple sites distributed throughout the USA. A variety of camera viewpoints and resolutions were included, and actions are performed by many different people.
- **Quantity:** diverse types of human actions and human–vehicle interactions are included, with a large number of examples (> 30) per action class.
- **Wide range of resolution and frame rates:** many applications, such as video surveillance, operate across a wide range of spatial and temporal resolutions. The data set is designed to capture these ranges, with 2-30 Hz frame rates and 10-200 pixels in person-height. The data set provides the original videos with HD quality as well as the down-sampled versions, both spatially and temporally.
- **Ground and aerial videos:** both ground camera videos and aerial videos are collected and released as part of the VIRAT video data set.

The VIRAT video data set will contain two broad categories of activities (single object and two objects) which involve both humans and vehicles. Details of the included activities and annotation formats may differ per release. Relevant information can be found from each release information.

The main characteristics of this new version are as follows:

- all videos are stationary ground videos;
- large amounts of data; a total ~ 8.5 hours of HD videos;
- a total of 12 event types annotated, from videos from 11 different outdoor scenes;
- includes suggested evaluation metrics and methodologies (data folds for cross-validation, etc.).





A.6. Comprehensive cars (CompCars)

References	Size	Link	Citing works (17)
(Yang, 2015)	N.A.	http://mmlab.ie.cuhk.edu.hk/datasets/comp_cars/index.html	(Liu, 2016) (Sochor, 2016)

The comprehensive cars data set (CompCars) contains data from two scenarios, including web-nature and surveillance-nature images. The web-nature data contain 163 car makes with 1 716 car models. There are a total of 136 726 images capturing the entire cars and 27 618 images capturing the car parts. The full car images are labelled with bounding boxes and viewpoints. Each car model is labelled with five attributes, including maximum speed, displacement, number of doors, number of seats and type of car. The surveillance-nature data contain 50 000 car images captured in the front view. Please refer to our paper for the details.

The data set is well prepared for the following computer vision tasks:

- fine-grained classification;
- attribute prediction;
- car model verification.



A.7. INRIA person data set

References	Size	Link	Citing works (14785)
(Dalal, 2005)	970 MB	http://pascal.inrialpes.fr/data/human/	(Felzenszwalb, 2010) (Wang, 2010) (Felzenszwalb, 2008)

This data set was collected as part of research work on the detection of upright people in images and videos. The data set was divided into two formats: (a) original images with corresponding annotation files; and (b) positive images in normalised 64 x 128 pixel format (as used in the CVPR paper) with original negative images.

The data set contains images from several different sources.

- Images from the GRAZ 01 data set, though annotation files are completely new.
- Images from personal digital image collections taken over a long period. Usually the original positive images were of very high resolution (approx. 2 592 x 1 944 pixels), so we have cropped these images to highlight persons. Many people are bystanders taken from the backgrounds of these input photos, so ideally there is no particular bias in their pose.
- Some images are taken from the web using Google.

Note

- Only upright persons (with person height > 100) are marked in each image.
- Annotations may not be right; in particular portions of annotated bounding boxes may at times be outside or inside the object.

