Pradeep K. Atrey
Mohan S. Kankanhalli
Andrea Cavallaro *Editors*

Intelligent Multimedia Surveillance

Current Trends and Research

Springer

Intelligent Multimedia Surveillance

Pradeep K. Atrey • Mohan S. Kankanhalli •
Andrea Cavallaro
Editors

Intelligent Multimedia Surveillance

Current Trends and Research

Springer

Editors
Pradeep K. Atrey
Dept. of Applied Computer Science
University of Winnipeg
Winnipeg, Canada

Mohan S. Kankanhalli
Dept. of Computer Science
National University of Singapore
Singapore, Singapore

Andrea Cavallaro
School of Electronic Engineering
and Computer Science
Queen Mary University of London
London, UK

ISBN 978-3-642-41511-1 ISBN 978-3-642-41512-8 (eBook)
DOI 10.1007/978-3-642-41512-8
Springer Heidelberg New York Dordrecht London

Library of Congress Control Number: 2013956037

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

Preface

Intelligent multimedia surveillance concerns the analysis of multiple sensing inputs including video and audio streams, radio-frequency identification (RFID) and depth data. These data are processed for the automated detection and tracking of people, vehicles and other objects. The goal is to locate moving targets, to understand their behavior and to detect suspicious or abnormal activities for crime prevention. Despite numerous benefits of this technology, there is a natural societal apprehension regarding the use of intelligent multimedia surveillance to infringe privacy. An important challenge in this research area is therefore to balance two contradictory goals: public safety and privacy. This book presents in nine chapters recent findings in the field of intelligent multimedia surveillance and covers various aspects such as privacy, surveillance as a service, crowded scene understanding, performance evaluation, and active vision.

In the chapter "Intelligent Video Surveillance as a Service", Prati et al. present a paradigm called *VSaaS* that considers video surveillance technology as a service. Distributed cloud resources are used to handle the storage and processing of large amounts of video data. The authors also describe a case study on the integration of computer vision algorithms in a VSaaS platform.

Current solutions, for video analysis of crowds are discussed by Thida et al. in the chapter "A Literature Review on Video Analytics of Crowded Scenes". A systematic comparison and critical review of existing methods and technologies for the automated analysis of complex and crowded scenes are presented. The authors divide the literature into two broad categories, namely the macroscopic and microscopic modeling approaches. The merits and weaknesses of these approaches are discussed and a recommendation for how existing methods can be improved is finally provided.

The next three chapters cover privacy issues in intelligent multimedia surveillance. In the chapter "Privacy and Security in Video Surveillance", Winkler and Rinner motivate the need for the integration of security and privacy features in video surveillance systems. The authors first present a comprehensive review of the state of the art and then describe a prototype system, the TrustCAM, where a dedicated hardware security module is integrated in a camera system to achieve a high-level

of security. A summary of open research issues and an outlook to future trends conclude the chapter. Privacy is also addressed by Qureshi in the chapter "Object Video Streams: A Framework for Preserving Privacy in Video Surveillance". The author introduces a framework that decomposes raw video footage into background and one or more object-video streams. The framework is used to preserve privacy (i.e., identity of people) in the video by representing object-video streams as blobs, by coding foreground objects in different colors, and by rendering the scene partially (i.e., revealing the identities of only some individuals). The approach is evaluated in a virtual train station environment and on real video footage. In the chapter "Surveillance Privacy Protection", Gulzar et al. further investigate privacy and present an evaluation of various aspects, such as what types of protection measures are being implemented in surveillance systems, how information is being used, and what rights individuals have over them. In addition, the authors also emphasize the importance of tools, data sets and databases that are being developed to give protection to surveillance privacy.

Next, in the chapter "RFID Localization Improved by Motion Segmentation in Multimedia Surveillance Systems", Ljubojević et al. discuss the use of passive RFID technology for localization of objects indoors. The authors describe the use of motion segmentation algorithms on the region of interest extracted using the information collected from RFID, which allows the reduction of the position estimation error and variance compared to the conventional RFID-based position estimation methods. A related topic is covered by Mahapatra and Saini in the chapter "A Particle Filter Framework for Object Tracking Using Visual-Saliency Information". The authors use neurobiology-saliency for object detection and tracking using particle filters. In this work, low-level features such as color, luminance and edge information along with motion cues are used to track a person under varying lighting conditions.

These concepts are extended by Kumar et al. in the chapter "Multiresolution Depth Map Estimation in PTZ Camera Network". In this chapter, the authors propose an active stereo vision system composed of two pan-tilt-zoom (PTZ) cameras. The proposed system is used for estimating the multiresolution depth map for a large and complex scene.

Finally, in the chapter "Performance Evaluation in Video-Surveillance Systems: The EventVideo Project Evaluation Protocols", SanMiguel et al. emphasize the need to automate the performance evaluation process for video surveillance systems. The authors describe the evaluation protocols for various analysis stages such as video object segmentation, people detection, video object tracking and event recognition, within the scope of the EventVideo project.

The editors of this book extend their sincere thanks to the authors and reviewers of the chapters and very much appreciate their contribution and support, without which this book would not have been possible. Prof. Pradeep K. Atrey has been supported by funding from the Natural Sciences and Engineering Research Council of Canada (NSERC) (Discovery Grant 408206). Prof. Andrea Cavallaro has been supported in part by funding from the European Union (Project CENTAUR, 324359, FP7-PEOPLE-2012-IAPP). Prof. Mohan Kankanhalli's work has been carried out

at the SeSaMe Centre which is supported by the Singapore NRF under its IRC@SG Funding Initiative and administered by the IDMPO.

Winnipeg, Canada — Pradeep K. Atrey
Singapore, Singapore — Mohan S. Kankanhalli
London, UK — Andrea Cavallaro

Contents

Intelligent Video Surveillance as a Service

Andrea Prati, Roberto Vezzani, Michele Fornaciari, and Rita Cucchiara

Abstract Nowadays, intelligent video surveillance has become an essential tool of the greatest importance for several security-related applications. With the growth of installed cameras and the increasing complexity of required algorithms, in-house self-contained video surveillance systems become a chimera for most institutions and (small) companies. The paradigm of Video Surveillance as a Service (VSaaS) helps distributing not only storage space in the cloud (necessary for handling large amounts of video data), but also infrastructures and computational power. This chapter will briefly introduce the motivations and the main characteristics of a VSaaS system, providing a case study where research-lab computer vision algorithms are integrated in a VSaaS platform. The lessons learnt and some future directions on this topic will be also highlighted.

1 Introduction

Video surveillance is an important application field of computer engineering, involving multidisciplinary studies, ranging from sensors and storage systems, display interfaces and networks to algorithms and software development.

A. Prati (✉)
DPDCE, University IUAV of Venice, Santa Croce 1957, 30135 Venice, Italy
e-mail: andrea.prati@iuav.it

R. Vezzani · M. Fornaciari · R. Cucchiara
DIEF, University of Modena and Reggio Emilia, Via Vignolese 905, 41125 Modena, Italy

R. Vezzani
e-mail: roberto.vezzani@unimore.it

M. Fornaciari
e-mail: michele.fornaciari@unimore.it

R. Cucchiara
e-mail: rita.cucchiara@unimore.it

A. Prati · R. Vezzani · M. Fornaciari · R. Cucchiara
SOFTECH-ICT, Via Vignolese 905, 41125 Modena, Italy

P.K. Atrey et al. (eds.), *Intelligent Multimedia Surveillance*,
DOI 10.1007/978-3-642-41512-8_1,

Since the first video surveillance installations in the mid-60s, only the hardware architecture has shown a monotonic growth, leading from research to the market. In 1969, the first system was installed at the Municipality Building in New York, while in 1993 the first digital system was installed at the World Trade Center following the arrival, in 1985, of the first Digital Video Recorder (DVR), which reached the market to create digital CCTV systems. The scenario has changed into networks of camera systems, translating simple platforms for building automation security to very large implementations such as those of the Chicago Virtual Shield (2006) from IBM, initially involving 3000 connected cameras in a single network, going through the 2-million-camera system of All-Seeing-Eye in Shenzen China (2009) [12], up to the new IoT (Internet of Things) video-surveillance project for the Chongqing Municipality, which comprises millions of cameras and other RFID, infrared, smoke detector sensors.

Conversely, software components are still unsuited to current needs, and from video processing initially used for coding and data transfer in the middle of 80s, in the 90s commercial systems started to include simple software modules for motion detection. Several industries (often in collaboration with research labs) have put great efforts in building real working video surveillance software systems (such as IBM [14], Object Video [8], or Sarnoff Corporation [19], just to mention a few). In designing and developing these systems, the balance between advanced, lab-tested features and stable yet simple ones has often given priority to the latter. In fact, customers and above all security officers are typically disappointed by false alarms and missed detections [18]: they would actually enjoy an automatic system, even with limited functions only.

The term "video analytics" indicates software tools that provide automatic video processing, computer vision and pattern recognition modules to extract knowledge from the observed scene. More recently, a new paradigm has gained attention in the field of video surveillance: this paradigm represents a "fourth generation" of Intelligent Video Surveillance (IVS) [15] systems, which configures "Video Surveillance as a Service" (*VSaaS*).

The VSaaS market and the commercial solutions offered have grown significantly in the last years. It has been estimated that since 2010 about 300,000 cameras have been connected to VSaaS systems worldwide, with an increase of about 100,000 in 2010 only [10]. This market is estimated to produce an average income per year of around 100 Millions of US dollars [10]. Based on these figures, several companies are now offering VSaaS solutions. The list in Table 1 is just a partial one, but gives an idea of the high interest on this topic.

The differences among the various VSaaS available on the market can be summarized in the following points:

- *Camera support*: some systems only support proprietary cameras, some others also allow third-party cameras with "open connectivity" to be used;
- *Degree of installation complexity*: ease of installation varies from system to system;

Table 1 List of some commercial VSaaS solutions

Name	Home page
Archerfish	http://www.myarcherfish.com/
ByRemote	http://www.byremote.net/
Alarm.com	https://www.alarm.com/video/
Axis AVHS (Axis Video Hosting System)	http://www.axis.com/products/avhs
Brivo	http://www.brivo.com/
CameraManager	http://www.cameramanager.com
Connexed	http://www.connexed.com/
Dropcam	https://www.dropcam.com/
DvTel NVMS (Network Video Management Software)	http://www.dvtel.com/products/isoc/network-video-management-system
Envysion	http://www.envysion.com/

- *Quality of the Video Management System (VMS)*: most systems come with a live and recorded video manager, but only a few of them also provide video analytics functions;
- *Service scalability*: some products are theoretically scalable on any number of cameras, while some others have limitations.

This chapter presents the integration of a commercial video surveillance system with a cloud-based architecture, an open-source video management system (VMS) and stable research-lab-built computer vision algorithms.

As such, the novelty of this chapter does not rely on the algorithms (which are rather simple, well-assessed and robust, and already published [5]), but on the architectural VSaaS viewpoint. The "as-a-service" paradigm is now very widespread and refers to three levels: "Application as a service" (*AaaS*), "Platform as a service" (*PaaS*) and "Infrastructure as a service" (*IaaS*). The VSaaS solution proposed covers all these three levels, since it proposes, mainly to public administration bodies (municipalities, local police stations, etc.), to remotely move every piece composing a video surveillance systems, except the cameras: using IP connections, video feeds are transferred via a high-bandwidth fiber channel to a data center where sophisticated video surveillance algorithms can be used. Nowadays, talking about "a sa-service" architecture also means talking about the *cloud*. Figure 1 shows the cloud-based architecture for VSaaS, where the most innovative part is the central one: exploiting remote server capabilities as a common IaaS is now common practice in many applications (mailing, document storage, etc.), on the other hand, the concept of PaaS extends the horizon to common services, where interaction between different content providers and content users must be tight.

The main motivations and advantages of a VSaaS system are highlighted in Sect. 2: following them, we developed a prototypical VSaaS implementation within the scope of a project called ViSERAs. The implemented architecture is described in Sect. 3, which describes in detail both the Video Management System (Sect. 3.1)

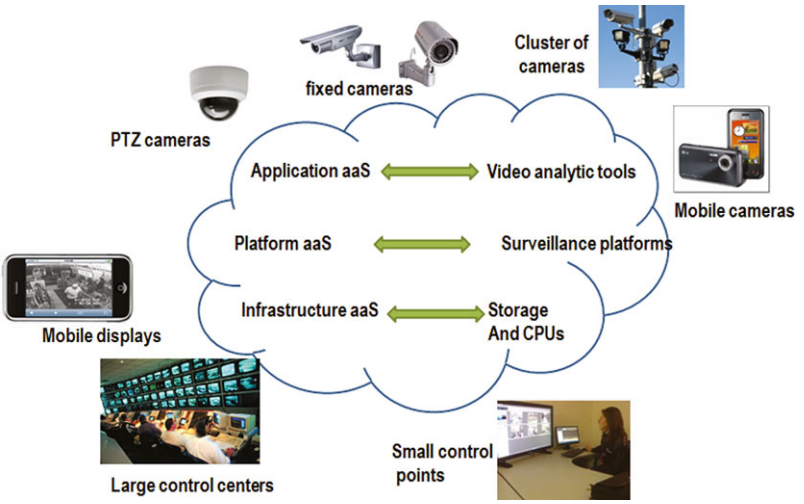


Fig. 1 Sketch of a cloud-based architecture for VSaaS

and the Commercial Video Analytics modules (Sects. 3.2, 3.3 and 3.4). To improve the platform and provide an extensive service, a plethora of surveillance plug-ins for the Video Analytics framework has been considered. Some of these plug-ins are described in Sect. 4, which also provides some visual examples of the Shadow removal plug-in.

2 VSaaS: The Motivations

If video surveillance system cameras are simply watched over monitors by human operators ("passive" video surveillance), the challenges are somehow limited to hardware installation, cable deployment and people hiring. Whenever video surveillance goes "active" (or, more properly, "intelligent"), in addition to these challenges the scalability of computational resources becomes a tough issue to be dealt with.

At the same time, telecommunication and computer engineering companies make their powerful data centers available to customers for storing and managing large video repositories, analyzing videos to detect interesting events, logging them and alerting operators. As a consequence, the huge computational power provided as a service (which is typically not affordable by individual customers) brings several advantages for the customers: (i) reduced costs in terms of hardware and software; (ii) layered services which can be adaptively activated/deactivated based on actual needs and costs; (iii) almost unlimited computational power which not only improves the system responsiveness, but also allows the implementation of services (e.g., in terms of advanced video analysis algorithms) which would not be feasible using their own hardware resources.

3 Architecture Description

Based on the foregoing premises, in 2011 we started a project called VISERAS ("VIdeo-Surveillance in Emilia Romagna as A Service"). Despite its limited geographical extension (Region Emilia-Romagna in Italy) and its short-term (nine months), VISERAS is a very good example of a timely industrial research project which combines previous experiences in the field of surveillance systems with the new trends of computation distribution within the cloud.

The project has aimed at defining a good architecture for providing not just video-surveillance systems but rather services applied to data remotely acquired and remotely stored. The availability of a cloud architecture makes new solutions available which can spread surveillance capabilities also to public bodies (e.g., small villages) for which the installation of a complete multi-camera, even distributed but proprietary system would not be affordable.

3.1 Video Management System

The VISERAS Video Management system comes from the open-source Video Management System (VMS) called ZoneMinder,[1] a video-management system for network cameras organized so as to operate on a single server. The VNS is responsible for handling all digital videos from cameras or encoders. This handling process includes initial registration and configuration of the video devices, receiving video streams from the same devices, recording this video, proxying live video streams to clients, and streaming recorded video to clients. During the project, the basic, publicly-available release of Zoneminder was amended and extended allowing management of multiple ZM instances in order to create a system capable of controlling hundreds of distributed cameras and clients.

A key issue related to the VMS chosen for our project was its scalability, which had to take the following aspects into account:

- greatest flexibility to adapt the video surveillance system to heterogeneous scenarios;
- use of tools from consolidated data centers to maximize data security and simplify regular maintenance;
- minimize any architectural changes made to the ZM software in order to avoid possible bugs, speed up the scalability process and have the system more easily aligned with the new releases of the official ZM source code.

For these reasons, we chose to implement an architecture based on virtual machines. Each ZM instance will run on a virtual machine and will refer to a single entity.

[1] http://www.zoneminder.com/

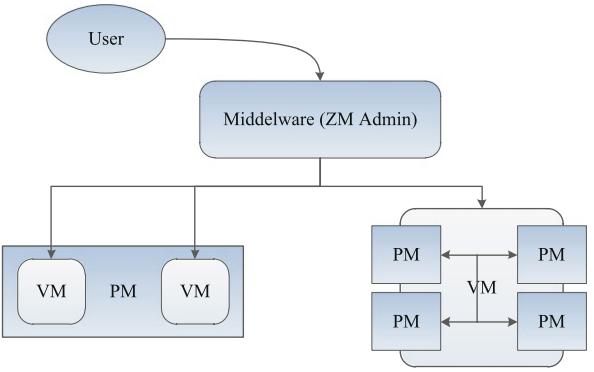


Fig. 2 Illustration of middleware introduction in the ZoneMinder data flow (VM = Virtual Machine, PM = Physical Machine)

In addition to the changes made to the basic ZM module, we must also consider a further complication, given our target "users" (mainly public entities, such as local police stations and municipalities in small towns). This requires the creation of a middleware level to take into account *federated users*. A user is defined "federated" when he/she is granted access to multiple ZoneMinder instances potentially dislocated on different virtual machines and geographically distributed. Figure 2 shows a common case, showing that a single physical machine (PM) can host multiple virtual machines (VM), and that a single VM can be hosted on multiple PMs.

The middleware role is to make this access clear and allow the federated user (for instance, local police authorities monitoring several neighboring cities) to be unaware of the underlying camera distribution on different instances and machines.

The general access architecture for federated users is shown in Fig. 3. This figure clearly shows that the middleware is mainly used for three tasks: (a) clearly

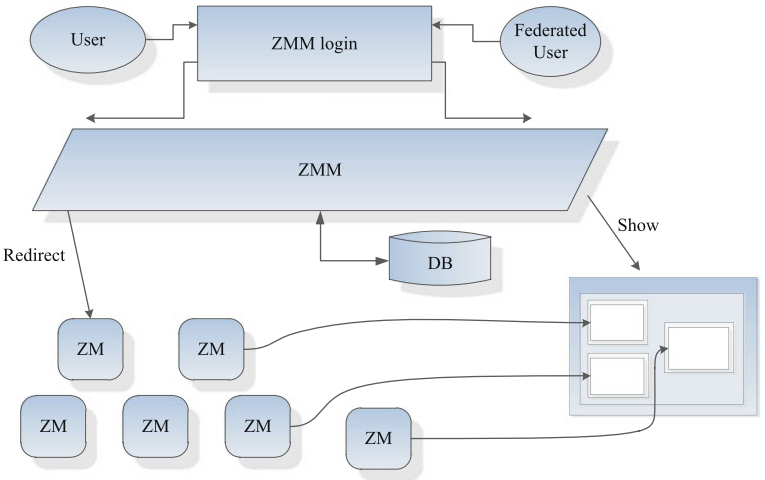


Fig. 3 Layout of access management with federated users (ZM = single instance of ZoneMinder, ZMM = ZoneMinder Middleware)

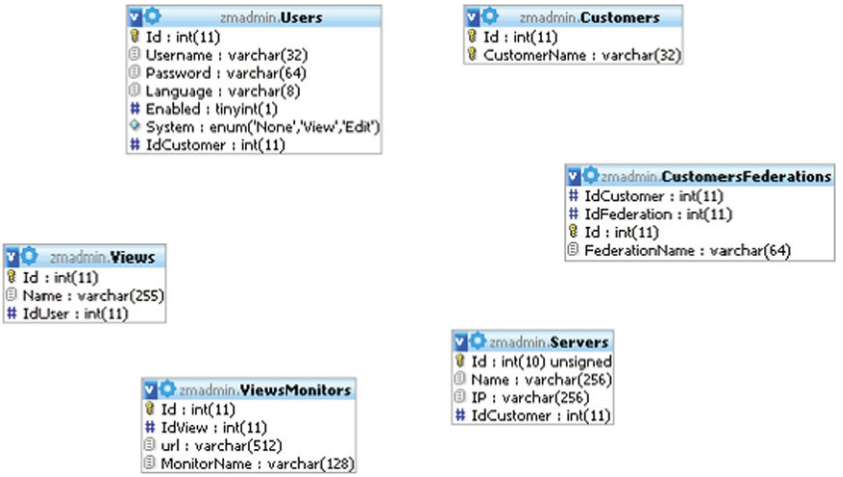


Fig. 4 Layout of the database used in ZoneMinder Middleware

redirecting the request from the users to the correct ZoneMinder instance; (b) once the required result is provided by ZoneMinder, feed it back through the middleware and showing it to the user; (c) maintaining a database with information about users and federated users, as well as a mapping to cameras, virtual machines and physical machines. The layout of such a database is illustrated in Fig. 4.

Some snapshots of the resulting interface for added (federated) functions are shown in Fig. 5.

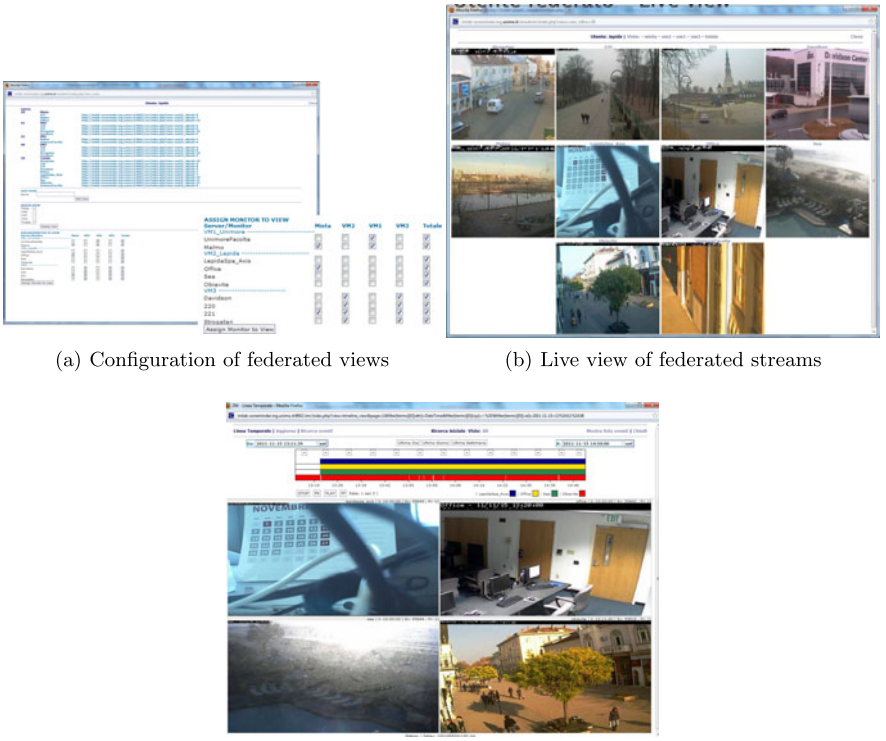
3.2 Commercial Video Analytics Engine

As mentioned in the Introduction, our project relies on a commercial video surveillance system, specifically the IBM Smart Vision Suite (SVS) [9]. Our choice has been dictated by the completeness of this solution, which embodies a complex architecture and allows plugging VMS and external video analytics algorithms with the existing algorithms developed by IBM.

The IBM SVS delivers two primary functions:

1. the ability to observe digitally encoded videos and detect events happening in the video in near-realtime;
2. the ability to index and store detected events to support search and correlation after occurrence.

These functions are intended to be applied to live and recorded digital videos provided by various cameras, encoders, digital video recorders (DVRs), network video recorders (NVRs) and VMSs.



(a) Configuration of federated views

(b) Live view of federated streams

(c) Timeline of federated cameras

Fig. 5 Snapshots of resulting variant of ZoneMinder VMS

This integrated solution can be divided into three distinct subsystems: VMS, Video Analytics, Metadata Engine/Interfaces. Each subsystem has distinct functions and interfaces as briefly described in the following and shown in Fig. 6.

Our VMS solution has been described in Sect. 3.1. It can be highlighted that, with reference to Fig. 6, (1) represents the Smart Surveillance Engine (SSE) using a DirectShow filter (described in Sect. 3.3) to access live streaming video while (3) represents web clients using an embedded ActiveX control (or applet) to access both live and recorded video streams from the VMS Server.

3.3 Video Analytics (SSE)

The IBM Video Analytics engine acts as a VMS client, accessing streams of live, proxied video from the VMS. While the SSE could access these streams directly from certain cameras, using the video proxy via VMS offers two advantages:

- a single video interface regardless of the variety of cameras deployed,

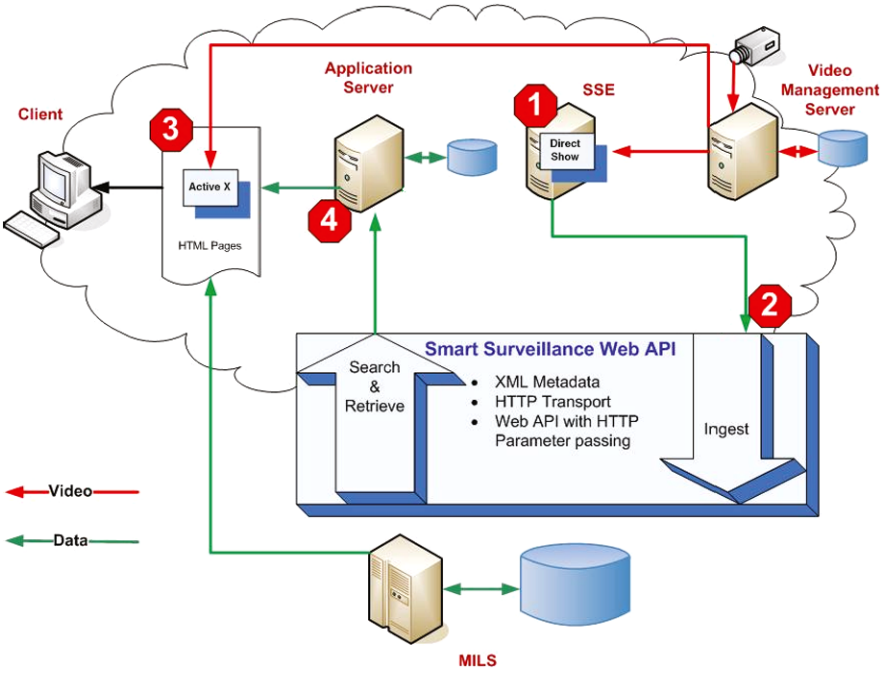


Fig. 6 MILS architecture

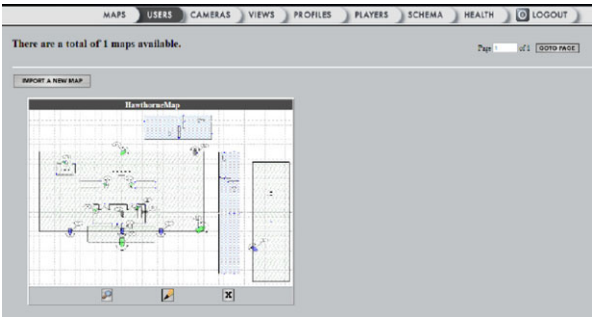
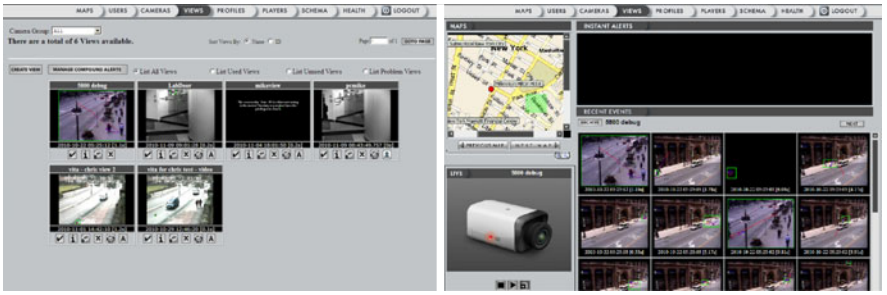


Fig. 7 Examples of operation of SSE: map setting

- mitigation of potential problems in the event that multiple clients saturate the network bandwidth or the camera streaming capacity when simultaneously accessing live videos.

Among the different system functions is the creation of an area map (Fig. 7) showing where the cameras are located, as well as the definition of views containing a subset of the available cameras, associating alarms with specific views (Fig. 8). The alerts are detected and visually highlighted; they are stored in a structured DB and can be searched by the user based on several features (visual and non-visual). Figure 9 shows some snapshots of alert management.



(a) (b)

Fig. 8 Examples of operation of SSE: view setting

(a) (b)

Fig. 9 Examples of operation of SSE: alerts

The SSE uses DirectShow (1) in order to view live video streams from the VMS server. To allow DirectShow Filters (DSF) to be used by the SSE, these must meet the following minimum requirements:

- The video must be requested using a single URL which meets all the necessary parameters for the video provider to start the desired stream. The SSE does not support user interaction so there is no way to enter any parameters interactively.
- The DSF parameters must be *majortype* = "MEDIATYPE_Video".
- The DSF must be requested via a specific protocol.

In order to link ZoneMinder video streams with IBM SSE, we have implemented a DirectShow source filter satisfying these requirements. The filter retrieves MJPEG video streams from ZoneMinder, which are accessed with the parameters embedded in the URL request. This is recalled when the protocol specified in the URL is ILP (ImageLab Protocol), which has been developed for this specific purpose.

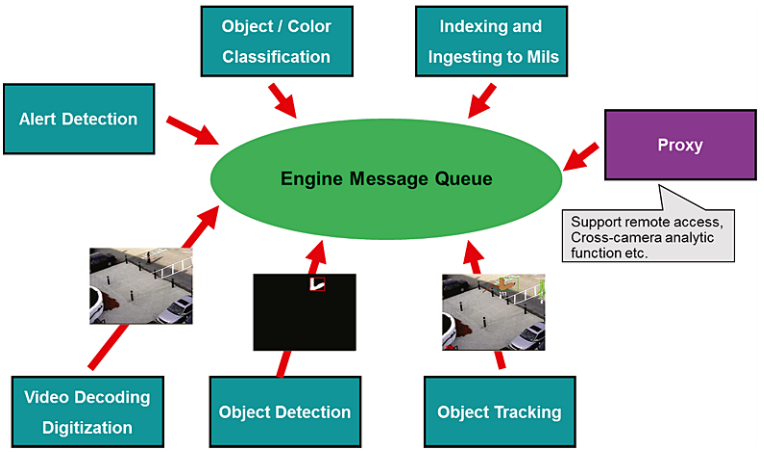


Fig. 10 SSE engine message queue

3.4 Metadata Engine (MILS)

The IBM Metadata Engine is the core of the Smart Vision Suite. It provides a massively scalable data repository that is optimized for rapid ingestion of video-associated metadata with real-time indexing. This optimization supports extremely efficient searches of the aggregated metadata from hundreds or thousands of video sources.

The SSE posts metadata describing events observed on videos to the MILS using a pre-standard MILS interface (2). The MILS Web UI has just two instances where videos are accessed from a video provider (3): one for live videos and one for recorded videos. Each of these instances may be loaded dynamically to specific frames within the broader MILS Web UI, which accesses relevant videos based on event metadata (4).

The integration of a video provider involves implementing these interfaces in the specified HTML files, as we did for the ZoneMinder integration. In order to access both live and recorded video streams from the VMS Server, we used the "ActiveX Axis Media Control". Starting from the information embedded in the URL as specified in the configuration files, we access the ZoneMinder database to get the correct video sequence and its URL. The request is completed forwarding the original request on the video sequence URL.

4 Plugging-in New Algorithms

The above described framework analyzes the video stream in order to extract meaningful information. The video analysis task is divided into several sub-tasks, each of which is performed by a specific component, called *plug-in*. This sequence of plug-ins is implemented using an Engine Message Queue (Fig. 10): each plug-in is

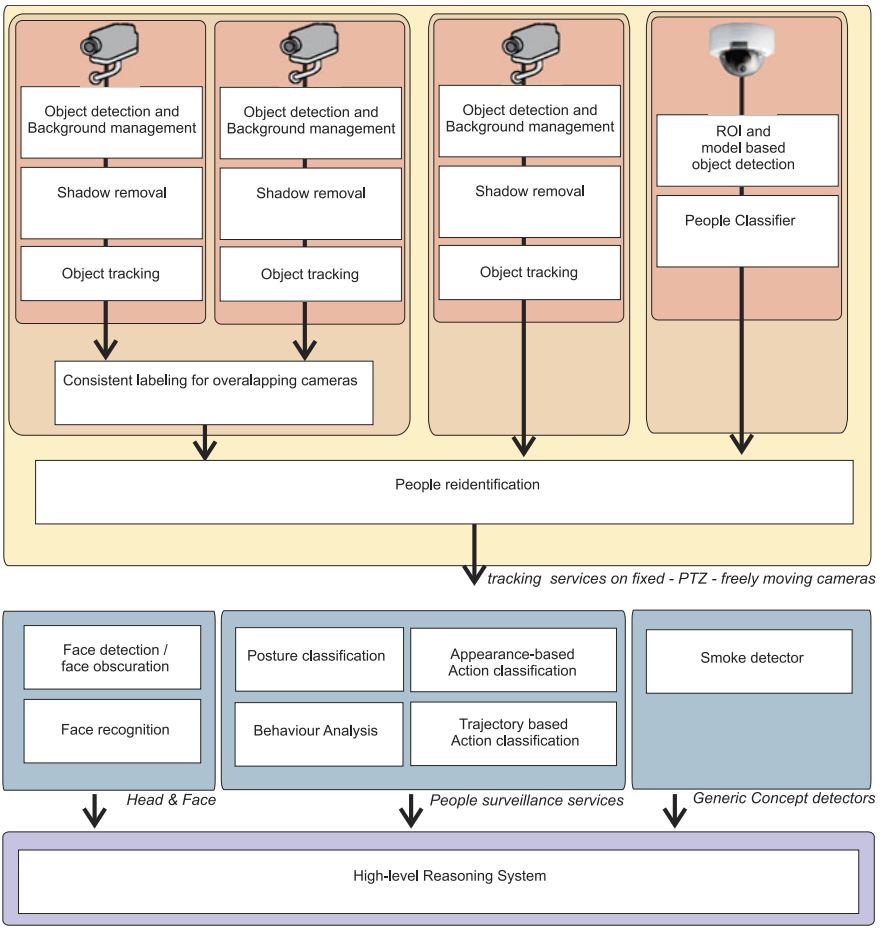


Fig. 11 Architecture and dependencies of the proposed surveillance plug-ins

attached to the queue where intermediate results are stored and then processed in a predefined order.

A set of low and high-level surveillance tasks have been studied [16] and are now available as modules. The availability of a plethora of plugins which can be combined to create a specific and customized surveillance application is one of the major advantages of the *VSaaS* framework. Among others, face detection and recognition, posture and action classification, crowd detection and analysis modules are available. A schema of the overall architecture and dependencies of the proposed surveillance plug-ins is depicted in Fig. 11. The following list contains a brief description of the selected plug-ins:

- *Shadow removal*: the algorithm reported in [5] has been implemented as a SSE plug-in and it removes the shadows from the foreground objects, feeding the

tracking step with the correct input. A detailed description of this plug-in together with some visual examples are reported in Sect. 4.1.

- *Fast background initialization*: a new and fast technique for background estimation from cluttered image sequences. Most background initialization approaches collect a number of initial frames and then require a slow estimation step that introduces a delay. Conversely, the proposed technique redistributes the computational load among all the frames by means of a patch-by-patch preprocessing, which makes the overall algorithm more suitable for real-time applications [2].
- *People classifier*: the HoG-based people classifier [7] is implemented as a service to detect people among the set of tracks, whenever they appear in the scene.
- *Face detector*: Two different face detectors are implemented in the framework: the well known Viola-Jones of the OpenCV library and the face detection library by Kienzle et al. [11].
- *Posture classifier*: the frame-by-frame posture of each person can be classified by means of the visual appearance. The implemented posture classification is based on projection histograms and select the most likely posture among Standing, Sitting, Crouching and Laying Down [6].
- *Appearance based action recognition*: the action in progress is detected using features extracted from the appearance data. Walking, waving, pointing are some examples of the actions considered. Two different approaches have been selected and implemented: the first is based on Hidden Markov Models [17] and the second on action signature [3].
- *Trajectory-based action recognition*: people trajectories (i.e., frame-by-frame positions of the monitored tracks) embed information about the people's behavior; in particular they can be used to detect abnormal paths that can be related to suspicious events. A trajectory classifier has been added to the system replicating the algorithm described in [4].
- *Smoke detector*: the smoke detection algorithm proposed in [13] has been integrated in the system. The object color properties are analyzed according to a smoke reference color model to detect if color changes in the scene are due to a natural variation or not. The input image is then divided into fixed size blocks and each block is evaluated separately. Finally, a Bayesian approach detects whether a foreground object is smoke.
- *People re-identification with 3D body models*: people appearance is the most useful source of information if we need to match images of people captured by spatially or temporally disjoint cameras, that is, geometrical relations are not available. Even if partially solved using region-based features, one of the main limitations of the available solutions for people re-identification is the fact that these depend on the point of view: for example, the specific location of characteristic patterns is usually lost and cannot be used for people matching. We therefore propose to create a simplified 3D body model which allows us to map appearance features to their 3D location in the body model [1].

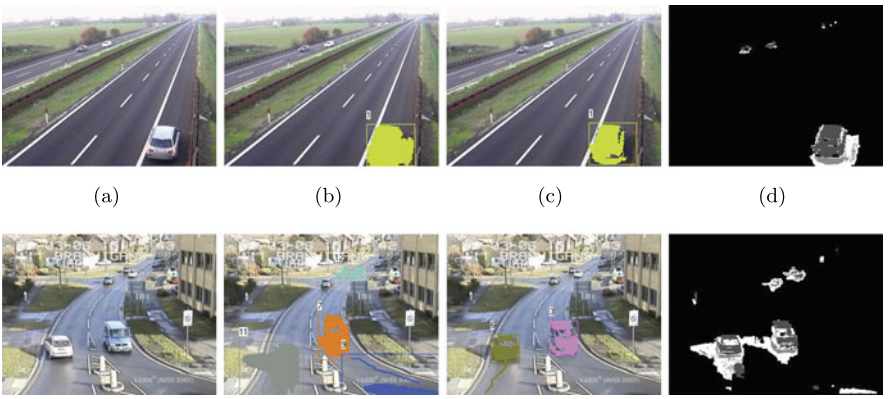


Fig. 12 Analysis with *shadow remover* plug-in. (**a**) source images, (**b**) object detection without shadow removal plug-in, (**c**) object detection with shadow removal plug-in, (**d**) pixel classification into background (*black*), object (*gray*) and shadow (*white*)

4.1 The Shadow Removal Plug-in

Since the information collected into MILS must be as accurate as possible in order to speed up the retrieval process, the analytics algorithms must assure the best performance in terms of computational time and accuracy. For instance, shadows cause nearby objects to merge and significantly change the shape and appearance of the object. As a consequence, the subsequent object classification (which is to some extent related to the object shape-detecting cars instead of people-and its appearance-finding red cars by the color average) can be affected by the inclusion of shadow pixels as belonging to the object. To address this problem, we created a new component with the aim of removing shadows. We implemented the algorithm reported in [5] as a SSE plug-in, which is positioned in the analytic sequence after the *background suppression* plug-in(thus removing the shadow from the foreground objects) and before the *tracking* plug-in (thus feeding it with the correct input).

Removing the shadows allows us to get the correct shape of the objects, as shown in Fig. 12. Given as input stream the left images, on the right we have the results of the shadow remover analysis: the background is black, the foreground is gray, and the white color identifies those pixels that were identified as part of the foreground object, but that are actually shadows and therefore belonging to the background. The central images clearly show how removing the shadow improves the shape of the cars. These enhancements allow us to obtain a better color and shape description for a given object.

5 Concluding Remarks

The purpose of this chapter is to propose a new paradigm for video surveillance systems which borrows concepts from cloud computing and distributed computer

systems. The idea of providing different services on an "as-a-service" platform is not new. The potential of VSaaS is enormous, but the tools available are still too limited. A key future development is the creation of an easy way for plugging in new algorithms for video analytics. The solution illustrated in this chapter represents a first good step, although subject to improvements, to reach this goal.

Another aspect still not fully explored by the solutions currently available is the possibility to use any type of available cameras, connections, and controls. The diffusion of VSaaS is actually still limited (even though growing fast) because it often requires major financial investment to buy new cameras, as the existing ones are not compatible with the VSaaS software.

Acknowledgements The ViSERaS project was funded by Lepida SpA, Italy and was carried out in collaboration with IBM Italia, Vitrociset SPA and CSP Scarl.

References

1. Baltieri, D., Vezzani, R., Cucchiara, R.: 3d body model construction and matching for real time people re-identification. In: Proc. of Eurographics Italian Chapter Conference 2010 (EG-IT 2010), Genova, Italy (2010)
2. Baltieri, D., Vezzani, R., Cucchiara, R.: Fast background initialization with recursive hadamard transform. In: Proc. of the 7th IEEE International Conference on Advanced Video and Signal-Based Surveillance, Boston, USA, pp. 165–171 (2010)
3. Calderara, S., Cucchiara, R., Prati, A.: Action signature: a novel holistic representation for action recognition. In: 5th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS 2008), Santa Fe, New Mexico (2008)
4. Calderara, S., Prati, A., Cucchiara, R.: Learning people trajectories using semi-directional statistics. In: Proc. of IEEE International Conference on Advanced Video and Signal Based Surveillance (IEEE AVSS 2009), Genova, Italy (2009)
5. Cucchiara, R., Grana, C., Piccardi, M., Prati, A.: Detecting moving objects, ghosts, and shadows in video streams. IEEE Trans. Pattern Anal. Mach. Intell. **25**(10), 1337–1342 (2003)
6. Cucchiara, R., Grana, C., Prati, A., Vezzani, R.: Probabilistic posture classification for human behaviour analysis. IEEE Trans. Syst. Man Cybern., Part A, Syst. Hum. **35**(1), 42–54 (2005)
7. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: Proc. of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR05), pp. 886–893. IEEE Comput. Soc., Washington (2005)
8. Haering, N., Venetianer, P., Lipton, A.: The evolution of video surveillance: an overview. Mach. Vis. Appl. **19**(5–6), 279–290 (2008)
9. Hampapur, A., Brown, L., Connell, J., Ekin, A., Haas, N., Lu, M., Merkl, H., Pankanti, S.: Smart video surveillance: exploring the concept of multiscale spatiotemporal tracking. IEEE Signal Process. Mag. **22**(2), 38–51 (2005)
10. Honovich, J.: Vsaas market size and state 2011. IPVM reports (2011). http://ipvm.com/report/vsaas_market_size_2011
11. Kienzle, W., Bakir, G., Franz, M., Scholkopf, B.: Face detection—efficient and rank deficient. Adv. Neural Inf. Process. Syst. **17**, 673–680 (2005)
12. Klein, N.: China's all-seeing eye. Rolling Stone Mag. (2008)
13. Piccinini, P., Calderara, S., Cucchiara, R.: Reliable smoke detection system in the domains of image energy and color. In: 6th International Conference on Computer Vision Systems, Vision for Cognitive Systems (2008)

14. Tian, Y.l., Brown, L., Hampapur, A., Lu, M., Senior, A., Shu, C.f.: IBM smart surveillance system (S3): Event based video surveillance system with an open and extensible framework. Mach. Vis. Appl. **19**(5–6), 315–327 (2008)
15. Valera, M., Velastin, S.: Intelligent distributed surveillance systems: a review. IEE Proc., Vis. Image Signal Process. **152**(2), 192–204 (2005)
16. Vezzani, R., Cucchiara, R.: Event driven software architecture for multi-camera and distributed surveillance research systems. In: Proc. of the First IEEE Workshop on Camera Networks—CVPRW, San Francisco, pp. 1–8 (2010)
17. Vezzani, R., Piccardi, M., Cucchiara, R.: An efficient bayesian framework for on-line action recognition. In: Proc. of the IEEE International Conference on Image Processing, Cairo, Egypt (2009)
18. Williams, B., Guin, A.: Traffic management center use of incident detection algorithms: findings of a nationwide survey. IEEE Trans. Intell. Transp. Syst. **8**(2), 351–358 (2007)
19. Zhao, T., Aggarwal, M., Germano, T., Roth, I., Knowles, A., Kumar, R., Sawhney, H., Samarasekera, S.: Toward a sentient environment: real-time wide area multiple human tracking with identities. Mach. Vis. Appl. **19**(5–6), 301–314 (2008)

A Literature Review on Video Analytics of Crowded Scenes

Myo Thida, Yoke Leng Yong, Pau Climent-Pérez, How-lung Eng, and Paolo Remagnino

Abstract This chapter presents a review and systematic comparison of the state of the art on crowd video analysis. The rationale of our review is justified by a recent increase in intelligent video surveillance algorithms capable of analysing automatically visual streams of very crowded and cluttered scenes, such as those of airport concourses, railway stations, shopping malls and the like. Since the safety and security of potentially very crowded public spaces have become a priority, computer vision researchers have focused their research on intelligent solutions. The aim of this chapter is to propose a critical review of existing literature pertaining to the automatic analysis of complex and crowded scenes. The literature is divided into two broad categories: the macroscopic and the microscopic modelling approach. The effort is meant to provide a reference point for all computer vision practitioners currently working on crowd analysis. We discuss the merits and weaknesses of various approaches for each topic and provide a recommendation on how existing methods can be improved.

M. Thida (✉) · H.-l. Eng
ZWEEC Analytics, Singapore, Singapore
e-mail: mthida81@gmail.com

H.-l. Eng
e-mail: howlungeng@zweec.com

Y.L. Yong · P. Climent-Pérez · P. Remagnino
Kingston University, London, UK

Y.L. Yong
e-mail: J.Yong@kingston.ac.uk

P. Climent-Pérez
e-mail: P.Climent@kingston.ac.uk

P. Remagnino
e-mail: P.Remagnino@kingston.ac.uk

P.K. Atrey et al. (eds.), *Intelligent Multimedia Surveillance*,
DOI 10.1007/978-3-642-41512-8_2,

1 Introduction

Automated video content analysis of a crowded scene has been an active research area in the field of computer vision in the last few years. This strong interest is driven by the increased demand for public safety at crowded spaces such as airports, train stations, malls, stadiums, etc. In such scenes, conventional computer vision techniques for video surveillance cannot be directly applied in the crowded scene due to large variations of crowd densities, complex crowd dynamics and severe occlusions in the scene.