A.8. TUGRAZ ICG long-term pedestrian data set

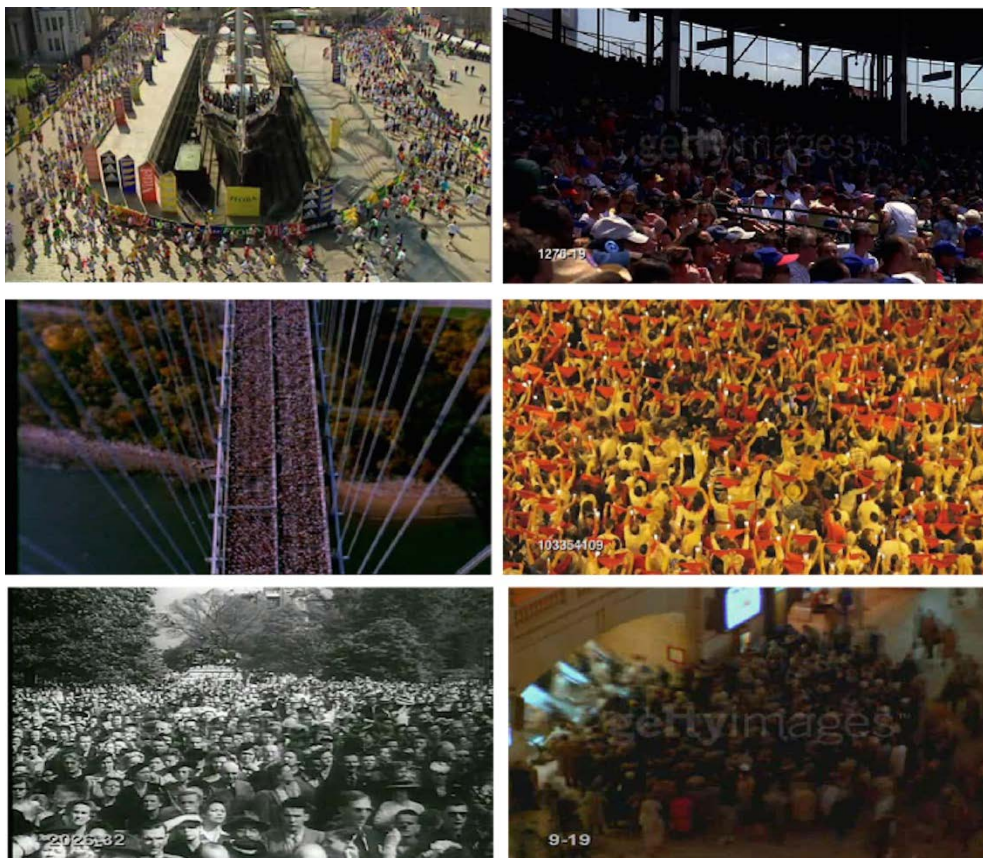
References	Size	Link	Citing works (61)
(Roth, 2009)	1.2 GB	http://lrs.icg.tugraz.at/datasets/longterm/	(Sternig, 2012) (Roth, 2011) (Htike, 2014)

The long-term pedestrian data set consists of images from a stationary camera running 24/7 at about 1 FPS. It is used for adaptive detection and background changes.

A.9. Crowd data set

References	Size	Link	Citing works (3)
(Lim, 2014)	98.25 MB	http://cs-chan.com/project4.htm	

The crowd data sets are obtained through a variety of sources, such as UCF and data-driven crowd data sets. The sequences are diverse, representing a dense crowd in public spaces in various scenarios such as pilgrimages, stations, marathons, rallies and stadiums. In addition, the sequences have different fields of view and resolutions and exhibit a multitude of motion behaviours that cover the obvious and subtle instabilities.



A.10. PEdesTrian attribute (PETA) data set

References	Size	Link	Citing works (20)
(Deng, 2014)	220 MB	http://mmlab.ie.cuhk.edu.hk/projects/PETA.html	(Tian, 2015) (Shi, 2015) (Ye, 2015)

The capability of recognising pedestrian attributes, such as gender and clothing style, at a far distance is of practical interest in far-view video surveillance scenarios where face and body close shots are hardly available.

The PETA data set consists of 19 000 images, with resolutions ranging from 17 x 39 to 169 x 365 pixels. Those 19 000 images include 8 705 persons, each annotated with 61 binary and four multi-class attributes. The detailed composition can be seen in the table below.



A.11. CUHK crowd data set

References	Size	Link	Citing works (30)
(Shao, 2014)	1.5 GB	http://www.ee.cuhk.edu.hk/~jshao/CUHKcrowd_files/cuhk_crowd_dataset.htm	(Yi, 2014) (Wang, 2014) (Shao, 2015)

This crowd data set includes the following.

- 474 video clips from 215 crowded scenes.
- Each clip with the extracted trajectories by gKLT tracker is pre-processed by deleting short trajectories, stationary points and some errors.
- Details of data sets can be found in dataset_info. It contains the video name, length, size and source, video_t0 (the frame for group-detection evaluation), group detection (300 group detection used in our CVPR paper), video_gt (video classes on ground truth) and scene number. (You can also choose any frame to do group detection. The frame list in video_info_t0 is what we use in our CVPR paper.)
- These data can only be used for academic research purposes.
- The copyright of the videos (with watermark) belongs to GettyImages and Pond5.