Algorithms for people detection, tracking and activity analysis which consider an individual in isolation (i.e., individual object segmentation and tracking) often face difficult situations such as the overlapping of pedestrians, complex events due to interactions among pedestrians in a crowd. For this reason, many papers consider the crowd as a single entity and analyse its dynamics. The status of crowd is updated as normal or abnormal based on the dynamics of the whole crowd. However, a crowded condition can also be unstructured where pedestrians are relatively free to move in many directions as opposed to a structured crowd where each individual moves coherently in one common direction. In an unstructured crowded scene, considering the crowd as one entity will fail to identify abnormal events which arise due to an inappropriate action of an individual in a crowd. For instance, a running person in a crowd can indicate an abnormal event if the rest of crowd are walking. Thus, considering the crowd as one entity can cause false detections.

Many paper works on modelling crowded scenes to identify different crowd events and/or to detect abnormal events. However, the definition of abnormal event or event of interest has been causing much confusion in the literature due to its subjective nature. Some researchers consider a rare and outstanding event as abnormal while some consider events that have not been observed are abnormal. The problem becomes more challenging as the density of people increases. As a result, more computer vision algorithms are being explored recently.

Despite the great interest and a large number of methods developed, there is a lack of a comprehensive review on crowd video analysis. As shown in Table 1, most current surveys focus on general human motion analysis [1, 5, 24, 75] of single or a small group of people, rather than addressing a crowded scenario. The survey paper by Zhan et al. [83], to the best of our knowledge, is the only one focusing on crowd video analysis. Zhan et al. reviewed some crowd density estimators and crowd modelling techniques, focusing on pedestrian detections, and tracking in a cluttered scene. However, they did not discuss the topic of crowd behaviour understanding and abnormality detection which is covered in this survey. We also present some advances on crowd motion modelling and multi-target tracking in a crowded scene which are not covered in the previous survey.

The goal of this survey is to review and organise the state-of-the-art methods in the domain of crowd video analysis such that their main focus becomes apparent. To achieve this, we have divided the research on crowd video analysis into three broad categories: macroscopic modelling, microscopic modelling and crowd event detection. The methods related to each task are further divided into sub-categories and a comprehensive description of representative methods is provided. In addi-

Table 1 A comparison of this chapter and previous surveys on human motion analysis and crowd video analysis

Year	Authors	Focus	Scenarios
1999	Aggarwal and Cai [1]	Motion analysis, tracking and recognising human activities	Individual or group
2003	Wang et al. [75]	Human detection, tracking and behaviour understanding	Individual or group
2004	Hu et al. [24]	Motion detection, tracking and behaviour understanding	Individual or group
2006	Yilmaz et al. [5]	Object tracking	Individual or group
2008	Zhan et al. [83]	Crowd information extraction and crowd modelling	Crowd
2013	This chapter	Macroscopic modelling, microscopic modelling, crowd event detection	Crowd

tion, we identify challenges and future directions for analysing a crowded scene. We believe this will help readers, especially newcomers to this area, to understand the major tasks of a crowded scene analysis system and hope to motivate for the development of new methods.

2 Macroscopic Modelling

In order to learn the typical motion patterns in a crowded scene, macroscopic observation-based methods utilise holistic properties of the scene such as motions in local spatio-temporal cuboid or instantaneous motion are utilised. It is also the preferred method in tracking and analysing the behaviour of both sparse and dense crowd using the following properties such as: density, velocity and flow [31]. Figure 1 depicts detailed various features available for use in macroscopic modelling and the techniques initialising those features.

2.1 Optical Flow Feature

Optical flow is a dense field of instantaneous velocities computed between two consecutive frames commonly used in extracting motion features [23]. Given a video of a crowded scene, the first step is to segment the input video into smaller video clips and compute pixel-wise optical flow between consecutive frames of each clip using the techniques in [11, 23, 49]. The extracted flow vectors may contain noise and redundant information. In order to reduce the computational cost and remove noise, researchers utilise unsupervised (Andrade et al. [6, 7] and Yang et al. [81]) or supervised (Hu and Shah [26, 27]) dimensional reduction techniques. Subsequently, the next step is to find the representative motion patterns of the scene by merging flow vectors from all video frames. Referring back to Fig. 1, it can be seen that

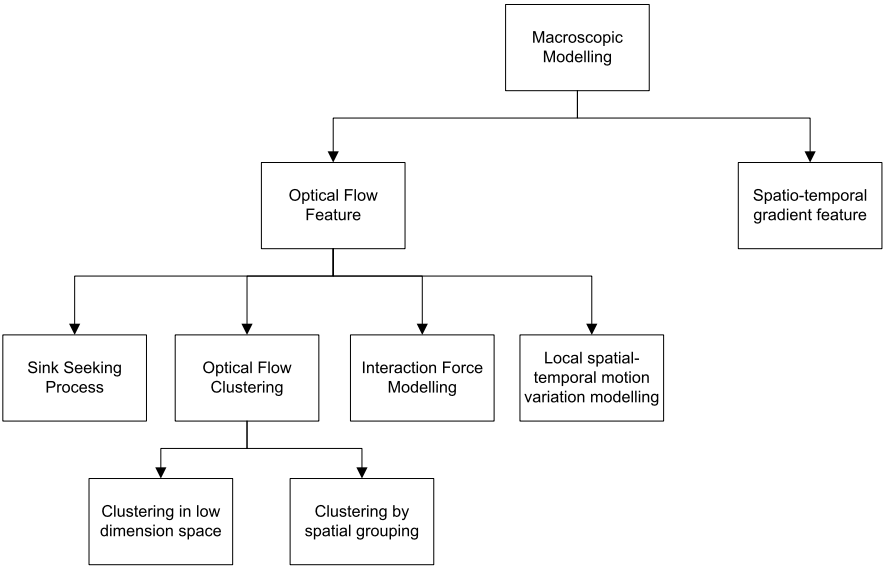


Fig. 1 A schematic illustration of the topics involved in macroscopic crowd video analysis

the motion features extracted from the optical flow can be utilised for motion pattern extraction such as: Sink Seeking Process, Optical Flow Clustering, Interaction Force Modelling, Local Spatio-temporal Motion Variation Modelling, and Spatio-temporal Gradient feature whereby the methods can be used separately or integrated with one another to obtain the desired crowd analysis.

2.1.1 Sink Seeking Process

In the sink seeking process, a grid of particles is overlaid on the first frame of the video clip and advected using a numerical scheme. The path taken by a particle to its final position is called a sink path and thus, the process of finding sinks (exits) and sink paths is called a sink seeking process. Hu and Shah [26, 27] carry out sink seeking process for each particle and thus generate one sink path per particle. These sinks and sink paths are later clustered to extract the dominant motion paths of the scene using an iterative clustering algorithm. On the other hand, Ali and Shah [3] generate a static floor field where each particle holds a value that represents the minimum distance to the nearest sink form its current location. Ali and Shah impose the static floor field together with dynamic and boundary floor field as constraints for tracking algorithm [4].

2.1.2 Optical Flow Clustering

Andrade et al. [6, 7] model the principal components of the optical flow vectors in each video clip using Hidden Markov Models. Then, video segments which have

similar motion pattern are grouped together using the spectral clustering method. The resulting clustered video segments are modelled using a chain of HMMs to represent the typical motion pattern of the scene. The emergency events in the monitored scene are detected by finding deviations from the obtained model.

Instead of the spatial segmentation of each video frame, the other approach is to cluster optical flow vectors by spatial grouping as in [64]. Imran et al. [64] proposed to cluster optical flow vectors in each video clip into N Gaussian mixture components. Then, these Gaussian components are linked over time using a fully connected graph. The connected component analysis of the graph is performed to discover different motion patterns. However, their method still faces the problem of having to determine how many components should there be in the mixture.

2.1.3 Interaction Force Modelling

In addition to learning dominant motion patterns, the optical flow vectors obtained can also be used to model interaction forces of a crowd, and then use the model to detect the stability of the crowd. For example, Mehran et al. [53] employ the optical flow vectors to model pedestrian motion dynamics using a social force model. Social force models [22] have been used in many studies in computer graphic fields for creating animations of the crowd [54]. In this model, the motions of pedestrians are modelled with two forces: a personal desire force and an interaction force. The interaction force is defined as an attractive and repulsive force between pedestrians. In [53], an interaction force between pedestrians is estimated based on optical flow computed over a grid of particles. The normal pattern of this force is later used to model the dynamics of a crowded scene and detect abnormal behaviours in crowds.

2.1.4 Local Spatio-Temporal Motion Variation Modelling

Optical flow data can also be used in modelling the variations of motions in local spatio-temporal volumes to describe the typical motion patterns of the scene [40–42, 50, 52, 79, 81]. In these approaches, an image space is usually divided into cells of a specific size (e.g., 10×10 in [81]) or cuboids (e.g., $30 \times 30 \times 20$ in [42]). Then, optical flow computed in each cell is quantised into different directions. For instance, Yang et al. [81], considered each quantised direction of a given location as a word and cluster these video words into different clusters using a diffusion embedding method. Each node in the graph corresponds to a word and the clusters extracted in the embedded space represent the typical motion patterns of the scenes. Kim and Grauman [40] used a space-time Markov Random Field (MRF) graph to detect abnormal activities in video. Each node in the graph corresponds to a local region in the video frames where the local motion is modelled using a mixture of probabilistic principle component analysis. Wu et al. [79] used Lagrangian framework to extract particle trajectories. These particle trajectories are later used for the modelling of regular crowd motion. The deviations of new motion from the learnt model indicates an abnormal event.

2.2 Spatio-Temporal Gradient Feature

In addition to optical flow information, other features such as spatio-temporal gradient are also used to model the regular movement of a crowd [42, 50]. In [42], the coupled HMM is trained based on the distribution of spatio-temporal motions to detect localised abnormalities in densely crowded scenes. Vijay et al. [52] combined motion information and appearance features to represent the local properties of a scene. The normality of a crowded scene is learned using a mixture of dynamic textures. Then, temporal and spatial abnormalities are separately detected by finding deviations from the normal pattern. Their method has been proved to achieve the better performance than the state-of-the-art methods, at a high computational cost. To address this limitation, Reddy et al. [61] proposed a simpler method using a set of similar features including shape, size and texture extracted from foreground pixels. The computational cost is reduced by removing background noise and considering each feature type individually. Compared to [52], the method proposed by Reddy et al. [61] achieved considerably better results.

2.3 Summary

To conclude the discussion on the macroscopic modelling, a summarisation of the strength and weaknesses of the various state-of-the-art implementation are provided in Table 2.

3 Microscopic Modelling

Microscopic analysis and modelling depends on the analysis of video trajectories of moving entities. This approach, in general, contains the following steps:

1. detection of the moving targets present in the scene;
2. tracking of the detected targets; and
3. analysis of the trajectories to detect dominant flows, and to model typical motion patterns.

Researchers have used different detection and tracking algorithms to generate reliable trajectories. Tracking people in crowds can be either used as a means to improve crowd dynamics analysis, using the tracks and mining trends out of these (*bottom-up* approach to crowd analysis); or, conversely, tracking methods can use cues obtained from the analysis of crowd dynamics, in order to improve accuracy (*top-down* approach). The complexity of tracking algorithms depends on the context and environment in which the tracking is performed. In the context of crowd video analysis, the problem of tracking individuals within a crowd introduces additional

Table 2 Summarisation of the macroscopic modelling techniques

Papers	Advantages	Disadvantages	Data-set and results
[26, 27]	• Instead of using the long term trajectories of the moving objects in learning the typical motion pattern, global motion flow field is used. • Not affected by density of objects within the image. • Does not require complete trajectory, therefore overcoming the problem of occlusion.	• More attention should be placed on learning motion patterns in crowds with less reliable tracking.	• Crowded scene. • Aerial vide from DARPA's VIVID dataset. • Hong Kong street scene.
[3]	• Lagrangian Coherent Structures (LCS) reveals underlying flow structures that are generally not evident from the raw velocity field of the object.	• If the changes in dynamics is not big enough to be detected, it will not be segmented out within the image.	• Videos from stock footage web sites. • National Geographic documentary footage from "Inside Mecca".
[4]	• Takes into consideration the crowd flow and scene layout for tracking. • Provides the shortest distance to a sink for each location.	• Tracking errors produced by the Static Floor field (SFF) is inconsistent and dependant on the trajectory of the tracked object and only works well when the object is moving in a straight path. • The Dynamic Floor Field produced error when there's noise and interference of other objects in the view.	• Marathon sequence from different perspective – Overhead cameras. – View from high rise building.
[53]	• Does not depend on tracked object in the analysis of crowd behaviour. • The incorporation of particle advection assist in capturing the crowd flow.	• Requires a set of goal destination for the scene.	• University of Minnesota dataset.
[40]	• Anomalies could be detected on both the global and local context.	• Posterior probabilities of all previous descriptors are not recalculated, and assumed not to change.	• Surveillance videos from subway station – Entrance Gate. – Exit Gate.

Table 2 (Continued)

Papers	Advantages	Disadvantages	Data-set and results
	• With the information from the global and local context available, anomalies are easily detected. • Ease of implementation.		
[52]	• Works extremely well in detecting anomalies in video scenes compared to other tracking based method such as optical flow.	• System requires training. • Long computation time making it not feasible for real-time system.	• Crowded Scenes.
[81]	• The weight of the Diffusion Map is unique to the application. • By manipulating the diffusion time, Diffusion Map can also be used for multi-scale analysis of the scene.	• Might be more suitable for scene understanding as low level feature have a better manifold structure.	• INGSIM dataset. • Far-field traffic scene.
[61]	• Low-level computation required. • Separate modelling and analysis of motion, size and texture feature make it commutation efficient.	• Requires training. • The density information obtained changes drastically due to the handling of the training data.	• UCSD anomaly detection dataset.

complexity due to the interactions and occlusions between people in the crowd. A number of tracking methods has been proposed to overcome the challenges encountered in a crowded scene. In this section, some popular human tracking methods in the context of crowd video analysis are discussed. The reader is referred to the survey by Yilmaz et al. [5] for a comprehensive review of various trackers. Figure 2 shows the different topics covered by this section.

3.1 The Particle Filter (PF) Framework

The most popular approach for tracking is the *Particle Filter*-based framework. Particle filtering framework was first introduced for visual tracking by Isard and Blake in [29]. Initially, particle filter approaches were only based on colour cues, and could only track one single target.

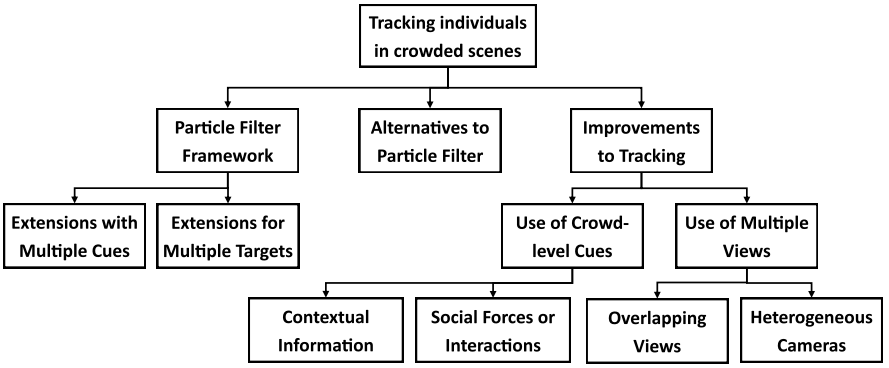


Fig. 2 The topics for microscopic approach, in which a focus is put on individual tracking in crowds

3.1.1 Additional Cues for Improved PF

The particle filter implementation based on appearance using colour information only does not perform well tracking more than one individual, specially when those wear similar clothing. In public demonstrations, sports matches and celebrations, it is normal that people's appearance is similar. Thus, a series of papers present alternatives to the plain 'colour-only' Particle Filter. Combinations include colour and contours, Harris, SIFT features [47, 59, 67, 78]; also Histograms of Oriented Gradients (HOGs) are used along with colour information in [69]; or Mean Shift and Joint Probabilities [10].

A completely different approach to improve tracking using particle filters is presented in [84]. The method proposed by the authors mines the interdependencies between particles in order to improve the results. Also different is the method proposed in [28], in which a new tracker is proposed which employs a particle filter tracking framework, where the state transition model is estimated by an optical-flow algorithm. That is, instead of using a pre-defined dynamic transition model.

There are also authors whose interest is in extending the particle filter to multiple cameras; in that case, particles are "shared" and "fused" among the views [57].

Others propose blob-based segmentation and tracking when no occlusions are present, and limit the use of Particle Filters as an occlusion resolution technique [70, 86]. The limitation of this techniques seems clear: blobs are needed and used as the main cue, which is not the case in most crowded scenes, although these techniques can be useful in sparse crowds.

Silhouettes or contours can be a useful cue for action recognition, or people counting in crowds; obviously, in the case of densely crowded scenes, only partial contours can be extracted, although those can be quite useful (e.g., as in 'Ω shape'-based methods). Since particle filter approaches work regardless of segmentation, reconstructing contours *a posteriori* to obtain shape cues might be of interest. Ma et al. [51] present this idea: Graph Cuts are applied to a particle filter method to obtain the silhouettes of tracked objects.

3.1.2 Alternative Cues for Tracking: Self-similarity

Schechtman and Irani [66] introduced the concept of self-similarity as a visual feature. Among the applications, they name object detection and recognition, action recognition, texture classification, data retrieval, tracking and image alignment and so on. BenAbdelkader et al. [12] seem to be the first to use image self-similarity plots (ISSPs) for gait recognition; according to the authors, some works state the ISSP of a moving person/object is a projection of its planar dynamics, and as such, these should encode much of gait information. Junejo et al. [34, 35] use a very similar descriptor as a means for action recognition, by using self-similarity matrices (SSMs) as descriptors of the action class. Dexter et al. [17] extend the SSM concept in order to apply it to the synchronisation of actions taken from multiple views. Rani and Arumugam [60] use it as a biometric signature in gait recognition as in [12]. Also, Walk et al. [74] introduced the self-similarity as a feature for pedestrian detection; and Cai et al. [14] have used it for person re-identification among different cameras or moments; the authors create a colour codebook and obtain the spatial occurrence distributions of colour self-similarities. To the best of our knowledge, as of today, no works seem to use self-similarities as a feature for tracking, although Gu et al. state it could be used as an alternative to other local descriptors such as SIFT or SURF.

3.1.3 Multiple Target Tracking Using PF

This framework has been extended in a series of papers [2, 15, 20, 39, 58] for tracking multiple targets. For example, Okuma et al. [58] extend a particle framework by incorporating a cascaded AdaBoost algorithm for the detection and tracking of multiple hockey players in a video. The AdaBoost algorithm is used to generate detection hypotheses of hockey players. Once the detection hypotheses are available, each hockey player is modelled with an individual particle filter that forms a component of a mixture particle filter. Similarly, Ali and Dailey [2] combine an 'AdaBoost cascade classifier'-based head detection algorithm and the particle filtering method for tracking multiple persons in high density crowds. The performance is further improved by a confirmation-by-classification method to estimate confidence in a tracked trajectory.

To conclude this subsection, a summarisation of the presented methods is shown in Table 3. Both single and multiple view methods are presented, as well as single and multiple target ones.

3.2 Handling Occlusions

Occlusions are one of the most important problems trackers need to face, since generalised models for them are not straightforward [44]. According to the survey in [82], occlusion can be classified into three categories: self-occlusion, which occurs while tracking articulated objects; inter-object occlusion (or dynamic occlu-

Table 3 Summarisation of the presented techniques

Main tracking	Additional information or method	Multi-view/-target	Works/papers
Particle Filter	Contour information	No	[47]
Particle Filter	SIFT, Harris-SIFT	No	[59, 67, 78]
Particle Filter	Histogram of Oriented Gradients (HOG)	No	[69]
Particle Filter	Mean Shift/Joint Probabilities	No	[10]
Particle Filter	None: Changes in the transition model	No	[28, 84]
Particle Filter	None: Particles are fused among views	Multi-view	[57]
Blob tracking	Particle filter for occlusion handling	Multi-target	[70, 86]
Particle Filter	None: GCs[a] used to recover contours	No	[51]
Particle Filter	AdaBoost, Cascaded AdaBoost	Multi-target	[2, 58]
Particle Filter	MRF[b], MCMC[c]	Multi-target	[39]
Particle Filter	NN Data Association, Mean-shift	Multi-target	[15]
Mean Shift/Kalman	Viterbi-style *tracklet* merge	Multi-target	[20]

[a]Graph Cuts

[b]Markov Random Field

[c]Markov Chain Monte Carlo

sion [72]), which arises when two tracked objects occlude each other; and occlusion by the background (or scene occlusion [72]), which occurs when structures in the scene (e.g., tree branches, pillars, etc.) occlude the object/s being tracked. Some approaches have already been presented in Sect. 3.1.1 [70, 86]. Yilmaz et al. [82] deal with occlusion handling from the lens of the tracking technique in use. A series of different tracker families are presented (point, 'geometric model'-based and silhouette); each tracking technique is then classified according to whether or not it can handle occlusions, and in the case it does, whether these can be full or only partial. Following this idea, trackers that respond well when occlusions are present, can be used for occlusion handling. In [85], the Kanade-Lucas-Tomasi (KLT) tracker is employed to resolve occlusions, while a particle filter is used as the main tracker. Similarly, a technique based on Mean-shift is used in [16].

Apart of exploiting the features of "occlusion-friendly" trackers, a series of occlusion handling techniques have also been devised, which can be found throughout the literature. Wang et al. [77], present a good historical review of such methods, which rely on the object's motion model, and keep predicting the object's location until it reappears. The authors state that serious long-term occlusions cannot be dealt with by this kind of techniques, since observations cannot be obtained while the object is occluded for a long period of time. Vezzani et al. [72] propose what they call the *non-visible regions model*, which deals with partial and full occlusions, whether these are inter-object or due to the scene. The object model is updated differently in a pixel-wise fashion: the appearance is updated only for the visible pixels; the probabilities associated to those are reinforced, while they remain unchanged for invisible pixels. Furthermore, in pixels with no correspondence due to changes in

the shape of the object (called appearance occlusions) probabilities are smoothed. Wang et al. [77], on the other hand, propose a means of modelling the occluder; once modelled, when objects disappear due to occlusion, a search is performed around the occluder in order to find the occluded object as it reappears.

In [37], the authors present a series of monocular approaches to occlusion handling, although this is only to conclude that single-view systems are intrinsically unable to handle occlusions correctly. The authors in [21, 37], use multiple oblique-view cameras to handle occlusions appropriately, and devise a common plane reconstruction, using communication among cameras. Approaches based on multiple views are designed to reduce the amount of hidden regions. Unfortunately, in the case of existing static camera networks, this is not always possible due to the restrictions of their infrastructures, which were not initially devised for automated surveillance. Another approach to occlusion handling is avoiding them in the first place. Occlusions can be reduced by placing the camera appropriately, as suggested by [82] (e.g., by placing a bird-eye view camera, no occlusions occur between the objects on the ground), but the problem of existing infrastructures persists.

Nevertheless, when dealing with occlusions under heavily crowded scenarios, full-body tracking is infeasible due to the continuous existence of partial occlusions, specially from side views [13]. Since the existing cameras tend to be placed above the heads of the people and tilted to face downwards looking at the scene, some authors suggest a good assumption is that heads and shoulders (often referred to as Omega-shape [46]) will be always visible, and that occlusions among subjects' heads is lower as compared to the rest of the body parts.

3.3 Improving Tracking Using Crowd-Level Cues

As stated in the introduction to this section (Sect. 3), tracking methods can use cues obtained from the analysis of crowd dynamics, in order to improve their accuracy, in a *top-down* approach. These higher-level cues can be either contextual or coming from the social interactions among the people in the crowd.

3.3.1 Higher-Level Contextual Information

The utility of high-level contextual information has demonstrated that exploiting contextual information improves the performance of human tracking significantly. Antonini et al. [9] use a discrete choice model (DCM) as motion priors to predict human motion patterns and then, fuse this model in a human tracker for improved performance. Similarly, Ali et al. [4] propose to exploit contextual information for tracking multiple people in a structured crowded scene. Assuming that all participants of the crowd are moving in one direction, Ali et al. learn the direction of motion as a prior information based on floor fields. The authors have demonstrated that a higher-level constraint greatly increases the performance of the tracker. However,

floor fields can be learned only when the scene has one dominant motion. As a result, the method proposed in [4] cannot be applied for unstructured crowded scenes where the motion of a crowd appears to be random with different participants moving in different directions over time. Some examples of unstructured crowded scenes include crowds at exhibitions, sporting events and railway stations. This shortcoming is addressed by Mikel et al. [63] where the authors employ a correlated topic model for modelling random motions in an unstructured crowded scene. Similarly, L. Kratz and K. Nishino [43] employ the normal motion pattern to predict tracking individuals in a crowd scene where the normal motion pattern is learnt based on local motion at fixed-size cells.

3.3.2 Social Interactions

Another interesting direction of tracking multiple targets is to integrate social interaction of targets in the tracking algorithm. This idea is motivated by the behaviour of targets in a crowd. In crowded scenarios, the behaviour of each individual target is influenced by the proximity and behaviour of other targets in the crowd. Several methods [8, 19, 39, 48, 80] have proposed to integrate the social interactions among targets in the tracking algorithms. This direction has shown promising performance to track multiple targets in crowded scenes. An early example which models the social interaction of targets is *Markov Chain Monte Carlo*-based (MCMC) particle filter [39]. Their method models social interactions of targets using Markov Random Field and adds motion prior in a joint particle filter. The traditional importance sampling step in the particle filter is replaced by a MCMC sampling step. French et al. [19] extended the method in [39] by adding social information to compute the velocity of particles. In [80], the authors formulated the tracking problem as a problem of minimising an energy function. The energy function is defined based on both the social information and physical constraint in the environment. Their preliminary results indicate that social information provides an important cue for tracking multiple targets in a complex scene. An overview of tracking algorithms that incorporate different high-level contextual information is illustrated in Fig. 3.

3.4 Tracking in Crowds from Multiple Views

Researchers have also explored the use of multiple cameras for tracking people under severe occlusion in a complex environment. Multiple camera tracking methods intend to expand the monitored area and provide complete information about interesting persons by gathering evidences from different camera views. Lee et al. [45] propose a multiple people tracking method for wide-area monitoring. An automated calibration method is introduced to find correspondences between distributed cameras. In their method, all camera views are calibrated to a global ground-plane view

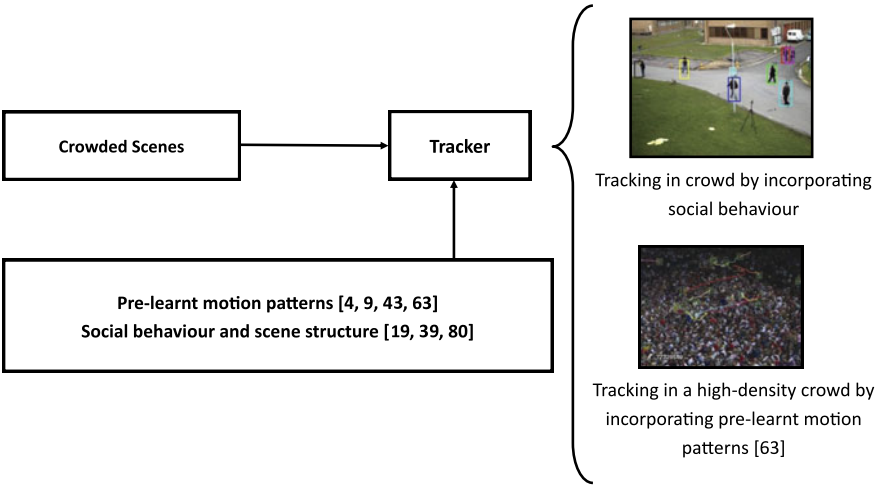


Fig. 3 An overview of different tracking algorithms that incorporate high-level contextual information

based on geometric constraints and tracking trajectories from each view. Another example in a similar context can be found in the papers by Khan and Shah [36, 38]. A planar homographic occupancy constraint that combines foreground likelihood information from different views is proposed for detection and occlusion resolution.

Another use of multiple cameras is to track people in an environment covered by multiple cameras with overlapping views. Mittal and Davis [55] use pairs of stereo cameras and combine evidences gathered from multiple cameras for tracking people in a cluttered scene. Foreground regions from different camera views are projected back into a 3D space so that the endpoints of the matched regions yield 3D points belonging to people. Dockastader and Tekalp [18] employ a Bayesian network for fusing 2D position information acquired from different camera views to estimate the 3D coordinate position of the interested person. Finally, a layer of Kalman filtering is used to update the position of people. A combination of static and pan-tilt-zoom (PTZ) cameras for multiple camera tracking is introduced in [65]. The static cameras are used to provide a global view of the interested persons when the PTZ cameras are used for face recognition of people.

The brief overview of the research literature indicates that multiple camera tracking methods provide an interesting mechanism to handle severe occlusion and to monitor large areas at public spaces (as seen in Sect. 3.2). However, advantages of the multiple cameras come together with additional issues such as camera calibration, matching information across the camera views, automated camera switching and data fusion. These challenges are still yet to be solved. On the other hand, integrating the social interaction among targets in the tracking algorithms has shown promising performance to track individual targets in a crowd.

4 Event Detection in Crowds

A series of surveys and reviews in this field [30, 32, 56, 62, 68, 73, 83] show there is a great interest in this area. Detecting anomalies or outstanding events in crowds has moved a lot of research efforts. Automatic systems would allow reducing the burden of manual video supervision, which makes is infeasible in most cases, given the enormous amounts of data, as compared to the manpower to process it [68, 73, 83]. Detection of anomalies in crowded scenes can be seen as a classification problem where only two classes are defined (i.e., "normal" versus "anomalous") [68].

The survey by Sodemann et al. [68], analyses the works in the literature across five aspects:

1. the target/s of interest (a person, a crowd);
2. the definitions of what is anomalous, and the assumptions taken;
3. the types of sensors involved, and the features used;
4. the learning methods; and
5. the modelling algorithms.

According to the authors, their survey is focused on the broader problem formulation and assumptions, rather than providing a review on specific pattern classification methods. In Revathi and Kumar [62], authors provide a categorisation of anomalies according to the number of people and other objects involved. Three categories are defined: anomalies involving a single person with a single object, multiple people with multiple objects, and group behaviour.

Vishwakarma and Agrawal [73] analyse human action recognition in more general terms in video surveillance, although, they present an interesting taxonomy to classify complexity of topic-related algorithms. From a completely different point of view, two other works review physics- and hydrodynamics-based techniques [32, 56] for anomaly and event detection in crowds. Moore et al. [56] present a review of techniques for crowd analysis that consider huge crowds as "fluids" or "liquids", which are bound to a series of rules and forces (e.g., repulsion and attraction) which explain the interactions among the particles that conform that fluid.

Jo et al. [32] further explore other physics-based techniques, and classify the works according to the categorisation presented in [30], which presents various "domains": the image space domain, based on the analysis at the pixel, texture or object levels; the sociological domain, which accounts for the social interactions or "crowd mentality"; the level of services, where different crowd conditions are provided; or the computer graphics domain, which deals with realistic crowd simulation. Jacques Junior et al. [30] also classify crowd event detection techniques as either object-based, in which individuals are tracked and these tracks are used to analyse the situations [25, 33, 76]; or holistic-based [7, 53, 71], where the crowd is considered as a whole, and events are detected by extracting the major crowd flows from the monitored scene.

5 Summary

This chapter presents a review and comparative study of various topics in the area of crowd video analysis. The advantages and disadvantages of the state-of-the-art methods related to video analytics in crowded scenes have been detailed.

Tracking individuals in a high-density crowd has been addressed in recent years, as opposed to previously tracking individuals in sparse or even *ad-hoc* scenarios. A major advance is the introduction of high-level crowd motion pattern as a prior into a general framework [4, 63]. However, the problem of tracking still remains as a challenging problem in the area of computer vision. One major challenge for tracking in a crowded scene is inter-object occlusion due to the interactions of participants in a crowd. There remains a gap between the state-of-the-art and robust tracking of people in a crowded scene. Most recent trackers for crowds use Particle Filters, using different kinds of features; the use of self-similarity measures for this particular application can be of interest and deserves further research, given the results it achieved in other Computer Vision fields.

During recent years there has been substantial progress towards understanding crowd behaviour and abnormality detection based on modelling crowd motion pattern. However, these approaches capture general movement of a crowd but do not accurately detect details of individual movements. As a result, the current literature in understanding crowd motion is not ready to capture the motion pattern of an unstructured crowd scene where the motion of the crowd appears to be random [63]. Future research in this area requires localised modelling of crowd motion to capture different behaviours in unstructured crowded scenes. On the other hand, the understanding and modelling of crowd behaviour remains immature despite the considerable advances in human activity analysis. Progress in this area requires further advances in modelling or representation of a crowd event and recognition of these events in a natural environment.

References

1. Aggarwal, J.K., Cai, Q.: Human motion analysis: a review. Comput. Vis. Image Underst. **73**(3), 428–440 (1999)
2. Ali, I., Dailey, M.N.: Multiple human tracking in high-density crowds. In: Advanced Concepts in Intelligent Vision Systems, pp. 540–549 (2009)
3. Ali, S., Shah, M.: A Lagrangian particle dynamics approach for crowd flow segmentation and stability analysis. In: Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition, Minneapolis, Florida, pp. 1–6. IEEE, New York (2007)
4. Ali, S., Shah, M.: Floor fields for tracking in high density crowd scenes. In: Proceedings of European Conference on Computer Vision, Marseille, France, pp. 1–14. Springer, Berlin (2008)
5. Alper, Y., Omar, J., Mubarak, S.: Object tracking: a survey. ACM Comput. Surv. **38**(4), 13–58 (2006)
6. Andrade, E., Fisher, R.: Simulation of crowd problems for computer vision. In: Proceedings of 19th International Conference on Pattern Recognition, vol. 3, pp. 71–80 (2005)

7. Andrade, E., Fisher, R., Blunsden, S.: Modelling crowd scenes for event detection. In: Proceedings of 19th International Conference on Pattern Recognition, vol. 1, pp. 175–178 (2006)
8. Andriyenko, A., Schindler, K.: Multi-target tracking by continuous energy minimization. In: Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition, Colorado Springs, CO, pp. 1265–1272. IEEE, New York (2011)
9. Antonini, G., Martinez, S.V., Bierlaire, M., Thiran, J.P.: Behavioral priors for detection and tracking of pedestrians in video sequences. Int. J. Comput. Vis. **69**(2), 159–180 (1998)
10. Bai, K.: Particle filter tracking with mean shift and joint probability data association. In: 2010 International Conference on Image Analysis and Signal Processing (IASP), pp. 607–612. IEEE, New York (2010)
11. Barron, J., Fleet, D.J., Beauchemin, S.: Performance of optical flow techniques. Int. J. Comput. Vis. **12**(1), 43–77 (1994)
12. BenAbdelkader, C., Cutler, R., Nanda, H., Davis, L.: Eigengait: motion-based recognition of people using image self-similarity. Technical report (2001)
13. Boltes, M., Seyfried, A.: Collecting pedestrian trajectories. Neurocomputing **100**, 127–133 (2013)
14. Cai, Y., Pietikäinen, M.: Person re-identification based on global color context. In: Asian Conference on Computer Vision 2010 Workshops (2011)
15. Cai, Y., de Freitas, N., Little, J.J.: Robust visual tracking for multiple targets. In: Proceedings of Eighth European Conference on Computer Vision, vol. 3954, pp. 107–118. IEEE, New York (2006)
16. Chen, A.h., Yang, B.q, Chen, Z.g.: A timely occlusion detection based on mean shift algorithm. In: Deng, W. (ed.) Future Control and Automation. Lecture Notes in Electrical Engineering, vol. 173, pp. 51–56. Springer, Berlin (2012)
17. Dexter, E., Pérez, P., Laptev, I.: Multi-view synchronization of human actions and dynamic scenes. In: Proceedings of the British Machine Vision Conference 2009, British Machine Vision Association, pp. 122.1–122.11 (2009)
18. Dockstader, S.L., Tekalp, A.M.: Multiple camera tracking of interacting and occluded human motion. Proc. IEEE **89**(10), 1441–1455 (2001)
19. French, A., Naeem, A., Dryden, I., Pridmore, T.: Using social effects to guide tracking in complex scenes. In: Proceedings of IEEE Conference on Advanced Video and Signal Based Surveillance, pp. 212–217 (2007)
20. Gilbert, A., Bowden, R.: Multi person tracking within crowded scenes. In: Proceedings of Workshop on Human Motion, pp. 166–179 (2007)
21. Haselhoff, A., Hoehmann, L., Nunn, C., Meuter, M., Kummert, A.: On occlusion-handling for people detection fusion in multi-camera networks. In: Dziech, A., Czyżewski, A. (eds.) Multimedia Communications, Services and Security. Communications in Computer and Information Science, vol. 149, pp. 113–119. Springer, Berlin (2011)
22. Helbing, D., Molnar, P.: Social force model for pedestrian dynamics. Phys. Rev. E **51**(5), 4282–4286 (1995)
23. Horn, B.K.P., Schunck, B.G.: Determining optical flow. Artif. Intell. **17**, 185–203 (1981)
24. Hu, W., Tan, T., Wang, L., Maybank, S.: A survey on visual surveillance of object motion and behaviors. IEEE Trans. Syst. Man Cybern., Part C, Appl. Rev. **34**(3), 334–352 (2004)
25. Hu, W., Xiao, X., Fu, Z., Dan, X., Tan, T., Steve, M.: A system for learning statistical motion patterns. IEEE Trans. Pattern Anal. Mach. Intell. **28**(9), 1450–1464 (2006)
26. Hu, M., Ali, S., Shah, M.: Detecting global motion patterns in complex videos. In: Proceedings of International Conference on Pattern Recognition, Tempa, Florida, pp. 1–5. IEEE, New York (2008)
27. Hu, M., Ali, S., Shah, M.: Learning motion patterns in crowded scenes using motion flow field. In: Proceedings of International Conference on Pattern Recognition, Tempa, Florida, pp. 1–5. IEEE, New York (2008)
28. Hu, N., Bouma, H., Worring, M.: Tracking individuals in surveillance video of a high-density crowd. In: Proceedings of SPIE, vol. 8399, p. 839909 (2012)

29. Isard, M., Blake, A.: CONDENSATION conditional density propagation for visual tracking. Int. J. Comput. Vis. **29**(1), 5–28 (1998)
30. Jacques Junior, J.C.S., Musse, S.R., Jung, C.R.: Crowd analysis using computer vision techniques: a survey. IEEE Signal Process. Mag. (September), 66–77 (2010)
31. Jiang, Y., Zhang, P., Wong, S., Liu, R.: A higher-order macroscopic model for pedestrian flows. Phys. A, Stat. Mech. Appl. **389**(21), 4623–4635 (2010)
32. Jo, H., Chug, K., Sethi, R.J., Rey, M.: A review of physics-based methods for group and crowd analysis in computer vision. J. Postdr. Res. **1**(1), 4–7 (2013)
33. Johnson, N., Hogg, D.: Learning the distribution of object trajectories for event recognition. Image Vis. Comput. **14**(8), 583–592 (1996)
34. Junejo, I., Dexter, E., Laptev, I., Pérez, P.: Cross-view action recognition from temporal self-similarities. In: Proceedings of the European Conference on Computer Vision 2008 (2008)
35. Junejo, I.N., Dexter, E., Laptev, I., Pérez, P.: View-independent action recognition from temporal self-similarities. IEEE Trans. Pattern Anal. Mach. Intell. **33**(1), 172–185 (2011)
36. Khan, S.M., Shah, M.: A multi-view approach to tracking people in dense crowded scenes using a planar homography constraint. In: Proceedings of Workshop on Human Motion, Graz, Austria, pp. 133–146 (2006)
37. Khan, S., Shah, M.: Tracking multiple occluding people by localizing on multiple scene planes. IEEE Trans. Pattern Anal. Mach. Intell. **31**(3), 505–519 (2009)
38. Khan, S.M., Shah, M.: Tracking multiple occluding people by localizing on multiple scene planes. IEEE Trans. Pattern Anal. Mach. Intell. **31**(3), 505–519 (2009)
39. Khan, Z., Balch, T., Dellaert, F.: MCMC-based particle filtering for tracking a variable number of interacting targets. IEEE Trans. Pattern Anal. Mach. Intell. **27**(11), 1805–1918 (2005)
40. Kim, J., Grauman, K.: Observe locally, infer globally: a space-time MRF for detecting abnormal activities with incremental updates. In: Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition, pp. 2921–2928 (2009)
41. Kratz, L., Nishino, K.: Spatio-temporal motion pattern modelling of extremely crowded scenes. In: The 1st International Workshop on Machine Learning for Vision-Based Motion Analysis, Marseille, France (2008)
42. Kratz, L., Nishino, K.: Anomaly detection in extremely crowded scenes using spatio-temporal motion pattern models. In: Proceedings of the Conference on Computer Vision and Pattern Recognition, Maimi Beach, Florida, pp. 1446–1453 (2009)
43. Kratz, L., Nishino, K.: Tracking with local spatio-temporal motion patterns in extremely crowded scenes. In: Proceedings of the Conference on Computer Vision and Pattern Recognition, San Francisco, USA, pp. 693–700 (2010)
44. Kwak, S., Nam, W., Han, B., Han, J.H.: Learning occlusion with likelihoods for visual tracking. In: 2011 IEEE International Conference on Computer Vision (ICCV), pp. 1551–1558 (2011)
45. Lee, L., Romano, R., Stein, G.: Monitoring activities from multiple video streams: establishing a common coordinate frame. IEEE Trans. Pattern Anal. Mach. Intell. **28**(8), 758–767 (2000)
46. Li, M., Zhang, Z., Huang, K., Tan, T.: Rapid and robust human detection and tracking based on omega-shape features. In: 2009 16th IEEE International Conference on Image Processing (ICIP), pp. 2545–2548 (2009)
47. Li, J., Lu, X., Ding, L., Lu, H.: Moving target tracking via particle filter based on color and contour features. In: 2010 2nd International Conference on Information Engineering and Computer Science (ICIECS), pp. 1–4. IEEE, New York (2010)
48. Luber, M., Stork, J.a., Tipaldi, G.D., Arras, K.O.: People tracking with human motion predictions from social forces. In: 2010 IEEE International Conference on Robotics and Automation, pp. 464–469. IEEE, New York (2010)
49. Lucas, B.D., Kanade, T.: An iterative image registration technique with an application to stereo vision. In: Proceedings of Image Understanding Workshop, pp. 121–130 (1981)
50. Ma, Y., Cisar, P.: Activity representation in crowd. In: Proceedings of the 2008 Joint IAPR International Workshop on Structural, Syntactic, and Statistical Pattern Recognition, Florida, USA, pp. 107–116. Springer, Berlin (2008)

51. Ma, L., Liu, J., Wang, J., Cheng, J., Lu, H.: A improved silhouette tracking approach integrating particle filter with graph cuts. In: 2010 IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP), pp. 1142–1145. IEEE, New York (2010)
52. Mahadevan, V., Li, W., Bhalodia, V., Vasconcelos, N.: Anomaly detection in crowded scenes. In: Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition, San Francisco, pp. 1975–1981 (2010)
53. Mehran, R., Oyama, A., Shah, M.: Abnormal crowd behavior detection using social force model. In: Proceedings of the Conference on Computer Vision and Pattern Recognition, Maimi Beach, Florida, pp. 935–942. IEEE, New York (2009)
54. Michel, B., Gianluca, A., Mats, W.: Behavioural dynamics for pedestrians. In: Lecture Notes in Computer Science, pp. 1–18 (2003)
55. Mittal, A., Davis, L.S.: M2Tracker: a multi-view approach to segmenting and tracking people in a cluttered scene. Int. J. Comput. Vis. **51**(3), 189–203 (2003)
56. Moore, B.E., Ali, S., Mehran, R., Shah, M.: Visual crowd surveillance through a hydrodynamics lens. Commun. ACM **54**(12), 64–73 (2011)
57. Ni, Z., Sunderrajan, S., Rahimi, A., Manjunath, B.: Distributed particle filter tracking with online multiple instance learning in a camera sensor network. In: 2010 17th IEEE International Conference on Image Processing (ICIP), pp. 37–40. IEEE, New York (2010)
58. Okuma, K., Taleghani, A., Freitas, N.D., Little, J.J., Lowe, D.G.: A boosted particle filter: multitarget detection and tracking. In: Proceedings of Eighth European Conference on Computer Vision, pp. 28–39. IEEE, New York (2004)
59. Qi, Z., Ting, R., Husheng, F., Jinlin, Z.: Particle filter object tracking based on Harris-SIFT feature matching. Proc. Eng. **29**, 924–929 (2012)
60. Rani, M., Arumugam, G.: An efficient gait recognition system for human identification using modified ICA. Int. J. Comput. Sci. Inf. Technol. **2**(1), 55–67 (2010)
61. Reddy, V., Sanderson, C., Lovell, B.: Improved anomaly detection in crowded scenes via cell-based analysis of foreground speed, size and texture. In: MLvMA Workshop, IEEE Conference on Computer Vision and Pattern Recognition, Colorado Springs, USA, pp. 57–63. IEEE, New York (2011)
62. Revathi, A., Kumar, D.: A review of human activity recognition and behaviour understanding in video surveillance. Comput. Sci. Inf. Technol. **2**, 375–384 (2012)
63. Rodriguez, M., Ali, S., Kanade, T.: Tracking in unstructured crowded scenes. In: Proceedings of IEEE International Conference on Computer Vision, Kyoto, Japan, pp. 1389–1396. IEEE, New York (2009)
64. Saleemi, I., Hartung, L., Shah, M.: Scene understanding by statistical modeling of motion patterns. In: Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition, San Francisco, pp. 2069–2076 (2010)
65. Scott, S.: A system for tracking and recognizing multiple people with multiple camera. Technical report GIT-GVU-98-25, Georgia Institute of Technology (1998)
66. Shechtman, E., Irani, M.: Matching local self-similarities across images and videos. In: 2007 IEEE Conference on Computer Vision and Pattern Recognition, pp. 1–8. IEEE, New York (2007)
67. Shu-hong, C., Chun-hai, H.: Particle filter tracking algorithm based on multi-information fusion. In: 2009 International Conference on Information Engineering and Computer Science, ICIECS 2009, pp. 1–4. IEEE, New York (2009)
68. Sodemann, A.A., Ross, M.P., Borghetti, B.J.: A review of anomaly detection in automated surveillance. IEEE Trans. Syst. Man Cybern., Part C, Appl. Rev. **42**(6), 1257–1272 (2012)
69. Sugano, H., Miyamoto, R.: Parallel implementation of pedestrian tracking using multiple cues on GPGPU. In: 2009 IEEE 12th International Conference on Computer Vision Workshops, ICCV Workshops, pp. 900–906. IEEE, New York (2009)
70. Tang, S.L., Kadim, Z., Liang, K.M., Lim, M.K.: Hybrid blob and particle filter tracking approach for robust object tracking. Proc. Comput. Sci. **1**(1), 2549–2557 (2010)
71. Thida, M., Eng, H.L., Monekosso, D.N., Remagnino, P.: Learning video manifold for segmenting crowd events and abnormality detection. In: Proceedings of 10th Asian Conference on Computer Vision, pp. 439–449. Springer, Berlin (2010)

72. Vezzani, R., Grana, C., Cucchiara, R.: Probabilistic people tracking with appearance models and occlusion classification: the ad-hoc system. Pattern Recognit. Lett. **32**(6), 867–877 (2011)
73. Vishwakarma, S., Agrawal, A.: A survey on activity recognition and behaviour understanding in video surveillance. Vis. Comput. (September) (2012)
74. Walk, S., Majer, N., Schindler, K., Schiele, B.: New features and insights for pedestrian detection. In: 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 1030–1037. IEEE, New York (2010)
75. Wang, L., Hu, W., Tan, T.: Recent developments in human motion analysis. Pattern Recognit. **36**(3), 585–601 (2003)
76. Wang, X., Tieu, K., Grimson, E.: Learning semantic scene models by trajectory analysis. In: Proceedings of European Conference on Computer Vision, vol. 3, pp. 110–123 (2006)
77. Wang, P., Li, W., Zhu, W., Qiao, H.: Object tracking with serious occlusion based on occluder modeling. In: 2012 International Conference on Mechatronics and Automation (ICMA) pp. 1960–1965 (2012)
78. Wu, P., Kong, L., Zhao, F., Li, X.: Particle filter tracking based on color and SIFT features. In: 2008 International Conference on Audio, Language and Image Processing, pp. 932–937. IEEE, New York (2008)
79. Wu, S., Moore, B.E., Shah, M.: Chaotic invariants of Lagrangian particle trajectories for anomaly detection in crowded scenes. In: Proceedings of the Conference on Computer Vision and Pattern Recognition, San Francisco, CA, USA, pp. 2054–2060. IEEE, New York (2010)
80. Yamaguchi, K., Berg, A.C., Ortiz, L.E., Berg, T.L.: Who are you with and where are you going? In: Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition, Colorado Springs, CO, pp. 1345–1352. IEEE, New York (2011)
81. Yang, Y., Liu, J., Shah, M.: Video scene understanding using multi-scale analysis. In: Proceedings of IEEE International Conference on Computer Vision, Kyoto, Japan, pp. 1669–1676. IEEE, New York (2009)
82. Yilmaz, A., Javed, O., Shah, M.: Object tracking: a survey. ACM Comput. Surv. **38**(4) (2006)
83. Zhan, B., Monekosso, D.N., Remagnino, P., Velastin, S.A., Xu, L.Q.: Crowd analysis: a survey. Mach. Vis. Appl. **19**(5–6), 345–357 (2008)
84. Zhang, T., Ghanem, B., Liu, S., Ahuja, N.: Robust visual tracking via structured multi-task sparse learning. Int. J. Comput. Vis. 1–17 (2012)
85. Zhang, C., Xu, J., Beaugendre, A., Goto, S.: A klt-based approach for occlusion handling in human tracking. In: Picture Coding Symposium (PCS), 2012, pp. 337–340 (2012)
86. Zhong, Q., Qingqing, Z., Tengfei, G.: Moving object tracking based on codebook and particle filter. Proc. Eng. **29**, 174–178 (2012)

Privacy and Security in Video Surveillance

Thomas Winkler and Bernhard Rinner

Abstract Video surveillance systems are usually installed to increase the safety and security of people or property in the monitored areas. Typical threat scenarios are robbery, vandalism, shoplifting or terrorism. Other application scenarios are more intimate and private such as home monitoring or assisted living. For a long time, it was accepted that the potential benefits of video surveillance go hand in hand with a loss of personal privacy. However, with the on-board processing capabilities of modern embedded systems it becomes possible to compensate this privacy loss by making security and privacy protection inherent features of video surveillance cameras. In the first part of this chapter, we motivate the need for the integration of security and privacy features, we discuss fundamental requirements and provide a comprehensive review of the state of the art. The second part presents the TrustCAM prototype system where a dedicated hardware security module is integrated into a camera system to achieve a high level of security. The chapter is concluded by a summary of open research issues and an outlook to future trends.