A.12. GRAZ-02

References	Size	Link	Citing works (378)
(Opelt, 2006)	1.0 GB	http://www.emt.tugraz.at/~pinz/data/GRAZ_02/	(Ramirez, 2010) (Opelt, 2006) (Leordeanu, 2007)

A database for object recognition or object categorisation containing images with objects of high complexity and high intra-class variability on highly cluttered backgrounds.

Three categories (bikes, persons and cars) and one counter-class (bg_graz) contain 365 images with bikes, 311 images with persons, 420 images with cars and 380 images not containing any of these objects.

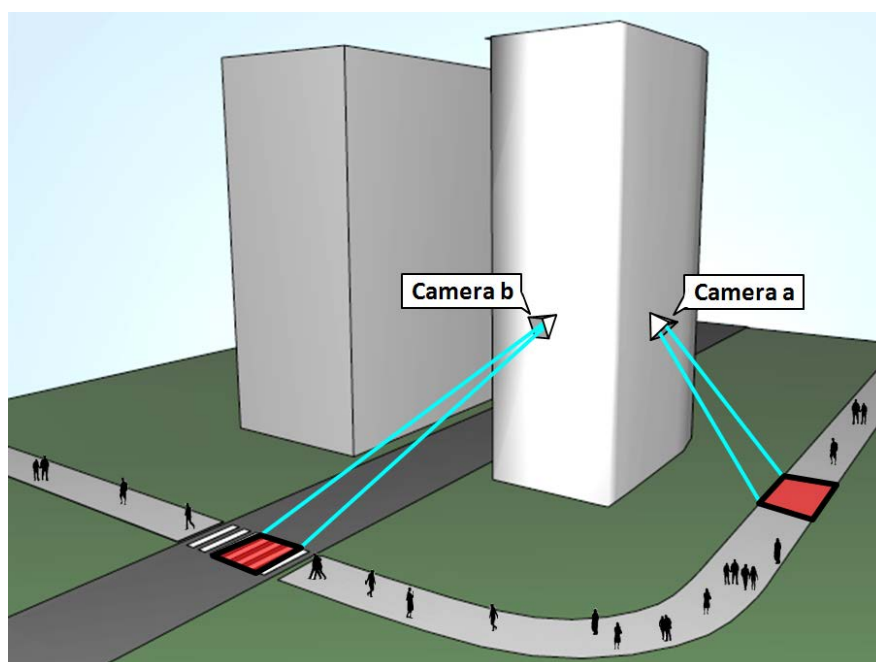
The ground truth for 300 images of each category is available and is given in terms of pixel segmentation masks with values between 0 and 255, where pixels with 0 denoting the object in the image.

A.13. Person Re-ID (PRID) 2011

References	Size	Link	Citing works (174)
(Hirzer, 2011)	1.0 GB	http://lrs.icg.tugraz.at/datasets/prid/index.php	(Hirzer, 2012) (Hirzer, 2012) (Ma, 2013)

This data set was created for the purpose of testing person re-identification approaches. The data set consists of images extracted from multiple-person trajectories recorded from two different static surveillance cameras. Images from these cameras contain a viewpoint change and a stark difference in illumination, background and camera characteristics. Since images are extracted from trajectories, several different poses per person are available in each camera view. 475 person trajectories were recorded from one view and 856 from the other, with 245 persons appearing in both views. Some heavily occluded persons, that is persons with less than five reliable images in each camera view, as well as corrupted images induced by tracking and annotation errors have been filtered out. This results in the following setup.

Camera view A shows 385 persons and camera view B shows 749 persons. The first 200 persons appear in both camera views, i.e. person 0001 of view A corresponds to person 0001 of view B, person 0002 of view A corresponds to person 0002 of view B, and so on. The remaining persons in each camera view (i.e. person 0201 to 0385 in view A and person 0201 to 0749 in view B) complete the gallery set of the corresponding view. Hence, a typical evaluation consists of searching the 200 first persons of one camera view in all persons of the other view. This means that there are two possible evaluation procedures, where either the probe set is drawn from view A and the gallery set is drawn from view B (A to B used in [1]) or vice versa (B to A). See the following figures for more detail.