1 The Need for Security and Privacy Protection

Reasons for deploying video surveillance systems are manifold. Frequently mentioned arguments are ensuring public safety, preventing vandalism and crime as well as investigating criminal offenses [40]. As part of that, cameras are often installed in public environments such as underground or train stations, in buses [39] or taxis [20], along roads and highways [8, 23], in sports stadiums or in shopping malls [30, 31]. But video surveillance is no longer deployed only in public but also in private and more intimate environments. For example, in assisted living applications [10, 25, 62] cameras are used to monitor the interior of people's homes to detect unusual behavior of residents.

T. Winkler (✉) · B. Rinner
Institute of Networked and Embedded Systems, and Lakeside Labs, Alpen-Adria Universität Klagenfurt, Lakeside Park B02b, 9020 Klagenfurt, Austria
e-mail: thomas.winkler@aau.at

B. Rinner
e-mail: bernhard.rinner@aau.at

P.K. Atrey et al. (eds.), *Intelligent Multimedia Surveillance*,
DOI 10.1007/978-3-642-41512-8_3,

A major driving factor for this widespread deployment of cameras is that video surveillance equipment has become increasingly cheap and simple to use. As part of this development, today's video surveillance systems are no longer the closed, single-purpose systems they used to be. Modern systems are highly flexible which is primarily achieved via software. Camera devices usually come with powerful operating systems such as Linux as well as a variety of software libraries and applications running on top of it. Furthermore, these systems frequently make use of wireless network interfaces and are part of larger, often public, networks such as the Internet. The increasing size of the software stack and the relative openness of the network infrastructure turn many of today's video surveillance systems into attractive targets for both casual as well as professional attackers.

With the performance of modern embedded camera systems, it is possible to make privacy protection an inherent feature of a surveillance camera. Sensitive data can be protected by various approaches including blanking, obfuscation or encryption. On-camera privacy protection is a clear advantage over server-side protection since it eliminates many potential attack scenarios during data transmission. When considering the software stack of an embedded camera system, privacy-protection is typically implemented at the application level. As a consequence, it is important to detect and avoid manipulations of the underlying software components such as the operating system or system libraries. Otherwise, an attacker might be able to manipulate the system and get access to sensitive data before privacy protection is applied. Depending on the application context, security guarantees such as integrity and authenticity are not only relevant for the system's software stack but also for delivered data. This is especially true for enforcement applications where captured images might serve as evidence at court.

1.1 Security and Privacy Requirements

This section discusses the main security requirements for video surveillance applications. Making a camera system more secure not only offers benefits for camera operators. It is of equal importance for monitored persons. While this is obvious for aspects such as confidentiality, this also holds for, for example, integrity of video data. If integrity is not protected, an attacker could modify video data in a way that intentionally damages the reputation of persons. The integration of the following basic security functionality is also a fundamental requirement for the design of high-level privacy protection techniques.

Integrity. Image data coming from a camera can be intentionally modified by an attacker during transmission or when stored in a database. Using checksums, digital signatures and watermarks, data integrity can be ensured. An often overlooked issue is that integrity protection is not only important for single frames but also for sequences. Simple reordering of images can substantially change the meaning of a video.

Authenticity. In many applications such as traffic monitoring and law enforcement, the origin of information is important. In visual surveillance, this is equivalent

to knowing the identity of the camera that captured a video stream. This can be achieved by explicitly authenticating the cameras of a network and embedding this information into the video streams.

Freshness and Timestamping. To prevent replay attacks where recorded videos are injected into the network to replace the live video stream, freshness of image data must be guaranteed. Even more importantly, in many areas such as enforcement applications, evidence is required when a video sequence was recorded. Explicit timestamping of images not only answers the question when an image was taken, but at the same time also satisfies the requirement for image freshness guarantees.

Confidentiality. It must be ensured that no third party can eavesdrop on sensitive information that is exchanged between cameras or sent from the cameras to a monitoring station. Confidentiality must not only be provided for image and video data transmitted over the network but also for videos that, for example, are stored on a camera to be transmitted at a later point in time. A common approach to ensure confidentiality is data encryption.

Privacy. In video surveillance, privacy can be defined as a subset of confidentiality. While confidentiality denotes the protection of all data against access by third parties, privacy means the protection of data against legitimate users of the system. For example, a security guard needs access to video data as part of her/his job. However, the identities of monitored persons are not required to identify unusual behavior. Privacy protection therefore can be interpreted as providing limited information to insiders while withholding sensitive, identity-revealing data.

Access Authorization. Access to confidential image data must be limited to persons with adequate security clearance. For access to highly sensitive data, involvement of more than one operator should be required to prevent misuse. If a video stream contains different levels of information (e.g., full video, annotations, ...), access should be managed separately for each level. Finally, all attempts to access confidential data should be logged.

Availability. A camera network should provide certain guarantees about availability of system services under various conditions. Specifically, reasonable resistance against denial of service attacks should be provided.

Clearly, these security properties are partially interdependent. It is, for example, meaningless to provide data confidentiality without implementing appropriate authorization mechanisms for accessing confidential data.

2 State of the Art

This section first presents an overview of the state of the art on security in video surveillance (Sect. 2.1). It is followed by a discussion of approaches towards privacy protection (Sect. 2.2). Section 2.3 summarizes our observations and outlines open issues for future research.

2.1 Video Surveillance Security

Serpanos and Papalambrou [52] provide an extensive introduction to security issues in the domain of smart cameras. They discuss the need for confidentiality, integrity, freshness and authenticity for data exchanged between cameras. The authors acknowledge that embedded systems might not have sufficient computing power to protect all data using cryptography. In such a situation, they propose concentrating on protecting the most important data. This work also recognizes the partial overlap of confidentiality and privacy protection and emphasizes the importance of data protection not only against external attackers but also against legitimate system operators.

Senior et al. [51] discuss critical aspects of a secure surveillance system including what data is available and in what form (e.g., raw images vs. metadata), who has access to data and in what form (e.g., plain vs. encrypted) and for how long it is stored. Data confidentiality is ensured via encrypted communication channels. Privacy protection is addressed by re-rendering sensitive image regions. The resulting, multiple video streams contain different levels of data abstraction and are separately encrypted.

Schaffer and Schartner [49] present a distributed approach to ensure confidentiality in a video surveillance system. They propose that the video stream is encrypted using a hybrid cryptosystem. Encryption is performed for full video frames without differentiating between sensitive and non-sensitive image regions. A single system operator is not able to decrypt a video but multiple operators have to cooperate. This property is achieved by the fact that every operator is in possession of only a part of the decryption key.

Integrity protection of image and video data is an important security aspect. It can be addressed by means of, for example, hash functions together with digital signatures or by embedding watermarks into the video content. An important design decision is whether the integrity protection technique is tolerant towards certain, acceptable image modifications or not. The work of Friedman [27] aims at "restoring credibility of photographic images" and therefore does not accept any image modifications. Specifically, authenticity and integrity of images taken with a digital still image camera should be ensured. This is achieved by extending the camera's embedded microprocessor with a unique, private signature key. This key is used to sign images before they are stored on mass storage. The public key required for verification is assumed to be made available by the camera manufacturer. Friedman suggests that the software required for signature verification should be made publicly available. This work can be seen as one of the earliest approaches towards a trustworthy, digital camera system.

Qusquater [43] et al. propose an approach for integrity protection and authentication for digital video stored on tape in the DV format. They use SHA-1 to compute the hash of the image. To be less sensitive to transmission or tape errors, the authors suggest that the images are divided into blocks that are hashed separately. Authenticity is ensured by signing the hash values. The hash of the previous image is also included in the signature to maintain correct ordering of video frames.

Atrey et al. [2, 3] present a concept to verify the integrity of video data. In their work, they differentiate between actual tampering and benign image modifications. In this context, operations that do not change the video semantically such as image enhancements or compression are defined as acceptable. Tampering of video data is divided into spatial and temporal modifications. Spatial tampering includes content cropping as well as removal or addition of information. Temporal tampering refers to dropping or reordering of frames which might result from, for example, network congestion. The authors argue that temporal tampering is acceptable as long as the semantic meaning of the video is not substantially affected. The proposed algorithm is based on a configurable, hierarchical secret sharing approach. It is shown to be tolerant to benign image modifications while tampering is detected.

He et al. [29] also discuss the design of a video data integrity and authenticity protection system. In contrast to other approaches, they do not operate on frames but on objects. Objects are separated from the video background using segmentation techniques. An advantage of this approach is that network bandwidth can be saved by transmitting primarily object data while background data is updated less frequently. Similar to Atrey et al. [2, 3], the authors require their integrity protection system to tolerate certain modifications such as scaling, translation or rotation. Considering these requirements, appropriate features are extracted from the detected objects as well as the background. A hash of these features together with error correction codes is embedded into the video stream as a digital watermark.

Digital watermarks are a popular technique to secure digital media content. A watermark is a signal that is embedded into digital data that can later be detected, extracted and analyzed by a verifier. According to Memon and Wong [36], a watermark can serve different purposes. This can be proof of ownership where a private key is used to generate the watermark. Other applications are authentication and integrity protection, usage control and content protection. Depending on the application domain, watermarks can be visible or invisible. When used for integrity protection, watermarks have the advantage that they can be designed such that they are robust against certain image modifications such as scaling or compression [1, 5]. An example where watermarking is used as part of a digital rights management system for a secure, embedded camera is presented by Mohanty [37]. He describes a secure digital camera system that is able to provide integrity, authenticity and ownership guarantees for digital video content. This is achieved using a combination of watermarking and encryption techniques. Due to the high computational effort, a custom hardware prototype based on an FPGA is used to meet the realtime requirements.

2.2 Privacy Protection in Video Surveillance

Cameras allow the field of view of observers to be extended into areas where they are not physically present. This “virtual presence” of an observer is not necessarily noticed by monitored persons. In the resulting, but misleading feeling of privacy, persons might act differently than they would in the obvious presence of other people. This example makes it apparent, that privacy in video surveillance is an issue

that needs special consideration. But when trying to identify what forms of privacy protection are appropriate, the picture becomes less clear. One reason is that there is no common definition of privacy. As discussed in [38, 51], the notion of privacy is highly subjective and what is acceptable depends on the individual person as well as cultural attitudes.

As pointed out by Cavallaro [12] or Fidaleo et al. [24], it is usually more important to be able to observe the behavior of a person than knowing the actual identity. This is achieved by identification and obfuscation of personally identifiable information such as people's faces [15, 35]. Only in situations where, for example, a law was violated, is this personal information is interesting and should be made available to authorized parties. The main challenge of such an approach is to determine which image regions are actually sensitive. As Saini et al. [45] argue, video data not only includes direct identifiers such as human faces but also quasi identifiers. These quasi identifiers are often based on contextual information and allow to infer the identity of persons with a certain probability. Such basic contextual information about an event includes, for example, what happened, where did it happen and when did it happen. Vagts et al. [59, 60] present an approach that addresses privacy protection not at the sensor level but at a higher abstraction level. As part of their task-oriented privacy enforcement system, data is only collected if it is required for a surveillance task. For that purpose, each task must be fully specified before data collection is started.

In the following paragraphs, we outline key aspects of privacy protection systems. They include basic protection techniques, multilevel approaches that support the recovery of unprotected data under controlled conditions and the need for involving monitored people by asking for their consent and giving them control over their personal data.

Privacy Protection Techniques. A common approach for privacy protection is the identification of sensitive image regions such as human faces of vehicle license plates. If this system component does not work reliably, privacy is at risk. A single frame of a video sequence where sensitive regions are not properly detected can break privacy protection for the entire sequence. Once the sensitive regions have been identified, different techniques can be applied to achieve de-identification. A very basic approach is blanking where sensitive regions are completely removed. An observer only can monitor the presence and the location of a person. Cheung et al. [16] apply video inpainting techniques to fill the blank areas with background. This way, an observer can no longer notice that information was removed from the video.

An alternative to simple blanking are obfuscation and scrambling where the level of detail in sensitive image regions is reduced such that persons can no longer be identified while their behavior remains perceptible. Researchers apply different techniques including mosaicing, pixelation, blurring [18, 61] or high, lossy compression. Work by Gross et al. [28] indicates the overall protection capabilities of such naive mechanisms are relatively low. A study by Boyle et al. [7] on the effects of filtered video on awareness and privacy indicates that pixelation provides better privacy protection than blurring. Another technique to protect

sensitive image regions is scrambling. In its basic form, JPEG compressed images are obscured by pseudo-randomly modifying the DCT coefficients [21] of sensitive regions.

Abstraction techniques replace sensitive image regions with, for example, bounding boxes or, in case of persons, with avatars, stick-figures and silhouettes [51]. Another form of abstraction is meta-information attached to a video. This can be object properties such as position and dimensions, but also names of identified persons [54]. Depending on the type of abstraction, either behavior, identity or both can be preserved. Identities should be protected using encryption.

Data encryption is used by many systems to protect sensitive regions. When encrypted, regions of interest can no longer be viewed by persons who do not have the appropriate decryption keys. Simple encryption not only protects the identity of monitored persons but also their behavior. Upon decryption, both—identity and behavior—are revealed. By using multiple encryption keys or split keys as described in [49], a system can be designed that requires multiple operators to cooperate to decrypt the original data. Such a design provides a certain degree of protection against operator misuse.

Multilevel Privacy Protection. Support for multiple privacy levels denotes that one single video stream contains different levels of information. These could range from the unmodified, sensitive image regions over obfuscated versions with blurred faces to abstracted versions. Depending on their sensitivity, these levels can be separately encrypted with one or more individual encryption keys. A multilevel approach allows a privacy protection system to be designed that presents different types of information to observers depending on their security clearance. While low-privileged operators can only access the version of the stream where behavioral data is visible, supervisors or government agencies could get access to the original data that contains the identity of monitored persons.

Consent and Control. Ideally, monitored people should first be asked for consent before they are captured by a video surveillance system. Today, installed cameras are often marked with signs or stickers that advertise their presence. User consent to video surveillance is given implicitly by acknowledging these signs when entering the area. As these signs are easily overlooked, consent should be sought more actively. Users could be automatically notified about presence and properties of cameras, for example, via their smartphone. Moreover, monitored people should remain in control of personal data captured by the system. If data is disclosed to a third party, explicit user permission should be required.

Some of these requirements have been addressed in research prototypes. By handing out dedicated devices or RFID tags to known and trusted users, a stronger form of awareness about video surveillance is realized [9, 61]. Users equipped with such devices are not only made aware of the installed cameras but even get a certain degree of control over their privacy. Cameras recognize them as trustworthy and remove or protect the corresponding image regions. The approach of Cheung et al. [17] goes even further. Using public key cryptography to

protect personal information, users get full control over their privacy-sensitive data since they have to actively participate in the decryption of this data.

Cavallaro [11, 12] emphasizes that digitalization of video surveillance introduces new privacy threats. Therefore, personal and behavioral data should be separated directly on the camera. While system operators only get access to behavioral data, a separate stream containing personal data is made available to law enforcement authorities. A benefit of this strict separation is prevention of operator misuse. Similar ideas are discussed in the already mentioned work of Senior et al. [51]. They suggest that privacy is protected by extracting sensitive information and re-rendering the video into multiple streams individually protected by encryption.

Fleck [25, 26] employs smart cameras from Matrix Vision in an assisted living scenario. The cameras are used to monitor the behavior of persons and detect unusual behavior such as a fall. For that purpose, the cameras create a background model which is the basis for detecting motion regions. Detected objects are tracked and their behavior is analyzed using support vector machines. Privacy protection is achieved by either transmitting only event information or replacing detected objects with abstracted versions. It is assumed that the camera's housing is sealed such that manipulation can be detected by the camera and leads to a termination of its services. Protection against software attacks such as integrity checks or data encryption is not part of the current system.

Boult [6] argues that many existing approaches are targeted at removing privacy-sensitive image data without providing mechanisms to reconstruct the original image. Based on this observation, he proposes a system called PICO that relies on cryptography to protect selected image regions such as faces. It allows the actions of a person to be monitored without revealing the person's identity. The faces are only decrypted if, for example, a crime was committed by the person. Encryption is performed as part of image compression and uses a combination of symmetric and asymmetric cryptography. Additionally, it is suggested that checksums of frames or sub-sequences are computed to ensure data integrity. In related work, Chattopadhyay and Boult present PrivacyCam [14], a camera system based on a Blackfin DSP clocked at 400 MHz, 32 MB of SDRAM and an Omnivision OV7660 color CMOS sensor. uClinux is used as operating system. Regions of interest are identified based on a background subtraction model and resulting regions are encrypted using an AES session key. Rahman et al. [44] also propose that regions of interest are encrypted. In their approach they do not rely on established crypto-systems but propose that chaos cryptography is used.

Moncrieff et al. [38] argue that most of the proposed systems rely on predefined security policies and are either too intrusive or too limited. Therefore, they suggest that dynamic data hiding techniques are applied. Via context-based adaptation, the system could remove or abstract privacy-sensitive information during normal operation while in case of an emergency, the full, unmodified video stream is automatically made available. This way, the system remains usable for the intended purpose but protects privacy during normal operation.

Dufaux and Ebrahimi [21] suggest scrambling of sensitive image regions. After detection of relevant areas, images are transformed using DCT. The signs of the co-

efficients of sensitive regions are then flipped pseudo-randomly. The seed for the pseudo-random number generator is encrypted. Decryption is only possible for persons who are in possession of the corresponding decryption key. According to the authors, the main benefits are minimal performance impact and that video streams with scrambled regions can still be viewed with standard players. A study by Dufaux and Ebrahimi [22] indicates that scrambling is superior to simple approaches such as pixelation and blurring.

A similar approach is discussed by Baaziz et al. [4] where, in a first step, motion detection is performed followed by content scrambling. To ensure data integrity, an additional watermark is embedded into the image which allows detection of manipulation of image data. Limited reconstruction of manipulated image regions is possible due to redundancy introduced by the watermark. Yabuta et al. [68] also propose a system where DCT encoded image data is modified. They, however, do not scramble regions of interest but extract them before DCT encoding and encrypt them. These encrypted regions are then embedded into the DCT encoded background by modifying the DCT coefficients. Li et al. [32] present an approach towards recoverable privacy protection based on discrete wavelet transform. Information about sensitive image regions together with their wavelet coefficients are protected with a secret key. Data hiding techniques are used to embed this information into the resulting image.

Qureshi [42] proposes a framework for privacy protection in video surveillance based on decomposition of raw video into object-video streams. Based on a segmentation approach, pedestrians are identified. Tracking is performed using color features. The privacy of detected persons is protected by selectively rendering the corresponding objects. Advanced protection mechanisms such as encryption are left as future work. Also the system presented by Tansuriyavong and Hanaki [54] is based on detection of sensitive entities. In an office scenario, the silhouettes of detected persons are blanked. Additionally, the system integrates face recognition to identify previously registered persons. Configuration options allow the choice of what information should be disclosed—full images, silhouettes, names of known persons or any combination thereof.

Troncoso-Pastoriza et al. [56] propose a generic video analysis system that is coupled with a Digital Rights Management (DRM) system. By exploiting the hierarchical structure of MPEG-4, the authors propose selective visualization of video objects either in clear or in obfuscated forms. Access to sensitive video objects is conditionally granted depending on the rights of the observer and the individual policies of monitored users. Sensitive content is protected by encryption. Intellectual Property Management Protection (IPMP) descriptors, as standardized in MPEG-4, are used to describe these encrypted streams. Access rights to protected video objects are formulated using the MPEG-21 Rights Expression Language (REL).

Finally, the Networked Sensor Tapestry (NeST) software architecture by Fidaleo et al. [24], represents a more generic privacy protection approach. Its design is not limited to videos and images but can handle arbitrary sensor data. The system uses a centralized architecture. An important component is the privacy buffer that is running on the server. Data received from the clients is fed into this privacy buffer. The buffer can be extended and configured by means of privacy filters and a privacy grammar. If incoming data is qualified as private by one of the privacy filters,

the data does not leave the privacy buffer. Non-private data is forwarded to a routing component that manages distribution of data to interested clients.

To protect the privacy of only selected users, systems have been presented that allow to remove known, trusted users from captured video. Due to the limited reliability of computer vision to detect personal image data, many researchers rely on portable devices carried by users for identification and localization. One such approach is presented by Brassil [9]. He proposes a Privacy Enabling Device (PED) that gives users control over their personal data. When activated, the PED records the location of the person together with timestamps. This data is uploaded to a clearinghouse. Before a camera operator discloses videos to a third party, the clearinghouse has to be contacted to check if an active PED was in the vicinity of the camera at the time in question. If so, video data has to be anonymized. Due to the absence of feedback, users have to trust camera operators to follow the advertised procedures.

Wickramasuriya et al. [61] perform realtime monitoring of the environment to increase user privacy. In particular, they suggest that motion sensors are used to monitor rooms or areas. If motion is detected, an RFID reader is triggered that tries to read the RFID tag carried by the person that entered the area. If no RFID tag can be found or the security level of the tag does not grant access to the area, a camera that oversees the region is activated. Image regions containing persons with valid RFID tags are blanked such that only potential intruders remain visible.

Chinomi et al. [18] also use RFID technology to detect known users. RFID readers, deployed together with cameras, are used to localize RFID tags carried by users based on signal strength. This location information is then mapped to motion regions detected by the cameras. As the RFID tag identifies the person, the individual privacy policy can be retrieved from a database. This policy defines the relationship between the monitored person and potential observers. Based on that, different forms of abstracted data are delivered by the system. Abstractions include simple dots showing only the location of a person, silhouettes as well as blurred motion regions. Also Cheung et al. [17] use RFID for user localization. Corresponding motion regions are extracted from the video and encrypted with the user's public encryption key. This key is retrieved from a database via the user ID from the RFID tag. The blanked regions in the remaining image are filled with background image data using video inpainting [16]. The encrypted regions are embedded into the compressed background image using data hiding techniques similar to steganography. Since decryption of privacy-sensitive image regions requires the user's private key, active user cooperation is necessary to reconstruct the original image. A dedicated mediator establishes contact between users and observers who are interested in the video data. In work from the same research group, Ye et al. [69] and Luo et al. [33] do not use RFID tags for identification but biometric information. As part of their anonymous biometric access control system, iris scanners are installed at the entrances of areas under video surveillance. Based on that, authorized individuals are then obfuscated in the captured video. Anonymity of authorized persons is maintained by using homomorphic encryption.

An approach that does not need electronic devices that are carried by users is presented by Schiff et al. [50]. Their "respectful cameras" use visual markers such as yellow hard hats worn by people to identify privacy-sensitive regions. Specifically,

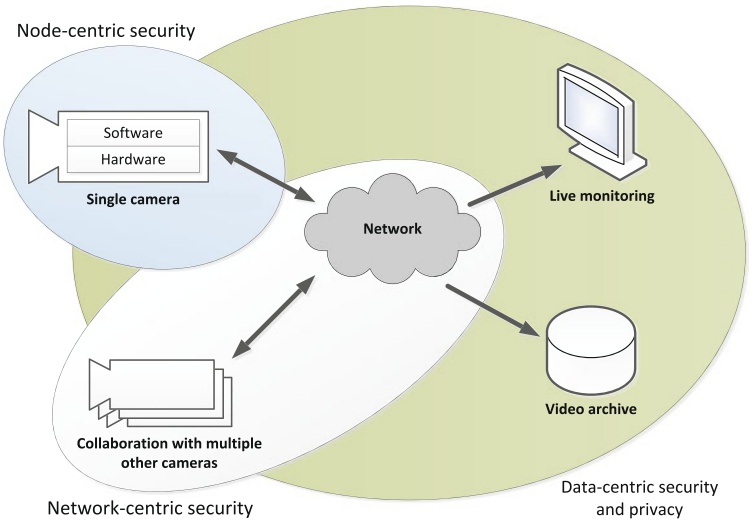


Fig. 1 The security requirements discussed in this chapter can be classified into three groups. First, node-centric security refers to security of the camera's hardware as well as its software stack. Second, network-centric security covers security of the communication channel and security aspects for inter-camera collaboration which include secure data sharing and aggregation techniques, camera discovery, topology control or time synchronization. The third group is data-centric security which denotes security (e.g., integrity, authenticity, etc.) and privacy protection for data from its creation to its deletion

they remove person's faces from images. For marker detection and tracking, probabilistic AdaBoost and particle filtering are used. Spindler et al. [53] apply similar ideas in the context of building automation and monitoring applications. Personal data is obfuscated based on individual privacy settings. For identification and localization, the authors suggest relying on computer vision. For the prototype, this was not implemented but replaced by manual selection of privacy-sensitive regions.

2.3 Observations and Open Issues

Most research on privacy and security in video surveillance is on selected and isolated topics. Figure 1 gives an overview of the three major areas. The majority of work addresses data-centric security and privacy issues which include authenticity and integrity of data, data freshness, timestamping as well as confidentiality. Ideally, data-centric security guarantees should be provided for the entire lifetime of data, i.e., from the moment an image is captured by the camera's sensor until the image and all derived data are deleted. As a consequence, data-centric security involves all components of a visual sensor network including monitoring stations as well as video archives. Adequate access authorization techniques must be integrated such that sensitive data can be accessed only by legitimate users.

When considering the architecture of a VSN node it is apparent that data-centric protection features are implemented typically as part of the camera's applications. To be able to provide meaningful security guarantees for captured and processed data the VSN device itself must be secured. This aspect, which is referred to as node-centric security in Fig. 1, is rarely addressed in related work. In a holistic approach, the security of both the VSN's hardware as well as its software stack must be taken into account. Otherwise, the protection achieved by application level security mechanisms must be questioned.

The third major group of security issues shown in Fig. 1 is network-centric security where a primary goal is a secure channel between two communication partners. This could be two cameras or one camera and a monitoring or archiving facility. A secure communication channel must provide basic non-repudiation and confidentiality properties. To a certain extent, there might be a redundancy between network channel security and data-centric security. The actual protection requirements depend on the specific application. An additional and equally important aspect is secure collaboration of multiple cameras. To facilitate secure collaboration, a range of topics must be considered such as secure data sharing and aggregation, localization and topology control, camera discovery and lookup mechanisms as well as inter-camera time synchronization.

In our review of related work, we identified some of the most important open issues.

Comprehensive Privacy Protection. The meaning of privacy in video surveillance is still a vague term. As discussed previously there is consensus that privacy protection denotes the protection of persons' identities while their behavior remains visible. However, it is not clear if the proposed protection techniques such as pixelation, blurring or scrambling are actually effective. Research by Dufaux and Ebrahimi [22] and Gross et al. [28] indicates that basic obfuscation techniques might provide less protection than previously thought. Additionally, object-based privacy protection mechanisms assume the availability of reliable detection algorithms for the identification of sensitive image regions. A mis-detection in a single frame of a video sequence can be sufficient to breach privacy for the entire sequence. Based on this observations, Saini et al. [47] suggest to rely on global protection techniques instead of object-based approaches. Global approaches apply uniform protection operations (e.g., downsampling, coarse quantization or edge detection) to the entire raw image and are therefore not prone to errors in the detection of sensitive regions.

But identity leakage does not result only from primary identifiers such as human faces. Contextual information [48] such as the location, the time and the observed action can also be sufficient to derive the identity of persons. The usefulness of this contextual information depends directly on the knowledge of the observer. One approach to reduce the likelihood of identity leakage via contextual information is to ensure that monitoring of video data is performed by randomly chosen persons without knowledge about the observed area and context [46]. The practical feasibility of such approaches is yet to be determined.

Regardless of the chosen approach—privacy protection reduces usually the amount of information that is available in a video and therefore privacy protection has a negative impact on system utility. An important aspect will be to explore the privacy vs. system utility design space and to determine a suitable and most probably application specific tradeoff.

Holistic Security Concept. There is still a lack of work that considers security and privacy in VSNs in a holistic way. It is apparent that most security solutions are situated at the application level and that node-centric security is not taken into account. Substantial work has been targeted at data- and network-centric security. But without addressing security of VSN nodes themselves, high-level protection mechanisms are literally built on sand. VSN designers will have to collaborate with engineers from other embedded system domains such as mobile handsets to promote the development of standardized node-centric security solutions.

Sensor-Level Security. Securing the VSN device is an important yet complicated task. On modern embedded camera systems a large amount of software is executed. This includes the operating system with all its subsystems such as the network stack as well as system libraries and middeleware components. Due to the substantial size of these software components it is impractical to fully verify them. As a consequence these components have to be implicitly trusted. One potential approach to address this issue would be to bring security and privacy protection closer to the sensor or even making them part of the sensor. If security and privacy are guaranteed at the sensor level, then the camera and its relatively large software stack would no longer have to be considered as trusted entities. This approach implies two major challenges: First, it is unclear what type of privacy protection is suitable and feasible at the sensor level. Second, sensor-level privacy protection means that image processing and analysis applications on the camera must be adapted to deal with pre-processed and pre-filtered data. A critical question is the identification of an appropriate tradeoff between sensor-level security and privacy protection and the remaining utility of the camera host system.

3 TrustCAM: A Camera with Hardware Security Support

This section describes an approach that specifically addresses two major issues outlined previously in Sect. 2.3: node-centric security and providing data-centric security guarantees for all data that is delivered by the camera. The presented TrustCAM prototype [63–66] puts a strong focus on node security to ensure that high-level data protection algorithms can be built on a solid basis. A fundamental question in computer security is whether a software solution can provide adequate levels of security or if an immutable hardware component is required that acts as a trust anchor. The later is assumed by an industry initiative called Trusted Computing Group (TCG). The main output of the group is a set of open specifications for a hardware chip—the Trusted Platform Module (TPM) [57]—and software infrastructure such as the TCG

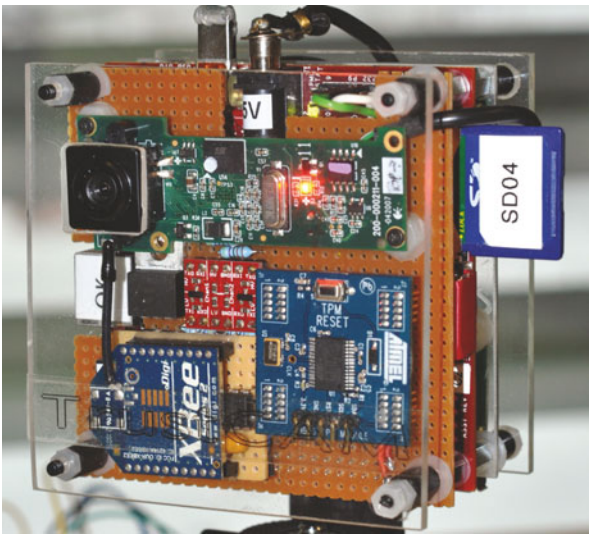


Fig. 2 The TrustCAM prototype. The image sensor, the XBee radio and the Atmel TPM can be seen on the *front circuit board*. Behind this board are the processing board and WiFi radio

Software Stack (TSS) [58]. The TPM chip implements a small and well defined set of core security functions which can not be altered by the TPM's host system. This approach of a hardware-based security solution has been adopted by the TrustCAM project for embedded smart cameras. The TrustCAM prototype as shown in Fig. 2 incorporates an Atmel AT97SC3203S TPM chip which is used to various security aspects including recording the boot process and software state of the camera device, securely storing cryptographic keys or digitally signing and encrypting outgoing data.

The system largely consists of commercial, off-the-shelf components. It is based on the BeagleBoard [55] (rev. C2) embedded processing platform. The board is equipped with an OMAP 3530 SoC from Texas Instruments. The OMAP SoC features a dual-core design and contains an ARM Cortext A8 processor which is clocked at up to 600 MHz and an additional TMS320C64x+ DSP that can run at speeds of up to 480 MHz. For stability reasons, the clock frequency of the TrustCAM's ARM core is set to 480 MHz. The DSP is not used in the current version of the prototype. The prototype is equipped with 256 MB of LPDDR RAM and 256 MB NAND flash memory. A CMOS image sensor (Logitech QuickCam Pro 9000) is connected via USB. Wireless connectivity is provided by an RA-Link RA-2571 802.11b/g WiFi adapter. An additional, low-performance wireless communication channel is implemented via an 802.15.4 based XBee radio connected to one of the platform's UARTs.

3.1 Trusted Computing Preliminaries

This section provides a brief overview of the most important Trusted Computing (TC) and TPM concepts. More detailed information can be found in the specifica-

tions of the TCG [57] and auxiliary sources [13, 34]. The TPM is typically implemented as a secure microcontroller (execution engine) with accelerators for RSA and SHA-1. Additionally, the TPM provides a random number generator and limited amount of volatile and non-volatile memory. With an Opt-In process, users can choose if they want to make use of the TPM.

RSA keys can be generated for different purposes such as encryption or signing. Upon creation, keys can be declared migratable or not. While migratable keys can be transferred to a different TPM, non-migratable keys can not. Regardless of key type and migratability, a private TPM key can never be extracted from the chip as plaintext but only in encrypted form. By definition, every key must have a parent key that is used to encrypt the key when it has to be swapped out of the TPM due to limited internal memory. At the top of this key hierarchy is the Storage Root Key (SRK) which never leaves the TPM. TC defines three roots of trust:

Root of Trust for Measurement (RTM). In TC, measuring is the process of computing the SHA-1 hash of an application binary before it is executed. Typically starting from an immutable part of the BIOS, a chain of trust is established where each component in the chain is measured before control is passed to it. The measurements are stored inside the TPM in memory regions called Platform Configuration Registers (PCRs). As available memory in the TPM is limited, a special operation called TPM_Extend is used to write to PCRs:

$$PCR[i] \leftarrow \mathit{SHA\text{-}1}\big(PCR[i]||measurement\big).$$

TPM_Extend computes the hash of the current PCR value concatenated with the new measurement. This accumulated value is written back into the PCR.

Root of Trust for Reporting (RTR). Reporting of the platform state is called attestation and is done with the TPM_Quote command. As part of that, PCR values get signed inside the TPM using a key unique to that TPM. In theory, this key could be the Endorsement Key (EK) which is inserted into the TPM upon manufacturing. For privacy reasons however, not directly the EK but alias keys are used. They are called Attestation Identity Keys (AIKs) and are generated with the help of an external, trusted third party.

Root of Trust for Storage (RTS). The RTS allows to use the TPM to securely store data. Binding of data refers to encrypting data with a TPM key and hence guaranteeing that this data only is accessible by this specific TPM instance. Sealing of data allows to specify a set of PCR values the data is associated with. Like unbinding, unsealing can only be done by the specific TPM instance that holds the private sealing key. Additionally, the plaintext is only released if the current PCR values match those specified upon sealing.

3.2 System Architecture and Setup

The primary goals of the TrustCAM system design are to provide authenticity, integrity, freshness and timestamping as well as confidentiality and multilevel pri-

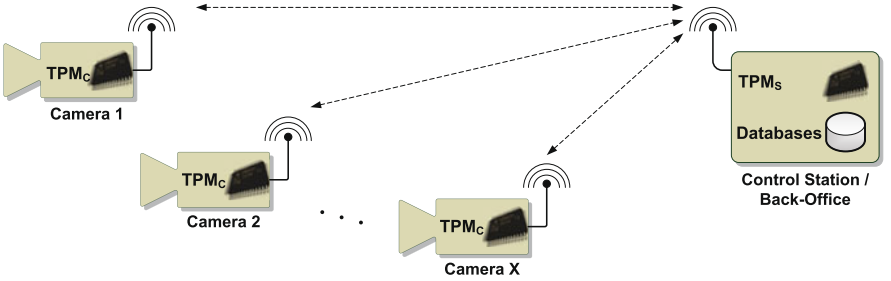


Fig. 3 A network of X TPM-equipped TrustCAMs which are managed by a central control station

vacy protection for streamed image and video data. As illustrated in Fig. 3, each TrustCAM of a visual sensor network (VSN) is assumed to be equipped with a TPM chip subsequently called TPM_C. Throughout the VSN, network connectivity is provided by wireless communication in single or multi-hop mode. For this work, cameras are assumed to be controlled and operated from a central facility subsequently called the Control Station (CS). A fundamental assumption is that the CS is a secure and trustworthy facility.

Figure 3 shows a network consisting of X TrustCAM nodes and one central control station. Not only the cameras, but also the control station is equipped with a TPM subsequently referred to as TPM_S. In addition to TPM_S, the CS also hosts several databases to store cryptographic keys generated during camera setup as well as data received from the cameras.

It is assumed that camera setup is done when cameras are under full control of the operating personnel. The main part of the setup involves the generation of TPM keys on the camera and at the control station. All keys are generated as 2048 bit RSA keys. The following setup steps and the key generation are done individually for each of the X cameras of the network.

TPM Ownership. Initially, the camera's TPM has to be activated. Calling the Take-Ownership operation of TPM_C sets an owner password and generates the Storage Root Key K_{SRK}. The owner secret is not required during normal operation of the camera and is set to a random value unique to every camera. For maintenance operations, the camera's owner secret is stored in the CS database.

Identity Key Creation. An Attestation Identity Key (K_{AIK}) serves as an alias for the TPM's Endorsement Key (K_{EK}) and is used during platform attestation. In contrast to a conventional PC, there are not multiple human users on a TrustCAM. The system software running on the camera takes the role of a single system user. Moreover, all cameras in the network are uniquely identified and well known by the operators. Consequently, there is no need for the anonymity gained by using multiple AIKs in conjunction with a PrivacyCA [41]. Therefore, only a single Attestation Identity Key K_{AIK} is generated during setup that serves for platform attestation. The public part $K_{AIK_{pub}}$ is stored in the CS database.

Signature Key Creation. For signing data such as events or images delivered by the camera, a non-migratable signing key K_{SIG} is created with K_{SRK} as its parent.

Table 1 The cryptographic keys generated during setup of a single camera. The "Control Station" and "TrustCAM" columns denote the storage location of the keys. Binding keys are generated by TPM_S while all other keys are generated by TPM_C. All keys are non-migratable, 2048 bit RSA keys. The *pub* subscript denotes the public RSA key

	Control Station	TrustCAM
Endorsement Key	$K_{EK_{pub}}$	K_{EK}
Storage Root Key	–	K_{SRK}
Attestation Identity Key	$K_{AIK_{pub}}$	K_{AIK}
Signature Key	$K_{SIG_{pub}}$	K_{SIG}
Binding Keys	K_{BIND_1}	$K_{BIND\_1_{pub}}$
	K_{BIND_2}	$K_{BIND\_2_{pub}}$
	...	...
	K_{BIND_N}	$K_{BIND\_N_{pub}}$

Being non-migratable ensures that the private key cannot leave the camera's TPM_C. This provides assurance that data signed with this particular key really originates from this specific camera.

Binding Key Creation. To ensure confidentiality and privacy protection, sensitive image data sent from the camera to the CS has to be encrypted. Encryption should be done at different levels including the full images as well as special regions of interest where, e.g., motion or faces have been detected.

To ensure confidentiality, at least one non-migratable binding key K_{BIND_1} is created by the control station's TPM_S. The public portion of this key, $K_{BIND\_1_{pub}}$, is exported from TPM_S and stored on the camera. Note that the private part of K_{BIND_1} cannot be exported from TPM_S and therefore, data encrypted with $K_{BIND\_1_{pub}}$ can only be decrypted at the CS and not by an intermediate attacker who interferes with the transmission. To decrypt data bound with $K_{BIND\_1_{pub}}$, the usage password of the key has to be supplied by the system operator. To avoid that a single operator who knows this usage password and has access to the control station can decrypt data, additional binding keys K_{BIND_2} to K_{BIND_N} are generated. Privacy sensitive data can be encrypted sequentially with multiple binding keys. Assuming that no single operator knows all the usage secrets for these binding keys, two or more operators have to cooperate to decrypt the data. The N binding keys can be used also to realize different security levels. Data at different abstraction levels (e.g., full images vs. images where people's faces have been removed vs. textual event descriptions) can be encrypted with different binding keys. Depending on security clearance, only certain abstraction levels can be accessed by an operator.

Table 1 summarizes the cryptographic keys that are generated as part of the setup procedure of a single camera.

Once the setup procedure is complete, the camera can be deployed. The boot process of the camera as well as its entire software state including all executed applications is recorded in the PCRs of its TPM_C. To monitor both the availability and the executed applications, we have previously proposed a trusted lifebeat. The involved tursted lifebeat protocols, the mapping of camera timestamps to world time as well as the trusted boot procedure of TrustCAM are fully detailed in [65].

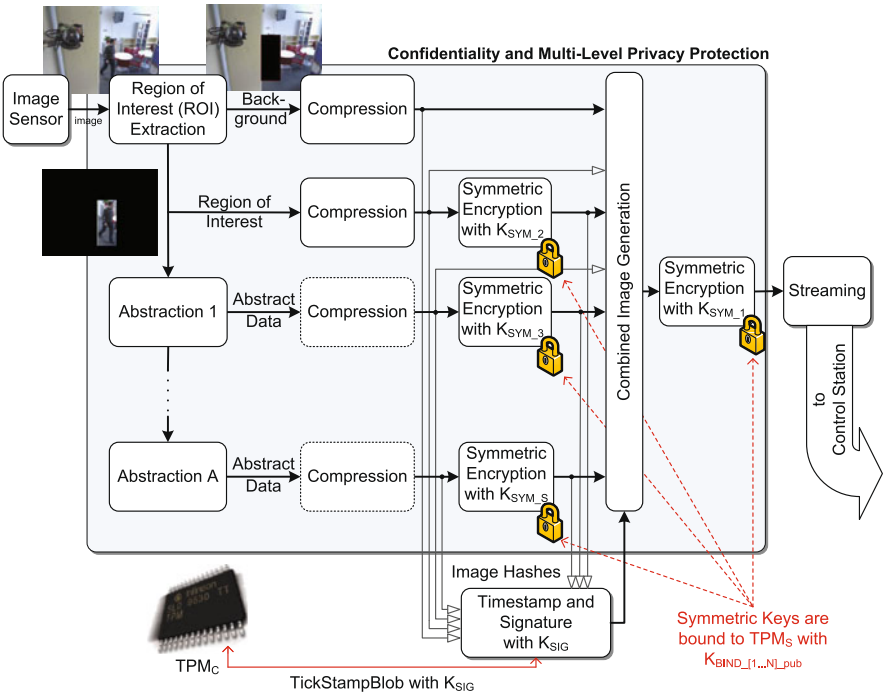


Fig. 4 Captured images are analyzed and regions of interest (ROI) are extracted. Abstracted versions of the ROI, the unmodified ROI as well as the remaining background are separately compressed. The ROI parts of the video stream are encrypted with symmetric session keys that are bound to TPM_S. Hash values of the compressed images and the encrypted ROI images are signed and timestamped by TPM_C. The background image, the encrypted ROI images, the ROI hashes and the signature are combined into a common container which is then encrypted. Subsequently, the video data is streamed to the control station

3.3 Video Confidentiality, Authenticity and Integrity

The TrustCAM system is designed to ensure (1) confidentiality of all image data as a protection against external attackers and (2) selective privacy protection to provide system operators with sufficient information to fulfill their duties without automatically revealing the identity of monitored persons. Furthermore, the proposed design provides (3) authenticity, (4) integrity and (5) timestamping guarantees for delivered data.