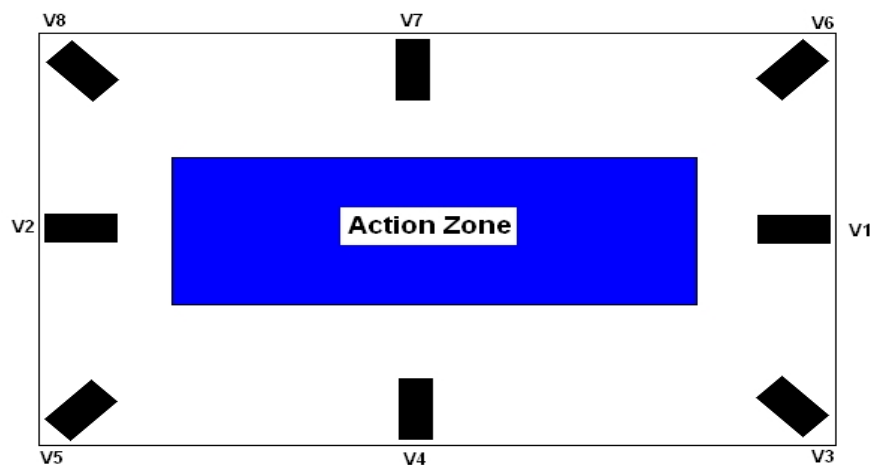


A.14. MuHAVi

References	Size	Link	Citing works (86)
(Singh, 2010)	50 GB	http://dipersec.king.ac.uk/MuHAVi-MAS/	(Chaaroui, 2013) (Cheema, 2011) (Eweiwi, 2011)

A large body of the multicamera human action video data (MuHAVi) using eight cameras have been collected in this data set. There are 17 action classes performed by 14 actors. Videos corresponding to seven actors were processed in order to split the actions and provide the JPG image frames. However, some image frames before and after the actual action are included for the purpose of background subtraction, tracking, etc.

Each actor performs each action several times in the action zone highlighted using white tapes on the scene floor. As actors were amateurs, the leader had to interrupt the actors in some cases and ask them to redo the action for consistency. As shown in Figure 1 and Table 1, we used eight CCTV Schwan cameras located at four sides and four corners of a rectangular platform. Note that these cameras are not necessarily synchronised. Camera calibration information may be included here in the future. Meanwhile, one can use the patterns on the scene floor to calibrate the cameras of interest.





A.15. GRAZ-01

References	Size	Link	Citing works (316)
(Opelt, 2004)	700 MB	http://www.emt.tugraz.at/~pinz/data/GRAZ_01/	(Lazebnik, 2006) (Boiman, 2008) (Opelt, 2006)

The image database contains four kinds of images of two categories: images containing bikes, persons, no bikes and no persons, and objects from both categories.

A.16. Mall data set

References	Size	Link	Citing works (41;19)
(Chen, 2012) (Loy, 2013)	90 MB	http://personal.ie.cuhk.edu.hk/~ccloy/downloads_mall_dataset.html	(Change Loy, 2013) (Zhang, 2015) (Xu, 2014)

The mall data set was collected from a publicly accessible webcam for crowd counting and profiling research.

- Ground truth: over 60 000 pedestrians were labelled in 2 000 video frames. We annotated the data exhaustively by labelling the head position of every pedestrian in all frames.
- Video length: 2 000 frames.
- Frame size: 640 x 480.
- Frame rate: < 2 Hz.