The basic concept is shown in Fig. 4. Image data grabbed from the camera's sensor is first analyzed and regions of interest (ROI) are detected. The definition of regions of interest depends on the application and can range from motion areas over vehicle license plates to people's faces. The ROI are then extracted from the image. The remaining background image Img_{BACK} as well as the extracted, original ROI Img_{ROI} are compressed. Additionally, one or more abstracted versions $Img_{ABST\_[1...A]}$ of the ROI are created. Abstracted versions can be images where,

for example, faces are blurred or persons are replaced with stick figures or generic avatars. Alternatively, the output of the abstraction process can be also non-image data such as a textual description. While compression of abstracted data is optional and depends on the actual data type, encryption is mandatory:

$$\begin{aligned} Img_{ROI_{enc}} &= Encrypt_{K_{SYM\_2}}(Img_{ROI}), \\ Img_{ABST\_1_{enc}} &= Encrypt_{K_{SYM\_3}}(Img_{ABST\_1}), \\ &\dots \\ Img_{ABST\_A_{enc}} &= Encrypt_{K_{SYM\_S}}(Img_{ABST\_A}). \end{aligned}$$

Upon startup of the streaming session, the symmetric session keys $K_{SYM\_[2...S]}$ are bound to the control station's TPM_S using the non-migratable binding keys $K_{BIND\_2_{pub}}$ to $K_{BIND\_N_{pub}}$:

$$\begin{aligned} K_{SYM\_2_{bound}} &= Bind_{K_{BIND\_3_{pub}}}\left(Bind_{K_{BIND\_2_{pub}}}(K_{SYM\_2})\right), \\ K_{SYM\_3_{bound}} &= Bind_{K_{BIND\_4_{pub}}}(K_{SYM\_3}), \\ &\dots \\ K_{SYM\_S_{bound}} &= Bind_{K_{BIND\_N_{pub}}}(K_{SYM\_S}). \end{aligned}$$

Binding K_{SYM_2} successively with two independent binding keys enforces the four-eyes principle for the original ROI at the control station where two operators have to cooperate to decrypt the data. Decryption at the control station requires knowledge of the usage passwords of the respective binding keys. Depending on individual security clearance, an operator might be able to, for example, decrypt the background image and an abstracted version of the regions of interest that reveals the behavior of monitored persons. ROI versions that contain a person's identity are reserved for, for example, supervisors with higher clearance. To prevent operator misuse, especially sensitive data can be protected by double-encryption of the symmetric session key such that two operators have to cooperate to decrypt the data. This is illustrated for K_{SYM_2} which is used to encrypt the original ROI. It is protected twice using K_{Bind_2} and K_{Bind_3}.

To couple data integrity and authenticity guarantees with data confidentiality, the encrypt/sign/encrypt approach discussed by Davis [19] is applied. As shown in Fig. 4, the hashes of the plain image regions Img_{BACK}, Img_{ROI} and $Img_{ABST\_[1...A]}$ as well as those of their encrypted equivalents are computed. Including both in the signature demonstrates that the plaintext as well as the ciphertext come from the same origin and provides protection against plaintext substitution attacks. Furthermore, by signing the plaintext, non-repudiation guarantees are given. Additionally, the system operator can correlate the inner encryption with the outer encryption by checking that the used binding keys all belong to the same camera. This protects against potential "surreptitious" forwarding attacks [19].

$$H_{BACK} = SHA\text{-}1(Img_{BACK}),$$

$$H_{ROI} = \texttt{SHA-1}(Img_{ROI}),$$
$$H_{ABST\_[1...A]} = \texttt{SHA-1}(Img_{ABST\_[1...A]}),$$
$$H_{ROI_{enc}} = \texttt{SHA-1}(Img_{ROI_{enc}}),$$
$$H_{ABST\_[1...A]_{enc}} = \texttt{SHA-1}(Img_{ABST\_[1...A]_{enc}}).$$

The individual hash sums are concatenated and a common hash sum H_{Img} is computed:

$$H_{Img} = \texttt{SHA-1}(H_{BACK}||H_{ROI}||H_{ABST\_[1...A]}||H_{ROI_{enc}}||H_{ABST\_[1...A]_{enc}}).$$

Due to performance limitations of current TPM implementations it is impossible to sign and timestamp every image hash H_{Img} individually. Instead, an accumulated hash sum for a group of F frames is computed:

$$AccSum_{Img[1...F]} = \texttt{SHA-1}(AccSum_{Img[1...(F-1)]}||H_{Img}).$$

This accumulated hash sum, the current tick values as well as the accumulated hash sum of the previous image group are then singed and timestamped by the camera's TPM_C:

$$TickStamp_{Res} = \texttt{TPM\_TickStampBlob}_{K_{SIG}}(TSN_{Img_F}||TCV_{Img_F}||TRATE_{Img_F}||$$
$$AccSum_{PrevGroup}||AccSum_{Img[1...F]}).$$

In the next step, the various components are combined into a common image container Img_{COMB}:

$$Img_{COMB} = [ImageParts, ImageHashes, K_{SYM\_[2...S]_{bound}}, Timestamp],$$

with:

$$ImageParts = [Img_{BACK}, Img_{ROI_{enc}}, Img_{ABST\_[1...A]_{enc}}],$$
$$ImageHashes = [H_{ROI}, H_{ABST\_[1...A]}],$$
$$Timestamp = [TickStamp_{Res}, TSN_{Img_F}, TSV_{Img_F}, TRATE_{Img_F}, start_{idx}, end_{idx}].$$

This combined image includes the background image, the encrypted original ROI as well as the encrypted abstracted ROI images. Additionally, it contains the hashes of the original and abstracted ROI images, the bound session keys and, in the case of the end of a frame group, the group's timestamp and signature together with start and end indices. Finally, the combined image Img_{COMB} is encrypted using K_{SYM_1} which, in turn, is bound to TPM_S:

$$Img_{COMB_{enc}} = \texttt{Encrypt}_{K_{SYM\_1}}(Img_{COMB}),$$
$$K_{SYM\_1_{bound}} = \texttt{Bind}_{K_{BIND\_1_{pub}}}(K_{SYM\_1}).$$

Since all image data including the background and the regions of interest as well as the derived abstracted versions are encrypted, confidentiality of all personal information is ensured. This also includes personal information that was accidentally missed by the ROI detection algorithm. Furthermore, using non-migratable signing keys for data singing guarantees data authenticity and integrity. Validation of associated timestamps and the mapping of local camera timestamps to world time is discusses in detail in [65]. In the last step, the encrypted, combined image data and the bound session key are streamed to the control station.

At the control station, a system operator can decrypt the individual image parts depending on the knowledge of the usage passwords of the camera's binding keys. Typically, an operator can only decrypt a subset of the included data. As a consequence, not all hash values of the ROI (H_{ROI}) and abstracted ROI ($H_{ABST[1...A]}$) images can be computed. To still be able to verify the signature of the frame group, the operator can substitute the missing hashes with those from the *ImageHashes* field included in the combined image. This approach allows verification of the overall signature of the frame group as well as the integrity and authenticity of those image parts which are accessible by the operator. The strategy used is based on the star chaining concept for hash values proposed by Wong and Lam [67] and has two main advantages. First, an operator can validate the integrity and authenticity of those image parts he actually sees and has legitimate access to. No decryption of additional image components is required. Second, on the camera only one single hash value (the accumulated H_{Img}) has to be sent to TPM_C for signing and timestamping despite the various individual parts the combined image might contain. This is an important advantage when considering the low performance of current TPM chips.

To illustrate the verification of the timestamp and signature, the following example is given. Operator 1 ($OP1$) at the control station knows the usage secrets for K_{Bind_1} and K_{Bind_4} which gives him access to the background image (Img_{BACK}) and the first abstracted ROI image (Img_{ABST_1}). For signature verification, the control station software computes the hashes of these two images:

$$H_{OP1\_BACK} = SHA\text{-}1(Img_{BACK}),$$
$$H_{OP1\_ABST\_1} = SHA\text{-}1(Img_{ABST\_1}).$$

Likewise, the hashes of the included encrypted image regions are computed:

$$H_{OP1\_ROI_{enc}} = SHA\text{-}1(Img_{ROI_{enc}}),$$
$$H_{OP1\_ABST\_[1...A]_{enc}} = SHA\text{-}1(Img_{ABST\_[1...A]_{enc}}).$$

Due to access limitations, operator 1 cannot compute the hashes H_{ROI} and $H_{ABST\_[2...A]}$ since the usage passwords for the binding keys required to decrypt the corresponding image parts are unknown. The missing hashes are substituted with H_{ROI} and $H_{ABST\_[2...A]}$ from the *ImageHashes* field of Img_{COMB}:

$$H_{OP1\_Img} = SHA\text{-}1(H_{OP1\_BACK}||H_{ROI}||H_{OP1\_ABST\_1}||H_{ABST\_[2...A]}||$$
$$H_{OP1\_ROI_{enc}}||H_{OP1\_ABST\_[1...A]_{enc}}).$$

The hash sum H_{OP1_Img} now serves as input for the computation of the expected accumulated hash sum which, in turn, is used for group signature verification.

Finally, it must be noted that the number of abstraction levels, the video compression algorithms, the container format for the combined image as well as the streaming format can be freely chosen by the application developer. Note that the discussed approach focuses on the protection of outgoing, sensitive image data. It does not cover control and status messages exchanged between cameras or the control station. For that purpose, additional mechanisms such as Transport Layer Security (TLS) can be considered.

3.4 Implementation Aspects

For the prototype, all image areas where motion is detected are defined as sensitive. From the extracted ROI, an abstracted version is created using edge-detection. The background image IMG_{BACK} allows the presence and position of persons to be observed, the edge-detected ROI IMG_{EDGE} gives access to behavioral information and the original ROI IMG_{ROI} reveals both behavior and identity of detected persons/moving objects. Next, the background and the two ROI images are compressed. JPEG compression is used for the background and the original ROI while the black and white edge-detected ROI is compressed using zlib. The compressed regions of interest Img_{EDGE} and Img_{ROI} are encrypted using AES 265 in CBC mode and the AES session keys are bound to CS's TPM_S using the binding keys that have been generated for this camera during setup:

$$K_{AES\_1_{bound}} = \mathtt{Bind}_{K_{BIND\_1_{pub}}}(K_{AES\_1}),$$

$$K_{AES\_2_{bound}} = \mathtt{Bind}_{K_{BIND\_2_{pub}}}(K_{AES\_2}),$$

$$K_{AES\_3_{bound}} = \mathtt{Bind}_{K_{BIND\_4_{pub}}}\left(\mathtt{Bind}_{K_{BIND\_3_{pub}}}(K_{AES\_3})\right).$$

The video format that was chosen for the prototype is Motion JPEG (MJPEG). As shown in Fig. 5, the encrypted image regions $Img_{ROI_{enc}}$ and $Img_{EDGE_{enc}}$ are embedded into the background image as custom EXIF data. Likewise, the bound AES keys $K_{AES\_2_{bound}}$ and $K_{AES\_3_{bound}}$ as well as the SHA-1 hashes of the unencrypted Img_{ROI} and Img_{EDGE} are included.

Subsequently, the SHA-1 hash of the concatenated hash sums of Img_{BACK}, Img_{EDGE}, Img_{ROI}, $Img_{EDGE_{enc}}$ and $Img_{ROI_{enc}}$ is computed and is fed into the previously described hash accumulation procedure of the frame group. The accumulated hash then is signed and timestamped by TPM_C once the end of the frame group is reached. The resulting signature and timestamp data as well as the start and end indices of the frame group are included in the EXIF data of combined image shown in Fig. 5.

At the control station, the streamed frames have to be decrypted before viewing. Note that access to the original ROI IMG_{ROI} requires the cooperation of two security guards since the corresponding AES session key K_{AES_3} is bound with the two

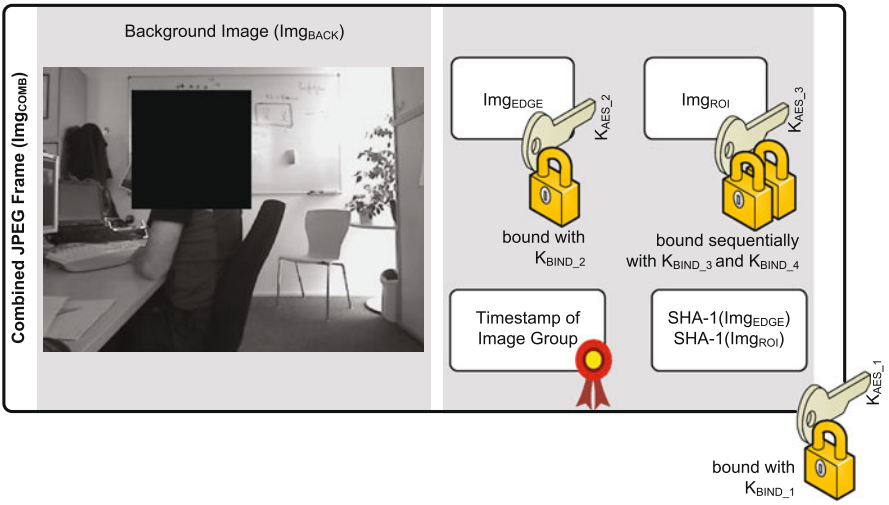


Fig. 5 The encrypted ROI image ($Img_{ROI_{enc}}$) as well as the encrypted edge image ($Img_{EDGE_{enc}}$) are embedded into the JPEG background image as custom EXIF data. The same is done for the bound AES keys as well as the SHA-1 hashes of Img_{ROI} and Img_{EDGE}. At the end of a frame group, the group's signature and timestamp are also included in the EXIF data. The combined image (Img_{COMB}) is then encrypted and streamed to the control station

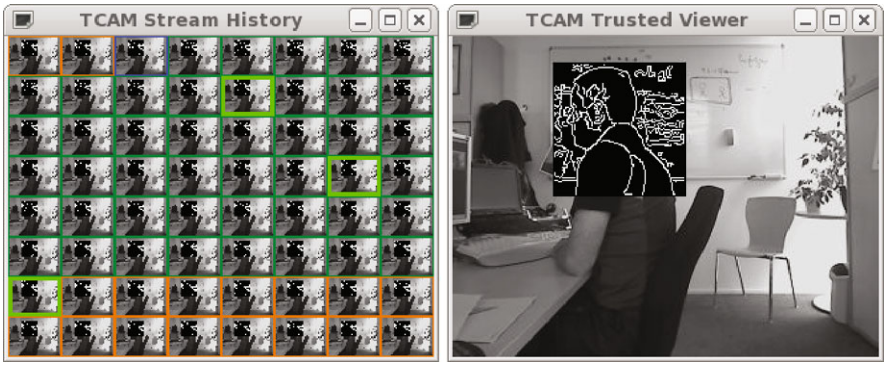


Fig. 6 The live viewer at the control station. *On the right* the current frame with the decrypted, edge-detected ROI is displayed. The *left window* shows the content of a circular buffer with the last 64 frames. The current frame is marked with a *blue border*. Frames with a signature that has not yet been verified have an *orange border* while successfully verified frames have a *dark green border*. The *last frame* of a group has a *light green border*

binding keys K_{BIND_3} and K_{BIND_4}. The right part of Fig. 6 shows the video stream at the control station where the background image is overlayed with the decrypted, edge-detected region of interest.

Accumulated image signatures and timestamps of frame groups are validated at the control station. Assuming that this validation is successful, the operator at the CS

has assurance that neither the individual images of a frame group nor their order was modified and the images of the group come from the expected camera. Freshness checks and world time mapping can be done as described in [65].

The left side of the live stream viewer example of Fig. 6 shows a circular buffer that contains thumbnails of the last 64 received frames together with their verification status. For frames with orange borders, the frame group signature was not yet received. Already verified frames have a green border and the last frame of a group is marked with a light green border. If authenticity and integrity of an image group cannot be verified before the circular buffer wraps around, a warning message is displayed and streaming is interrupted.

3.5 Performance Considerations

Table 2 presents the frame rates that are achieved on TrustCAM in different streaming modes. The sensor can deliver images either as uncompressed YUYV data or as a JPEG compressed version. Input images are delivered at a resolution of 320×240 or 640×480 pixels. For internal processing, input images are converted to either RGB or grayscale. The "Plain Streaming" column of Table 2 shows the achieved streaming frame rates if no ROI is extracted and neither encryption nor digital signatures are performed. Therefore, this column reflects the baseline streaming performance of the system without any additional security or privacy protection.

The second column, "Image Timestamping", shows the delivered frame rates if groups of outgoing images are timestamped. Overheads for the TPM TickStampBlob command are eliminated from the critical path by signing frame groups and executing the TPM operations asynchronously. As a consequence, the small performance impact that can be observed for some cases in the "Image Timestamping" column result from the additional computation of the accumulated SHA-1 hash for a frame group. Performance impacts on video streaming can be observed if YUYV input images are used. In this case, the images have to be JPEG compressed before being hashed and streamed. JPEG compression is computing intensive and puts a high load on the OMAP's ARM CPU. Therefore, even the small additional effort of the SHA-1 computation results in a reduction in the frame rate.

The "Image Encryption" column of Table 2 presents the achieved frame rates if a randomly placed region of interest is extracted from the input image, the ROI images are encrypted and embedded into the remaining background and, finally, the combined image is encrypted. For data encryption, AES 256 in CBC mode is used. Encryption runtimes for typical input sizes range from 1.6 ms (8 kB) to 15.4 ms (80 kB). Across all input format combinations, a considerable impact on the achieved streaming framerate can be observed. Another slight performance reduction can be perceived in the last column of Table 2 which presents the frame rates if both image timestamping and encryption (ROI size 200×200 pixels) are performed.

To investigate the cause for the substantial performance impact that is apparent in the "Image Encryption" column of Table 2, the involved processing steps have

Table 2 Frame rates (avg. over 1000 frames) for different types of video streaming between TrustCAM and CS via WiFi. In the "Plain Streaming" case, JPEG or YUYV frames are delivered by the sensor. JPEG frames are directly streamed as a MJPEG video stream. Note that JPEG images delivered by the sensor unit are in RGB. A conversion to grayscale would only add an extra overhead for decompression and recompression and is therefore omitted (cells marked with *n/a*). YUYV frames are converted to grayscale or RGB24 before they are JPEG compressed and streamed. The "Image Timestamping" column presents the achieved frame rates if groups of full, unmodified images are signed and timestamped. The "Image Encryption" column shows the frame rates that are achieved if a randomly placed region of interest of 200 × 200 pixels is extracted, an edge-detected version is created and the individual image parts (Img_{ROI}, Img_{EDGE} and Img_{COMB}) are encrypted before streaming. Finally, the last column shows the achieved frame rates when doing both—image timestamping/signing and encryption—before streaming

Input format		Internal Format	Plain Streaming	Image Timestamping	Image Encryption	Image Encryption and Timestamping
Resolution	type					
320 × 240	YUYV	Gray	25.0 fps	25.0 fps	20.5 fps	19.7 fps
	JPEG		n/a	n/a	13.5 fps	13.2 fps
	YUYV	RGB24	25.0 fps	24.4 fps	12.4 fps	12.0 fps
	JPEG		25.0 fps	25.0 fps	10.3 fps	10.1 fps
640 × 480	YUYV	Gray	13.1 fps	12.8 fps	9.6 fps	9.2 fps
	JPEG		n/a	n/a	5.1 fps	5.0 fps
	YUYV	RGB24	6.5 fps	6.4 fps	5.1 fps	5.0 fps
	JPEG		25.0 fps	25.0 fps	4.0 fps	3.9 fps

been analyzed in detail (see [65] for details). This analysis reveals that the runtime overheads for AES 265 encryption and SHA-1 computation are acceptable. AES encryption for the compressed ROI takes around 1.5 ms while only 1 ms is required for the compressed edge image. For the combined image, where the encrypted ROI and edge image are embedded as EXIF data, AES encryption requires between 4 and 9 ms. Binding of the AES session keys using the public binding keys of TPM_S takes about 5 ms and has to be done only at startup of the streaming application or when new session keys are created. Finally, SHA-1 computation requires between 2 and 3.1 ms. Overall, the direct performance impact of the added security and privacy functions is acceptable. The biggest bottleneck—the slow TPM—could be removed from the critical processing path. Additionally, TPM commands are executed in parallel to the main CPU and therefore this does not have an influence on the image processing blocks.

4 Concluding Remarks and Outlook

Security and privacy protection are crucial properties of video surveillance systems, since they capture and process sensitive and private information. In this chapter, we presented an overview of existing privacy protection and security solutions. A key

observation is that there is still a lack of approaches that consider security and privacy in video surveillance in a holistic way. It is apparent that most security solutions are situated at the application level and that node-centric security is not taken into account. A lot of work has been targeted at data- and network-centric security. But without taking the security of camera devices themselves into account, high-level protection mechanisms are literally built on sand.

With bringing security and privacy protection onto camera devices, one can achieve reasonable protection against attacks on data that is delivered by surveillance cameras. However, only limited protection is applied for data while it is on the camera. It is an open research topic to identify suitable approaches for on-device data protection. One potential approach is to bring security and privacy protection even closer to the data source by integrating dedicated security functions into the image sensor. If security and privacy are guaranteed at the sensor level, then the camera and its relatively large software stack would no longer have to be considered as trusted entities. This approach contains two main challenges: First, it is unclear what type of privacy protection is suitable and feasible at the sensor level. Second, sensor-level privacy protection means that image processing and analysis applications on the camera must be adapted to deal with pre-processed and pre-filtered data. A related question is if and how privacy protection can be objectively measured. Since privacy depends on personal as well as cultural attitudes, technical approaches alone will be insufficient. A thorough exploration of the privacy protection design space will also have to involve extensive user surveys to determine which privacy protection techniques are appropriate.

Acknowledgements This work was performed as part of the project *TrustEYE: Trustworthy Sensing and Cooperation in Visual Sensor Networks.*[1] The work was supported by funding from the European Regional Development Fund (ERDF) and the Carinthian Economic Promotion Fund (KWF) under grant KWF-3520/23312/35521.

References

1. Albanesi, M.G., Ferretti, M., Guerrini, F.: A taxonomy for image authentication techniques and its application to the current state of the art. In: Proceedings of the International Conference on Image Analysis and Processing, pp. 535–540 (2001)
2. Atrey, P.K., Yan, W.-Q., Chang, E.-C., Kankanhalli, M.S.: A hierarchical signature scheme for robust video authentication using secret sharing. In: Proceedings of the International Conference on Multimedia Modelling, pp. 330–337 (2004)
3. Atrey, P.K., Yan, W.-Q., Kankanhalli, M.S.: A scalable signature scheme for video authentication. Multimed. Tools Appl. **34**(1), 107–135 (2006)
4. Baaziz, N., Lolo, N., Padilla, O., Petngang, F.: Security and privacy protection for automated video surveillance. In: Proceedings of the International Symposium on Signal Processing and Information Technology, pp. 17–22 (2007)
5. Bartolini, F., Tefas, A., Barni, M., Pitas, I.: Image authentication techniques for surveillance applications. Proc. IEEE **89**(10), 1403–1418 (2001)

[1] TrustEYE website: http://trusteye.aau.at.

6. Boult, T.E.: PICO: privacy through invertible cryptographic obscuration. In: Proceedings of the Workshop on Computer Vision for Interactive and Intelligent Environments, pp. 27–38 (2005)
7. Boyle, M., Edwards, C., Greenberg, S.: The effects of filtered video on awareness and privacy. In: Proceedings of the Conference on Computer Supported Cooperative Work, pp. 1–10 (2000)
8. Bramberger, M., Brunner, J., Rinner, B., Schwabach, H.: Real-time video analysis on an embedded smart camera for traffic surveillance. In: IEEE Real-Time and Embedded Technology and Applications Symposium, pp. 174–181 (2004)
9. Brassil, J.: Using mobile communications to assert privacy from video surveillance. In: Proceedings of the Parallel and Distributed Processing Symposium, p. 8 (2005)
10. CARE Consortium: CARE – ambient assisted living: safe private homes for elderly persons. http://care-aal.eu/. Last visited March 2013
11. Cavallaro, A.: Adding privacy constraints to video-based applications. In: Proceedings of the European Workshop on the Integration of Knowledge, Semantics and Digital Media Technology, p. 8 (2004)
12. Cavallaro, A.: Privacy in video surveillance. IEEE Signal Process. Mag. **24**(2), 168–169 (2007)
13. Challener, D., Yoder, K., Catherman, R., Safford, D., van Doorn, L.: A Practical Guide to Trusted Computing. IBM Press, Raleigh (2008)
14. Chattopadhyay, A., Boult, T.E.: PrivacyCam: a privacy preserving camera using uClinux on the Blackfin DSP. In: Proceedings of the International Conference on Computer Vision and Pattern Recognition, pp. 1–8 (2007)
15. Chen, D., Chang, Y., Yan, R., Yang, J.: Tools for protecting the privacy of specific individuals in video. EURASIP J. Appl. Signal Process. **2007**(1), 107–116 (2007)
16. Cheung, S.-C.S., Zhao, J., Venkatesh, M.V.: Efficient object-based video inpainting. In: Proceedings of the International Conference on Image Processing, pp. 705–708 (2006)
17. Cheung, S.-C.S., Paruchuri, J.K., Nguyen, T.P.: Managing privacy data in pervasive camera networks. In: Proceedings of the International Conference on Image Processing, pp. 1676–1679 (2008)
18. Chinomi, K., Nitta, N., Ito, Y., Babaguchi, N.: PriSurv: privacy protected video surveillance system using adaptive visual abstraction. In: Proceedings of the International Multimedia Modeling Conference, pp. 144–154 (2008)
19. Davis, D.: Defective sign & encrypt in S/MIME, PKCS#7, MOSS, PEM, PGP, and XML. In: Proceedings of the USENIX Technical Conference, pp. 65–78 (2001)
20. Dawson, D., Derby, P., Doyle, A., Fonio, C., Huey, L., Johanson, M., Leman-Langlois, S., Lippert, R., Lyon, D., Pratte, A.-M., Smith, E., Walby, K., Wilkinson, B.: A report on camera surveillance in Canada (part two): surveillance camera awareness network. Technical report, The Surveillance Project (2009)
21. Dufaux, F., Ebrahimi, T.: Scrambling for video surveillance with privacy. In: Proceedings of the International Conference on Computer Vision and Pattern Recognition Workshop, pp. 160–166 (2006)
22. Dufaux, F., Ebrahimi, T.: A framework for the validation of privacy protection solutions in video surveillance. In: Proceedings of the International Conference on Multimedia and Expo, pp. 66–71 (2010)
23. Farmer, D., Mann, C.C.: Surveillance nation (part I). Technol. Rev. **4**, 34–43 (2003)
24. Fidaleo, D.A., Nguyen, H.-A., Trivedi, M.: The networked sensor tapestry (NeST): a privacy enhanced software architecture for interactive analysis of data in video-sensor networks. In: Proceedings of the International Workshop on Video Surveillance and Sensor Networks, pp. 46–53 (2004)
25. Fleck, S., Straßer, W.: Smart camera based monitoring system and its application to assisted living. Proc. IEEE **96**(10), 1698–1714 (2008)
26. Fleck, S., Straßer, W.: Towards secure and privacy sensitive surveillance. In: Proceedings of the International Conference on Distributed Smart Cameras, p. 7 (2010)

27. Friedman, G.L.: The trustworthy digital camera: restoring credibility to the photographic image. IEEE Trans. Consum. Electron. **39**(4), 905–910 (1993)
28. Gross, R., Sweeney, L., De Torre, F., Baker, S.: Model-based face de-identification. In: Proceedings of the International Conference on Computer Vision and Pattern Recognition Workshop, p. 8 (2006)
29. He, D., Sun, Q., Tian, Q.: A secure and robust object-based video authentication system. EURASIP J. Adv. Signal Process. **2004**(14), 2185–2200 (2004)
30. Helten, F., Fischer, B.: What do people think about CCTV? Findings from a Berlin survey. Technical report, Berlin Institute for Social Research (2004)
31. Krempl, S., Wilkens, A.: Datenschützer beanstanden Videoüberwachung in ECE-Einkaufszentren. http://heise.de/-1187205 (2011). Last visited March 2013
32. Li, G., Ito, Y., Yu, X., Nitta, N., Babaguchi, N.: Recoverable privacy protection for video content distribution. EURASIP J. Inf. Secur. **2009**, 11 (2009)
33. Luo, Y., Ye, S., Cheung, S.-C.S.: Anonymous subject identification in privacy-aware video surveillance. In: Proceedings of the International Conference on Multimedia and Expo, pp. 83–88 (2010)
34. Martin, A.: The ten page introduction to trusted computing. Technical report RR-08-11, Oxford University Computing Laboratory (December 2008)
35. Martínez-Ponte, I., Desurmont, X., Meessen, J., Delaigle, J.-F.: Robust human face hiding ensuring privacy. In: Proceedings of the International Workshop on Image Analysis for Multimedia Interactive Services, p. 4 (2005)
36. Memon, N., Wong P.W.: Protecting digital media content. Commun. ACM **41**(7), 35–43 (1998)
37. Mohanty, S.P.: A secure digital camera architecture for integrated real-time digital rights management. J. Syst. Archit. **55**(10–12), 468–480 (2009)
38. Moncrieff, S., Venkatesh, S., West, G.: Dynamic privacy in public surveillance. IEEE Comput. **42**(9), 22–28 (2009)
39. Ney, S., Pichler, K.: Video surveillance in Austria. Technical report, Interdisciplinary Centre for Comparative Research in the Social Sciences, Austria (2002)
40. Norris, C.: A review of the increased use of CCTV and video-surveillance for crime prevention purposes in Europe. Technical report, Department of Sociological Studies, University of Sheffield, United Kingdom (2009)
41. Pirker, M., Tögl, R., Hein, D., Danner, P.: A PrivacyCA for anonymity and trust. In: Proceedings of the International Conference on Trust and Trustworthy Computing, pp. 101–119 (2009)
42. Qureshi, F.Z.: Object-video streams for preserving privacy in video surveillance. In: Proceedings of the International Conference on Advanced Video and Signal-Based Surveillance, pp. 442–447 (2009)
43. Qusquater, J.-J., Macq, B., Joye, M., Degand, N., Bernard, A.: Practical solution to authentication of images with a secure camera. Storage Retr. Image Video Databases **3022**(1), 290–297 (1997)
44. Rahman, S.M.M., Hossain, M.A., Mouftah, H., El Saddik, A., Okamoto, E.: A real-time privacy-sensitive data hiding approach based on chaos cryptography. In: Proceedings of the International Conference on Multimedia and Expo, pp. 72–77 (2010)
45. Saini, M., Atrey, P.K., Mehrotra, S., Emmanuel, S., Kankanhalli, M.S.: Privacy modeling for video data publication. In: Proceedings of the International Conference on Multimedia and Expo, pp. 60–65 (2010)
46. Saini, M., Atrey, P.K., Mehrotra, S., Kankanhalli, M.S.: Anonymous surveillance. In: Proceedings of the International Workshop on Advances in Automated Multimedia Surveillance for Public Safety, p. 6 (2011)
47. Saini, M., Atrey, P.K., Mehrotra, S., Kankanhalli, M.S.: Hiding identity leakage channels for publication of surveillance video. In: Transactions on Data Hiding and Multimedia Security (2011)

48. Saini, M., Atrey, P.K., Mehrotra, S., Kankanhalli, M.S.: W3-privacy: understanding what, when, and where inference channels in multi-camera surveillance video. Springer Int. J. Multimed. Tools Appl. (August), 24 (2012)
49. Schaffer, M., Schartner, P.: Video surveillance: a distributed approach to protect privacy. In: Proceedings of the International Conference on Communications and Multimedia Security, pp. 140–149 (2007)
50. Schiff, J., Meingast, M., Mulligan, D.K., Sastry, S., Goldberg, K.Y.: Respectful cameras: selecting visual markers in real-time to address privacy concerns. In: Proceedings of the International Conference on Intelligent Robots and Systems, pp. 971–978 (2007)
51. Senior, A., Pankanti, S., Hampapur, A., Brown, L., Tian, Y.-L., Ekin, A., Connell, J., Shu, C.F., Lu, M.: Enabling video privacy through computer vision. IEEE Secur. Priv. **3**(3), 50–57 (2005)
52. Serpanos, D.N., Papalambrou, A.: Security and privacy in distributed smart cameras. Proc. IEEE **96**(10), 1678–1687 (2008)
53. Spindler, T., Wartmann, C., Hovestadt, L., Roth, D., van Gool, L., Steffen, A.: Privacy in video surveilled areas. In: Proceedings of the International Conference on Privacy, Security and Trust, p. 10 (2006)
54. Tansuriyavong, S., Hanaki, S.: Privacy protection by concealing persons in circumstantial video image. In: Proceedings of the Workshop on Perceptive User Interfaces, p. 4 (2001)
55. Texas Instruments. BeagleBoard website. http://www.beagleboard.org. Last visited March 2013
56. Troncoso-Pastoriza, J.R., Pérez-Freire, L., Pérez-González, F.: Videosurveillance and privacy: covering the two sides of the mirror with DRM. In: Proceedings of the Workshop on Digital Rights Management, pp. 83–94 (2009)
57. Trusted Computing Group. TCG Software Stack (TSS) Specification, Version 1.2, Level 1, Errata A. http://www.trustedcomputinggroup.org/resources/tcg_software_stack_tss_specification (March 2007). Last visited March 2013
58. Trusted Computing Group. TPM Main Specification 1.2, Level 2, Revision 116. http://www.trustedcomputinggroup.org/developers/trusted_platform_module (July 2007). Last visited March 2013
59. Vagts, H., Bauer, A.: Privacy-aware object representation for surveillance systems. In: Proceedings of the International Conference on Advanced Video and Signal Based Surveillance, pp. 601–608 (2010)
60. Vagts, H., Beyerer, J.: Security and privacy challenges in modern surveillance systems. In: Proceedings of the Future Security Research Conference, pp. 94–116 (2009)
61. Wickramasuriya, J., Datt, M., Mehrotra, S., Venkatasubramanian, N.: Privacy protecting data collection in media spaces. In: Proceedings of the International Conference on Multimedia, pp. 48–55 (2004)
62. Williams, A., Ganesan, D., Hanson, A.: Aging in place: fall detection and localization in a distributed smart camera network. In: Proceedings of the International Conference on Multimedia, pp. 892–901 (2007)
63. Winkler, T., Rinner, B.: A systematic approach towards user-centric privacy and security for smart camera networks. In: Proceedings of the International Conference on Distributed Smart Cameras, p. 8 (2010)
64. Winkler, T., Rinner, B.: TrustCAM: security and privacy-protection for an embedded smart camera based on trusted computing. In: Proceedings of the International Conference on Advanced Video and Signal-Based Surveillance, pp. 593–600 (2010)
65. Winkler, T., Rinner, B.: Securing embedded smart cameras with trusted computing. EURASIP J. Wirel. Commun. Netw. **2011**, 20 (2011)
66. Winkler, T., Rinner, B.: User centric privacy awareness in video surveillance. Multimed. Syst. **18**(2), 99–121 (2012)
67. Wong, C.K., Lam, S.S.: Digital signatures for flows and multicasts. IEEE/ACM Trans. Netw. **7**(4), 502–513 (1999)

68. Yabuta, K., Kitazawa, H., Tanaka, T.: A new concept of security camera monitoring with privacy protection by masking moving objects. In: Proceedings of the International Pacific-Rim Conference on Multimedia, pp. 831–842 (2005)
69. Ye, S., Luo, Y., Zhao, J., Cheung, S.-C.S.: Anonymous biometric access control. EURASIP J. Inf. Secur. **2009**, 18 (2009)

Object Video Streams: A Framework for Preserving Privacy in Video Surveillance

Faisal Z. Qureshi

Abstract Here we introduce a framework for preserving privacy in video surveillance. Raw video footage is decomposed into a background and one or more object-video streams. Such object-centric decomposition of the incoming video footage opens up new possibilities to provide visual surveillance of an area without compromising the privacy of the individuals present in that area. Object-video streams allow us to render the scene in a variety of ways: (1) individuals in the scene can be represented as blobs, obscuring their identities; (2) foreground objects can be color coded to convey subtle scene information to the operator, again without revealing the identities of the individuals present in the scene; (3) the scene can be partially rendered, that is, revealing the identities of *some* individuals, while preserving the anonymity of others, etc. We evaluate our approach in a virtual train station environment populated by autonomous, lifelike virtual pedestrians. We also demonstrate our approach on real video footage. Lastly, we show that Microsoft Kinect sensor can be used to decompose the incoming video footage into object-video streams.

1 Introduction

Video surveillance is ubiquitous. Recent advances in camera and communication technologies along with the decrease in deployment costs have made it possible to set up large video surveillance infrastructures relatively easily. The societal shift that has occurred during the first decade of the 21st century with its focus on the *war on terrorism* has all but removed any opposition to putting citizenry under video surveillance with the stated aim to enhance public safety and security. Many cities around the world are increasingly relying on video surveillance for crime prevention and community safety. Video footage captured through surveillance cameras is routinely used to identify suspects and as evidence in the courts. In addition to the video surveillance infrastructure controlled by city councils and government

F.Z. Qureshi (✉)
Faculty of Science, University of Ontario Institute of Technology, Oshawa, ON, Canada
e-mail: faisal.qureshi@uoit.ca

P.K. Atrey et al. (eds.), *Intelligent Multimedia Surveillance*,
DOI 10.1007/978-3-642-41512-8_4,

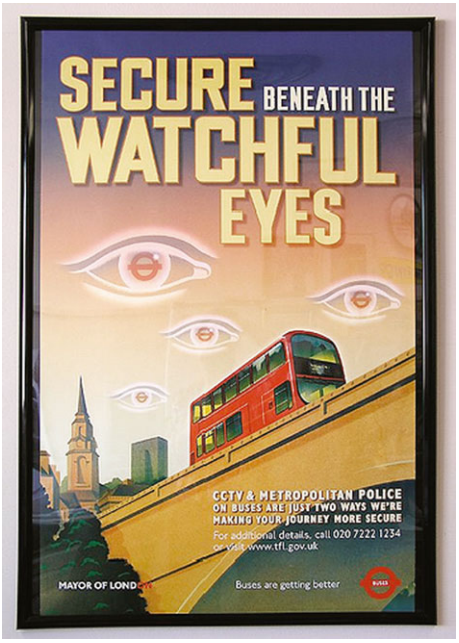


Fig. 1 British Government poster outside Metro station in London (*circa* 2007)

bodies, private sector has also invested heavily in video surveillance technologies. Retail stores, for example, are using video cameras to collect data needed to analyze and model consumer behavior [10, 13]. Video cameras are also quickly becoming an essential part of smart environments, for example, supporting home automation to enable elderly and disabled to safely remain in their own homes.

The panoptic effect of pervasive video surveillance (Fig. 1) raises many questions: (1) Who is collecting information about us? (2) How this information is being used? (3) What information is being collected? (4) Who has access to this information? and (5) What is the retention policy for the collected information? These issues have been studied by social and legal experts, and policies and best practices have been suggested. The use of video surveillance, however, is still largely unregulated. In 2001 Superbowl, law enforcement videotaped attendees without their knowledge, and then compared their faces against a database containing faces of known criminals [12]. Casinos, for example, also use biometric technology to identify cheaters and for "patron management" [11]. Experts agree that video surveillance undermines our "right to anonymity." Video surveillance augmented with biometric technology (e.g., face recognition) raises even more privacy concerns. Balancing the need for video surveillance against an individual's right to privacy is a challenge that needs to be addressed within social, legal, and technical contexts. A timely challenge for computer vision researchers is to develop video surveillance systems with built-in *privacy protection* capabilities. Such capabilities will help camera operators implement best practices and uphold laws regulating video surveillance.

Here we introduce a framework for privacy preserving video surveillance systems.[1] Captured video is decomposed into *object-video* streams. Each object-video stream contains visual information about a single object in the scene.[2] These streams can be recombined to visualize the area under surveillance in a variety of ways. For example, individuals present in the scene can be represented as color-coded blobs, hiding their identities. Selected individuals can be also blurred. Additionally some individuals can be removed from the video entirely. We also envision that these object-video streams are encrypted at source and can only be viewed by operators with the necessary authorization.

We embrace the *Virtual Vision* paradigm, exploiting visually and behaviorally realistic virtual environments to develop and empirically evaluate our video surveillance framework [17]. We employ a virtual train station environment populated by autonomous lifelike virtual pedestrians that is described in [24]. The vision pipeline for our prototype video surveillance system matches the performance of the vision pipeline (for real video) presented in [6]. Therefore, the obtained results are legitimate and valuable. We describe vision pipeline in Sect. 3. We also show object-stream construction and selective rendering using real video footage in Fig. 9. Furthermore, we show decomposing video into object-video streams using the Microsoft Kinect sensor [15].

The remainder of the chapter is organized as follows. We summarize relevant literature in the next section. Section 3 develops the vision pipeline: background learning, foreground detection, and pedestrian tracking. Then in Sect. 4, we describe how raw video is decomposed into a background stream and one or more object-video streams. Section 5 describes how object-video streams can be used to develop a privacy preserving video surveillance system. Preliminary results of our approach are presented in Sect. 6. While we have not deployed and tested our system in a real-world setting, the results presented here serve to demonstrate the applicability of the proposed strategy. We conclude our chapter with conclusions and future directions in Sect. 7.

2 Relevant Literature

Typically, sensory data gathered by a video surveillance system is monitored by human operators to detect events of interest. Computer vision technologies, such as pedestrian tracking, face recognition, and detection of unclaimed baggage, have

[1] This chapter is based upon our paper that appeared in the 6th International Conference on Advanced Video and Signal Based Surveillance in 2009 [16].

[2] This assumption sometimes breaks due to the limitations of video processing routines, such as background subtraction, object tracking, image segmentation, etc. Still under favorable conditions—good lighting, sparsely populated scenes, etc.—it is possible to decompose the video into object-video streams as we show later in the chapter.

been employed to increase the effectiveness of existing video surveillance systems and to develop the next-generation camera networks capable of perceptive coverage of large areas with little or no human supervision. These highly capable video surveillance systems shift the balance of power between intrusiveness and privacy, raising new privacy concerns. Clearly, these systems severely undermine the right to anonymity in public space.

The ability to visually track people present in the scene is necessary for camera networks capable of carrying out visual surveillance tasks autonomously. Face detection and recognition enable these networks to identify individuals [1, 3, 7, 26, 28]. Computer vision techniques also allow these video surveillance systems to compute soft and hard biometric signatures of individuals. In short, computer vision technologies will play a central role in developing the video surveillance systems of the future.

Interestingly computer vision technologies can also be used to develop camera networks that can uphold privacy policies and regulations [5, 23]. Pedestrian detection and tracking routines can identify individuals present in the scene and obscure them to hide their identities. The operator can still see the scene and know how many people are present in the scene without knowing the identities of those people. An activity recognition technique can reveal an individual if it detects an anomalous behavior.

Schiff et al. develop a video surveillance system capable of obscuring the faces of individuals present in the scene [21]. Individuals who do not want to be identified wear a visual marker, which allows the video surveillance system to locate the face of the individual and obscure it with an ellipse, while allowing observation of his or her actions in full detail. This allows the operator to observe the activities taking place in the scene without knowing the identities of the people present.

Sony patented a privacy mode for camcorders that replaces the skin color of individuals so as to avoid race-based discrimination [2]. [27] patented a system capable of obscuring a privacy region in a pan-tilt-zoom camera. [8] develops a system that is able to locate and obscure people in a video, thereby preventing statistical inferences from the video. Chattopadhyay and Boult developed a privacy preserving smart camera, called *PrivacyCam* [5]. PrivacyCam uses on-board digital signal processor to locate and encrypt human faces in the image. The original image can be recovered given the correct decryption key.

Saini et al. have carefully studied privacy leakage in video surveillance systems [18]. They correctly identify that an individual's identity can be learned through other channels even when that individual is not identifiable within a video. Consequently, obscuring/blurring an individual in a video footage alone is not sufficient to ensure that the privacy of that individual is not compromised. Object-video streams might alleviate this problem somewhat, since it is possible to make an individual disappear from the video by simply removing the object-video stream corresponding to that individual from the mix. Saini et al. have studied adaptive video blurring to protect the privacy of individuals present in the scene [19].

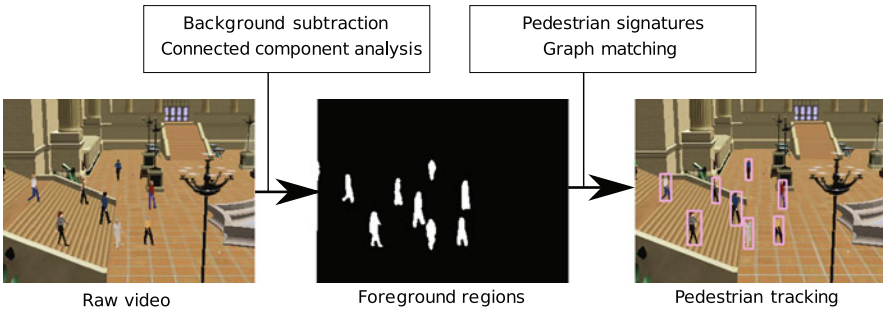


Fig. 2 Vision pipeline: We have adapted well-understood computer vision algorithms for our purposes. The vision routines operate upon both synthetic video captured by virtual cameras and real video captured through physical cameras. Background subtraction is used to identify foreground pixels. Pedestrians signatures that encode pedestrian color distribution in HSV space are matched in successive frames to perform tracking

3 Vision Pipeline

The performance of the proposed surveillance system is ultimately tied to the capabilities of the vision pipeline that is responsible for segmenting raw video into object video streams. We have adapted well-understood computer vision algorithms, including background subtraction, blob detection, and pedestrian tracking, to construct a vision pipeline that works equally well on both synthetic video captured within our virtual vision simulator and real video captured by physical cameras. Recently, we have also used the Microsoft Kinect RGBD sensor to construct object video streams from raw videos. In Fig. 2 we briefly explain the various components of the vision pipeline.

3.1 Background Subtraction

During an initial training phase, when no pedestrian is visible, each camera learns a background model of the scene. We model the variation in each pixel using the codebook method that was developed in [9]. We use the implementation of codebook method for background learning provided in the Open Computer Vision Library (OpenCV) [4]. Background subtraction step involves comparing the current frame against the learnt background model and constructing a (in general, noisy) foreground mask. In our case, the foreground mask constructed through background subtraction is cleaner due to lack of shadows, however, this does not invalidate our vision pipeline. Many techniques exist in the literature to account for shadows and other artifacts, such as camera motion, during background subtraction [6]. In a real system, we would also need a mechanism to update the background model to account for changes in the background. It is straightforward to incorporate this capability into our background model.