A.17. KTH action

References	Size	Link	Citing works (2341)
(Schuldt, 2014)	1.2 GB	http://www.nada.kth.se/cvap/actions/	(Laptev, 2008) (Wang, 2011) (Jhuang, 2007) (Zhang, 2012)

The current video database containing six types of human actions (walking, jogging, running, boxing, hand waving and hand clapping) was performed several times by 25 subjects in four different scenarios, as illustrated below: outdoors s1, outdoors with scale variation s2, outdoors with different clothes s3, and indoors s4. The database currently contains 2 391 sequences. All sequences were taken over homogeneous backgrounds with a static camera with a 25 FPS frame rate. The sequences were down-sampled to the spatial resolution of 160 x 120 pixels and have a length of 4 seconds on average.

All sequences are stored using AVI file format and are available online (DIVX — compressed version). The uncompressed version is available on demand. There are $25 \times 6 \times 4 = 600$ video files for each combination of 25 subjects, six actions and four scenarios. Each file contains about four subsequences used as a sequence in experiments. The subdivision of each file is into sequences in terms of start_frame and end_frame as well as the list of all sequences is given in a simple file that can be downloaded from the data set website.



A.18. Weizmann actions

References	Size	Link	Citing works (1334)
(Blank, 2005)	N.A.	http://www.wisdom.weizmann.ac.il/~vision/SpaceTimeActions.html	(Jhuang, 2007) (Gorelick, 2007) (Ferrari, 2008) (Zhang, 2012) (Brendel, 2011)

A database of 90 low-resolution (180 x 144, deinterlaced 50 FPS) video sequences showing nine different people, each performing 10 natural actions such as run, walk, skip, jumping-jack (or 'jack' for short), jump forward on two legs (or 'jump'), jump in place on two legs (or 'pjump'), gallop sideways (or 'side'), wave two hands (or 'wave2'), wave one hand (or 'wave1') or bend.

A.19. UT-Interaction

References	Size	Link	Citing works (369)
(Ryoo, 2009)	N.A.	http://cvrc.ece.utexas.edu/SDHA2010/Human_Interaction.html	(Zhang, 2012) (Brendel, 2011) (Ryoo, 2011)

The UT-Interaction data set contains videos of continuous executions of six classes of human-human interactions: shake hands, point, hug, push, kick and punch. Ground truth labels for these interactions are provided, including time intervals and bounding boxes. There is a total of 20 video sequences of around 1 minute each. Each video contains at least one execution per interaction, providing us with eight executions of human activities per video on average. Several participants with more than 15 different clothing conditions appear in the videos, which are taken with the resolution of 720*480, 30 FPS and the height of a person in the video is about 200 pixels.

Videos are divided into two sets. The first set is composed of 10 video sequences taken in a parking lot. The videos of the first set are taken with a slightly different zoom rate, and their backgrounds are mostly static with minimal camera jitter. The second set (i.e. the other 10 sequences) is taken on a lawn on a windy day. The background is moving slightly (e.g. tree moves) and they contain more camera jitters. From sequences 1 to 4 and from 11 to 13, only two interacting persons appear in the scene. From sequences 5 to 8 and from 14 to 17, both interacting persons and pedestrians are present in the scene. In sets 9, 10, 18, 19 and 20, several pairs of interacting persons execute the activities simultaneously. Each set has a different background, scale and illumination.

Types of activities in the interaction challenge

Hand shaking	Hugging	Kicking	Pointing	Punching
				

A.20. i-LIDS

References	Size	Link	Citing works
N.A.	N.A.	http://tna.europarchive.org/20100413151426/sciencelandresearch.homeoffice.gov.uk/hosdb/cctv-imaging-technology/i-lids/index.html	(Bayona, 2010)

This is a publicly available training data set designed to allow system manufacturers to concentrate on their development. It allows manufacturers to self-test their systems' performance, whose score can be submitted to HOSDB to be considered for inclusion into one of the annual evaluations; a privately held evaluation data set that HOSDB uses to benchmark the performance of VA systems in annual evaluations.

The data sets for the event-detection scenarios each contain approximately 24 hours of footage. Each of these data sets are filmed to represent all weather, times of day and scene densities expected within the scenario. The multiple camera-tracking scenario data sets each contain approximately 50 hours of real-world footage.

Each data set consists of two or three camera views referred to as stages and is further segmented into shorter video clips of 30 to 60 minutes. The training data set is further split into individual events. Each data set is supplied with a user guide detailing the library structure, user interface and procedure used to evaluate the systems against the relevant scenario.