3.2 Pedestrian Tracking

The foreground mask obtained through background subtraction is cleaned up through connected component analysis and blobs representing foreground objects are extracted. In our case, each blob represents one or more pedestrians. We employ an appearance-based pedestrian tracker that is able to detect and track pedestrians in both synthetic and real video footage. Pedestrian appearance signatures are matched across frames to track pedestrians. Specifically pedestrian tracking is performed by setting up a bipartite graph matching problem as suggested in [6]. The optimal solution to the matching problem resolves pedestrian identities across multiple frames. We refer the reader to [6] for more details. Pedestrian tracker assigns each blob to one or more pedestrians. If an appropriate blob is not found in a frame, the pedestrian is matched to the entire frame.

The tracker maintains a list of pedestrians that are currently being tracked. In each frame, each pedestrian is either matched to a blob (using pedestrian signature matching) or to the background. The tracker is robust to short-duration occlusions.

3.3 Microsoft Kinect RGBD Sensor

It turns out that Microsoft Kinect Red-Green-Blue-Depth (RGBD) sensor is able to perform background subtraction, blob detection, pedestrian tracking, and pose estimation in real-time (around 15 frames per second). Furthermore, Microsoft Kinect also estimates the 2.5D structure of the scene by associating a depth value with each pixel. The depth information makes it much easier to identify the blobs belonging to different individuals present in the scene, which is the first step towards constructing object video streams from raw videos. In other words Microsoft Kinect already includes the vision pipeline that we require. It is, however, important to bear in mind that the Kinect sensor's operational range is limited to roughly 2.5 m. Consequently we still need our vision pipeline in order to be able to use generic cameras that have much larger operational ranges.

4 Object-Video Streams

This section describes the process of decomposing captured video into object-video streams. Let F_t be the video frame and M_t be the (binary) foreground mask at time t. We begin by extracting background pixels:

$$F_t^B(\mathbf{x}) = \begin{cases} [F_t(\mathbf{x}), 1] & \text{if } M_t(\mathbf{x}) = 0; \\ \mathbf{0} & \text{otherwise.} \end{cases} \tag{1}$$

Here, $\mathbf{x}$ is defined over the domain of F_t. $[F_t(\mathbf{x}), 1]$ denotes an RGBA vector and $\mathbf{0}$ denotes a zero vector. F_t^B is an RGBA image. Next, assume that the foreground

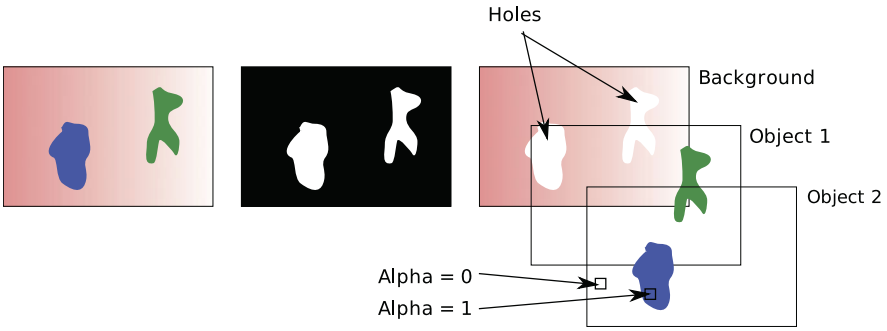


Fig. 3 Cleaned up foreground mask decomposes a video frame into a background component and two foreground components. Pedestrian to blob mapping information maintained by the pedestrian tracker links each foreground component to one (or more) pedestrians

mask M_t contains n blobs. Then for each blob C_i identified in the foreground image F_t, perform the following steps,

1. Construct blob mask M_t^i.

$$M_t^i(\mathbf{x}) = \begin{cases} 1 & \text{if } \mathbf{x} \in A(C_i); \\ 0 & \text{otherwise.} \end{cases} \tag{2}$$

   $A(C_i)$ denotes the area enclosed by blob C_i.
2. Construct an RGBA color image F_t^i.

$$F_t^i(\mathbf{x}) = \begin{cases} [F_t(\mathbf{x}), 1] & \text{if } M_t^i(\mathbf{x}) = 1; \\ \mathbf{0} & \text{otherwise.} \end{cases} \tag{3}$$

   Here, $\mathbf{x}$ is defined over the domain of F_t. $[F_t(\mathbf{x}), 1]$ is an RGBA vector. $\mathbf{0}$ denotes a zero vector.

The above process, which is illustrated in Figs. 3 and 4, partitions frame F_t into a background image, F_t^B (with holes in places of foreground objects), and n object images F_t^i, where $i \in [1, n]$. Each object image contains pixel data for one (or more) foreground objects. We note that this is a loss-less operation by observing that

$$F_t = F_t^B \cup \left(\bigcup_i F_t^i \right).$$

We define a *Partition*$(\cdot)$ operator that partitions a frame into background and foreground components as described above:

$$\mathit{Partition}(F_t) = \left\{ F_t^B, F_t^i | i \in [1, n] \right\}.$$

Given a sequence of video frames F_t, we construct the object-video stream O^k for a particular object k as follows. Let O^k be an empty sequence. Then for each frame F_t:

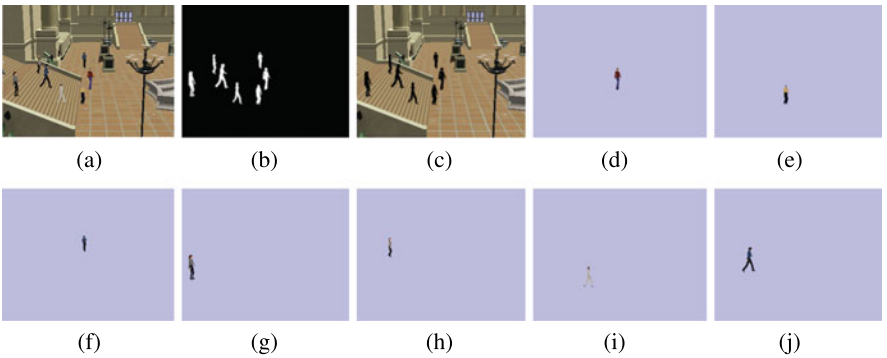
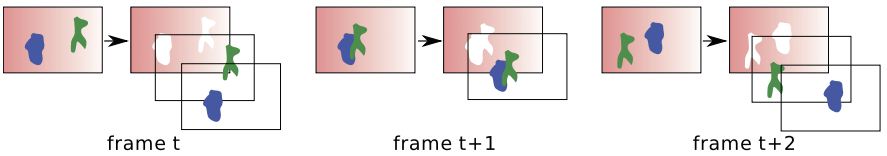
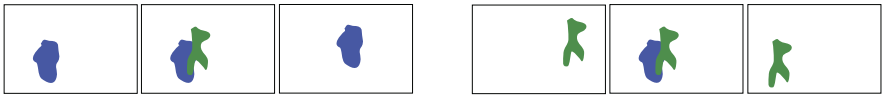


Fig. 4 Decomposing video into a background component and 7 foreground components. Each foreground component encodes visual data for a particular pedestrian. (**a**) Raw video. (**b**) Foreground mask. (**c**) Background image containing holes. (**d**)–(**j**) RGBA frames containing color data for 7 pedestrians visible in the frame



(a) Raw video frame is decomposed into a background stream and foreground stream(s)



(b) Object-video stream for Blue object (c) Object-video stream for Green object

Fig. 5 Constructing object-video streams

1. Construct $Partition(F_t)$.
2. Extend the sequence O^k by appending F_t^i at the end, if the tracker maps object k to blob i at time t. If the tracker does not map object k to any blob in the current frame, extend the sequence O^k by appending F^t.

Pedestrian crossover, proximity or occlusions can lead to poor blob segmentation and tracking errors. Multiple pedestrians can be mapped to the same blob. Consider, for example, the scenario shown in Fig. 5. The two objects represented as Green and Blue blobs are correctly segmented in frame t, so frame t is correctly decomposed into three components: background, Blue object, Green object. In frame $t+1$, however, the two objects are seen as a single blob, and the frame is incorrectly decomposed into two components. The pedestrian tracker assigns both objects to Blue/Green blob. Next, the two objects are correctly segmented in frame $t+2$, so frame $t+2$ is correctly decomposed into three components.

5 Privacy

Decomposing raw video into object streams opens up new possibilities for implementing privacy policies. At the most basic level, it allows the video surveillance system to obscure the identities of individuals present in the scene. An operator can still see scene activity without knowing the identities of individuals present in the scene. Object-video streams can be used to render the scene for a variety of purposes. We employ Laplacian pyramid blending to combine different object-video streams for rendering purposes [14]. Laplacian pyramid blending is also used to fill the holes in the rendered scene by using the stored average background image F^B.

- Object-video streams can be used to enhance the situational awareness of the operator. Objects can be color coded to convey qualitative scene information to the operator. This can be a powerful scheme for drawing operator's attention to events of interest. Sophisticated video analytics or simple image-space heuristics can assign unique colors to pedestrian blobs. For example, any pedestrian who enters a prohibited zone can be drawn as a red blob. Similarly, poorly segmented blobs, which map to multiple pedestrians, can be color coded to indicate pedestrian interactions (or simply overlap).
- Object-video streams also enable selective scene rendering. An operator can render the scene showing only some of the pedestrians present in the scene, without disclosing the identities of other individuals.
- Object centric decomposition of surveillance video has the potential to give more control to the individual. For example, a person might be able to find a lost item by sifting through an appropriate rendering of the scene that hides the identity of other individuals. Presently individuals are not allowed the access to the surveillance video as it might violate the privacy of others present in the scene.

We will be remiss to not point out that similar ideas of leveraging computer vision to obscure the identity of individuals present in the scene have been explored by others [22]. It is envisioned that in a real video surveillance system, object-video streams will be encrypted. Access control mechanisms will determine how the scene is rendered providing a way to strike a balance between the need-to-know on the part of an operator and the right-to-privacy on the part of an individual.

6 Results

We evaluate our approach on a *virtual* video surveillance system deployed in a virtual train station. The video surveillance system comprises 4 passive, wide field-of-view cameras with overlapping fields-of-view. It is assumed that the camera setup is fully calibrated, which simplifies pedestrian identity management across multiple cameras. Decomposing raw video into object-video streams does not require the camera network to be calibrated. We also report results on real video footage, further demonstrating the validity of our approach. Last, we demonstrate how Microsoft Kinect RGBD sensor can be used to construct object-video streams.

(a) Video (b) Foreground Mask (c) Tracking output

(d) Rendering 1 (e) Rendering 2 (f) Rendering 3

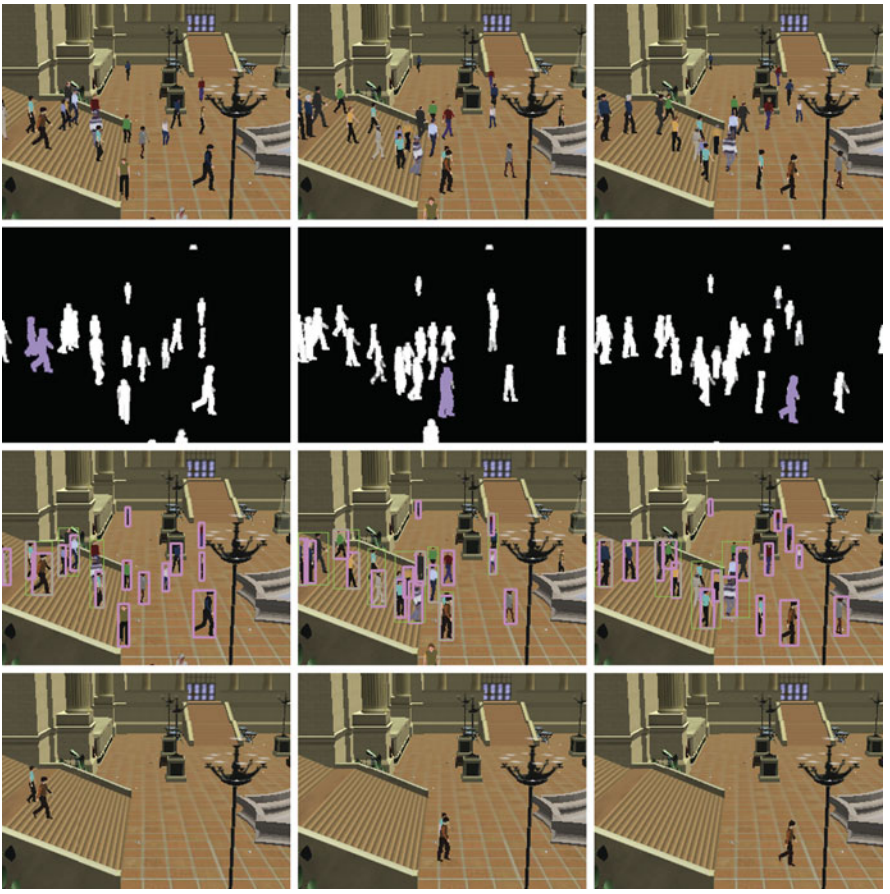
Fig. 6 Decomposition into object-video streams presents new possibilities to view the scene

6.1 Synthetic Footage

We show different rendering possibilities in Fig. 6. Figure 6(d) shows a privacy preserving rendering where each pedestrian is seen as a color blob. Single person blobs are Green; whereas, multi-person blobs are colored Blue. Pedestrian tracker selects an appropriate color for the blob. Figure 6(e) shows a rendering where the identities of two individuals (the man in Red shirt and the man in Orange shirt) have been revealed. All other individuals are still shown as blobs. Figure 6(f) is showing the scene with only two persons. In this case, the viewer can know the identity of these persons; however, he can not tell how many people were present in the scene.

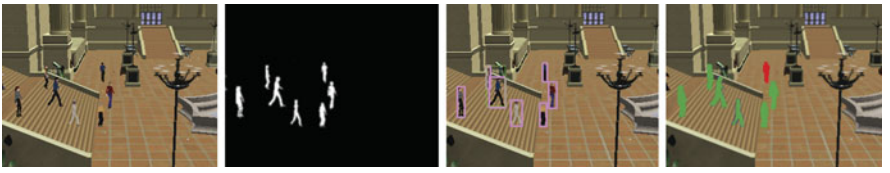
Figure 7 shows selective rendering. The top row contains original video frames. The second row shows foreground mask. Tracking output is shown in the third row, and the fourth row shows a rendering of the scene using the object-video stream associated with the person in Brown shirt. Notice that frames 266 and 308 (row 4) also show a woman in a Blue top. This is an artifact of poor segmentation. Foreground detection erroneously merged blobs for the two individuals in frames 266 and 308. The blobs associated with the person in Brown shirt are shown in Violet.

Figure 8 shows how blob coloring can improve scene awareness of an operator, while still preserving the privacy of individuals present in the scene. The Red blob shows a pedestrian who has crossed a virtual trip wire. Virtual trip wires, which are typically defined in pixel space, are routinely used in video surveillance systems to raise alarms.



Frame 266 Frame 308 Frame 325

Fig. 7 This sequence shows the effects of poor foreground segmentation on the object-video stream for the pedestrian wearing a Brown shirt. Pedestrian tracker maps the pedestrian of interest to *Violet blobs* in the shown frames



(a) (b) (c) (d)

Fig. 8 Event based color coding is also possible. The *Red blob* indicates a person who has tripped a virtual wire (defined in pixel space). Such wires are routinely used in video surveillance systems. (**a**) Video frame, (**b**) foreground mask, (**c**) tracking output, and (**d**) privacy preserving color coded rendering

(a) (b) (c) (d) (e) (f)

Fig. 9 Bootstrapping sequence from the Wallflower dataset [25]. (**a**) Raw frame, (**b**) mean image estimated using 2000 frames, (**c**) foreground mask, (**d**) pixel data for foreground objects, (**e**) showing all pedestrians as color blobs, and (**f**) re-imagining the scene with only two pedestrians

(a) (b) (c) (d)

Fig. 10 Using Microsoft Kinect RGBD images to construct object video stream. (**a**) Captured video and (**b**)–(**d**) object video streams constructed corresponding to the three individuals present in the scene

6.2 Real Video Footage

Figure 9 shows object-stream decomposition and subsequent selective rendering on real video footage. Figure 9(e) renders pedestrians as colored blobs: multi-person blobs are shown in red and single person blobs are shown in blue. Tracker is unable to resolve the green blob in the top-left corner of the frame. Figure 9(f) combines mean image estimated by observing 2000 frames and object-video streams for the two pedestrians in the bottom-right corner of the frame to render the scene showing only these two pedestrians. A closer look reveals ghosting artifacts in the rendered frame as the estimated mean frame is used to close the holes left by other pedestrians. Ghosting artifacts can be reduced by providing a reference background frame.

6.3 Microsoft Kinect RGBD Sensor

Figure 10 shows object-stream decomposition and subsequent selective rendering using Microsoft Kinect RGBD sensor. The captured video containing 3 individuals is decomposed into 3 object video streams, each containing only a single individual. In this case, both color and depth information available through the Kinect sensor is used to construct the object video streams. Figure 11 illustrates a situation where Kinect shines. The foreground mask shown in Fig. 11(a) shows a situation discussed in Sect. 4 where sometimes a single (connected) foreground region is associated to two or more individuals present in the scene. These situations are difficult to deal with in a general setting. Kinect sensor, however, can easily deal with these situ-

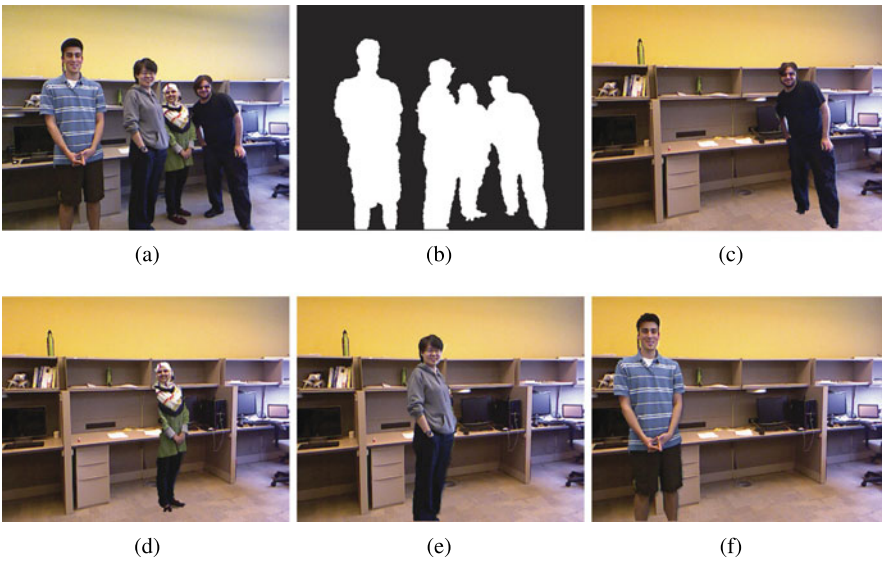


Fig. 11 Using Microsoft Kinect RGBD images to construct object video stream. (**a**) Captured video, (**b**) foreground mask, and (**c**)–(**f**) object video streams constructed corresponding to the three individuals present in the scene

ations by relying upon the depth value associated with each pixel. In the example shown in Fig. 11, the foreground region (Fig. 11(b)) is decomposed into four individuals.

6.4 Limitations

The work on privacy preserving video surveillance systems, including the work presented here, is focused on technical challenges related to obfuscating individuals present in the captured video stream. The underlying assumption is that the privacy of an individual is not violated if an operator is unable to see that person. While obfuscating individuals in captured video streams is a necessary first step towards realizing privacy preserving video surveillance system, this capability alone does *not* address the privacy issues surrounding pervasive video surveillance. This is not only true for the system presented here, but is also true for any system that attempts to hide the identity of an individual in the surveillance video.

Saini et al. [20] have developed privacy leakage models that attempt to quantify the loss of privacy due to video surveillance even when an individuals is never visually identified in any of the video streams. They cogently argue that privacy is compromised even in the presence of an obfuscation mechanisms that never fails. One a more practical note, however, it is worthwhile remembering that error toler-

ance for any obfuscation scheme is nearly zero. If the obfuscation scheme fails even for a single frame, the privacy of an individual is compromised.

7 Conclusions

We have proposed a novel framework for preserving privacy in video surveillance. Raw video data is decomposed into object-video streams. Such object-centric decomposition of the raw video presents new alternatives for upholding privacy policies and regulations in video surveillance. Object-specific privacy policies can be implemented. Object-video streams can be combined to recreate the original video, when warranted. Selective scene rendering, which focuses on a single aspect of the scene, is also supported.

The quality of object-based video decomposition is closely tied to the performance of low-level vision processing—poor segmentation leads to poor, or worse useless, video decompositions. Recent advances in background segmentation and pedestrian tracking suggest that the proposed approach is useful for scenes with low to medium crowd density. Pedestrian segmentation is still difficult in crowded scenes. It is conceivable that a privacy preserving scheme, such as ours, can be easily implemented in RGBD sensors similar to Microsoft Kinect. Many technical challenges, however, need to be addressed before such RGBD sensors can be used for video surveillance in general.

We are currently investigating encryption and access control mechanisms to develop secure rendering modules for video surveillance systems. These modules will combine object-video streams to present a mediated view of the scene to the operator. Such rendering modules are needed to gain the benefits of video surveillance technologies while preserving individual privacy. In closing, we need to pay more attention to privacy implications of pervasive video surveillance. More work is needed to develop robust computer vision routines capable of stripping identifiable information from surveillance footage without compromising the usefulness of the captured footage. Furthermore, any privacy preserving video surveillance system must also take into account the privacy leakage channels inherent in pervasive video surveillance systems.

Acknowledgements We thank Wei Shao and Mauricio Plaza-Villegas for their invaluable contributions to the implementation of the Penn Station simulator. We also thank Jordan Stadler for his work on constructing object-video streams using Microsoft Kinect sensor. This work is supported in part by the UOIT Startup Fund. We also acknowledge the NSERC Discovery Grant program.

References

1. Arandjelovic, O.D., Cipolla, R.: Face recognition from video using the generic shape-illumination manifold. In: Proc. European Conference on Computer Vision (ECCV06), Graz, Austria, vol. 4, pp. 27–40 (2006)

2. Berger, A.M.: Privacy mode for acquisition cameras and camcorders. US Patent 6,067,399 to Sony Corp., Patent and Trademark Office, 2000
3. Bourdev, L., Brandt, J.: Robust object detection via soft cascade. In: Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR05), San Diego, CA, vol. 2, pp. 236–243 (2005)
4. Bradski, G., Kaehler, A.: Learning OpenCV: Computer Vision with the OpenCV Library. O'Reilly Media, Beijing (2008)
5. Chattopadhyay, A., Boult, T.E.: PrivacyCam: a privacy preserving camera using uCLinux on the Blackfin DSP. In: Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR07), Minneapolis, MN, pp. 1–8 (2007)
6. Chen, H.-T., Lin, H.-H., Liu, T.-L.: Multi-object tracking using dynamical graph matching. In: Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR01), Hawaii, vol. 2, pp. 210–217 (2001)
7. Dornaika, F., Ahlberg, J.: Fast and reliable active appearance model search for 3-d face tracking. IEEE Trans. Syst. Man Cybern., Part B, Cybern. **34**(4), 1838–1853 (2004)
8. Fan, J., Luo, H., Hacid, M.-S., Bertino, E.: A novel approach for privacy-preserving video sharing. In: Proc. 14th ACM International Conference on Information and Knowledge Management (CIKM05), pp. 609–616. ACM, New York (2005)
9. Kim, K., Chalidabhongse, T.H., Harwood, D., Davis, L.: Real-time foreground-background segmentation using codebook model. Real-Time Imaging **11**, 167–256 (2005)
10. Kirkup, M., Carrigan, M.: Video surveillance research in retailing: ethical issues. Int. J. Retail Distrib. Manag. **28**(11), 470–480 (2000)
11. Moser, T., Nelson, D., Williams, R., Rowe, R.: Casino patron tracking and information use. US Patent WO/2008/067212, June 2008
12. Nieto, M., Johnston-Dodds, K., Wear Simmons, C.: Public and Private Applications of Video Surveillance and Biometric Technologies. California Research Bureau, California State Library, Bureau (2002)
13. Norris, C., McCahill, M., Wood, D.: The growth of CCTV: a global perspective on the international diffusion of video surveillance in publicly accessible space. Surveill. Soc. **2**, 110–135 (2004)
14. Ogden, J.M., Adelson, E.H., Bergen, J.R., Burt, P.J.: Pyramid-based computer graphics. Technical report RCA Engineer 30-5, RCA Corporation (September 1985)
15. OpenKinect. http://openkinect.org/wiki/Main_Page. Last accessed 28 May 2012
16. Qureshi, F.Z.: Object-video streams for preserving privacy in video surveillance. In: Proc. 6th International Conference on Advanced Video and Signal Based Surveillance (AVSS09), Genovo, Italy, pp. 1–8 (2009)
17. Qureshi, F.Z., Terzopoulos, D.: Smart camera networks in virtual reality. Proc. IEEE (Special Issue on Smart Cameras) **96**(10), 1640–1656 (2008)
18. Saini, M., Atrey, P.K., Mehrotra, S., Emmanuel, S., Kankanhalli, M.: Privacy modeling for video data publication. In: Proc. IEEE International Conference on Multimedia and Expo (ICME), Singapore, pp. 60–65 (2010)
19. Saini, M., Atrey, P.K., Mehrotra, S., Kankanhalli, M.: Adaptive transformation for robust privacy protection in video surveillance. Adv. Multimed. **2012**, 1–14 (2012)
20. Saini, M., Atrey, P., Mehrotra, S., Kankanhalli, M.: W3-privacy: understanding what, when, and where inference channels in multi-camera surveillance video. Multimed. Tools Appl. 1–24 (2012)
21. Schiff, J., Meingast, M., Mulligan, D.K., Sastry, S., Goldberg, K.: Respectful cameras: detecting visual markers in real-time to address privacy concerns. In: Proc. IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS07), San Diego, CA, pp. 971–978 (2007)
22. Senior, A., Pankanti, S., Hampapur, A., Brown, L., Tian, Y.-l., Ekin, A.: Blinkering surveillance: enabling video privacy through computer vision. Technical report, IBM, NY (2003)
23. Senior, A., Pankanti, S., Hampapur, A., Brown, L., Tian, Y.-L., Ekin, A., Connell, J., Shu, C.F., Lu, M.: Enabling video privacy through computer vision. IEEE Trans. Secur. Priv. **3**(3), 50–57 (2005)

24. Shao, W., Terzopoulos, D.: Autonomous pedestrians. Graph. Models. **69**(5–6), 246–274 (2007)
25. Toyama, K., Krumm, J., Brumitt, B., Meyers, B.: Wallflower: principles and practice of background maintenance. In: Proc. IEEE International Conference on Computer Vision (ICCV99), Kerkyra, Greece, vol. 1, pp. 255–261 (1999)
26. Viola, P., Jones, M.: Rapid object detection using a boosted cascade of simple features. In: Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR01), Hawai, pp. 1–8 (2001).
27. Wada, J., Wakiyama, K., Kogane, H., Takada, N.: Monitor camera system and method of displaying pictures from monitor camera thereof. European Patent EP 1 081 955 A3 to Matsushita Electric Industrial, European Patent Office, 2001
28. Wu, B., Nevatia, R.: Detection of multiple, partially occluded humans in a single image by Bayesian combination of edgelet part detectors. In: Proc. Tenth IEEE International Conference on Computer Vision (ICCV05), Beijing, China, vol. 1, pp. 90–97 (2005)

Surveillance Privacy Protection

Nikki Gulzar, Basra Abbasi, Eddie Wu, Anil Ozbal, and WeiQi Yan

Abstract Surveillance Privacy Protection (SPP) is a realistic issue in the world we are living in today. Due to the massive progress in technologies and systems, surveillance is becoming quite impossible to avoid. More information is being handed out without realizing the risks involved. The objective of this chapter is to evaluate what types of surveillance, privacy and protection measures are being implemented, how information is being used and what rights individuals have over this. In addition, this chapter also emphasizes the importance of tools, data sets and databases that are being developed to enable surveillance privacy.

1 Introduction

Surveillance can be defined as close observation over an object or a person for an undefined period of time, especially when one is under suspicion. The word "surveillance" has commonly been associated with police and intelligence agencies. Previously, there was a set purpose behind "surveillance", however since the late 1980s with the emergence of cyberspace technology, this has changed [9]. Today, the public is being monitored without having their consent or without having prior knowledge about how these activities have become a breach in privacy [33, 52].

This chapter introduces three types of surveillance: digital surveillance, audio surveillance and video surveillance. The main focus of this chapter is video surveillance since it is the most popular type of surveillance that is being deployed in society. Section 1 of this chapter introduces three sub-sections: surveillance, privacy and protection. This is followed by the types of technology for digital surveillance and the techniques that are being used for video and images in Sect. 2. Section 3 describes the systems and tools for surveillance privacy protection, it also highlights recent study on W^3 privacy: understanding *what*, *when* and *where* inference channels in multi-camera surveillance video. Section 4 summarizes what surveillance privacy protection is, how it is being deployed and what the future expectations are.

N. Gulzar · B. Abbasi · E. Wu · A. Ozbal · W. Yan (✉)
Auckland University of Technology, Auckland, New Zealand
e-mail: wyan@aut.ac.nz

P.K. Atrey et al. (eds.), *Intelligent Multimedia Surveillance*,
DOI 10.1007/978-3-642-41512-8_5,

1.1 Digital Surveillance

One of the cutting-edge surveillance technologies is digital surveillance. It involves the monitoring data and traffic on internet [64]. Government agencies such as Information Awareness Office, NSA and the FBI spend billions of dollars each year to develop systems so as to intercept and analyze the data transmitted, and extract the data that is useful to law enforcements [14].

1.1.1 Biometric Surveillance

With the advancement of digital technology, we are seeing significant advancement is biometrics. Biometric surveillance measures and analyzes human physical and behavioral characteristics for authentication, identification or screening purpose [68]. Physical characteristics include fingerprints, facial recognition and DNA. Behavior characteristics include voice or gait. The September 11 tragedy has given biometric surveillance massive attention [68].

Biometric technologies are being marketed as a "silver bullet" for terrorism [68]. The FBI is spending $1 billion to build a new biometric database, which will store DNA, facial recognition, fingerprints and other biometric data [59]. Biometrics cannot fully identify whether the person under surveillance is a terrorist or not, no matter how accurately the person is identified, it cannot determine it all alone [99], but with introduction of facial thermographs, it will evolve the biometrics technology even further [3, 50]. Facial thermographs allow machines to identify certain emotions such as fear or stress by measuring the temperature generated by blood flow to different parts of their face [77]. This system will help the law enforcements if the suspect is worried, nervous, lying or hiding [77]. As well as biometrics, RFID surveillance is making headway in digital surveillance [32]. RFID tags can be applied to animals, humans and products to keep track using radio waves [82]. It can be read from several meters away. There are companies who are already using RFID tags on employees, it helps employers to monitor them on their job. There are concerns that RFID will soon allow people to be tracked and scanned everywhere they go [18, 59].

1.1.2 Audio Surveillance

Audio surveillance is one of the oldest forms of surveillance technology. Audio surveillance is used to keep tracking phone conversations, tracking the location and monitoring the data [62]. Wiretapping is one of the most common and simple forms of audio surveillance. Wiretapping is highly inconspicuous and is able to clearly record conversations from both sides [103]. Small audio devices which are commonly referred to as bugs, are attached to a telephone circuitry, then signals are transmitted from wireless to another device that records the conversation [103], but with introduction of mobile phones, wiretapping has been replaced by software that

keeps tracking all mobile phone users. It also gives geographical locations of a mobile phone even when it is not in use [25].

Another audio surveillance that is usually utilized is a room microphone. This usually involves placing wireless microphones in a room to pick up on conversations. The microphone can be planted in common places such as clocks, pens and stuffed toys [103]. Room microphone works in a similar fashion to wiretapping. The microphone sends signals to a receiver and the conversation can be directly recorded.

Just like room microphone, long distance microphones are another means of audio surveillance. A parabolic microphone has the ability to pick up conversations up to 91.4 meters away [76]. Parabolic microphone is also referred to as a shotgun microphone because of its long shape [103]. The disadvantage of parabolic microphone is that it is highly sensitive. While picking up conversations, it can also pick up other noises and if there is obstruction between the microphone and the conversation, then functionality of the microphone will be affected [103].

Conceivable transmitters, also known as body wires which are a very well-known type of audio surveillance. Small microphones are worn by a person, and the signals are sent back to the receiver for recording [10]. It is a portable device and allows the person wearing the device to engage in conversations and get specific details [89].

1.1.3 Video Surveillance

Video surveillance uses video cameras to view a wide range of areas. The footage is recorded and can be viewed by a security guard or by members of the law enforcement. Before Closed-Circuit Television (CCTV) would only be installed in places such as banks, casinos, airport, military installation and convenience stores but now a day CCTVs are located everywhere [61], U.K. has the largest CCTV network in the world [9]. In January 2000, Prime Minister Tony Blair funded 150 million pounds for the expansion of CCTV network [104].

Another form of video surveillance is aerial surveillance which is mostly used by military to gather visual imagery or video from airborne vehicle [68]. Military aircrafts use a range of sensors to monitor battlefields. Digital imaging technology and miniaturized computers are some of the technologies that have contributed to rapid advances in aerial surveillance hardware such as micro-aerial vehicles and high resolution imagery capable of identifying objects at extremely long distances [59, 91]. MQ-9 Reaper is a U.S. drone plane used by the Department of Homeland Security, it carries cameras that are capable of identifying an object from altitude of 60,000 feet, and it has infrared devices that can detect the heat emitted from a human body at a distance of up to 60 kilometers [10].

In 2007, state and domestic federal agencies were able to access imagery from military intelligence satellites and aircraft sensors which are now being used to observe activities of U.S. citizens [59, 97]. Software such as Google Earth provides similar information but the satellite imagery provides real-time video with higher

resolution, it will also be able to identify objects in buildings and also detect chemical traces no matter what type of weather is (cloudy, rainy or stormy) [14]. In 1928, U.S. Departments of Defence launched the Navstar Global Positioning System (GPS), which is composed of 24 geo-rotational satellites that orbit the earth at a distance of 12,660 miles [14]. No matter where we are positioned on the earth, there are several satellites above us and no movements on the ground or in orbit of the satellite will cause a temporary blind spot [34].

1.2 Surveillance Privacy

Whenever we are under surveillance, privacy becomes an issue. Privacy has been commonly used within western society, however, it was not a general concept and to many cultures it was virtually unknown, until recently [95]. Privacy is an individual right to control what happens with personal data [71, 102]. The meaning of privacy may differ throughout cultures but the general conception is that privacy means wanting to keep information unnoticed or unidentified from the general public. Privacy can be categorized in different contexts [35, 65].

Personal privacy is one of the first issues that are being violated. Personal privacy allows an individual to keep their body or beliefs private. Physical privacy can be defined as preventing intrusion into one's personal space or solitude. The concerns may be [59]:

- *Not allowing personal possessions searched by an unwelcome party*
- *Not allowing access to people's homes and vehicle without authorization*

Most countries have trespassing and property rights which help to determine the right to physical properties [44].

Data privacy is the second most important issue when it comes to privacy. We all want to secure our personal data but data privacy is about an evolving relationship between technology and legal rights, which makes it harder to keep data private. The data storage causes some privacy issues such as who will access to the data, how the data is stored and the user's rights for protection [67]. There are some web sites that ask for more data than necessary but it is unclear as to what they share. Privacy issues especially to personal data include insecure, electronic transmission, data trails and logs of email messages, and the tracking of web pages visited [102]. Nowadays every kind of organization is marketing online users, which means that we are putting more of our personal data online and sometimes it becomes hard to keep track of all the information. Without our knowledge, the data could be sold to make profits [39, 47]. Therefore data privacy has become very important, it gives us a little control over the information we share, and the penalty for privacy violation has become more severe. There are four ways that threats to privacy [6, 96]:

- *Phishing may be used for private information*. How this is done is that usually cybercriminals send emails or maybe instant messages that look like they are from trusted organization and may require personal information or mobile num-

bers. Someone who is not familiar with technology and the risks that come with handing out personal information may fall for one of these traps. It is important to educate family, friends, and colleagues about the risks of disclosing sensitive information. When it comes to personal information, we should always provide bare minimum.

- *Using malware and spyware has become quite common when extracting personal information*. Today cyber criminals just use malicious web sites to download programs through security holes in software on the PCs [96]. Anti-virus is important for protecting PCs, especially in today's digital world, where most of our personal data are stored electronically.
- *Storing data electronically has become very popular*. With electronically stored data, risk of privacy breach has increased. Law has been tightened for medical storage, however there are still some loopholes that have been identified. It is important to talk to companies and organizations about privacy concerns and understand how our data is being protected.
- *Wireless hotspots are just about everywhere around the globe*. Public Wi-Fi connection can make it easy for hackers to gain access. It is important to have a strong and operating firewall and avoid entering financial information because the data that is being carried over the public Wi-Fi networks may not be encrypted [96].

Under the common law if an individual's privacy has been violated, (s)he has the right to sue [2, 81]. People have their own right of privacy. The privacy act is there to control how individual's information is collected, used, stored or disclosed [36].

Organizational privacy allows government agencies or organizations to keep their activities private and prevent it from being leaked to other organizations. Each organization has its own privacy policy, which helps them to maintain the privacy of personal data. An example of the organizational privacy is internet privacy. When organizations have web interactions for their customers, they must take customers' right into consideration when it comes to their personal data [26]. Data Privacy Day is about empowering people to protect their privacy and control their digital footprints to ensure the protection of data privacy [4].

1.3 Surveillance Privacy Protection

Protection is a very broad term, we use protection in our daily routines, sometimes we are aware of it and sometimes we are not [43, 83]. We tend to protect the information that are important and precious to us. In regard to surveillance, "protection" is mostly emphasized on databases. Database security is an essential part of surveillance privacy protection.

Database security holds a range of security topics, there is physical and network security, encryption and authentication, and also focus more on securing data concepts and mechanisms aspects. Database security is constructed upon a framework surrounding three concepts: Confidentiality, Integrity and Accessibility (CIA) [7].

Confidentiality from privacy aspects looks into protection of data against illegal access, consign to the avoidance and recovery from both hardware and software blunder as well as from nasty data access follow-on in denial of data accessibility [63]. Permitting to these three constructs, a database security factor in any sequence will be required to cover access control to databases, application access, vulnerability to the system, outside inference, and auditing mechanisms as well.

The main method used to protect data is limiting the access. This action can be performed through authentication, authorization and access control. All three mechanisms are relatively different but they can be used in a combination alongside the access control for granularity by handing rights to specialized objects and users. For example, generally a database system uses some forms of authentication, like a username and secret password, to control unauthorized access to the database system. In addition, usually users are authorized or privileges are delegated to specific recourses [16].

Majority of the users do not access company database directly by simply logging into the database system. As an alternative, they log into the database via an application program. Currently a tool that is being used as a security (or CRUG) matrix can also be used to plainly identify the necessary access rights that are required by an application program. Particularly, the security matrix supplies a visual depiction of the connection among operation or authorizations can be required for database entities and input/output sources like documentation and reports [16].

Database security breaches have increased dramatically in the past decade. Possibly the most commonly well-known database vulnerability would be the SQL injection. SQL which present superb illustration for examining security as they represent a very important database security matter, risks intrinsic to non-authorized user contribution. SQL injections can take place when SQL statements are selected dynamically and created by taking user input [101]. The vulnerability happens mainly due to features of the SQL language which lets user do such things as implanting comments like double hyphens "(- -)", concatenating SQL statements detached by semicolons, and the capability to question meta data taken from database data dictionaries as well [105].

Database auditing can be used to trace database access and client activity like where and when a database was logged on. Auditing is initially used to detect who accessed database items, which activities were executed, and what data was altered during that time. One of the down sides is that it does not avoid security breaches, but still provides I.T. administrator with enough information to identify if a breach has cropped up [16].

2 Technologies and Techniques

Surveillance is a distinctive concept of the modern world. Surveillance technologies and techniques highlight how information is gathered, stored, retrieved and processed.

2.1 Digital Database Technology

Digital database technology is a pivot part of numerous computing systems. Data is permitted to be taken and distributed electronically and the level of data enclosed in these systems keeps rising at an exponential rate. That is why it is important to ensure the integrity of collected data and secure the private information from unauthorized access [42].

Technology such as smart phones is a great invention. It's more practical than a laptop but what we don't realize is that tracking someone's mobile phone has become very easy with the help of applications that likes to use locations. Turning off those applications is only the first step in securing our mobile phone [29]. Usually, we store information on our mobile phones in case we don't forget it. "Safe Note" is the application that allows personal information that we store on the phone to be encrypted using a pin. Also, all the "Safe Note" that are entered use a 128-bit encryption, which will make it difficult to get access to personal information stored on individual's phone.

Another application that is used to encrypt and protect text and email messages is called encrypted messages. This application allows us to give a password to all the messages that we want to transmit and receive the message in encrypted form. If the message is intercepted, then the only way a person will be able to decrypt the message is whether they had the application and also knew the password that is used to encrypt the message [29]. Even having a conversation via a mobile phone is not safe, it is prone to interception. HeyTell VoIP application allows all data and audio in transit to be encrypted [29]. It is very easy to use, just set up the privacy level before we dial a number. It basically converts the mobile phones into a digital walkie-talkie with encrypted messages [29]. The only downfall of these applications is that the person we want to correspond with must have these applications installed on their phones, but with the popularity of smart phones nearly everyone has, thus this problem can be easily fixed.

2.2 Audio Recording Technology

One of the important aspects of audio surveillance technology is recording. With the digitized recording technology, audio recorder, also known as voice recorder has become smaller and easier to use [11]. Digital audio recording can be classified in three ways [40]:

- *Compression*. Compressed audio (data) is where the amount of data recorded in a waveform is reduced for transmission. The two types of compression are lossless, where compression exploits statistical redundancy to represent data more concisely without losing information [40]. Another type is lossy. With lossy, some loss of information is acceptable. Depending on the application, details can be dropped from the data to save storage space [40]. It may not be a good option for

portable storage because it takes up too much space but it is a commonly used format for uncompressed archiving.

- *Long play*. With development of video disk system, it has made long play digital audio disk system possible [28]. The bandwidth requirement for the channels of digital audio signal is less than that of the video signals, and combines with a reduction of the revolution, it makes the longer playing time possible [64]. The 3PM (three-position modulation) code is implemented to improve the packing density. The packing density can attain 150 % of MFM coding for the same minimum wavelength to be recorded, which is achieved at the expense of decreasing the jitter margin [27].
- *Storage*. Hard disk recording system uses a high-capacity hard disk to record digital audio. This system represents an alternative to more traditional reel-to-reel tape or cassette multi-track systems, it provides editing capabilities which is unavailable to tape recording [66].

2.3 Video Imaging Techniques

Nowadays video surveillance systems are widespread used in many strategic places such as public transportation, airport, banks, they also use in public place such as elevators, stores, and hallways. Video surveillance systems are becoming more and more ubiquitous. However, people usually feel safe, thanks to the sense of increased security by using surveillance systems but there are people who fear the content that has been obtained via surveillance which can cause identity leakage and privacy loss [20]. In particular, there could be five ways where surveillance systems would likely be abused [63]:

- Criminal abuse or identity thieves
- Institutional abuse or bad law enforcement use
- Discriminatory targeting or find disproportionately on people color
- Voyeurism or stalking women
- Abuse for personal purpose or track estranged spouses

Due to privacy issue, the concern about the conflict between security and privacy for video surveillance is on rise [90]. Therefore, over the past few years, many researches to protect privacy in surveillance systems have been made. They have proposed various approaches for the privacy-protecting goal, including using distortion filters to pixelize, blur, black out, silhouette, transparency, replaced by generic object and mask the object contained sensitive information which may explore privacy [93].

Although these approaches and techniques in some degree fulfil the purpose of protecting privacy in surveillance systems, but they also have many kind of flaws. Five criteria for developing effectiveness privacy protection approaches are [49, 51, 70]:

- Intelligibility of video
- Cryptographic technologies

- Compression efficiency
- Computational complexity
- Ease of integration

In addition, there are also five main parts for privacy protection technologies in video surveillance system [106]:

- *Privacy protection techniques and technologies*. Detect the Region of Interest (ROI) need to protect and then scrambling.
- *Coding and encryption efficiency*. Encryption by using JPEG, MPEG, H.264/AVC etc.
- *Security management*. Prevent outside attack such as brute-force attack, error concealment attack etc.
- *Selective storage*. Secure database.
- *Access control*. Authorized access to the original copy of video.

Furthermore, with video and images, following four techniques can be used to enhance the security:

- *Discrete Wavelet Transform* (*DWT*). DWT is becoming popular in many image applications due to its multi-resolution representation feature [16]. DWT based two-dimensional image follows pyramid-structured wavelet transform. The original image will encounter different combinations of a low-pass filter and a high-pass filter, and then based on the convolution with these filters to generate the low-low (LL), low-high (LH), high-low (HL) and high-high (HH) sub-bands [9]. To obtain the next coarser scaled wavelet coefficients, the sub-band LL is further decomposed and critically sub-sampled. This process can be repeated several times. With the pyramid-structured wavelet transform, size of the original image is equivalent to summing all the decomposed sub-images up. Using this decomposition structure, there will be no information lost when the decomposed pieces are reconstructed. This reconstruction process is called the Inverse Discrete Wavelet Transform (IDWT) [41]. For example, Motion JPEG 2000 (JPEG XR) uses DWT based transformation, it used as a global transformation which performs at the level of a tile [100].
- *Discrete Cosine Transform* (*DCT*). DCT is widely used transform technique in image processing [58, 92], which is commonly used in image compression such as JPEG, Motion JPEG, MPEG, and DV etc. [78, 105]. In 2D blocked DCT, $N \times N$ blocks are computed, N is normally 8. Typically DCT is limited to this size. Instead of converting the image as whole, DCT used 8×8 blocks separately to the image [73].
- *Pixelization*. Pixelization is a technique used for modifying image or video for privacy protection, it is achieved by noticeably lower resolution in ROI, using a square block of pixels with its average [22, 37]. The primary purpose is to use for censorship. It is commonly used in television news to obscure the object contained sensitive information such as nudity, the proper name of people, locations or any other inappropriate discourse [21]. The advantage of using pixelization in video surveillance system for privacy protection is very simple and easy to integrate in existing system. On the other hand, the disadvantage is that the process is

irreversible and the privacy information is lost [88]. The equation of pixelization of image $I(x, y)$ is shown in Eq. (1).

$$\bar{I}(x, y) = \frac{1}{b^2} \sum_{i=0}^{b-1} \sum_{j=0}^{b-1} I\left(\left\lfloor \frac{x}{b} \right\rfloor \cdot b + i, \left\lfloor \frac{x}{b} \right\rfloor \cdot b + j\right) \tag{1}$$

where image pixel coordinates are x and y, block size is b and $\lfloor \cdot \rfloor$ indicates the floored division.