The following five scenarios are currently within i-LIDS:

- sterile zone monitoring;
- parked vehicle detection;
- abandoned baggage detection;
- doorway surveillance;
- multiple camera-tracking scenario.

These scenarios are made up of three data sets each.

A.21. NIST digital video 1

References	Size	Link	Citing works
N.A.	N.A.	https://catalog.data.gov/dataset/nist-digital-video-1-nist-special-database-26 http://www.nist.gov/srd/nistsd26.cfm	N.A.

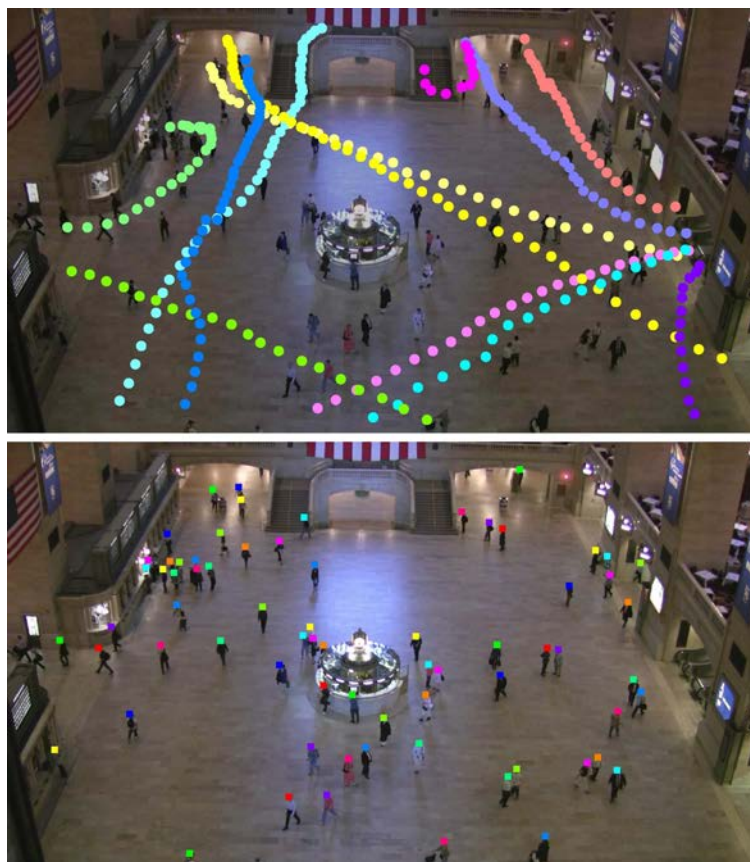
NIST digital video 1 is a public domain collection of digital video created to encourage more researchers to address real-world problems and support the scientific comparison of solutions of digital video search, retrieval and display. This collection consists of eight videos, totalling over 2 hours in length selected from NIST's public domain archive of marketing, technical and educational material. The characteristics of these videos include, but are not limited to, different levels of motion (static to fast-moving objects), close-up figures (talking heads, moving arms and moving hands), outdoor shots (laboratory, auditorium and conference room environments) and various levels and quality of audio.

In addition to the base data, pre- or post-production transcripts are included as reference data. It is our intention to gather feedback on the use of this collection, the need for additional base data and further requirements for reference data (or 'truth').

A.22. Pedestrian walking path data set (Grand Central data set)

References	Size	Link	Citing works (11)
(Yi, 2015)	2.69 GB	http://www.ee.cuhk.edu.hk/~syi/	(Yi, 2015) (Yi, Pedestrian Travel Time Estimation in Crowded Scenes)

A 1-hour surveillance video, together with the exact walking paths of all 12 684 pedestrians, is included in this data set.



A.23. PETS 2007

References	Size	Link	Citing works
N.A.	7.2 GB	http://www.cvg.reading.ac.uk/PETS2007/data.html	(Bayona, 2010) (Arsié, 2008) (Leach, 2014)

The data sets are multi-sensor sequences containing the following three scenarios, with increasing scene complexity: loitering, attended luggage removal (theft) and unattended luggage.