- *Gaussian Blurring* (*Smoothing*). Gaussian blurring is an approach widely used for privacy protection in video surveillance, it removes details in ROI by using a Gaussian low-pass filter [5]. It is also commonly used with edge detection [37, 46]. The equation of Gaussian function in one dimension is shown in Eqs. (2) and (3).

$$g(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \tag{2}$$

where μ is the mean, σ is the variance, $x \sim N(\mu, \sigma)$.

In multi-dimension, it is shown as Eq. (3):

$$G(X) = \frac{1}{\sqrt{(2\pi)^k |\Sigma|}} e^{-\frac{1}{2}(X-\mu)^T \Sigma^{-1} (X-\mu)} \tag{3}$$

where $|\Sigma|$ is determinant of the matrix Σ, $k = |X|$, $X = (x_1, x_2, \ldots, x_k)$, $X \sim N_k(\mu, \Sigma)$.

Scalable Video Coding (SVC) approach is an extension of standard H.264/MPEG-4 AVC for video coding which provides greater coding flexibility and support by three forms of scalability which are spatial, temporal and SNR (Signal-to-Noise Ratio) scalability [30]. It also allows video bit stream to be broken into multiple layers of resolution, frame rate and quality, called Flexible Macroblock Ordering (FMO). The encoded bit stream includes Video Coding Layer (VCL) and Network Abstraction Layer (NAL). The VCL and NAL units contain coded slice data, and the non-VCL and NAL units contain associated additional information like Supplemental Enhancement Information (SEI), Sequence Parameter Sets (SPSs), and Picture Parameter Sets (PPSs). The picture parameter set transmits the parameters that indicate an ROI, like the slice group map type, the slice group IDs, and the top-left and bottom-right address of the slice groups [94]. The ROI can be taken out of JPEG XR by extracting spatial tiles in the compressed domain area, in cooperation of spatial and frequency mode. This feature of JPEG XR is also known as fast tile extraction [86].

There are two different types of tile layouts that are being used, they are a uniform and a non-uniform tile grid. The uniform tile layout where every tile holds identical width and height, while the non-uniform tile layout allows the use of tiles with dissimilar widths and heights (tiles located on the same row still require have the identical height, whereas tiles on the same column will still require having the identical width). The non-uniform tile outline is illustrated in Fig. 1 [48, 69]. For the purpose of updating location of an ROI, extra picture parameter set NAL units

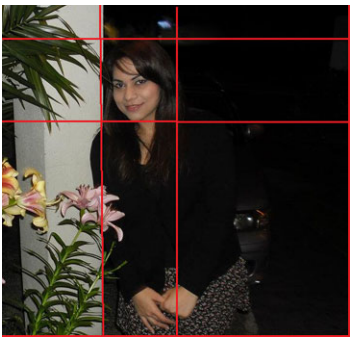


Fig. 1 ROI illustration in JPEG XR

that can be added into a coded bit stream, which includes the updated ROI coordinates. The new picture parameter set NAL unit is used to indicate each move of the ROI [94].

The main advantage of scalable video coding in video surveillance system depends on the circumstances that can create a One-Source Multi-Use (OSMU) service. Such a path is able to provide adapted video content in best condition to law-enforcement authorities and authorized clients through heterogeneous network environments [17, 85, 94].

The AVC scrambling approach is sourced from a Pseudo Random Number Generator (PRNG), which is initialized by a seed value. Multiple seeds can be used to strengthen the security. The seed values are forwarded to the correct area so it can be encrypted, usually done by asymmetric encryption, which is then transmitted to the decoder either via private data or a different channel [30]. Authorized personal, in custody of the secret encryption key, can then manage to recover the seed values and consequently reproduce an identical pseudo-random series to descramble the coefficients [30].

Normally scrambling process doesn't pose any negative impact on coding efficiency. A likely option is therefore to apply scrambling to the AC coefficients. Additionally, the amplitude of AC coefficients is linked, but their signs do not correlate. Per consequent, it has been suggested to scramble the quantized AC coefficients of all 4×4 block of the MB in the forefront slice group by pseudo-randomly tossing their sign [30]. This sort of technique involves negligible computational complexity. On the contrary, second scrambling method takes a random permutation to reorganize the order of AC coefficients in 4×4 blocks relating to MB in the forefront slice.

In Fig. 2, the Hallway camera test sequences in CIF format, with ground truth added to segmentation marks. Different types of experiments have been tested with JM13.2 reference software [30]. The capability of the presented scrambling is to conceal information in ROI. Figure 2 represents the result for both random sign inversion as well as permutation methods. The result clearly shows that both approaches are efficient at masking the ROI so that a person can no longer be recognized. It can also be noted that in spite of scrambling the image, it can still be accurately comprehended [30]. The effect of the two planned scrambling technique on coding efficiency can be compared to regular AVC format.

Fig. 2 Scrambling for "hallway camera". (**a**) Random sign inversion. (**b**) Random sign permutation

The data clearly shows that the scrambling has nominal impact on coding efficiency. Even though the bit rate increase of random sign inversion scrambling data is of 1 % at high level and 8 % at low level. The random permutation scrambling generated a somewhat big penalty with a rate increase of 4 % at the elevated end and 11 % was at the down end [30].

3 Systems and Tools for Surveillance Privacy Protection

Surveillance systems not only provide full security for the system and the physical property but also ensure safety of the employees and the public. There are many systems that have been implemented to monitor threats and possibly prevent criminal activities.

3.1 Systems

In the past, getting a hold of audio and video used to be a simple matter. An absurd amount of audio visual data is becoming accessible in digital form, in digital archives, on internet, in live transmission data streams and in confidential and professional databases. The importance of information normally depends on how effortlessly data can be obtained, sorted out and managed at the end [15].

Video surveillance systems are being commonly used due to high-speed network connections, and they are able to hold huge amount of sensitive data, at a high computational authority [8]. Additionally, thanks to constantly developing computer vision algorithms, video surveillance systems managed to analyze more data and understand events of security. When it comes to video surveillance system, spatial resolution and visual quality are key factors for the performance of computer vision algorithms [23]. Obviously use of high-resolution and high-quality video content can then enhance the overall performance of computer vision algorithms by directing object detection, recognition and tracking [15].

Along with video surveillance systems, privacy has rapidly become a vital issue. Although video surveillance systems can assist in limiting law-breaking and

criminal activities, on the other hand, extensive use of security cameras has led to well written political campaigns [31, 98]. Current research results have revealed that new Privacy Enabling Technologies (PET) are promising with the prospective to successfully protect individual privacy, with no major hindering video surveillance tasks. The final results confront the common surmise that increased security may overcome a failure of privacy [74, 75].

The issue of privacy protection has been catching a lot of interests though performance investigation is still missing. It is also very important to authenticate planned PET against client and system requirements for privacy use. Furthermore, we are unsure whether these approaches will be easily incorporated into existing system and whether surveillance architecture can be organized at a wider platform [53, 75]. We can undertake this issue by assessing the competence of PET to make facial identity jumbled, in future, this will render face recognition techniques useless and conceal the individual identity from the public.

At present, this is a key risk to surveillance privacy protection especially in video surveillance department. Basically, a face de-identification algorithm is explained in [107], which holds a big amount of key facial characteristics but still manage to make the individual face unidentifiable. Needless to say that PET does not hinder successful face detection [87]. In the past, companies managed to explain a framework to assess the performance of face recognition algorithm that were used on pictures that has been changed by PET and supported by the Face Identification Evaluation System (FIES) [12]. Carried out trials on the Facial Recognition Technology (FERET) records reflect the incompetence of raw PET, for instance, pixelization and blurring the image, and established the usefulness of more complicated scrambling methods to hide face recognition from unauthorized personal [19].

Nowadays, extensive experiments are being done, including with PSNR and SSIM (Structural Similarity) objective quality measures to recognize facial features [19, 101]. Simple incorporation of PET in present existing video surveillance infrastructure is one more imperative decisive factor that can lead to quick implementation of the technology.

Concurrently, with legacy operating systems make sure that it will present a broader and lucrative relate back to PET [57, 107]. Thus, methods which solely depend on extensively used video cryptogram standards (e.g., MPEG-4, H.264/AVC, Motion JPEG), in place of proprietary demonstration, should be favored. PET also holds conserve syntax layout compliance that present a significant benefit. In this instance, standard decoders can precisely decode and demonstrate the CCTV video stream, while some areas may be obscured.

Protecting the stream syntax code and permits content modification stand on either scalability as well transcoding throughout network broadcast [107, 108]. One more helpful characteristic is to broadcast the same guarded video stream to the entire end-users and their identification is certainly not necessary either. Ultimately, video surveillance evidence is normally used in examination forensic analysis by law authorities like CIA or FBI [53]. Therefore, it is very important that PET is completely reversible in case of emergency [50, 66]. This will eventually help authorized user possibly to solve or even recover the unaffected privacy-sensitive footage informative. In some instances, inconsequential PET methods purely applying pix-

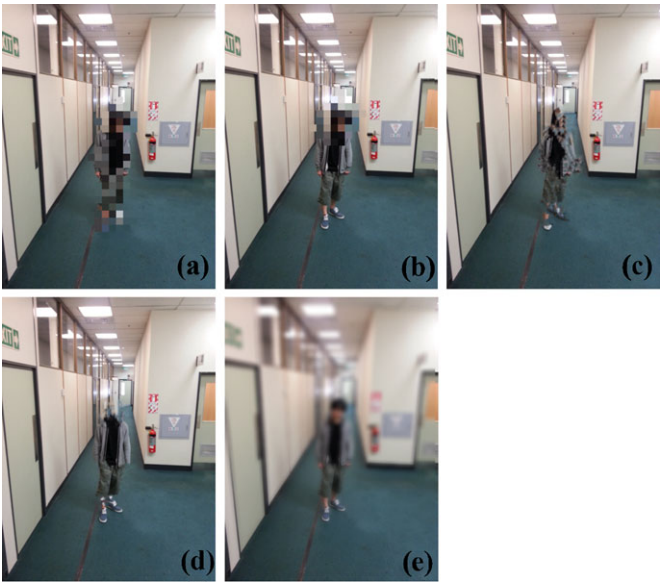


Fig. 3 The result of the selective privacy protection approaches. (**a**) Pixels with whole object. (**b**) Pixels with the facial. (**c**) Scrambling by random. (**d**) Facial removed permutation. (**e**) Blurring

elization, noise disruption, or black box disguise to cover confidential data that tend not to fulfil this vital obligation [80, 87].

The gray scale Facial Recognition Technology (FERET) [30] database can be used for training and testing [75]. The standard training is performed using a training set of images. The training set includes 30 images taken from different scenarios. This data set is used to train the surveillance system, and then carry out experiments for detecting, intra-coding, and scrambling functions for privacy protection purpose.

The face recognition algorithms like Principal Components Analysis (PCA) [23] and Linear Discriminant Analysis (LDA) [24, 57] techniques are used. PCA is basically a mathematical procedure that involves orthogonal transformation to alter a set of reflection of possibly connected variable into linearly uncorrelated variable which are called as principle components. LDA is basically related to Fisher's linear discriminate that is a method used to discover a linear combination of characteristics which can be exemplify or separate two or more classes of objects or procedures [24]. Figure 3 shows the result of the selective privacy protection approaches. The Region Of Interest (ROI) in the training set is expected to detect with minimum failure, and hide by using the selective scrambling approach. Whereas Facial Identification Evaluation Systems (FIES) consists of four key components [12]:

- Image pre-processing
- Initial training
- Testing quality
- Performance analysis

These steps are designed to minimize detrimental variations between the facial images. Individual facial feature is initially geometrically stabilized and line up with the eye coordinates, then an elliptical mask is used to trim the facial images. More exclusively, the face area between the forehead and chin along with left side cheek and right side cheek is held, meanwhile the rest of the data is disposed. Histogram equalization is executed, contrast and brightness of the image are normalized [12, 108]. Finally, face recognition piece is examined. More explicitly, a collective match curve is produced. For this particular reason, the recognition rank is calculated. After that cumulative match curve is taken by adding the number of correct matches from each ranked individual [38, 45, 53].

Another system that can be used is PRISURV. PRISURV was designed for a small community such as school area or office place. PRISURV adopted a mechanism named visual abstraction to control disclosure of visual information [13]. PRISURV is able to generate several types of images according to viewer's authority level, in other way it enables video surveillance system to manage and control privacy of the object shown in the video regarding to different viewer [72]. PRISURV consists of six main components [60]:

- *Analyzer*. The two main functions of analyzer are image stratification and subject identification. Image stratification is used to produce stratified images of surveillance video, every image represents one subject. The stratification is accomplished by background subtraction and subject region extraction by projecting the foreground image vertically. Subject identification is used to distinguish subject's identity. Every subject in the stratified images is identified by video analysis.
- *Profile Generator*. Profile generator is used to setup the profile for registered member, it contains member's privacy information such as name, gender, age, address and relationships. Profile generator also connected with privacy policy to determine the outcome of the video according to the relationship between viewer and the object shown in the video. Each profile can only be modified and updated by member themselves, it is inaccessible by non-members.
- *Profile Base*. Profile base is a secure database server used to store all the profiles.
- *Access Controller*. Access controller is used to match up viewers' information to subjects' privacy policies to determine what kind of abstraction is needed to make and then send the command to the Abstractor.
- *Abstractor*. Abstractor is used to process the video for visual abstraction according to the command received from access controller. It adopts visual abstraction approach which can generate 12 different abstractions on video.
- *Video Database*. Video database is a secure database used to store past video, and play them to viewers through visual abstraction when needed.

Hidden inference channels of *what*, *when* and *where* can initiate considerable level of privacy loss when an adversary gets a hold of several-camera surveillance video footage. The privacy loss that was carried out through these inference channels is modeled as W^3-Privacy [84]. The privacy loss calculated by the presented model

Table 1 Specifications for digital surveillance

Specifications			
Capsa Free	Device Monitoring Studios	PRTG Network Monitor	Verilook Surveillance SDK
Real time network video capturing	Software solution for monitoring	Device monitoring application	Performs searching and detection of faces
Traffic monitoring	Logging network activity	Lightweight application	Can track multiple faces simultaneously
Expert network diagnosis	Analyzing data coming through pc parts	Can be configured to find a subnet	Can be run on multiple PCs
Network activity logging	Analyzing data through physical connection media	Uses Graphical User Interface	New faces saved to database either manually or automatically

is nearer to the user perceived privacy loss rather than prior models. In addition, privacy loss can only take place when sensitive information and identity leakage co-exist at the same time. That is why any of these can be managed separately to minimize privacy loss. For example, in a surveillance setting, the tenants of the surveyed area can present sensitive information and the person who has authorization to this surveillance footage can be measured as an adversary.

The W^3 survived assessment model is essentially the foremost and very useful step towards privacy protection of individuals in multi-camera video footage. This work does help to set up directions for future research, for instance, to investigate ways to lessen the privacy loss with minimum loss of efficacy in video quality [84]. However, our surveillance privacy protection research covers multiple types of surveillance, privacy and protection measures which are presently being used in order to guard an individual's privacy at all time.

3.2 Tools

With every technology software plays an important role. The following tables show software tools that can be used for digital surveillance and highlights the features that are supported by the software. Table 1 highlights that software tools that are being used for digital surveillance [54].

Audio surveillance also plays an important role in terms of surveillance. Audio surveillance devices can also be described as listening devices. Some of the audio surveillance softwares available have been showcased in Table 2 [76].

Ableton Live is an audio recording software program. As well as audio surveillance, it can also be used as editing tools for audio [76]. Pro Tools LE has been

Table 2 Audio surveillance features

Audio features	Ableton Live	Pro Tools LE	Komplete	Sound Forge	Acid Pro	Audio Mulch
Audio conversion tools	√	√	√	√	√	×
Stabilized performance mode	√	√	×	×	×	×
Volume maximizer	√	√	×	×	×	√
External control compatibility	√	√	×	√	√	×
Auxillary port	√	√	√	√	√	×
Supports XP, Windows 7, Windows 8	√	√	√	√	√	√

Table 3 Tools and features for video surveillance

Tools features	Photos editing	Videos recording	Images blurring	Code editing debuging
OpenCV	√	√	√	√
Luxand FaceSDK	√	√	√	×
Keylemon	√	√	√	×
ImageGraphicsVideo	√	√	√	×
Logitec Face Recognition software	√	√	√	×

one of the greatest software for audio surveillance tool for many years. As Ableton, Pro Tools LE also has the functionality for recording and editing audio [76]. Komplete is another audio recording software. It is compatible with all major audio types [76]. Sound Forge is a type of audio surveillance software that can record multi-channel audio at the same time. It is also able to reduce noise and repair sound quality. Acid Pro is one of the leading technologies in audio surveillance software. It also allows users for multi-channel recording. AudioMulch is modular audio recording software that can record and playback multiple sound files [76].

Table 3 highlights five popular video surveillance software tools.

- Intel OpenCV
- Luxand FaceSDK
- Keylemon
- ImageGraphicsVideo
- Logitech Face Recognition software
- 3vr

Face detection, infrared and blurring images are the most significant specifications that cameras should have [56]. Face detection is really significant function in surveillance camera. Face detection employs sophisticated algorithms to detect facial features [79]. Luxand FaceSDK is an example of this kind of software. The FaceSDK processes an image, detects human face within it, and returns the facial feature points such as eyes, eye contours, eyebrows, lip contours, nose tip, and so on [105]. Some of the other face recognition software that is available for users are: keylemon, ImageGraphicsVideo, Logitech Face Recognition software [78].

Intel OpenCV is defined as a computer vision library that sets its focus on the image processing [93]. Upgraded versions of OpenCV integrate with the programming languages such as C, C++, Python and Android [106]. Some of the currently available functions for OpenCV are:

- Eighenfaces (createEighenFaceRecognizer())
- Fisherfaces (createFisherFaceRecofnizer())
- Local Binary Patterns Histograms (createLBPHFaceRecognizer())

There are numerous face recognition databases [17]. The choice of the appropriate database to be used based on the task given (aging, expression, lighting, etc.). Some of the face recognition databases are: The Colour Feret Database, SCFace, Multi-Pie, The Yale Face Database, and Face in Action (FIA) Face Video Database, AT&T, Cohn–Kanade AU Coded Facial Expression Databases, NIST Mugshot Identification Database, NLPR Face Database, The AR Face Database, Caltech Faces, and Georgia Tech Face Database. Although there could be hundreds of available face recognition database, there are two very useful databases:

- *AT&T face database*. In this database, the images are taken at different times varying the lightning and facial expressions such as open/closed eyes, smiling as well as the facial details such as glasses/no glasses [60].
- *Yale face database*. The database is a fairly simple database to use. It catches the images at different times when the person is: happy, sad, angry, or sleepy [13].

Since infrared camera has the ability to pick up movement in dark scenarios, it will be difficult to obtain images properly without infrared cameras [88]. The camera should pick all the faces that enter the surveillance area and blur all other images. If the camera does not provide blurring function, it will break this community privacy rule [94]. 3vr is another good example for blurring surveillance cameras images. 3vr is software that catches all suspicious behaviors that occur in vision range of the cameras. As well as catching the suspicious people, it also catches all other innocent people and tweaks its software to automatically blur the faces of these innocent individuals [85]. Therefore, it protects the privacy of the innocent individuals [55]. Some of the 3vr's video analytics include:

- Facial surveillance
- Advanced object tracking
- People counting
- Queue line analysis

4 Conclusion

Privacy is one of the major concerns in this technological society. Surveillance technology has reached a place where it is impossible to avoid being caught under surveillance. There are growing trends towards privacy and data protection acts in the world [78]. Today, we are seeing advancements in the technology and the approaches of databases [1]. Since the September 11, 2001 attacks, surveillance technology has leapfrogged into the 21st century [27]. Public protection has become fundamental, in future we will see more enriched surveillance systems as well as evolving database systems.

In conclusion, this chapter has evaluated different types of surveillance, privacy and protection measures which are currently being implemented in order to protect an individual's privacy. In addition, this chapter also covered how relevant information is being used and what sort of rights individuals might have over them. Furthermore, this chapter has emphasized the importance of tools, data sets and also the databases which have been developed to provide protection for surveillance privacy.

References

1. Adams, A., Ferryman, J.M.: The future of video analytics for surveillance and its ethical implications. Secur. J. **3**(1), 1–22 (2012)
2. Alderman, E., Kennedy, C.: The Right to Privacy. Vintage Book, New York (1995)
3. Baker, B.D., Gunter, W.D.: Surveillance: concepts and practices for fraud, security and crime investigation. Int. Found. Prot. Off. **2**, 1–17 (2005)
4. Barrett, T.: Data privacy day is just around the corner—are you respecting privacy, safeguarding data and enabling trust? Nat. Cyber Secur. Alliance **4**, 105–114 (2010)
5. Baym, N.K.: A call for grounding in the face of blurred boundaries. J. Comput.-Mediat. Commun. **14**(3), 720–723 (2009)
6. Beagle, T.: Search and surveillance act threatens privacy. Tech Lib. NZ **3**, 77–85 (2009)
7. Bertino, E., Sandhu, R.: Database security-concepts, approaches, and challenges. IEEE Trans. Dependable Secure Comput. **2**(1), 2–19 (2005)
8. Beveridge, R., Draper, B., Givens, G., Fisher, W.: Introduction to the statistical evaluation of face recognition algorithms. In: Face Processing, Advanced Modeling and Methods. Elsevier, Amsterdam (2005)
9. Boult, T.E.: PICO: privacy through invertible cryptographic obscuration. In: Proceedings of the Computer Vision for Interactive and Intelligent Environment, pp. 27–38 (2005)
10. Bowyer, K.W.: Face recognition technology: security versus privacy. IEEE Technol. Soc. Mag. **23**(1), 9–19 (2004)
11. Calvel, C., Ehrette, T., Richard, G.: Event detection for audio-based surveillance system. IEEE Multimed. Expo. **1**, 1306–1309 (2005)
12. Capdevila, G.: Communications: Technologies for Expression and for Censorship. Inter. Press Service, Geneva (2003)
13. Carrillo, P., Kalva, H., Magliveras, S.: Compression independent object encryption for ensuring privacy in video surveillance. In: IEEE ICME, pp. 273–276 (2008)
14. Carter, M., DeMolay, J., Kuszai, J.: Global positioning satellites as surveillance devices. Prog. Astronaut. Aeronaut. **5**, 20–28 (2006)
15. Cavallaro, A.: Privacy in video surveillance. IEEE Signal Process. Mag. **24**(2), 168–169 (2007)

16. Chen, D., Chang, Y., Yan, R., Yang, J.: Tools for protecting the privacy of specific individuals in video. EURASIP J. Adv. Signal Process. **1**(1), 107 (2007)
17. Chinomi, K., Nitta, N., lto, Y., Babaguchi, N.: PriSurv: privacy protected video surveillance system using adaptive visual abstraction. Adv. Multimed. Model. Lect. Notes Comput. Sci. **4903**(2), 144–154 (2008)
18. Citron, D.K., Gray, D.: Total surveillance's privacy harms: a reply to Professor Neil Richards. Harvard Law Rev. **1**, 2–10 (2012)
19. Clarke, R.: Internet privacy concerns confirm the case for intervention. Commun. ACM **42**(2), 60–67 (1999)
20. Collins, R.T., Lipton, A.J., Kanade, T., Fujiyoshi, H., Duggins, D., Tsin, Y., Tolliver, D., Enomoto, N., Hasegawa, O., Burt, P., Wixson, L.: A system for video surveillance and monitoring. United States of America: Robotics Institute at Carnegie Mellon University (CMU) (2000). http://www.ri.cmu.edu/publication_view.html?pub_id=3325
21. Cotton, W.D., Uson, J.M.: Image pixelization and dynamic range. Nat. Radio Astron. Obs. **1**(1), 3–10 (2007)
22. Cotton, W.D., Uson, J.M.: Pixelization and dynamic range in radio interferometry. Astron. Instrum. **490**(1), 455–460 (2008)
23. Cullen, R.: Identify and information privacy in the age of digital government. Online Inf. Rev. **33**(3), 405–421 (2009)
24. DeBeer, J.F., Clemmer, C.D.: Global trends in online copyright enforcement: a non-neutral role for network intermediaries? Jurimetrics **49**(4), 375–409 (2009)
25. Dempsey, J.S.: Electronic surveillance and interception. In: Introduction to Private Security, vol. 4, p. 257 (2010)
26. Desurmont, X., Bastide, A., Chaudy, C., Parisot, C., Delaigle, J., Macq, B.: Image analysis architectures and techniques for intelligent surveillance systems. IEE Proc., Vis. Image Signal Process. **152**(2), 224–231 (2005)
27. Doi, T., Itoh, T., Ogawa, H.: A long-play digital audio disk system. J. Audio Eng. Soc. **27**(12), 975–981 (1979)
28. Dong, X.: Data hiding via phase manipulation of audio signals. In: IEEE ICASSP'04, pp. 377–380 (2004)
29. Dube, R.: How to protect yourself from government cellphone surveillance. N.Y. Times **2**, 87–95 (2005)
30. Dufaux, F., Ebrahimi, T.: H.264/AVC video scrambling for privacy protection. In: IEEE ICIP, pp. 1688–1691 (2008)
31. Dufaux, F., Ebrahimi, T.: A framework for the validation of privacy protection solutions in video surveillance. In: IEEE ICME, pp. 66–71 (2010)
32. Fickes, M.: Automated eye in the sky. Gov. Secur. **1**, 22 (2004)
33. Flaherty, D.: Protecting Privacy in Surveillance Societies, vol. 22(17), pp. 75–89. University of North Carolina Press, Chapel Hill (1989)
34. Flemming, J.: Privacy: the menace of satellite surveillance. USA Evid. Lawsuit Filed **3**, 82–105 (2003)
35. Fraser, D.T.: Privacy law and video surveillance: guidance from the ontario courts. Mcinnes Cooper **3**(1), 10–13 (2004)
36. Gilens, N.: New justice department documents show huge increase in warrantless electronic surveillance. ACLU Speech, Priv. Technol. Proj. **1**, 4–7 (2012)
37. Gouaillier, V.: Intelligent video surveillance: promises and challenges technological and commercial intelligence report. CRIM Technopôle Def. Secur. **3**(2), 9–68 (2009)
38. Graves, L.: The right to privacy in light of presidents' programs: what project MINARET's admissions reveal about modern surveillance of Americans. Tex. Law Rev. **88**(7), 1855–1904 (2010)
39. Gray, M.: Urban surveillance and panopticism: will we recognize the facial recognition society? Surveill. Soc. **1**(3), 314–330 (2003)
40. Grigoras, C.: Digital audio recording analysis: the electric network frequency criterion. Diam. Cut Product. Inc. **4**, 1–6 (2003)

41. Gu, G.S., Han, G.Q.: The application of chaos and DWT in image scrambling. In: Machine Learning and Cybernetics, pp. 3729–3733 (2006)
42. Hampapur, A., Brown, L., Connell, J., Ekin, A., Haas, N., Lu, M., Merkl, H., Pankanti, S.: Smart video surveillance: exploring the concept of multiscale spatiotemporal tracking. IEEE Signal Process. Mag. **22**(2), 38–51 (2005)
43. Hier, S.P., Walby, K.: Privacy pragmatism and streetscape video surveillance in Canada. Int. Sociol. **26**(6), 844–861 (2011)
44. Hill, M.: Government funds chat room surveillance research. USA Today Tech **2**, 30–38 (2004)
45. Hooper, T., Vos, M.: Establishing business integrity in an online environment: an examination of New Zealand web site privacy notices. Online Inf. Rev., 343–361 (2009)
46. Hummel, R.A., Kimia, B., Zucker, S.W.: Deblurring Gaussian blur. Comput. Vis. Graph. Image Process. **38**(1), 66–80 (1987)
47. Julie, R.S.: High-tech surveillance tools and the fourth amendment: reasonable expectations of privacy in the technological age. Am. Crim. Law Rev. **37**(1), 192–222 (2000)
48. Junga, K., Kimb, K.I., Jainc, A.K.: Text information extraction in images and video: a survey. Pattern Recognit. **37**(5), 977–997 (2004)
49. Karimaa, A.: Efficient video surveillance: performance evaluation in distributed video surveillance systems. In: Video Surveillance, pp. 17–26. InTech, Rijeka (2011)
50. Koskela, H.: 'The gaze without eyes': video-surveillance and the changing nature of urban space. Prog. Hum. Geogr. **24**(2), 243–265 (2000)
51. Kumar, G.S., Ragu, S., Kumar, N.S.: Embedded video surveillance with real time monitoring on web. Int. J. Math. Trends Technol. **2**(1), 46–49 (2011)
52. Labaton, S.: Learning to live with big brother, from civil liberties: surveillance and privacy. In: The Economist Intelligence, vol. 384(8548), pp. 62–64 (2007)
53. Lace, S.: The Glass Consumer: Life in a Surveillance Society. Policy Press, Bristol (2005)
54. Lambert, T.A.: Digital surveillance system with pre-event recording. U.S. Patent WO/2001/035668 (2001)
55. Langheinrich, M.: Privacy by design—principles of privacy-aware ubiquitous systems. In: Proceedings of UbiComp, pp. 273–291 (2001)
56. Li, G., Ito, Y., Yu, X., Nitta, N., Babaguchi, N.: Recoverable privacy protection for video content distribution. EURASIP J. Multimed. Inf. Secur. **3**(4), 2–9 (2009)
57. Lyon, D.: Identifying citizens: ID cards as surveillance. Br. J. Sociol. **62**(4), 749–750 (2011)
58. Maheshwari, S., Gunjan, R., Laxmi, V., Gour, M.S.: A DCT based permuted image digital watermarking method. In: IEEE TENCON, pp. 2264–2268 (2010)
59. Mann, S., Nolan, J., Wellman, B.: Sousveillance: inventing and using wearable computing devices for data collection in surveillance environments. IEEE Surveill. Soc. **1**(3), 331–355 (2003)
60. Martin, K., Plataniotis, K.N.: Privacy protected surveillance using secure visual object coding. IEEE Trans. Circuits Syst. Video Technol. **18**(8), 1152–1162 (2008)
61. McCarthy, A.: Closed circuit television. Mus. Broadcast Commun. **2**, 101 (2012)
62. McCullagh, D.: Privacy bill requires search warrants for e-mail. Cell Track. **1**, 8–13 (2012)
63. Milan, P., Jonker, W.: Security, privacy, and trust. In: Modern Data Management. Springer, New York (2007)
64. Miller, M.: Understanding digital audio formats. Que Publ. **1**, 1–4 (2005)
65. Millett, T.: Copyright guidelines for research student. Libr. Consort. N.Z. **2**, 3–14 (2012)
66. Mynatt, E., Back, M., Want, R., Baer, M., Ellis, J.: Designing audio aura. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp. 566–573 (2010)
67. Nabbali, T., Perry, M.: Computer law & security report. Elsevier Sci. Ltd. **20**(2), 84–94, (2004)
68. Nagendran, A., Harper, D., Shah, M.: New system performs persistent wide-area aerial surveillance. SPIE Newsroom **5**, 20–28 (2010)
69. Newton, E.M., Sweeney, L., Malin, B.: Preserving privacy by de-identifying face images. IEEE Trans. Knowl. Data Eng. **17**(2), 232–243 (2005)

70. Norris, C., McCahill, M., Wood, D.: The growth of CCTV: a global perspective on the international diffusion of video surveillance in publicly accessible space. Surveill. Soc. **2**(2/3), 110–135 (2004)
71. Orwell, G.: Go on, watch me: people are voluntarily surrendering their privacy. The Economist **75**(4), 1125–1192 (2002)
72. Pankanti, S., Hampapur, A., Brown, L., Tian, Y.L., Ekin, A., Connell, J., Shu, C.F., Lu, M.: Enabling video privacy through computer vision. IEEE Secur. Priv. **3**(3), 50–57 (2005)
73. Paruchuri, J.K., Cheung, C.S., Hail, M.W.: Video data hiding for managing privacy information in surveillance systems. EURASIP J. Multimed. Inf. Secur. **8**(3), 18 (2009)
74. Pentland, A.: Face recognition using eigenfaces. In: Proceedings of IEEE CVPR, pp. 586–591 (1991)
75. Phillips, P.J., Moon, H., Rauss, P.J., Rizvi, S.A.: The FERET evaluation methodology for face recognition algorithms. IEEE Trans. Pattern Anal. Mach. Intell. **22**(10), 252–274 (2000)
76. Pollock, D.A.: Method of electronic audio surveillance. Law J. Libr. **5**(12), 380–385 (2002)
77. Poulsen, K.: FBI's secret spyware tracks down teen who made bomb threats. Priv. Secur. Crime Online **1**, 40–44 (2007)
78. Promyarut, I., Suvonvorn, N., Limsiroratana, S.: Video scrambling for privacy protection in surveillance system. In: Proceedings of International Conference on Circuits, System and Simulation, pp. 177–182 (2011)
79. Rao, K.R., Yip, P.C., Britanak, V.: Discrete Cosine Transform: Algorithms, Advantages, Applications. Academic Press, San Diego (2007)
80. Raths, D.: Video surveillance: all eyes turn to IP. In: Campus Technology (2011)
81. Richmond, R.: 12 ways technology threatens your privacy (and how to protect yourself). Inf. Manag. Comput. Secur. **2**, 12–13 (2009)
82. Roberti, M.: Two stories higlight the RFID debate. RFID J. **2**, 15–22 (2005)
83. Rubin, R.I., Stempler, M.J.: Video surveillance in personal injury cases. Fla. Bar J. **85**(6), 98 (2011)
84. Saini, M., Atrey, P.K., Mehrota, S., Kankanhalli, M.S.: W3-privacy: understanding what, when and where inference channels in multi-camera surveillance video. Multimed. Tools Appl., 1–24 (2012)
85. Schwarz, H., Marpe, D., Wiegand, T.: Overview of the scalable video coding extension of the H.264/AVC standard. IEEE Trans. Circuits Syst. Video Technol. **17**(9), 1103–1120 (2007)
86. Senior, A.: Protecting Privacy in Video Surveillance. Springer, London (2009)
87. Shamsi, H., Abdo A, A.: Privacy and surveillance post-9/11. Hum. Rights **38**(1) (2011)
88. Sigal, L., Balan, A.O., Black, M.J.: Synchronized video and motion capture dataset and baseline algorithm for evaluation of articulated human motion. Int. J. Comput. Vis. **87**(1–2), 4–27 (2010)
89. Simone, F.D., Naccari, M., Tagliasacchi, M., Dufaux, F., Tubaro, S., Ebrahimi, T.: Subjective quality assessment of H.264/AVC video streaming with packet losses. AEURASIP J. Image Video Process. **2**, 1–12 (2011)
90. Slobogin, C.: Public privacy: camera surveillance of public places and the right to anonymity. Miss. Law J. **72**(2), 30–60 (2002)
91. Slobogin, C.: Privacy at Risk: The New Government Surveillance and the Fourth Amendment. University of Chicago Press, Chicago (2007)
92. Smithsimon, M.: Private lives, public spaces: the surveillance state. Dissertation, University of Pennsylvania Press, USA (2003)
93. Socek, D., Kalva, H., Magliveras, S.S., Marques, O., Culibrk, D., Furht, B.: New approaches to encryption and steganography for digital videos. Multimed. Syst. **13**(3), 191–204 (2007)
94. Sohn, H., AnzaKu, E., Neve, W., Ro, Y.M., Plataniotis, K.: Privacy protection in video surveillance systems using scalable video coding. In: IEEE AVSS, pp. 424–429 (2009)
95. Sohn, H., Lee, D., Neve, W.D., Plataniotis, K.N., Ro, Y.M.: An objective and subjective evaluation of content-based privacy protection of face images in video surveillance systems using JPEG XR. In: Effective Surveillance for Homeland Security. Balancing Technology and Social Issues, vol. 3, pp. 111–140 (2013)

96. Solove, D.J.: The Digital Person: Technology and Privacy in the Information Age. NYU Press, New York (2004)
97. Strassfeld, R.N., Ough, C.: Somebody's watching me: surveillance and privacy in an age of national insecurity. Case West. Reserve J. Int. Law **42**(3), 110–543 (2010)
98. Thomas, D., Loader, B.D.: Cybercrime: Law Enforcement, Security and Surveillance in the Information Age. Routledge, London (2000)
99. Tien, L., Abernathy, W.: Biometrics: who's watching you? Electron. Front. Found. **1**, 9–11 (2003)
100. Trent, B.: Technology and tomorrow: a challenge to liberty. Humanist **64**(6), 21–24 (2004)
101. Tunick, M.: Privacy in public places: do GPS and video surveillance provide plain views? Soc. Theory Pract. **35**(4), 597–622 (2009)
102. Vaile, D.: Law in the information age 2.0. Masters Thesis, The University of New South Wales, Australia (2012)
103. Valenzise, G., Gerosa, L., Tagliasacchi, M., Antonacci, F., Sarti, A.: Scream and gunshot detection and localisation for audio surveillance. In: IEEE AVSS, pp. 21–26 (2007)
104. Walczak, R.: Network video surveillance technology tools expectations. Real World Video Surveill. **5**(2), 10–16 (2011)
105. Waterson, A.B.: Image compression using the discrete consine transform. Math. J. **4**(1), 81–88 (1994)
106. Wee, S.J., Apostolopoulos, J.G.: Secure scalable streaming enabling transcoding without decryption. In: IEEE ICIP, pp. 437–440 (2001)
107. Zhao, W., Chellappa, R., Krishnaswamy, A., Wen, J.: Discriminant analysis of principal components for face recognition. In: Proceedings of IEEE International Conference on Automatic Face and Gesture Recognition, pp. 336–341 (1998)
108. Zhou, N., Zhou, M.L., Xu, Y.P.: BACnet for video surveillance. BACnet Today, 18–23 (2004)

RFID Localization Improved by Motion Segmentation in Multimedia Surveillance Systems

Miloš Ljubojević, Zdenka Babić, and Vladimir Risojević

Abstract An important issue in multimedia surveillance systems is determining the physical location of moving objects. Due to features like contactless communications, high data rate, non-line-of-sight readability, compactness and low cost, passive Radio Frequency Identification technology is very attractive for indoor localization. Technologies and techniques can be employed in combination, aimed to improve accuracy and precision of localization by heterogeneous data fusion. Object recognition, moving object localization and tracking can be successfully implemented using integration of RFID technology and digital image processing techniques. The block matching algorithm based on region of interest can be efficiently used in image processing analysis for motion segmentation and object tracking. By using regions of interest we eliminate the influence of other large moving objects and avoid unnecessary computations. In this chapter, the improvement of RFID localization using motion segmentation applied on the region of interest extracted using RFID is described. The presented solution shows significant reduction of the position estimation error and variance in comparison to the conventional passive RFID position estimation.

1 Introduction

Integration of multiple technologies, using the advantages of different tools and a multidisciplinary view, leads to a new generation of systems for automatic scene analysis, surveillance, object localization and tracking, robotics and many other ap-

M. Ljubojević (✉)
Academic and Research Network of RS, Banja Luka, Bosnia and Herzegovina
e-mail: ljubojevic.milos@gmail.com

Z. Babić · V. Risojević
Faculty of Electrical Engineering, University of Banja Luka, Banja Luka, Bosnia and Herzegovina

Z. Babić
e-mail: zdenka@etfbl.net

V. Risojević
e-mail: vlado@etfbl.net

P.K. Atrey et al. (eds.), *Intelligent Multimedia Surveillance*,
DOI 10.1007/978-3-642-41512-8_6,

plications. Using different media (images, video, audio, sensor signals) in one complex system can help scene understanding, but at the same time generates new issues that have to be solved.

The fundamental problem in the above applications is determining the physical location of a moving object. In contrast to outdoor environments where precise location can be easily derived by using GPS data, indoor environments require a different approach to estimate physical location. Fusion of information obtained by several sensors is one of the possible methods for scene analysis and indoor localization of moving objects [1]. For precise localization, technologies and techniques can be employed in combination, because heterogeneous data fusion improves the accuracy and precision of localization. In multimedia systems, great attention has to be given to adequate selection of media types and information fusion.

1.1 Localization Technologies

The problems of localization of objects or persons are closely related to the choice of the applied technologies and techniques [2, 3]. The main characteristics used as criteria for choosing appropriate technology during localization system design are [4, 5]:

- physical position and symbolic location information,
- absolute versus relative location,
- accuracy and precision,
- size of localization area,
- recognition capability and object identification,
- cost realization.

The often used localization technologies in indoor and outdoor applications are:

Satellite technologies. Geostationary collocated satellites are used for positioning of moving objects. Among several satellite technologies, the Global Positioning System (GPS) is the most popular technology for localization and tracking of moving objects. Technical aspects of the GPS enable outdoor localization and tracking of moving objects with high precision and accuracy, but this technology is not suitable for indoor localization due to the poor GPS signal in indoor environment.

IEEE 802.11x technology. The existing wireless LAN (WLAN) infrastructure based on IEEE 802.11x standard can be easily and efficiently used for indoor localization. Accuracy and precision of localization is satisfactory and depends on the applied technique and chosen hardware. The signal propagation depends on the surrounding conditions, that is, the number of walls and obstructions, and the necessary infrastructure regarding the number of reference points per localization area. It directly affects the localization precision.

Bluetooth. Due to its presence in almost every mobile commercial device and short-range wireless connectivity, the Bluetooth technology represents a prom-ising

solution for indoor localization. Robustness, low complexity and low power consumption are the key features for choosing Bluetooth based localization system.

Infrared (IrDA). The IrDA technology is compact, low power, inexpensive and ubiquitous technology. On the other hand, if this technology is employed, ultimate line of sight presence, direct sunlight influence and short communication range restrictions have to be solved.

Ultrasound. Localization systems based on ultrasound technology have good precision and represent a simple and inexpensive solution. It is well known that environmental factors have substantial effects, so the large number of elements is necessary within the system which increases the costs of the installation.

Radio-frequency (RF) technology. The main characteristics such as non-contact communication, communication without the line of sight, short read time and simple maintenance make the RF based localization systems the most used in almost every area of implementation. Except localization information, those systems can provide valuable information for other systems, so they can be used as support systems within more complex, integrated systems [6].

Scene analysis. Important characteristic and big advantage of scene analysis is that the location of objects can be inferred using passive observation without any interaction with objects of interest. Observing the features of a scene, represented with different types of sensor signals, the conclusions about the location of the observer or objects in the scene can be drawn.

1.2 Localization Techniques

Very important aspect of localization and tracking of moving objects is a technique that is used for automatic positioning. For indoor localization, most frequently used techniques are: geometry based (trilateration and triangulation), proximity technique and visual scene analysis. The choice of technique depends directly on the technology used for localization.

Trilateration and triangulation. The technique of trilateration is based on measuring the distances between the sensors and the object. Position is determined as the point of intersection for at least three circles or spheres in terms of localization in two or three dimensional space. The centers of these circles or spheres are at the sensor locations and radiuses are the distances from the object. These distances can be determined based on several parameters: the time of arrival (TOA), time difference of arrival (TDOA), time of flight (TOF) and received signal strength (RSS). Trilateration technique is based on angle of arrival (AOA) as a parameter for determining the position of the reader. Position is calculated from the angles formed by at least two reference points and the object whose position is determined [2].

Proximity technique. With proximity technique, relative position of the object to the known, reference, location is determined by finding the moment when the object is near the reference locations. The presence of the object can be detected

by physical contact, wireless networks, or object identifiers. The physical contact approach uses appropriate sensors (pressure sensors, touch sensors, etc.). The localization based on wireless networks examines whether the mobile object is in the zone of supervised access points. The localization based on object identifiers uses identification systems to determine the presence and location of objects. This technique is often used in the localization based on RF technology.

Image and video processing techniques. Passive observation and data collection about the scene generate the necessary information that can be used for object localization. Scene analysis for object localization is usually based on visual information and methods of digital image and video processing. Beside visual, other types of information such as RF signals can be used for image generation and scene analysis.

1.3 Localization Systems

There are many systems for localization of moving object or persons based on different technologies and techniques. Up to date, indoor localization systems using one technology and one technique dominate. The RADAR system [7] measures the signal strength and signal-to-noise ratio of signals that wireless devices send in order to find the distance between the transmitting and receiving base stations. The location estimation is performed using triangulation, proximity or scene analysis. The Active Badge system [8] uses infrared technology and the system locates each person that wears a small infrared badge emitting the globally unique number. Appropriately arranged infrared sensors collect this data and provide information to a central server where the location of moving person is calculated. Cricket system [9] uses both the RF and ultrasound technologies for location estimation. The ultrasound emitters are used to create the infrastructure and the receiver is embedded in the moving object, while the RF signal is used for synchronization and delineation of the time interval during which the receiver should acquire ultrasound signals. APIT and DV HOP localization systems [10] are based on the heterogeneous networks, and beside the primary wireless sensors for localization, they use reference RF transmitters and triangulation technique for localization improvement. SpotON system [11] enables three-dimensional localization using radio signal strength. SmartFloor system [12] provides high precision localization using sensor network on the test floor with the known sensors' locations. Easy Living [13] localization system is based on scene analysis using image processing techniques. Bluepass system [14] is an example of localization system in indoor environment based on the Bluetooth technology.

Radio Frequency Identification (RFID), based on RF technology and proximity technique that uses identification systems to determine the presence and location of passive tags, is a very attractive technology for indoor localization because of its features like contactless communications, high data rate and security, non-line-of-sight readability, compactness and low cost. Important issues, such as privacy protection in a video surveillance system, also can be accomplished using RFID.

But, information obtained from RFID readers with proximity technique leads to coarse localization. In RFID systems, accuracy and resolution of location estimation depend directly on arrangement and density of tags and readers. On the other hand, scene analysis with motion based localization is a very precise localization technology, but computationally very expensive, and often requires suitable features of objects as a prior knowledge. However, under some assumptions, it is shown that the simple and fast block matching algorithm can be used with high success for motion segmentation and object tracking [15–17]. Indoor localization of moving objects based on RFID localization can be improved by motion segmentation in video sequence. Information acquired by an RFID reader is used for region of interest extraction and motion is estimated within the region of interest only. In this way, significant reduction of the position estimation error and variance in comparison to the conventional RFID position estimation can be obtained with reasonable computational complexity.

1.4 Organization of the Chapter

The remaining sections of the chapter are organized as follows. Section 2 briefly describes main principles and related work important for indoor positioning and localization of moving objects using the RFID technology. Section 3 describes localization of moving objects using image and video processing. The emphasis is on the moving object segmentation as a main task. Motion estimation is analyzed in details, especially estimation based on the block matching algorithm (BMA). The region of interest (ROI) extraction for the redundancy reduction in the visual scene analysis was explained also in this section. Section 4 presents a method for RFID localization and motion segmentation integration in order to improve accuracy and precision of moving object localization. Moving object segmentation, consisting of block matching algorithm followed by morphological postprocessing, based on the region of interest extracted by RFID data, is described in detail in this section. Finally, Sect. 5 concludes the chapter and outlines possible lines of future work.