A.24. PETS 2006

References	Size	Link	Citing works
N.A.	7.2 GB	http://www.cvg.reading.ac.uk/PETS2006/data.html	(Hattori, 2015) (Godec, 2010)

The data sets are multi-sensor sequences containing left-luggage scenarios with increasing scene complexity.



A.25. PETS 2009

References	Size	Link	Citing works
(Shahrokni, 2009) (Ferryman, 2009)	10 GB	http://www.cvg.reading.ac.uk/PETS2009/data.html	(Yang, 2013) (Shu, 2013) (Wang, 2014) (Milan, 2015)

The data sets are multi-sensor sequences containing different crowd activities.



A.26. PETS 2015

References	Size	Link	Citing works
(Li, 2015)	N.A.	http://www.cvg.reading.ac.uk/PETS2015/a.html	(Bastani, 2015) (Nawaz, 2015)

The PETS 2015 challenge uses two data sets that are ARENA and P5 data sets. The ARENA data set was used in its full form in the [PETS 2014 challenge](#); here we use a more restricted set of ARENA data sets by including a few scenarios that are more relevant to the PETS 2015 challenge. The selected scenarios from the ARENA and P5 data sets are grouped into 'Normal', 'Warning' and 'Alarm' categories. 'Normal' alludes to activities that do not pose any threat, 'Warning' refers to abnormal activities that may potentially develop into a threat, and 'Alarm' refers to activities that cause a threat in the scene and hence require immediate action. Below is a description of the two data sets. The ARENA data set contains sequences with different activities around a parked vehicle in a parking lot, while the P5 data set contains sequences with different activities staged at the OKG nuclear plant outside Oskarshamn, Sweden.

A.27. UCF aerial action data set

References	Size	Link	Citing works
N.A.	N.A.	http://crcv.ucf.edu/data/UCF_Aerial_Action.php	N.A

This data set features video sequences that were obtained using an R/C-controlled blimp equipped with an HD camera mounted on a gimbal. The collection represents a diverse pool of actions featured at different heights and aerial viewpoints. Multiple instances of each action were recorded at different flying altitudes, ranging from 400-450 feet, and were performed by different actors.

The actions collected in this data set include walking, running, digging, picking up an object, kicking, opening a car door, closing a car door, opening a car trunk and closing a car trunk.

A.28. Mini-drone video data set (DronesProtect data set)

References	Size	Link	Citing works (6)
(Bonetto, 2015)	N.A.	http://mmspg.epfl.ch/mini-drone	(Ruchaud, 2015)

The created data set consists of 38 different contents captured in full HD resolution, with a duration of 16 to 24 seconds each and shot with the mini-drone Phantom 2 Vision+ in a parking lot. The data set contents can be clustered into three categories: normal, suspicious and illicit behaviours. Normal content depicts people walking, getting in their cars and parking their vehicles. In suspicious content, nothing wrong happens a priori but people act in a questionable way. Contents with illicit behaviours show people mis-parking their vehicles, stealing items and cars, or fighting. All participants read and signed a consent form, stating that they agree to appear with their vehicles in the video.



Europe Direct is a service to help you find answers to your questions about the European Union
Free phone number (*): 00 800 6 7 8 9 10 11
(*) Certain mobile telephone operators do not allow access to 00 800 numbers or these calls may be billed.

A great deal of additional information on the European Union is available on the internet.
It can be accessed through the Europa server <http://europa.eu>

How to obtain EU publications

Our publications are available from EU Bookshop (<http://bookshop.europa.eu>),
where you can place an order with the sales agent of your choice.

The Publications Office has a worldwide network of sales agents.
You can obtain their contact details by sending a fax to +352 2929-42758.

JRC mission

As the Commission's in-house science service, the Joint Research Centre's mission is to provide EU policies with independent, evidence-based scientific and technical support throughout the whole policy cycle.

Working in close cooperation with policy directorates-general, the JRC addresses key societal challenges while stimulating innovation through developing new methods, tools and standards, and sharing its know-how with the Member States, the scientific community and international partners.

*Serving society
Stimulating innovation
Supporting legislation*