2 RFID Localization

A system for indoor positioning and localization of moving objects based on the RFID is determined by several important elements: the process of reading of tags, selection of the technique and metric for distance measurement, algorithms for positioning and localization, as well as representation and display of the determined location in accordance to the chosen coordinate system. The block scheme of a RFID system for positioning and localization is presented in Fig. 1. Output parameters of RFID system (tag identification data—ID, TOA, TDOA, TOF, AOA and RSS) provide information for localization of objects of interest. The most used localization techniques with RF technology are triangulation, trilateration and proximity [18].

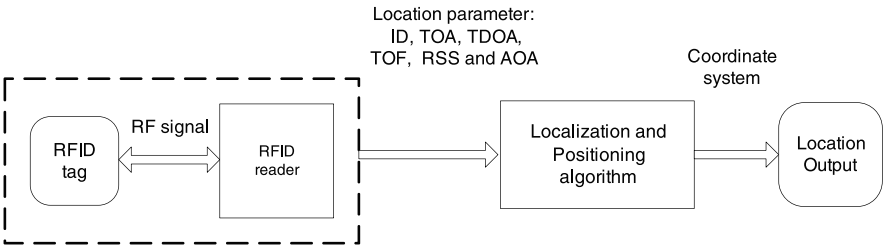


Fig. 1 Block scheme of RFID system for positioning and localization

Three most important characteristics which should be taken into account during the design of the RFID localization system have direct influence to architecture of the system are:

- type of the tag (active, passive or semi-active),
- frequency range (LF, HF, UHF, microwave),
- localization principle (RFID tag localization or RFID reader localization) [19].

The RFID localization systems often determine location of the RFID reader attached to the moving object, while the active or passive RFID tags are at known positions.

Due to the simplicity and low cost, indoor localization can be done using HF passive RFID tags arranged on the test floor at the known positions and using proximity localization technique. Having in mind that RFID systems based on HF passive tags acquire only the information about the presence of the tag in the RF field of the reader antenna, it is important to precisely define density and arrangement of the tags on the test floor.

The localization error is directly related to the arrangement and the density of the tags. In order to reduce the number of tags, while keeping localization error below a certain limit, it is very important to optimally arrange the RFID tags. Relation between the minimum number of the tags which RFID reader reads at the moment, the RFID tags location coordinates and the predefined maximum of localization error determines the optimal density and arrangement of the tags. In that way, localization error is limited [20]. Better localization results can be achieved using the same number of the tags with the triangular, hexagonal or diamond arrangement instead of the rectangular arrangement of the passive RFID tags [20–22].

Knowing the start position of the moving object and time needed for reading the tag, it is possible to significantly reduce localization error. Authors in [23] found that localization error is determined by the radius of the RF field and dimensions of the rectangular passive HF RFID tag.

It is also shown that using known start position of the moving object and arrangement of tags in a triangular pattern the localization error can be reduced and accuracy of localization increased [24]. The additional benefit of this solution is that accuracy improvement is realized without increasing the number of tags.

Depending on frequency range, type of the RFID tags used or applied localization technique, different approaches for localization of moving objects are pro-

posed [25]. In order to get the best possible results authors in [7–12] combined different types of RFID systems and localization techniques. A good example of a moving object or person localization with active RFID tags is based on the set of reference active tags placed at previously defined fixed positions and active tags carried by persons or moving object [26]. The signal strength of a reference RFID tags is used for system calibration. Position of the moving object is calculated using the strenght of signal from reference and moving tags. Due to the low price and simple maintenance, passive RFID tags are intensively used, especially in conjunction with the proximity localization technique and predefined arrangement and density of tags [24].

RFID localization systems are often used for improving some other tasks. The evaluation of the wireless communication protocols using the node position information achieved using the RFID localization system is a good example [23]. In [27] it is shown that systems based on RFID can be easily and efficiently embedded in the existing information and communication infrastructure, and offering up new possibilities for the RFID technology applications.

3 Localization of Moving Objects Using Image and Video Processing

Many intelligent video surveillance systems, traffic monitoring and control systems, as well as access control systems are based on scene analysis and video motion detection [28]. Localization systems based on visual scene analysis use information about size, shape, color, texture and other object features, as well as information about shadows, space and geometry information that characterize observed scene and objects of interest. Systems for visual information analysis and processing involve complex and intensive mathematical calculations due to the complex structure of the visual information.

The main problems in the localization systems based on image and video processing are:

- presence of noise that masks objects and movements,
- features of objects of interest similar to background or noise,
- two or more objects interpreted as one object,
- one object interpreted as two or more objects,
- actual 3D motion of objects in a scene can be estimated only from their 2D projections (images),
- projected motions that do not generate optical flow exist (for example, rotation of uniform color sphere),
- optical flow that does not correspond to projected motion (for example, illumination changes cause changes in optical flow),
- occlusion problem,
- aperture problem.

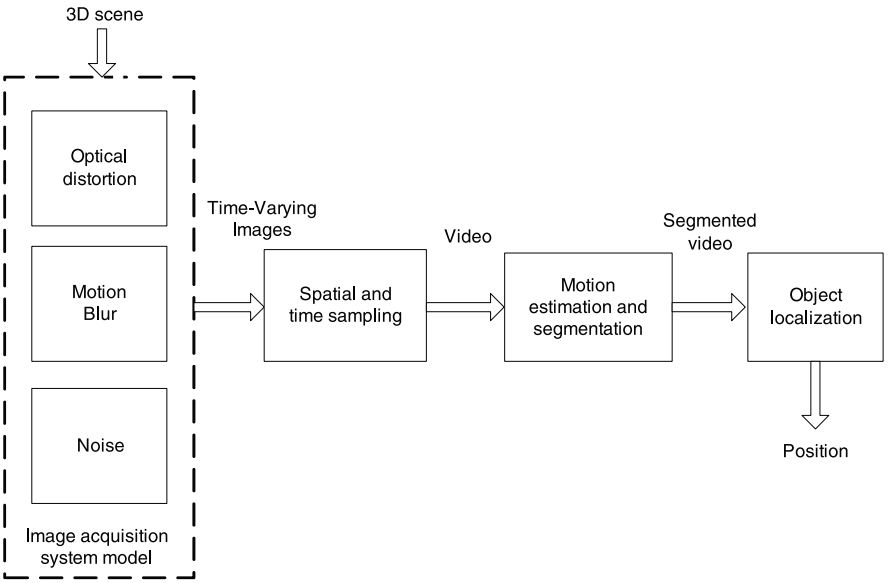


Fig. 2 Localization of moving object using digital image analysis

Moving object segmentation is one of the main tasks in localization systems based on digital image and video processing, and it can be performed using different algorithms depending on application and project design [29]. Figure 2 shows the block diagram of the localization system based on digital image analysis. The time-varying image is formed by the projection of time-varying three-dimensional scene in a two-dimensional image plane. Variations of the 3D scene are usually caused by movements of objects presented in the scene. The time-varying image can be represented as a function of two spatial variables (x_1, x_2) and time variable t.

3.1 Motion Estimation

In object tracking applications, the primary goal is object segmentation based on common motion in a video sequence, grouping 3D pixels (pels) into the most prominent moving groups. Motion estimation enables detection of a moving object of interest although multiple moving objects are present in the scene [30].

Algorithms for moving object segmentation depend on analyzed elements (pixels, regions, blocks, angles or lines), movement representation (2D or 3D movement) and criteria for segmentation. 2D motion estimation methods can be classified in optical flow methods and change detection methods. Further, optical flow methods can be based on differential methods, block matching or statistical methods [29, 31]. Basic classification of motion estimation methods are given in Fig. 3.

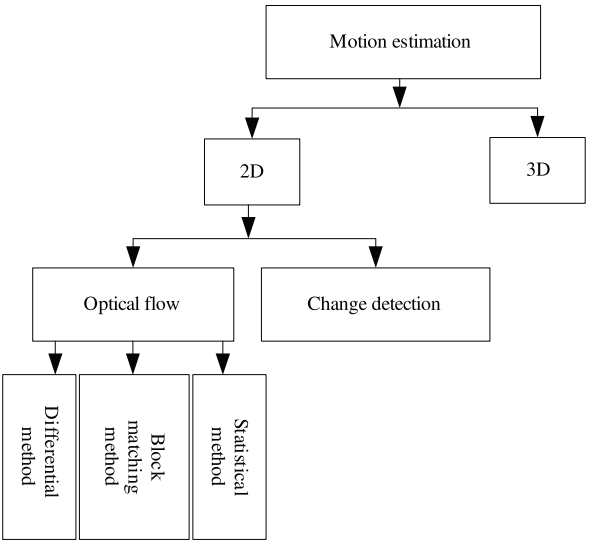


Fig. 3 Motion estimation methods

If point correspondences between frames are considered, dense motion vectors fields are obtained, but these methods fail at large displacements. Instead of that, block-based motion segmentation methods try to group pels into regions moving coherently in space and time using uniformity of space descriptors [32–34], or spatio-temporal image gradient, which results in sparse vector fields. Most of the proposed algorithms work under certain assumptions, such as small motion, unchanged illumination, etc. [35]. For robust motion segmentation, the cost of high computational complexities must be paid [29] and therefore such a complicated approach is inappropriate for real-time applications.

Simple and often used technique for 2D image analysis and motion estimation is the Block matching algorithm (BMA). Pixel displacements are represented by motion vectors. The motion vector for the block of pixels in the actual frame is determined by searching for the most similar block within the reference frame. Applicability of simple and fast BMA for tracking is investigated in [15]. It is shown that, under the assumption of motionless camera and unchanged illumination, the BMA can be used with high success for motion segmentation. Under assumption of images with different levels of activity, sub-optimal search algorithms, which increase the speed at the expense of some accuracy, can be employed [36]. BMA is often used in video compression, moving object detection and localization [15, 32]. BMA is also used in video surveillance systems, traffic control and monitoring, and other similar systems based on the scene analysis [28, 34].

Two-dimensional motion estimation is analyzed in the context of different influences. The most important problems are occlusion, aperture problem and motion estimation sensitivity to the presence of the noise in the video [37]. Each motion estimation method has some drawbacks. By choosing the appropriate parameters of the algorithms, it is possible to reduce some of their negative effects. The most important parameters in BMA implementation are: block size and search region size,

Table 1 Block matching criteria and search strategies

Block matching algorithm (BMA)	
Matching criteria	Search strategy
(a) Mean Squared Error	(a) Exhaustive Search/Full search
(b) Mean Absolute Difference	(b) Three Step Search
(c) Matching Pel Count	(c) Four Step Search
	(d) Diamond Search
	(e) Adaptive Road Pattern Search

matching criteria, and search strategy. The determination of block size and search region size depends on the application, the size of the moving object, the amount of noise in the video frames, the texture of the object and the background. Often used matching criteria are: Mean Squared Error (MSE), Mean Absolute Difference (MAD), and Matching Pel Count (MPC) [16, 37]. BMA is a computationally intensive, and can be improved by appropriate search strategy. Instead of Exhaustive Search/Full search [16], other strategies can be used: Three Step Search [16], New Three step Search [38], Simple and Efficient Search [39], Four Step Search [40], Diamond Search [41], and Adaptive Road Pattern Search [42]. Block matching criteria and search strategies are summarized in Table 1.

3.2 Region of Interest Extraction

Important aspect of scene analysis for moving object localization and tracking based on image and video processing is redundancy reduction. Because of that, region of interest (ROI) extraction, capturing the objects of interest within the ROI, is an important issue. Color and shape information are usually used for ROI based motion estimation process [36, 37].

The full search BMA depends on ROI extraction more than other algorithms, because of its computational complexity [43]. Different methodologies for ROI extraction are presented in previous research. Two main approaches are often proposed. The first method analyzes the objects within the predefined and fixed ROI, which is in the same position for all frames in a video sequence [44]. The second approach takes into account dynamics of the scene and automatically adapts the ROI position within the frame. The moving object location is calculated analyzing the ROI whose position is continually updated from frame to frame of the video sequence [45, 46].

The simplest way of forming the region of interest in analysis of a static image is to divide the frame into blocks grouped according to previously defined rules. The disadvantage of this approach is that fixed and predefined regions of interest are located at the same position within each frame, so it cannot be used in the case of moving cameras and dynamic scene changes. The above described approach is often used in security video surveillance systems where the main task is to control and monitor important, fixed location in the scene [44].

It is known that different factors influence user's visual attention. Some of them are movement, contrast, size of the object, shape, color, location, scene background, presence of the multiple moving objects at the scene and context of the object presence. Those factors are estimated and combined in order to create the map of key influence factors that attracts user attention to the specific region or object [45]. If only image processing techniques are used, the first step in the process of the ROI extraction is frame segmentation into homogenous regions. Second step represents combining the factors that influence visual attention with the spatial factors and features (size and shape of the region, foreground and background characteristics etc.). Based on the obtained attention model, automatic ROI extraction can be performed.

The physical object of interest appears in several consecutive frames. Usually, the extracted ROI is the tightest rectangle surrounding the object of interest. The methods of ROI tracking in video sequences are often based on similarity of visual features in successive frames. As previously described, BMA has been widely used in motion estimation and tracking. Authors in [47] proposed efficient human motion tracking method using human figure model. They introduce a ROI extraction using the object of interest modeling with object masking. In general, with known object model and the BMA search region limited by the ROI, whose size and position are continually refreshed, the improvement of the moving object tracking can be achieved.

4 RFID Localization of Moving Objects Improved by Motion Segmentation

The scene analysis is not a simple task, especially when several objects with different characteristics are present in a scene. Navigation systems, systems for localization of moving objects and robot auto-localization include object recognition, ROI extraction, position determination, etc. Having in mind rapid development of mobile and Internet communications, the integrated systems can be used.

Complex systems that achieve significant improvement in scene analysis integrate the RFID and image processing techniques. Integration of visual information and RFID data reduce localization problems caused by high interaction between persons and objects, and eliminate the redundant scene information. The object recognition, localization and tracking can be successfully employed using integration of RFID technology and digital image processing techniques [48]. Protecting and managing privacy information is important an task in modern video surveillance systems. Authors in [49, 50] propose integration of RFID technology with video surveillance system for masking regions in acquired image, so the person's privacy protection is provided. An example of RFID system integrated with the image and video processing is remote surveillance of kid's activity in the kindergarten [51].

Methods of integration depend on specific systems, characteristics and tasks expect to be improved. The possible improvements of visual scene analysis systems are related to:

- image or video archive size,
- person privacy protection in video surveillance systems,
- real time moving object or person tracking using RFID data as a corrective element,
- extracting image frames from video in face recognition systems,
- object or person identification in surveillance and tracking systems.

On the other hand, RFID systems can also be improved by digital image processing and analysis. Many improvements are possible, and some of them are:

- increasing reliability of identification by integrating biometrical data,
- reducing active RFID reader energy consumption by triggering the reader only if the object of interest is detected at the scene,
- accuracy and precision of localization.

Besides the ID data, RFID tag can store other useful data. Information about absolute location, image histogram of the object or person wearing the tag, object images and other important data can be stored depending on the RFID tag type. Due to the possibility of using different data types and solving complex tasks, RFID technology presents more than identification technology [52–56]. Different tasks can be solved using the data stored on the RFID tag and image processing techniques. 3D analysis, surveillance and object classification, activity recognition are examples of use [52, 53, 57, 58].

In this chapter, a method for RFID localization and motion segmentation integration in order to improve accuracy and precision of moving object localization will be described. Reduction of information redundancy is an important task in modern multimedia applications, and this method reduces redundancy by extracting and analyzing only the ROI that represents the core interest of the user, containing the essence of the frame information.

RFID technology is employed for ROI detection. ROI is extracted based on a priori knowledge about passive RFID tags arrangement and the RF field determined by the RFID reader performances. After that, a moving object in ROI is segmented based on estimated motion vectors and morphological postprocessing. Motion vectors are estimated using BMA. To decrease the impact of illumination changes, only chromatic image components are used. Centroid of the segmented object determines the object position.

The integrated system consists of the three functional segments: information acquisition segment based on RFID devices and image sensors usage, image processing segment and global information acquiring based on RFID ID data. General architecture block scheme of a system for localization and identification using integration of RFID and image processing is presented in Fig. 4.

The proposed framework could be applicable in applications based on the use of industrial robot. Well-known problems in robotics, such as initial position determination, self localization, and robot calibration can be solved using the proposed framework. High precision localization in multi robot systems is necessary in order to avoid possible collisions.

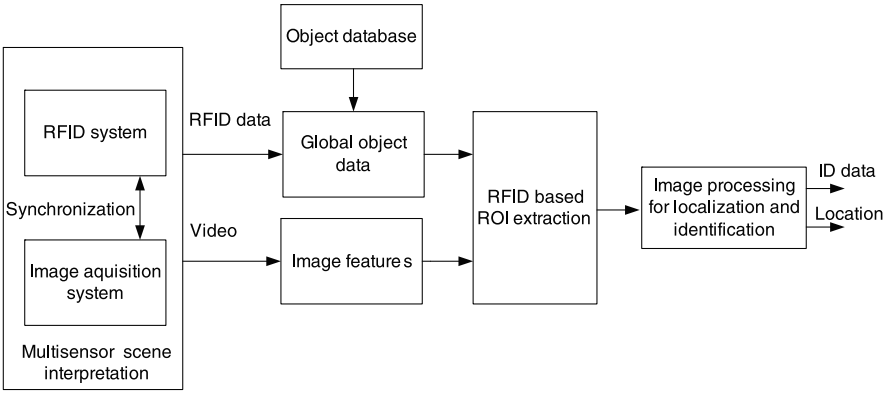


Fig. 4 General architecture block scheme for localization and identification system using integration of RFID and image processing

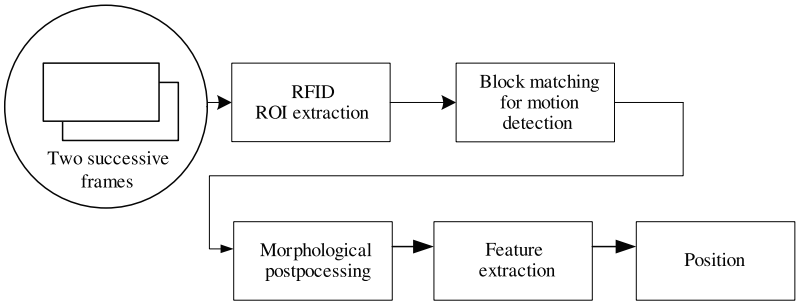


Fig. 5 Moving object localization method

4.1 Moving Object Segmentation and Localization

The block scheme of the moving object localization system is shown in Fig. 5. It is assumed that the object of interest moves across an indoor field of passive RFID tags. In order to decrease the error of RFID position estimation of moving objects of interest, motion segmentation followed by morphological postprocessing and feature extraction is employed.

The aim of the proposed method is to achieve smaller absolute distance error without increasing the number of tags.

4.2 RFID Position Estimation

The RFID localization system based on proximity technique and passive RFID tags [21, 24] consists of a moving object with an RFID reader attached to the bottom,

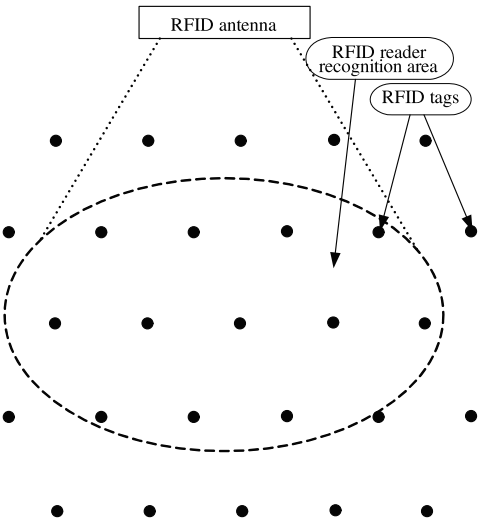


Fig. 6 RFID reader recognition area (RRA)

and $N \times M$ passive RFID tags arranged on the floor in either a square or triangular pattern. The RFID reader antenna forms the RF field so the tags under the effective area of the antenna (reader recognition area—RRA) are detected (see Fig. 6). Each tag, $T_{k,l}$, sends its identification number related to its two-dimensional coordinates (x_k, y_l), $k = 1, 2, \ldots, N$, $l = 1, 2, \ldots, M$.

Let us assume the triangular tags pattern. Coordinate information of RFID tags detected inside the circular RRA with radius R are $x_k, \ldots, x_{k+n}$ and $y_l, \ldots, y_{l+m}$.

The center of RRA is the true RFID reader position $(x_{\text{true}}, y_{\text{true}})$. The estimated position of the RFID reader is represented as [24]

$$x_{\text{est}} = \frac{x_k + x_{k+n}}{2}, \tag{1}$$

$$y_{\text{est}} = \frac{y_l + y_{l+m}}{2}. \tag{2}$$

The RFID absolute position estimation error represents the difference between the true and estimated positions of the RFID reader. The maximum estimation error is

$$\begin{aligned} e_{RF} &= \sqrt{(x_{\text{est}} - x_{\text{true}})^2 + (y_{\text{est}} - y_{\text{true}})^2} \\ &\leq \sqrt{\left(\frac{1}{4}d_{\text{tag}}\right)^2 + \left(\frac{1}{4}d_{\text{tag}}\right)^2} = \frac{\sqrt{5}}{4}d_{\text{tag}}. \end{aligned} \tag{3}$$

The error is proportional to the gap between the tags. Estimation error decreases with d_{tag} decreasing, for example, with increasing the number of tags detected under the same RRA. Authors in [24] conducted an experiment with $d_{\text{tag}} = 10$ cm, using square and triangular patterns, and showed that average absolute distance errors are 2.0 cm and 1.6 cm, respectively.

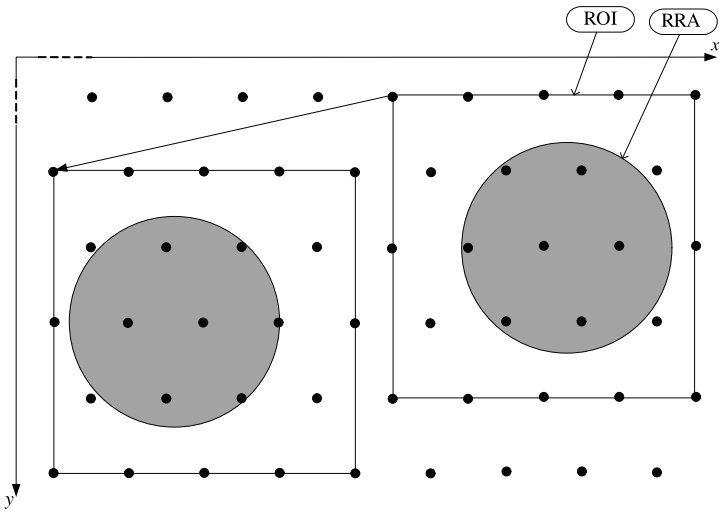


Fig. 7 Relations between ROI and RRA in successive frames and roughly estimated motion vector

4.3 Rough RFID Localization

The position of the moving object is roughly estimated using the RFID localization. When the RFID reader finishes collecting data, the minimum x and y tags coordinate values inside the RRA are used to determine the reference pixel in each frame for ROI extraction. We used the triangular tag pattern where the reference position (x_{k-2}, y_{l-1}) has to be the upper left corner of ROI in order to include the whole RRA in ROI, for every possible RRA position regarding the detected tags. Roughly estimated motion vector is determined by two ROI reference pixels in successive frames (see Fig. 7).

4.4 Motion Estimation

Roughly estimated motion vectors can be used for PTZ (pan-tilt-zoom) video camera control in such a way that video camera follows the RFID reader, so the moving object of interest is always captured. Another way is to use a static high resolution video camera that captures the whole field of passive RFID tags. When motion estimation between whole frames is performed, motion vectors give information about all motions, including camera motion, background and illumination changes, and this approach suffers from aperture and occlusion problems. Motion estimation is an ill-posed problem and requires enormous number of operations if it is calculated for large number of pixels.

In order to simplify motion estimation and find only the motion of the object with attached RFID reader, information acquired by the RFID reader is used to extract

regions of interest within the frames captured by the video camera. The size of extracted ROI depends on the size and shape of RRA, the tags read within the RRA, and the geometry of the moving object, so the object of interest occupies most of the ROI area. ROI modeling is based on the assumption that the gap between the tags is smaller than the diameter of the RRA. It is assumed that the center of the RRA is equal to the centroid of the moving object.

Motion vectors are estimated for the blocks in extracted ROIs of the successive frames. Due to its simplicity, BMA is appropriate for motion vectors calculation. When ROI extraction is not used, large search windows are needed to find matching blocks for fast moving objects (i.e. having large displacement). RFID based extraction of ROI enables block matching using small search windows regardless of the magnitude of displacement.

4.5 Moving Object Segmentation

For moving object segmentation it is not necessary to find exact motion vectors. It is enough to find blocks moving in the similar way. The whole object should be included in ROIs extracted by RFID in both frames. It does not matter if ROIs of frames represent different physical scenes determined by different sets of tags.

Block matching algorithm is not robust to illumination changes. Shadows of objects can significantly distort segmentation, because shadows move together and in a similar way as objects and they are segmented as parts of objects. The most information about illumination and shadows by its nature belong to the luminance component. To reduce the impact of shadows and changes in illumination, only chromatic components of CIE L*a*b* color space are used for motion analysis. Each chromatic component of frames is analyzed separately.

Since the object of interest occupies most of the ROI area, the most of the motion vectors in ROI belong to the object of interest, and have similar orientation and intensity. In contrast to these motion vectors, the rest of the motion vectors result from camera movements, random changes in illumination, noise or small moving objects. Motion vectors are segmented if the appropriate bin in the histogram is above the predefined percent of the histogram maximum. Blocks with such vectors in at least one of the chromatic frame components belong to the moving object.

Due to the complex structure of the moving object, camera movements, changes in illumination, shadows, noise and, eventually, small moving objects that are near the object of interest, all pixels from the object of interest are not segmented correctly and, additionally, some isolated small regions from background or other moving objects are also segmented. All of these inaccuracies have significant influence to extraction of object features, especially to centroid calculation. Morphological postprocessing improves initial segmentation obtained by motion vectors. The most of segmented blocks belonging to the object of interest are connected.

Fig. 8 RFID reader attached to the object

4.6 Absolute Position Estimation

Since the RFID tags are at the predefined absolute positions, the absolute position of the ROI is also known. As we mentioned before, the upper left corner of ROI is determined by tag position (x_{k-2}, y_{l-1}), where (x_k, y_l) are the coordinates of the upper left tag sensed by the RFID reader. The relative pixel based position of the moving object is determined by the vector connecting the upper left corner of ROI and the centroid of the segmented object. With known relation between the image plane and the real scene, relative pixel based position is transformed to the real relative position. The final absolute location of the moving object is obtained by correcting the RFID based roughly estimated absolute location with this relative position.

4.7 Experimental Results

The aim of this experiment is to show the improvement of RFID localization using motion segmentation with BMA and morphological postprocessing. It is clear that only the relative position of the segmented object centroid to the reference ROI pixel (upper left corner) makes difference to the RFID localization. Because of that, it is enough to analyze motion in the ROIs of successive frames that represent the physical scene determined with the same set of tags, to show improvements of the proposed method.

This experiment is conducted in realistic environment (varying illumination, presence of shadows and noise) using a moving object with irregular shape and texture. An object with an RFID reader attached to its bottom (see Fig. 8) is moving on the floor with triangle pattern of passive RFID tags. Process of RFID data acquiring is synchronized with frame capturing, i.e. at the exact moment of frame capturing new information from RFID reader about sensed tags is available.

Calibration between the RFID and the camera is performed as explained in Sect. 4.3. Common coordinate system is established in accordance with Fig. 7. With known video camera parameters and the reference linear measure presented on the scene, relation between the image plane and the real scene plane was established,

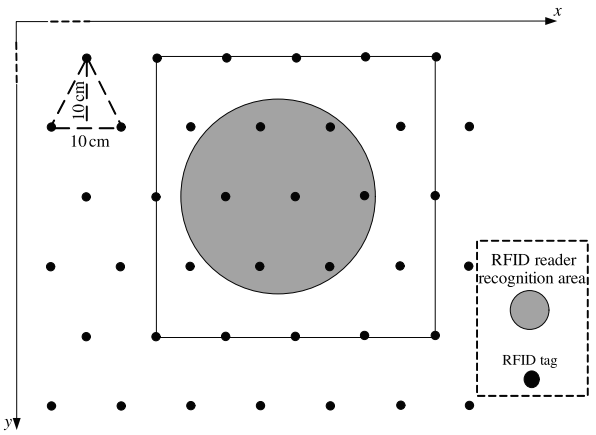


Fig. 9 RRA and tags' grid

namely p image pixels correspond to 1 cm ($p = 6$ in this example). In order to use as few tags as possible, the diameter of RRA should be smaller than $3d_{\text{tag}}$ (see Fig. 9).

In this way, at least two or at most three tags both in horizontal and vertical direction are detected simultaneously. The size of the moving object is chosen to be smaller than RRA.

Under the given assumptions, the dimension of ROI is chosen to be $p \times 4d_{\text{tag}}$ in each dimension, in order to cover the entire object by ROI. Considering that the diameter of RRA is 28 cm, tags are located at every 10 cm using triangular pattern. In this experiment ROIs with dimensions 240×240 pixels are cropped from high resolution video frames.

Motion vectors are calculated for 8×8 blocks in ROI, in CIE L*a*b* color space for chromatic components only, and the search area of 7 pixels in each of the four directions. Figure 10 shows two successive frames of the real scene, Figs. 11 and 12 show the corresponding motion vectors and segmented blocks of the moving object, respectively, calculated for chromatic components in CIE L*a*b* color space.

The two-dimensional cumulative histogram of the motion vectors is generated (see Fig. 13). The height of each bar in this histogram represents number of motion vectors that have the same real and imaginary parts. The motion vectors corresponding to the most significant peaks in the histogram have similar orientation and intensity. They determine the blocks belonging to the moving object equipped with the RF reader.

Motion analysis, performed separately for both chromatic components, gives the vectors with similar orientation and the intensities. Those vectors belong to object of interest and they are clearly identified within ROI. 2D histogram of the motion vectors for chromatic "a*" component of the CIE L*a*b* color space is shown in the Fig. 13.

The segmented moving object obtained after morphological postprocessing is shown in Fig. 14.

All objects whose area is smaller than the area of the largest object are removed. Small holes inside the segmented blocks should be filled before centroid calculation.

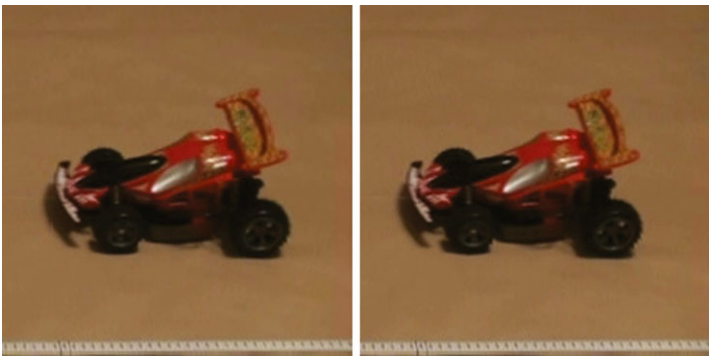
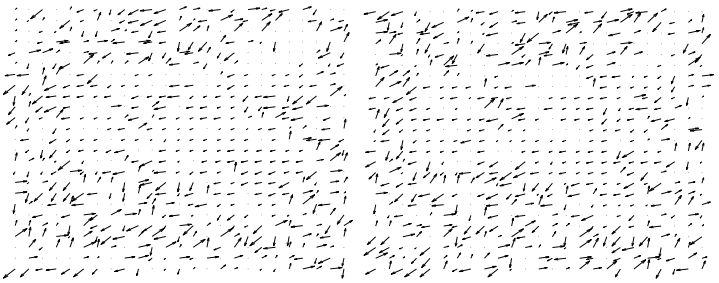
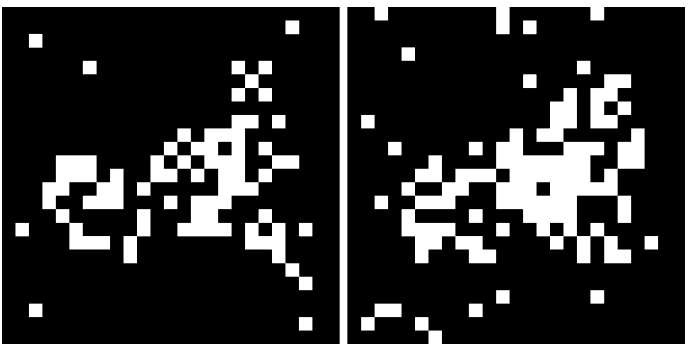


Fig. 10 Two successive frames



(a) (b)

Fig. 11 Motion vectors computed using: (**a**) a* component, (**b**) b* component of CIE L*a*b* color space



(a) (b)

Fig. 12 Segmented blocks using: (**a**) a* component, (**b**) b* component of CIE L*a*b* color space

It can be implemented using conditional dilatation with the edge of the ROI image as a seed and the inverted segmentation obtained beforehand as a mask image. Fi-

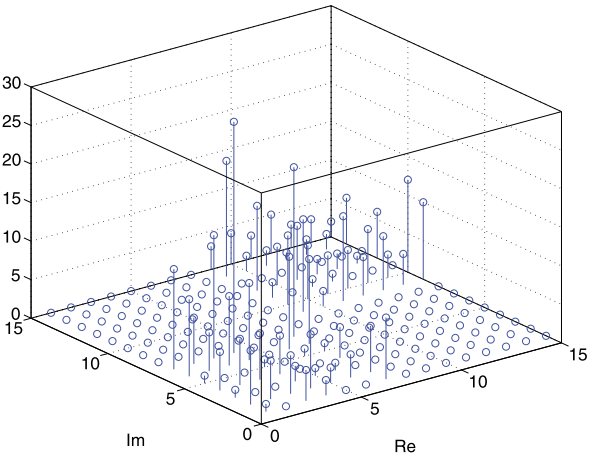


Fig. 13 2D histogram of the motion vectors for chromatic "a*" component of the CIE L*a*b* color space

Fig. 14 Moving object segmented using chromatic components

nally, binary morphological closing with a flat, disk-shaped structuring element with 2 pixels radius smoothes the edges of the segmented object.

In order to reduce the influence of illumination changes and segmentation of shadows as a part of the object, luminance component is omitted. Figure 15(a) shows the segmented blocks of luminance obtained by the same procedure. If the final result is obtained with all of three CIE L*a*b* components, shadows of the object are also segmented, compare Fig. 15(b) with Fig. 14.

After segmentation and morphological postprocessing, the absolute distance errors are calculated. Two separate tests were performed, with uniform and textured background. Typical scenes with uniform and textured background are shown in Figs. 10 and 16, respectively. The locations in 48 positions were estimated, while an object was crossing over the whole area covered by the same set of tags. The test was performed four times, twice for video with uniform and twice with textured background, each time in 12 randomly chosen locations. The true location of the object was manually determined in each captured frame based on centroid marker presented at the object and the reference linear measure presented in the test scene.

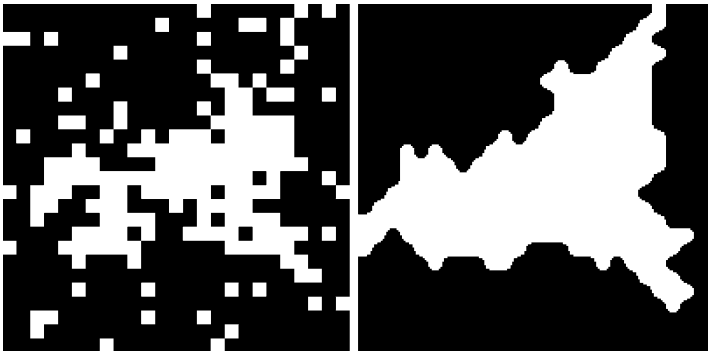


Fig. 15 (**a**) Segmented blocks using luminance component, (**b**) moving object segmented using all CIE L*a*b* components

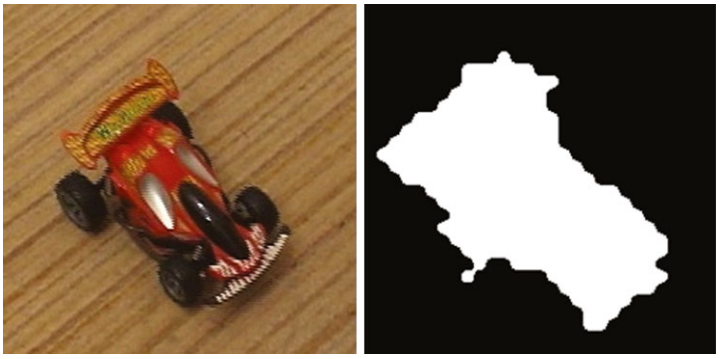


Fig. 16 Segmentation of moving object on textured background (best viewed in color)

The comparison of the results obtained for these 48 randomly chosen locations with the coarse RFID position estimation and the proposed method are given in Table 2. The estimation errors are related to the RFID localization and the RFID localization improved with motion object segmentation (RFID+MS) for uniform and textured background. Better results are obtained with textured background, because block matching on static uniform background contaminated with noise gives more motion vectors that point to non existing background motion, than on textured background.

Using the coarse RFID localization method the average absolute distance error of 1.37 cm is obtained with variance of 0.51 cm^2. The estimation error and variance are reduced significantly using this method.

Calculated for all 48 positions, the average absolute distance error decreased to 0.72 cm with variance of 0.16 cm^2.

Table 2 Comparison of the RFID localization and RFID localization improved by moving object segmentation

	RFID		RFID+MS	
	Uniform background	Textured background	Uniform background	Textured background
average	1.37	1.36	0.76	0.68
variance	0.50	0.54	0.17	0.15
average	1.37		0.72	
variance	0.51		0.16	

5 Conclusion

In this chapter, the possibilities and advantages of integration of multiple technologies in multimedia surveillance systems are described. It is shown that heterogeneous data fusion improves the accuracy and precision of localization and represents a promising research area.

The method for moving object localization using integration of a passive RFID indoor localization system and scene analysis techniques is presented. Moving object segmentation, based on the region of interest extracted by RFID data, eliminates the influence of other large moving objects and avoids unnecessary image processing computations. Due to its simplicity, the block matching algorithm followed by morphological postprocessing is used for moving object segmentation. The presented method eliminates the aperture problem because the extracted scene of interest has sufficient structure to capture the whole moving object. The implementation shows significant reduction of the position estimation error and variance in comparison to the conventional RFID position estimation. The simplicity and error reduction are the main advantages of this solution.

Several relevant issues that deserve further investigation are also identified. Object modeling can improve object segmentation and localization. Also, the influence of complex environments (texture, shadow, noise, etc.) on the block matching algorithm in order to perform automated ROI extraction can be investigated. Attention has to be paid to computational requirements of the proposed solution.

References

1. McCall, R., Snaider, J., Franklin, S.: Sensory and perceptual scene representation. J. Cogn. Syst. Res. (2010)
2. Muthukrishnan, K., Lijding, M., Havinga, P.: Towards smart surroundings: enabling techniques and technologies for localization. In: Proc. of the First International Workshop on Location-and Context-Awareness (LoCA). Springer, Berlin (2005)
3. Misra, P., Burke, B.P., Pratt, M.M.: GPS performance in navigation. Proc. IEEE **87**(1), 65–85 (1999)
4. Yang, P., Wu, W., Moniri, M., Chibelushi, C.C.: RFID tag infrastructures for camera tracking in virtual studio environment. In: Proc. 4th European Conference Visual Media Production IETCVMP, pp. 1–8 (2007)
5. Hightower, J., Borriello, G.: Location systems for ubiquitous computing. IEEE Comput. **34**(8), 57–66 (2001)

6. Tesoriero, R., Tebar, R., Gallud, J.A., Lozano, M.D., Penichet, V.M.R.: Improving location awareness in indoor spaces using RFID technology. Expert Syst. Appl. **37**(1), 894–898 (2010)
7. Bahl, P., Padmanabhan, V.: RADAR: an in-building RF-based user location and tracking system. In: Proc. IEEE Infocom 2000, pp. 775–784 (2000)
8. Hopper, A., Harter, A., Blackie, T.: The active badge system. In: Proc. of the INTERACT'93 and CHI'93 Conference on Human Factors in Computing Systems (1993)
9. Priyantha, N.B., Chakraborty, A., Balakrishnan, H.: The cricket location-support system. In: Proc. of the 6th Annual International Conference on Mobile Computing and Networking, pp. 32–43 (2000)
10. He, T., Huang, C., Blum, B.M., Stankovic, J.A., Abdelzaher, T.: Range-free localization schemes for large scale sensor networks. In: Proc. of the 9th Annual International Conference on Mobile Computing and Networking (MobiCom'03), pp. 81–95 (2003)
11. Hightower, J., Want, R., Borriello, G.: SpotON: an indoor 3D location sensing technology based on RF signal strength. Technical report UW-CSE 00-02-02, University of Washington, Seattle (2000)
12. Orr, R.J., Abowd, G.D.: The smart floor: a mechanism for natural user identification and tracking. In: Proc. CHI'00 Extended Abstracts on Human Factors in Computing Systems, pp. 275–276 (2000)
13. Krumm, J., Harris, S., Meyers, B., Brumitt, B., Hale, M., Shafer, S.: Multi-camera multi-person tracking for easyliving. In: Proc. of the Third IEEE International Workshop on Visual Surveillance (VS'2000), pp. 3–10 (2000)
14. Diaz, J.J.M., Maues, R.A., Soares, R.B., Nakamura, E.F., Figueiredo, C.M.: Bluepass: an indoor bluetooth-based localization system for mobile applications. In: Proc. IEEE Symposium on Computers and Communications ISCC, pp. 778–783 (2010)
15. Gyaourova, A., Kamath, C., Cheung, S.C.: Block matching for object tracking. Project report UCRL-TR-200271 (2003)
16. Barjatya, A.: Block matching algorithms for motion estimation. Technical report, Utah State University (2004)
17. Babic, Z., Ljubojevic, M., Risojevic, V.: Indoor RFID localization improved by motion segmentation. In: Proc. of the 7th International Symposium on Image and Signal Processing and Analysis ISPA (2011)
18. Boontrai, D., Jingwangsa, T., Cherntanomwong, P.: Indoor localization technique using passive RFID tags. In: Proc. of the 9th International Conference on Communications and Information Technologies ISCIT, pp. 922–926 (2009)
19. Sanpechuda, T., Kovavisaruch, L.: A review of RFID localization: applications and techniques. In: Proc. 5th International Conference Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology (ECTI-CON 2008), vol. 2, pp. 769–772 (2008)
20. Ku, W.S., Sakai, K., Sun, M.T.: The optimal k-covering tag deployment for RFID-based localization. J. Netw. Comput. Appl. **34**(3), 914–924 (2011)
21. Choi, B.S., Lee, J.W., Lee, J.J.: An improved localization system with RFID technology for a mobile robot. In: Proc. 34th Annual Conference on Industrial Electronics IECON, pp. 3409–3413 (2008)
22. Park, S., Hashimoto, S.: Indoor localization for autonomous mobile robot based on passive RFID. In: Proc. of the 2008 IEEE International Conference on Robotics and Biomimetics, pp. 1856–1861 (2009)
23. Munishwar, P., Singh, S., Mitchell, C., Xiaoshuang, W., Gopalan, K., Abu-Ghazaleh, N.B.: RFID based localization for a miniaturized robotic platform for wireless protocols evaluation. In: Proc. IEEE International Conference on Pervasive Computing and Communications, PerCom, pp. 1–3 (2009)
24. Lim, H.S., Choi, B.S., Lee, J.M.: An efficient localization algorithm for mobile robots based on RFID system. In: Proc. of International Joint Conference SICE-ICASE, pp. 5945–5950 (2006)

25. Bouet, M., Santos, A.L.: RFID Tags: positioning principles and localization techniques. In: IFIP Wireless Days—2nd International Home Networking Conference IHN (2008)
26. Lionel, M.N., Yunhao, L., Lau, Y.C., Patil, A.P.: LANDMARC: indoor location sensing using active RFID. Wirel. Netw. **10**(6), 701–710 (2004)
27. Ahson, S., Ilyas, M.: RFID Handbook Applications, Technology, Security, and Privacy. CRC Press/Taylor & Francis, Boca Raton/London (2008)
28. Cheng, P.G., Yong, J.H.: Study on tracking of moving object in intelligent video surveillance system. Adv. Mater. Res. **433**, 6583–6588 (2012)
29. Zhang, D., Lu, G.: Segmentation of moving objects in image sequence: a review. Circuits Syst. Signal Process. **2**(2), 142–183 (2001)
30. DiStefano, L., Viarani, E.: Vehicle detection and tracking using the block matching algorithm. In: Proc. of 3rd IMACS/IEEE, vol. 1, pp. 4491–4496 (1999)
31. Tekalp, A.M.: Digital Video Processing. Prentice Hall, New York (1995)
32. Brox, T., Bregler, C., Malik, J.: Large displacement optical flow. In: Proc. IEEE Conference on Computer Vision and Pattern Recognition, pp. 41–48 (2009)
33. Altunbasak, Y., Eren, P., Tekalp, A.M.: Region-based parametric motion segmentation using color information. Graph. Models Image Process. **60**(1), 13–23 (1998)
34. Jianbo, S., Malik, J.: Motion segmentation and tracking using normalized cuts. In: Proc. of the 6th International Conference on Computer Vision, pp. 1154–1160 (1998)
35. Cremers, D., Soatto, S.: Motion competition: a variational approach to piecewise parametric motion segmentation. Int. J. Comput. Vis. **62**(3), 249–265 (2005)
36. Rani, T.J., Priyadharsini, S.S.: Region of interest tracking in video sequences. Int. J. Comput. Appl. **3**(7), 32–36 (2010)
37. Yang, F., Li, J., Zhang, Z.W., Pan, G.F.: Motion estimation algorithm based on the region of interest. Appl. Mech. Mater. **20–23**, 581–587 (2010)
38. Li, R., Zeng, B., Liou, M.L.: A new three-step search algorithm for block motion estimation. IEEE Trans. Circuits Syst. Video Technol. **4**(4), 438–442 (1994)
39. Lu, J., Liou, M.L.: A simple and efficent search algorithm for block-matching motion estimation. IEEE Trans. Circuits Syst. Video Technol. **7**(2), 429–433 (1997)
40. Po, L.M., Ma, W.C.: A novel four-step search algorithm for fast block motion estimation. IEEE Trans. Circuits Syst. Video Technol. **6**(3), 313–317 (1996)
41. Zhu, S., Ma, K.K.: A new diamond search algorithm for fast block-matching motion estimation. IEEE Trans. Image Process. **9**(2), 287–290 (2000)
42. Nie, Y., Ma, K.K.: Adaptive rood pattern search for fast block-matching motion estimation. IEEE Trans. Image Process. **11**(12), 1442–1448 (2002)
43. Mahmoud, I.I., Hashimaa, S.M., Elazm, A.A.: Proposed one point Pentagon inner search fast block matching algorithm. In: Proc. Radio Science Conference NRSC, pp. 1–9 (2009)
44. Hirzallah, N.: Automated camera monitoring system for selective areas of interest. J. Comput. Sci. **3**, 62–66 (2007)
45. Osberger, W.M., Rohaly, A.M.: Automatic detection of regions of interest in complex video sequences. In: Proc. SPIE Human Vision and Electronic Imaging VI, vol. 4299, pp. 361–372 (2001)
46. Cheng, R.W.H., Chu, W.T., Wu, J.L.: A visual attention based region-of-interest determination framework for video sequences. IEICE Trans. Inf. Syst. **E88-D**(7), 1578–1586 (2005)
47. Yun, B.J., Cho, J.H., Jeong, J.W.: Real-time object tracking method by using HFM (human figure model) in moving camera. In: Proc. International Conference on Intelligent Computing (2005)
48. Kamol, P., Nikolaidis, S., Ueda, R., Arai, T.: RFID based object localization system using ceiling cameras with particle filter. In: Proc. of the 2nd Int. Symposium on Smart Home (SH'07) (2007)
49. Venkatesh, M.V., Cheung, S.-C.S., Paruchuri, J.K., Zhao, J., Nguyen, T.: Protecting and managing privacy information in video surveillance systems. In: Senior, A. (ed.) Protecting Privacy in Video Surveillance. Springer, Berlin (2009)

50. Byungkwan, J., Kyoungkeun, K., Youngwoog, Y., Yeongseog, L.: Combined RFID with sensor of motion detect for security systems. In: Proc. of the WSEAS Int. Conference on Circuits, Systems, Signal and Telecommunications, pp. 221–225 (2007)
51. Huang, R., Ma, J.: Homelog based kid's activity awareness. In: Proc. of the 2009 Computation World: Future Computing, Service Computation, Cognitive, Adaptive, Content, Patterns, pp. 591–596 (2009)
52. Boukraa, M., Ando, S.: Tag-based vision: assisting 3D scene analysis with radio-frequency tags. In: Proc. Int. Conf. Image Processing, pp. 269–272 (2002)
53. Shirasaka, Y., Yairi, T., Kanazaki, H., Shibata, J., Machida, K.: Supervised learning for object classification from image and RFID data. In: Proc. International Joint Conference SICE-ICASE, pp. 5940–5944 (2006)
54. Boukraa, M., Ando, S.: A computer vision system for knowledge-based 3D scene analysis using radio-frequency tags. In: Proc. IEEE International Conference Multimedia and Expo, vol. 2, pp. 245–248 (2002)
55. Nankaghaw, S., Soh, K., Mine, S., Saito, H.: Image systems using RFID tag positioning information. NTT Tech. Rev. **1**(7), 79–83 (2003)
56. Cerrada, C., Salamanca, S., Perez, E., Cerrada, J.A., Abad, I.: Fusion of 3D vision techniques and RFID technology for object recognition in complex scenes. In: Proc. IEEE International Symposium Intelligent Signal Processing WISP, pp. 1–6 (2007)
57. Park, S., Kautz, H.: Hierarchical recognition of activities of daily living using multi-scale, multi-perspective vision and RFID. In: Proc. IET 4th International Conference Intelligent Environments, pp. 1–4 (2008)
58. Cucchiara, R., Fornaciari, M., Haider, R., Mandreoli, F., Prati, A.: Identification of intruders in groups of people using cameras and RFIDs. In: Proc. of ACM/IEEE International Conference on Distributed Smart Cameras, pp. 1–6 (2011)

A Particle Filter Framework for Object Tracking Using Visual-Saliency Information

Dwarikanath Mahapatra and Mukesh Saini

Abstract Automated processing of video streams is core to current surveillance systems. The basic building blocks of video processing techniques are object detection and tracking. Tracking results are further analyzed to detect various events and activities for situation assessment. Several approaches to object detection and tracking are based on background modeling. These approaches are generally vulnerable to noise, illumination changes etc. Further, the object may not look similar in an image sequence over time due to changes in orientation, lighting, occlusion, etc. In this chapter, we explore application of neurobiology-saliency for object detection and tracking using particle filters. We use low-level features such as color, luminance and edge information along with motion cues to track a single person. Experimental results show that this approach is illumination invariant and can track persons in varying lighting conditions.

1 Introduction

Current surveillance systems employ a large number of cameras capturing huge amounts of video. Since it is difficult and expensive to monitor these videos using humans, there is a need to automatically process them for situation assessment and anomaly detection. The main components of automatic situation assessment are activity and behavior analysis, which in turn depend on object detection and tracking. Numerous algorithms have been developed to track humans and other objects. A comprehensive review of these methodologies can be found in [47]. Most of these method rely on a single image cue for tracking. Spengler et al. [42] proposed integration of different useful image cues for robust tracking. Triesch et al. in [44] introduced a *Democratic Integration* method. This is one of the initial approaches

D. Mahapatra (✉)
Department of Computer Science, ETH Zürich, Zürich, Switzerland
e-mail: dwarikanath.mahapatra@inf.ethz.ch

M. Saini
School of Computing, National University of Singapore, Singapore, Singapore
e-mail: mksaini@comp.nus.edu.sg

P.K. Atrey et al. (eds.), *Intelligent Multimedia Surveillance*,
DOI 10.1007/978-3-642-41512-8_7,

where the cues produce a resulting state that serves as the basis for adaptations of individual cues.

Tracking in varying environments is essential for the purpose of security. Approaches such as mean shift tracking have shown good results in many scenarios [8]. Background subtraction approaches, where the objects in the neighborhood are more or less static, are also used to detect humans and objects [7]. However, even a small change in the intensity of the surroundings can alter the reference background and lead to erroneous tracking. Hence, it is necessary to use features that are invariant to intensity changes.

In this chapter, we explore the use of features derived from neurobiology-saliency maps for the purpose of robust object detection and tracking. A saliency based approach identifies the regions of the image that are probabilistically interesting to the viewers based on the workings of the human visual system (HVS). Since the HVS can efficiently track objects in the presence of noise and intensity changes, it is worth exploring the use of saliency models to improve the robustness of existing automatic detection and tracking methods. As a case study, we propose a *neurobiology-saliency based particle filter* method for illumination invariant tracking in office environments. The multi-feature-based saliency measure, when combined with motion information, produces results that are unaffected by intensity changes. It is shown in [16] that an entropy-based saliency measure remains consistent over time. Walther et al. use saliency to initialize tracks for objects, and track them using a Kalman filter. Saliency inspired features have been used in various applications like registration [22, 23, 26, 29] and segmentation [24, 25, 27, 28]. In our method, we fuse two saliency maps, a *static saliency map* of a single scene obtained using purely low-level features (using the neurobiology based approach explained in [13]), and a *motion saliency map* generated using motion cues across successive frames.

Low-level features like pixel intensity, color and edge orientation in static saliency map calculation are robust individual cues for tracking purposes. A combination of these features may provide a more robust tracking framework. We chose the particle filter because of the following reasons: (1) it can be applied to nonlinear systems; (2) the noise need not follow a Gaussian distribution; (3) it can work for multimodal distributions; and (4) the particle filter predicts multiple possible states for each object being tracked.

The rest of the chapter is organized as follows. In Sect. 2, we give details on a standard neurobiology based saliency map as well as our modifications to adapt it to our data. Section 3 briefly explains the particle filter and our implementation. Experimental results are shown in Sect. 4 and we conclude with Sect. 5.

2 Saliency

Saliency defines how different a region is from its surroundings based on various features, thus attracting our attention. Visual attention models (or saliency models) refer to computational models that determine a saliency map (conspicuity map) based on the interaction of different features. Saliency models may consider two

types of features, that is, bottom-up (e.g., intensity, color, texture, edge orientation) or top-down (e.g., prior knowledge of the desired task). We shall review some models of bottom-up feature-based visual attention. Bajcsy and Gelade in [2] proposed a feature integration theory, one of the earliest hypothesis for attention. It suggests that attention must be directed serially to each stimulus in a display whenever conjunctions of more than one separable feature are needed to characterize or distinguish the presented objects. Another model was proposed by Mozer in [32] which modeled an object for recognition tasks. The work by Itti and Koch [15] proposes the popular neurobiological attention model based on saliency maps and has been found to have a high correlation with human fixations [14]. Soto and Blanco in [41] explored the role of space-based and object-based visual attention within a cueing paradigm. Participants had to discriminate the orientation of a line that appeared within one of four moving circles differing in color. A cue appearing close to one of the four circles indicates the location or circle where the target stimulus was likely to appear. Results suggest that object and space-based attention interact with selection-by-location over object-based selection. Logan in [19] proposed a theory integrating space-based and object-based approaches to visual attention.

Saliency is defined by local image features at various scales. Salient regions are those where feature strength is greater than its neighbors. For example, the edges are salient because the difference in intensity between edge pixels and its neighbors is high, hence human visual system (HVS) is strongly attracted to the edges. We now look at some works that use local features for object detection or salient region detection. Scale is an important factor in these methods leading to robust identification of salient regions. Kadir and Brady [16] use entropy in a scale-space model to detect salient regions in an image. Entropy gives a measure of information content in a neighborhood, and different scales are used for robust identification of salient regions. Lowe in [20] introduces a scale-invariant feature descriptor that identifies salient points irrespective of rotation and the scale at which features are selected. This technique has been used in many object matching tasks [16]. Serre et al. in [39] propose a biological model for object detection which is inspired by the working of the HVS. It uses a feature set combining position and scale-tolerant edge detectors over neighboring locations and multiple orientations for an object detection task.

Apart from detecting salient regions in static images, many works have focused on detecting salient regions in videos. Wixson [46] integrates optical flow cues to determine objects that are motion salient. In [11, 12], the authors have used a Bayesian framework to predict surprising regions in video while in [5] the problem of detecting salient regions in videos is posed as an inference process in a probabilistic graphical model. The computational model in [30] uses entropy for identifying salient regions in numerous short-duration video clips.

2.1 Neurobiology Based Saliency Model

Primates have a remarkable ability to interpret complex scenes in real time. It is believed that intermediate and higher visual processes select a subset of available sen-

sory information for processing [45]. This is most likely to reduce the complexity of scene analysis [33]. The selection of visual information appears to be in the form of a spatially circumscribed region of the visual field which is also called the focus of attention. It scans the scene both in a rapid, bottom-up, saliency-driven, and task independent manner and in a slower, top-down and task dependent manner [33]. Models of visual attention include "dynamic routing models" where information from a small region of the visual field can progress through the cortical visual hierarchy. The attended region is selected through dynamic modifications of cortical connectivity or by establishing specific temporal patterns of activity [33, 36, 45].

The saliency model by Itti and Koch builds on biologically plausible architecture proposed in [17] and is at the basis of several other models [3, 31]. The model is related to the "feature integration theory" that explains human visual search strategies [2]. Visual input is first decomposed into a set of topographic feature maps and different spatial locations compete for saliency within each map such that only locations that stand out from their surroundings persist. The feature maps serve as input to a saliency map that determines the conspicuity over the entire visual scene. It is believed that such a map is located in the posterior parietal cortex of primates [38]. The model represents a complete account of bottom-up saliency and does not require any top-down guidance to shift attention. Such a framework allows for parallel processing for fast selection of a small number of interesting image locations.

From the input image, nine spatial scales are created using dyadic Gaussian pyramids [10]. They progressively low-pass filter and subsample the input image yielding horizontal and vertical reduction factors ranging from 1 : 1 to 1 : 256 in eight octaves. Each feature is computed by a set of linear center-surround operations akin to visual receptive fields. Typical visual neurons are most sensitive in a small region of the visual space (the center). Stimuli presented in a broader, weaker antagonistic region concentric with the center (referred as the surround) inhibit the neuronal response. Such an architecture is sensitive to local spatial discontinuities, and is particularly well suited to detecting locations that stand out from their surroundings. This is a general computational principle in the retina [18]. Center-surround is implemented in the model as the difference between fine and coarse scales. The center is a pixel at scale $c \in \{2, 3, 4\}$, and the surround is the corresponding pixel at scale $s = c + \delta$, $\delta \in \{3, 4\}$. The across-scale difference between the two maps is obtained by interpolating to finer scales and point-by-point subtraction. Several scales lead to multiscale feature extraction by including different size ratios between center and surround regions.

2.1.1 Extraction of Early Visual Features

Let R, G and B be the red, green and blue channels of the input image and an intensity image I is obtained as $I = (R + G + B)/3$. I is used to create a Gaussian pyramid $I(\sigma)$ where $\sigma \in [0 \ldots 8]$ is the scale. Center-surround difference (denoted as $\ominus$) between a "center" fine scale c and "surround" coarse scale s yields the feature maps. The first set of feature maps for intensity contrast is given by

$$I(c, s) = \left| I(c) \ominus I(s) \right|. \tag{1}$$

A second set of maps is constructed for the color channels, which, in cortex are represented using a "color double-opponent system". In the center of their receptive fields, the neurons are excited by one color and inhibited by another while the converse is true for the surround. Such spatial and chromatic opponency exists for red/green, green/red, blue/yellow and yellow/blue color pairs in human primary visual cortex [9]. Thus, maps $RG(c,s)$ is created to simultaneously account for red/green and green/red double opponency and $BY(c,s)$ for blue/yellow and yellow/blue opponency.

$$\begin{aligned} RG(c,s) &= \big|\big(R(c) - G(c)\big) \ominus \big(G(s) - R(s)\big)\big|, \\ BY(c,s) &= \big|\big(B(c) - Y(c)\big) \ominus \big(Y(s) - B(s)\big)\big|. \end{aligned} \tag{2}$$

Local orientation information is obtained from I using oriented Gabor pyramids $O(\sigma,\theta)$, where $\sigma \in [0 \ldots 8]$ represents scale and $\theta \in \{0°, 45°, 90°, 135°\}$ are the preferred orientations [10]. Orientation feature maps are obtained as

$$O(c,s,\theta) = \big|O(c,\theta) \ominus O(s,\theta)\big|. \tag{3}$$

In total 42 feature maps are computed: 6 for intensity, 12 for the three color channels and 24 for orientation. The individual features are robust for object tracking and their combination is expected to give better results.

2.1.2 The Saliency Map

The purpose of the saliency map is to represent the conspicuity (or saliency) at every location in the visual field by a scalar quantity, and to guide the selection of attended locations based on the spatial distribution of saliency. A combination of the feature maps provides bottom-up input to the saliency map modeled as a dynamic neural network. The different feature maps represent different modalities with different dynamic ranges and extraction mechanisms. When all feature maps are combined salient objects appearing strongly in a few maps may be masked by noise or less salient objects in other maps. Therefore, a normalization operator $N(\cdot)$ is used to globally promote maps having a small number of strong peaks of activity (conspicuous locations), while globally suppressing maps containing numerous comparable peak responses. $N(\cdot)$ consists of the following steps:

1. Normalize the values in the map to a fixed range $[0 \ldots M]$, in order to eliminate modality-dependent amplitude difference;
2. Find the location of the map's global maximum M and compute the average $\bar{m}$ of all its other local maxima; and
3. Globally multiply the map by $(M - \bar{m})^2$.

Comparing the maxima of the entire map to the average overall activation measures how different the most active location is from the average. When this difference is large, the most active location stands out and the map is strongly promoted. When the difference is small the map contains nothing unique and is suppressed. The biological motivation behind the design of $N(\cdot)$ is that it coarsely replicates

cortical lateral inhibition mechanisms, where neighboring similar features inhibit each other via specific anatomically defined connections [6]. The feature maps are combined into "conspicuity" maps, $\bar{I}$ for intensity, $\overline{O}$ for orientation at the scale $\sigma = 4$ of the saliency map and $\overline{C}$ for color. The final saliency map obtained as the combination of the two normalized conspicuity maps is

$$SM = \frac{1}{3}\left[N(\bar{I}) + N(\overline{O}) + N(\overline{C})\right]. \tag{4}$$

At any given time the maximum of the saliency map (SM) defines the most salient image location where the focus of attention (FOA) is directed. However in a neuronally plausible implementation, the SM is modeled as a 2D layer of leaky integrate-and-fire neurons at scale $\sigma = 4$. These model neurons consist of a single capacitance that integrates the charge delivered by synaptic input, of a leakage conductance and a voltage threshold. When the threshold is reached, a prototypical spike is generated and the capacitive charge is shunted to zero. The SM feeds into a biologically plausible 2D "winner-take-all" (WTA) neural network [17, 45] at scale $\sigma = 4$, where synaptic interactions among units ensure that only the most active locations are suppressed.

The neurons receiving excitatory input from SM are all independent. The potential of SM neurons at more salient locations increases faster and each SM neuron excites its corresponding WTA neuron. All the WTA neurons also evolve independently of each other, until one (the winner) first reaches threshold and fires. This triggers three simultaneous mechanisms:

1. The FOA is shifted to the location of the winner neuron;
2. The global inhibition of the WTA is triggered and completely inhibits (resets) all WTA neurons;
3. Local inhibition is transiently activated in the SM, in an area with the size and new location of the FOA; this not only yields dynamical shifts of the FOA, by allowing the next most salient location to subsequently become the winner, but also prevents the FOA from immediately returning to a previously attended location.

Such an inhibition of return has been demonstrated in human visual psychophysics [37]. As no top-down attentional component is modeled, the FOA is a simple disk with radius fixed to one-sixth of the smaller of the input image width or height. The time constants, conductances and firing thresholds of the simulated neurons were chosen so that the FOA jumps from one salient location to another in approximately 30–70 ms of simulated time. The attended area is inhibited for approximately 500–900 ms. The difference in the relative magnitude of these delays is sufficient to ensure thorough scanning of the image and prevent cycling through a limited number of locations.

2.1.3 Strengths and Limitations

Despite its simple architecture and feed-forward structure the model is capable of strong performance with complex natural scenes. It can quickly detect salient points

in different kinds of images [15]. Another strength of the model is the parallel implementation of the computationally expensive early feature extraction stages and the attention focusing system. This allows for real time operation on dedicated hardware. A critical part of the model is the implementation of the normalization operator $N(\cdot)$ which provides a general mechanism for computing saliency. The resulting saliency measure is closer to human saliency as it implements spatial competition between salient locations. The feed-forward implementation of $N(\cdot)$ is faster and simpler than iterative schemes. The efficiency of the proposed saliency model depends upon the features used and can be tailored to arbitrary tasks through the implementation of dedicated feature maps.

2.2 Our Modifications

We have described the saliency map as developed in the original work by Itti et al. [15]. For our tracking algorithm, we modify the method to make it suitable for our datasets. Since we worked on grayscale images, only we do not use the color channel for saliency map calculation. Therefore, our saliency map is based on a combination of intensity and orientation conspicuity maps to get the final saliency map. The second change is that we do not implement the winner take all step to get the most salient region. Instead all regions are assigned with a saliency value which is combined with the motion intensity map for subsequent analysis.

2.3 Motion Saliency Map

We define *motion saliency* as attention due to motion. Since tracking involves video clips, motion is undoubtedly a strong factor in capturing the viewers attention. Abrams et al. [1] have shown that onset of motion captures attention. Objects that accelerate, such as those that have just begun to move, are more likely to be seen than objects that undergo deceleration [43]. Models have been developed that emulate the response of the middle temporal (MT) area of the primate cortex which is selective to velocity in visual stimuli [40]. Attention models have been used to not only sense and analyze eye movements, but also guide them by using a special kind of gaze-contingent information display [4].

A *motion saliency map* is a representation of regions that are moving and salient. It is calculated by combining spatial coherency and temporal coherency [21]. It is based on the concept of motion vector fields (MVF). The MVF is analogous to the retina of the eye and motion vectors are the perceptual response of optic nerves. This approach results in the calculation of three maps—Intensity map, the Spatial Coherency map, and the Temporal Coherency map, each corresponding to 3 inductor fields of a MVF. These three maps are then combined to get the final motion attention map highlighting regions that are moving and are visually salient. We shall describe each component of the motion saliency map.

2.3.1 Motion Vectors

Motion vectors are an integral part of many video compression algorithms, where they are used for motion compensation or block matching. The idea behind block matching is to divide the current frame into a matrix of blocks which are then compared with the corresponding block and its neighbors in the previous frame to determine a motion vector. The motion vector estimates the movement of a block from one frame to another. We calculate the motion vectors using the Adaptive Rood Pattern Search (ARPS) [34]. The ARPS algorithm makes use of the fact that the general motion in a frame is usually coherent, that is, if the blocks around the current block moved in a particular direction then there is a high probability that the current block will also have a similar motion vector. This algorithm uses the motion vector of the block to its immediate left to predict its own motion vector.

Motion Intensity The motion intensity I_t, a measure of induced motion energy or activity, is computed as the normalized magnitude of motion vectors:

$$I_t(x, y) = \frac{\sqrt{dx^2{}_{x,y} + dy^2{}_{x,y}}}{MaxMag}, \tag{5}$$

where $dx_{x,y}$, $dy_{x,y}$ denote the components of motion vectors at location x, y, and $MaxMag$ is the maximum magnitude in the motion vector field.

2.3.2 Spatial Coherency

Spatial coherency indicates the blocks of pixels that belong to the same rigid object. This is achieved by calculating the entropy over a block of pixels. The higher the entropy the smaller the probability of that particular block belonging to the same object. Lower entropy implies greater relationship between pixels and thus a higher probability of the group of pixels belonging to the same object. The spatial coherency at pixel x, y, considering a window of size 8×8, is given by

$$C_s(x, y) = -\sum_{i=1}^{N} p_s(i) \log p_s(i), \tag{6}$$

where $p_s(i)$ is the probability of occurrence of pixel intensity i and $N = 8 \times 8$.

2.3.3 Temporal Coherency

The motivation behind the temporal coherency map is similar to that of the spatial coherency map. Instead of a block of pixels we have a group of pixels over different image frames. Higher entropy implies greater motion and hence a higher measure of

saliency. We analyze a maximum of 7 frames prior to the current frame and compute the entropy for all the pixel points:

$$C_t(x, y) = -\sum_{i=1}^{M} p_t(i) \log p_t(i), \tag{7}$$

where $p_t(i)$ is the probability of occurrence of pixel intensity i at the corresponding location for different frames. For the temporal coherency map $M = 7$, as the effect of motion in one frame on the scan path of the eye lasts for 5–7 frames.

2.3.4 Combining the Maps

The three maps are combined into the final motion saliency map as follows:

$$B = I_t \times C_t \times (1 - I_t \times C_s), \tag{8}$$

which is used for all further analysis. In the temporal coherency map higher entropy implies greater motion over that particular area. Since our aim is to determine *motion salient* regions, this is a direct indicator of interesting regions. For this reason the intensity map is multiplied with the temporal coherency map ($I_t \times C_t$).

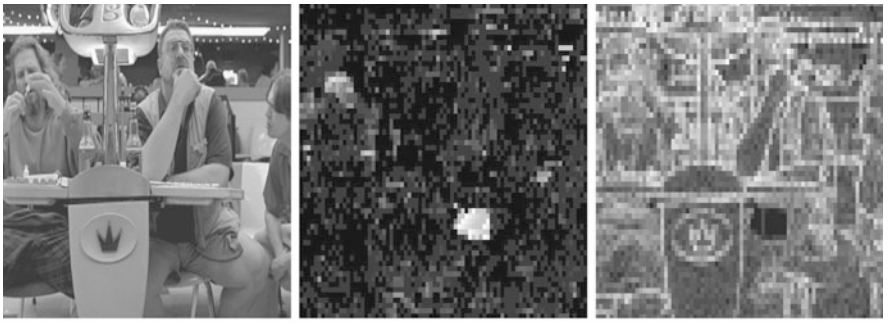
However, in the spatial coherency map greater entropy indicates disparate objects. Our aim is to group objects together. This is the justification for the third term ($1 - I_t \times C_s$), which in essence assigns higher value to pixels belonging to the same object. Thus, the output of the motion saliency map are regions in the image that belong to one object and are moving.

An example of the various maps for a frame from one of our test clips is shown in Fig. 1. The motion saliency map highlights those regions that are motion salient (Fig. 1(e)). Motion saliency of a region is quantified by the value of the corresponding pixel in the motion saliency map.

3 Particle Filter

Let $\mathbf{X}_t$ denote the state of a tracked object, and $\mathbf{Z}_t = \{z_1, \ldots, z_t\}$ denote observations up to t time instances. The use of particle filters is popular in scenarios where the posterior density $p(X_t|Z_t)$ and observation density $p(Z_t|X_t)$ are non-Gaussian. Particle filtering approximates the probability distribution with a weighted sample set $S = \{(s^{(n)}, w^{(n)}), n = 1 \ldots N\}$. Each sample s represents a hypothetical state of the object with a corresponding discrete sampling probability w, where $\sum_{n=1}^{N} w^{(n)} = 1$.

The samples' evolution is described by propagating them with the help of a system model. The elements of the set are weighted in terms of the observations. N samples are drawn with replacement with a particular sample chosen with the



(a) first frame (b) motion intensity map (c) spatial coherency map

(d) temporal coherency map (e) final motion saliency map

Fig. 1 (**a**) first frame of a movie sequence; (**b**)–(**e**) different components of the motion saliency map for the frame shown in (**a**)

probability $\pi(n) = p(z_t|X_t = s_t^{(n)})$. A object's mean state at each step is estimated by (9). Further details of the principles of particle filter can be found in [35].

$$E[S] = \sum_{n=1}^{N} w_{(n)} s_{(n)}. \tag{9}$$

3.1 Implementation

We use 100 particles which are propagated over time. The final state of the object is determined by a weighted combination of these samples based on their likelihood. The tracking algorithm can be summarized in the following steps: (1) Initialize a target model defining the characteristics of the object to be tracked. (2) Generate saliency map of each frame. (3) Add random noise to the samples and determine their weights. (4) Determine the new state, resample the particles and update centroid of the template to the new value. The template features were defined as the centroid of object, its normalized motion saliency value and the average motion

saliency value within a bounding box over the object. Normalizing the saliency map ensures that the most salient location in a frame has a value 1. The updated centroid at each step 'tracks' the object of interest.

3.1.1 Assigning Weights

Each sample is assigned a weight that depends on similarity with the template. The assigned weights are calculated in the following manner:

(1) The Euclidean distance between sample point and model is calculated which is denoted as *dist*.
(2) The absolute difference in saliency values at sample point and model is denoted as S_d. The corresponding difference between the average saliency of bounding box is denoted as $\widehat{S}_d$.
(3) Saliency coefficient *sc* is calculated as

$$sc = (1 + dist)(1 + S_d)(1 + \widehat{S}_d). \tag{10}$$

(4) The assigned weight of the sample is given by

$$w^n = \frac{1}{\sqrt{2\pi}\sigma} e^{\frac{1-sc^2}{2\sigma^2}}. \tag{11}$$

Equation (11) assigns lower weights to samples that greatly differ from the model. The formulation of (10) is such that all three attributes contribute to the similarity criterion and an accidental perfect match for one attribute does not bias the algorithm to the particular location.

The algorithm was implemented in MATLAB on a Pentium D 2.8 Ghz machine having 2 GB RAM. The movie clips used for the experiment were of resolution 240×320 and had a frame rate of 15 fps. The saliency values were normalized to lie between 0 and 1.

4 Results and Discussion

We test our algorithm on different test clips containing various scenarios like tracking a person indoor, tracking a person in an outdoor environment, and sports videos. For each clip the initial object to be tracked is determined by the static saliency map, that is, the most salient object in the first frame is tracked over the entire clip.

4.1 Comparison with Background Subtraction

We implement a simple background subtraction method using Kalman filters to make a qualitative comparison of the performance of our algorithm. A scene's background is determined from the average image over a prolonged period with all lights

Fig. 2 Frames of output sequence involving illumination change using our algorithm

Fig. 3 Frames of output sequence involving illumination change using our background subtraction

in the room switched on. A background subtraction method works well for cases where the intensity does not vary much. Switching off a light can greatly affect the tracking procedure. However our saliency based algorithm, combining motion and low-level features, greatly increases the robustness of the algorithm.

Figure 2 shows the results of tracking in indoor environments under different illumination conditions using our saliency-based algorithm. Figure 3 shows the corresponding results using a background subtraction method. Our algorithm performs better than the background subtraction approach thus showing that saliency is a useful approach for tracking under changing ambient illumination.

Figures 4 and 5 show results of our algorithm on outdoor videos and in sports videos. For the outdoor video only one person needs to be tracked (as indicated by the static saliency map), and the overall tracking is fairly accurate. In the sports video there are multiple athletes and the saliency map denotes the middle athlete as the most salient object and hence begins to track him. In the second image of Fig. 5, we observe that the bounding box which tracks the object is not entirely on the middle athlete. This is because the background color matches that of the athlete's jersey and hence deviates the tracking algorithm. Because of our method's reliance on motion information it quickly recovers and tracks the athlete in subsequent frames.

In terms of the number of frames where tracking was accurate, our saliency based algorithm, with 97 % accuracy, performed better than the background subtraction algorithm which had an accuracy rate of 89 %. For our algorithm, we used a particle filter for updating the states whereas for the background subtraction method a Kalman filter was used. As a result, one erroneous prediction of state using the Kalman filter made it difficult for the method to keep track of the original object.

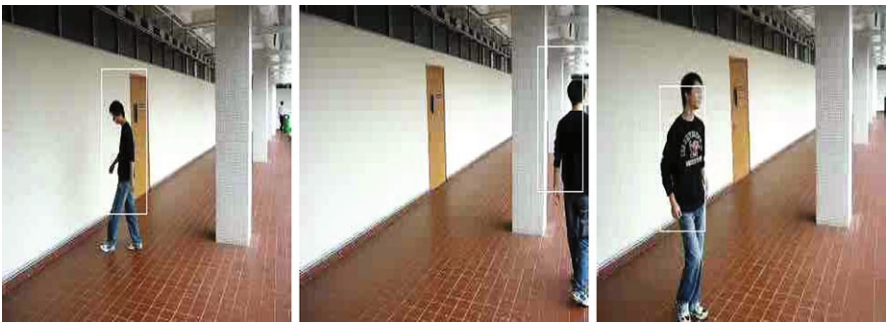


Fig. 4 Results of tracking in an outdoor scene

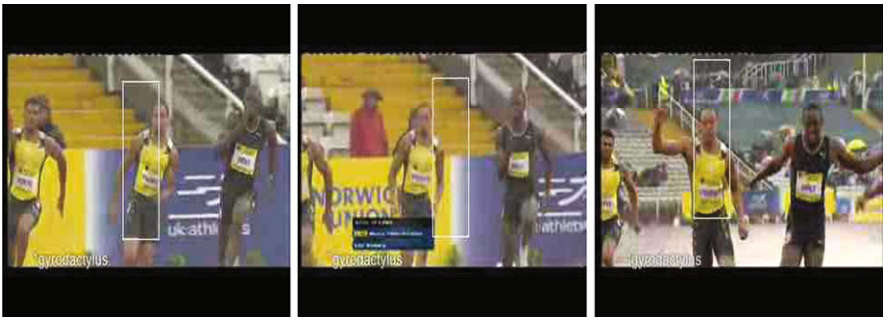


Fig. 5 Results of tracking a person in a sports video

But the use of a particle filter proved to be more robust and any erroneous prediction of states could be rectified in subsequent frames.

5 Conclusion

In this chapter, we have proposed a saliency-based particle filter approach for object tracking that incorporates motion cues. The algorithm was tested on different videos and found to perform better than a background subtraction method using Kalman filters. Our algorithm had a 97 % accuracy rate compared to the background subtraction algorithm which had an accuracy rate of 89 %. From the observed results, we conclude that saliency has great promise for use in object tracking. Depending upon the scenario in question, appropriate saliency maps can be generated using motion cues and appropriate low level features. It is expected that such an approach will be more effective than conventional approaches using only intensity and color features. In the future, we would like to investigate robust approaches to detecting motion saliency maps for noisy videos or videos with insufficient illumination.

References

1. Abrams, R.A., Christ, S.E.: Motion onset captures attention. Psychol. Sci. **14**(5) (2003)
2. Bajcsy, A., Gelade, G.: A feature integration theory of attention. Cogn. Psychol. **12**, 97–136 (1980)
3. Baluja, S., Pomerleau, D.: Expectation based selective attention for visual monitoring and control of a robot vehicle. Robot. Auton. Syst. **22**(3–4), 329–344 (1997)
4. Barth, E., Dorr, M., Böhme, M., Gegenfurtner, K.R., Martinetz, T.: Guiding the mind's eye: improving communication and vision by external control of the scanpath. In: Proc. SPIE Human Vision and Electronic Imaging, San Jose, CA, USA, vol. 6057 (2006)
5. Boiman, O., Irani, M.: Detecting irregularities in images and in video. In: IEEE Intl Conf. on Computer Vision, pp. 1–8 (2005)
6. Cannon, M., Fullenkamp, S.: A model for inhibitory lateral interaction effects on perceived contrast. Vis. Res. **36**(8), 1115–1125 (1996)
7. Chen, C., Wolf, W.: Background modeling and object tracking using multi-spectral sensors. In: 4th ACM International Workshop on Video Surveillance and Sensor Networks, pp. 27–34 (2006)
8. Comaniciu, D., Ramesh, V., Meer, P.: Kernel-based object tracking. IEEE Trans. Pattern Anal. Mach. Intell. **25**(5), 564–577 (2003)
9. Engel, S., Zhang, X., Wandell, B.: Color tuning in visual cortex measured with functional magnetic resonance imaging. Nature **388**(6637), 68–71 (1997)
10. Greenspan, H., Belongie, S., Goodman, R., Perona, P., Rakshit, S., Anderson, C.H.: Overcomplete steerable pyramid filters and rotation invariance. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 222–228 (1994)
11. Itti, L., Baldi, P.: Bayesian surprise attracts human attention. In: Neural Information Processing Systems (NIPS), pp. 1–8 (2005)
12. Itti, L., Baldi, P.: A principled approach to detecting surprising events in video. In: IEEE Intl. Conf. Computer Vision and Pattern Recognition, pp. 631–637 (2005)
13. Itti, L., Koch, C.: A saliency-based search mechanism for overt and covert shifts of visual attention. Vis. Res. **40**, 1489–1506 (2000). citeseer.ist.psu.edu/itti00saliencybased.html
14. Itti, L., Koch, C.: A saliency-based search mechanism for overt and covert shifts of visual attention. Vis. Res. **40**, 1489–1506 (2000)
15. Itti, L., Koch, C., Niebur, E.: A model of saliency-based visual attention for rapid scene analysis. IEEE Trans. Pattern Anal. Mach. Intell. **20**(11), 1254–1259 (1998)
16. Kadir, T., Brady, M.: Saliency, scale and image description. Int. J. Comput. Vis. **45**(2), 85–105 (2001)
17. Koch, C., Ullman, S.: Shifts in selective visual attention: towards the underlying neural circuitry. Hum. Neurobiol. **4**, 219–227 (1985)
18. Leventhal, A.: The Neural Basis of Visual Function. Vision and Visual Dysfunction, vol. 4. CRC Press, Boca Raton (1991)
19. Logan, G.: The CODE theory of visual attention: an integration of space-based and object-based attention. Psychol. Rev. **103**, 603–649 (1996)
20. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. Int. J. Comput. Vis. **60**, 91–110 (2004)
21. Ma, Y., Lu, L., Zhang, H., Li, M.: A user attention model for video summarization. In: Proceedings of ACM Multimedia (2002). citeseer.ist.psu.edu/ma03user.html
22. Mahapatra, D., Sun, Y.: Nonrigid registration of dynamic renal MR images using a saliency based MRF model. In: Proc. MICCAI, pp. 771–779 (2008)
23. Mahapatra, D., Sun, Y.: Registration of dynamic renal mr images using neurobiological model of saliency. In: Proc. ISBI, pp. 1119–1122 (2008)
24. Mahapatra, D., Sun, Y.: Using saliency features for graphcut segmentation of perfusion kidney images. In: 13th International Conference on Biomedical Engineering, pp. 639–642 (2008)
25. Mahapatra, D., Sun, Y.: Joint registration and segmentation of dynamic cardiac perfusion images using MRFs. In: Proc. MICCAI, pp. 493–501 (2010)

26. Mahapatra, D., Sun, Y.: Mrf based intensity invariant elastic registration of cardiac perfusion images using saliency information. IEEE Trans. Biomed. Eng. **58**(4), 991–1000 (2011)
27. Mahapatra, D., Sun, Y.: Orientation histograms as shape priors for left ventricle segmentation using graph cuts. In: Proc: MICCAI, pp. 420–427 (2011)
28. Mahapatra, D., Sun, Y.: Integrating segmentation information for improved mrf-based elastic image registration. IEEE Trans. Image Process. **21**(1), 170–183 (2012)
29. Mahapatra, D., Saini, M., Sun, Y.: Illumination invariant tracking in office environments using neurobiology-saliency based particle filter. In: IEEE ICME, pp. 953–956 (2008)
30. Mahapatra, D., Winkler, S., Yen, S.C.: Motion saliency outweighs other low-level features while watching videos. In: Proc. SPIE Human Vision and Electronic Imaging, San Jose, CA, vol. 6806 (2008)
31. Milanese, R., Gil, S., Pun, T.: Attentive mechanisms for dynamic and static scene analysis. Opt. Eng. **34**(8), 2428–2434 (1995)
32. Mozer, M., Sitton, M.: Computational modeling of spatial attention. In: Pashle, H. (ed.) Attention, pp. 341–393. UCL Press, London (1998)
33. Neibur, E., Koch, C.: Computational architectures for attention. In: Parasuraman, R. (ed.) The Attentive Brain, pp. 163–186. MIT Press, Cambridge (1998)
34. Nie, Y., Ma, K.H.: Adaptive rood pattern search for fast block-matching motion estimation. IEEE Trans. Image Process. **11**(12), 1442–1448 (2002)
35. Nummiaro, K., Koller-Meier, E., Gool, L.V.: An adaptive color-based particle filter. Image Vis. Comput. **21**(1), 99–110 (2003). citeseer.ist.psu.edu/nummiaro02adaptive.html
36. Olsahausen, B., Anderson, C.H., van Essen, D.: A neurobiological model of visual attention and invariant pattern recognition based on dynamic routing of information. J. Neurosci. **13**(11), 4700–4719 (1993)
37. Posner, M., Cohen, Y.: Components of visual orienting. In: Bouma, H., Bouwhuis, D. (eds.) Attention and Performance, pp. 531–556. Erlbaum, Hilldale (1984)
38. Robinson, D., Peterson, S.: The representation of visual salience in monkey parietal cortex. Nature **391**(6,666), 481–484 (1998)
39. Serre, T., Wolf, L., Poggio, T.: A new biologically motivated framework for robust object recognition. Technical report AI Memo 2004-026, Computer Science and Artificial Intelligence Laboratory, Massachussets Institute of Technology (2004)
40. Simoncelli, E.P., Heeger, D.J.: A model of neuronal responses in visual area MT. Vis. Res. **38**(5), 743–761 (1998). http://www.cns.nyu.edu/~eero/ABSTRACTS/simoncelli96-abstract.html
41. Soto, D., Blanco, M.: Spatial attention and object-based attention: a comparison within a single task. Vis. Res. **44**, 69–81 (2004)
42. Spengler, M., Schiele, B.: Towards robust multi-cue integration for visual tracking. ACM Comput. Surv. **14**(1), 50–58 (2003)
43. Tremoluet, P., Feldman, J.: Perception of animacy from the motion of a single object. Perception **29**, 943–951 (2000)
44. Triesch, J., Malsburg, C.: Self-organized integration of adaptive visual cues for face tracking. In: International Conference on Automatic Face and Gesture Recognition, pp. 102–107 (2000)
45. Tsotsos, J., Culhane, S., Hai, W., Lai, Y., Davis, N., Nuflo, F.: Modeling visual attention via selective tuning. Artif. Intell. **78**(1), 507–545 (1995)
46. Wixson, L.: Detecting salient motion by accumulating directionally-consistent flow. IEEE Trans. Pattern Anal. Mach. Intell. **22**(8), 774–780 (2000)
47. Yilmaz, A., Javed, O., Shah, M.: Object tracking: a survey. ACM Comput. Surv. **38**(4) (2006)

Multiresolution Depth Map Estimation in PTZ Camera Network

Sanjeev Kumar, Christian Micheloni, and Balasubramanian Raman

Abstract In this chapter, an active stereo vision system composed of two pan-tilt-zoom (PTZ) cameras is proposed for estimating multiresolution depth map for a large and complex scene. The rectification of stereo images is performed based on the sigmoid interpolation with a set of neural networks. The orientation parameters (pan and tilt values) and the rectification transformations of corresponding images are used as the input-output pairs for network training. The input data is read from cameras directly, whereas the output data is computed offline. The trained neural network is used to interpolate rectification transformations in real time for the stereo images captured at arbitrary pan and tilt settings. The correspondence between the stereo images is obtained using a chain of homographies based scheme. Non-homogeneity between the intrinsic parameters of two cameras is treated by means of zoom compensation to improve the quality of stereo rectification. Experimental results are given for estimating multiresolution depth map for a scene.

1 Introduction

The development of modern surveillance systems has attracted a lot of interest in recent years [1, 4, 7, 18, 19]. Recently, the concept of stereo vision has been implemented in surveillance systems to make the latter more efficient. Stereo vision can estimate the 3D position of an object in a given coordinate system from its two perspective images [2]. Traditional stereo vision research uses static cameras for their low cost and relative simplicity in modeling. A pan-tilt-zoom (PTZ) camera is

S. Kumar (✉)
Department of Mathematics, Indian Institute of Technology Roorkee, Roorkee 247 667, India
e-mail: malikfma@iitr.ernet.in

C. Micheloni
Department of Mathematics and Computer Science, University of Udine, Udine 33100, Italy
e-mail: christian.micheloni@uniud.it

B. Raman
Department of Computer Science and Engineering, Indian Institute of Technology Roorkee, Roorkee 247 667, India
e-mail: balarfma@iitr.ernet.in

P.K. Atrey et al. (eds.), *Intelligent Multimedia Surveillance*,
DOI 10.1007/978-3-642-41512-8_8,

a typical and the simplest active camera, whose pose can be fully controlled by pan, tilt and zoom parameters. As PTZ cameras are able to obtain multi-angle-views and multiresolution information (i.e., both global and local image information), these are used for wide area monitoring [9]. Therefore, a PTZ camera based stereo vision system is able to cover a large environment. However, such type of active stereo vision systems are much more challenging when compared to the traditional stereo vision system. PTZs, on purpose (e.g., zoom on face, zoom on license plate etc.), can vary both the intrinsic and the extrinsic parameters thus changing the stereo properties. In this context, it is not an easy task to perform some operations like stereo image rectification in an active stereo vision system.

Recently, a novel image rectification algorithm has been proposed for a dual-PTZ-camera based stereo system [23]. In such a system, the problem related to inconsistency of intensities in two camera images is solved by addressing a two-step stereo matching strategy. An interesting approach in the case of the active stereo vision system by means of rotating cameras has been proposed with analytic formulation in [8]. To do this, an off-line initialization process has been performed to initialize the essential matrix using known calibration parameters. During on-line operations the rotation angles of the cameras are retrieved and exploited to compute the current essential matrix. However, when the zoom is considered, the calibration for any adopted zoom level is required for both cameras. Moreover, in a Dual PTZ camera system, the discrepancies [3] in the field-of-view (FOV) and in the resolution levels of the two cameras lead to difficulties not only in the stereo rectification but also in the depth estimation.

This work introduces a new method to perform the image rectification process in the case of two PTZ camera based active stereo vision systems. The idea is to model pan and tilt values for any setting of cameras as independent variables and a number of parameters (rotation parameters here) depend on these. In this way, the online rectification problem can be modeled as a nonlinear function approximation problem. A LUT can be constructed offline having the corresponding values for all these parameters on different orientation and a fixed zoom. Then, the function approximation problem can be solved using supervised learning of a neural network. In other words, the rotation parameters are interpolated with respect to given pan and tilt values for computing the required rectification transformations. Neural networks have been used rarely in video surveillance application. Here, a few properties of neural networks such as the function approximation property in the case of highly nonlinear data and fast simulation make it suitable for such applications. In the case of zoom operation in any PTZ camera, a focal ratio based approach is used to compensate the effect of unequal zoom levels between the two cameras [13].

To show the effectiveness of the proposed approach, a video surveillance application is considered which shows the importance of the zoom settings of PTZ cameras in scene understanding. In the case of the static cameras based stereo system, the images are captured with the same resolution. However, in the case of the PTZ cameras, we can consider two cases: (1) if a region has small depth variations, that is, almost flat in nature, low resolution images can be used for obtaining the depth map, and (2) when large depth variations occur in a region, high resolution images are required. Based on these two facts, the PTZ cameras based stereo vision system

provides a multiresolution depth map that can be used for better scene understanding and requires low computational cost. In this context, another application of the dual PTZ camera based stereo system is to grab images with the above facts in an automatic manner and create a multiresolution depth map mosaic of a wide area.

In brief, the main advantages of the proposed PTZ camera based stereo vision system are:

- There is no need to assume a fixed center of projection for the PTZ camera during rotations.
- The depth-map computations can be achieved with wide baseline stereo systems.
- Only limited a priori information (i.e., information provided by a static camera) is required to compute multiresolution depth maps for a large environment.

In particular, concerning the last advantage, contrarily to [24], there is no need to have the coarse depth map and the precise FOV of the left camera. In addition, instead of using wide baseline feature matching techniques [6, 15], that even though efficient are computationally expensive, an approach based on a chain of 2D homographies is addressed to find corresponding points between wide baseline images in real-time.

The rest of the chapter is organized as follows: Sect. 2 describes the system architecture. Section 3 introduces a method for obtaining wide baseline stereo correspondence. The offline steps, such as computation of rectification transformations on sampled values of pan and tilt angles, training of ANN and zoom to focal length fitting, are given in Sect. 4. The online steps are described in Sect. 5. A detailed process for a constructing high resolution depth map mosaic is presented in Sect. 6. Various results and discussions about them are given in Sect. 7. Finally, Sect. 8 concludes the chapter.

2 System Architecture and Description

The proposed system contains mainly two different units of cameras. The first unit, called static camera unit (SCU), is composed of multiple static cameras. These static cameras have wide FOV and cover a large environment with limited overlapping FOV. The second unit contains two different PTZ cameras placed at a wide distance (7 meters) from each other and considered as a dual-PTZ based stereo system. This unit is called active stereo unit (ASU). The main functionalities of SCU are object detection [16], behavior understanding and anomalous event detection [20]. Once a region of interest is detected by the SCU, the system delivers the information to the ASU for focusing the two PTZ cameras towards the selected region. The ASU starts the stereo task as soon as the selected target appears in the FOVs of both cameras. The handover of scene information [21, 25] between the different cameras allows a cooperative tracking of the objects within the monitored environment. The sequences acquired by the two PTZ cameras in ASU are transmitted to a central node. The information of the orientation and resolution of these cameras also transmitted with these sequences. A communication system based on a multi-cast

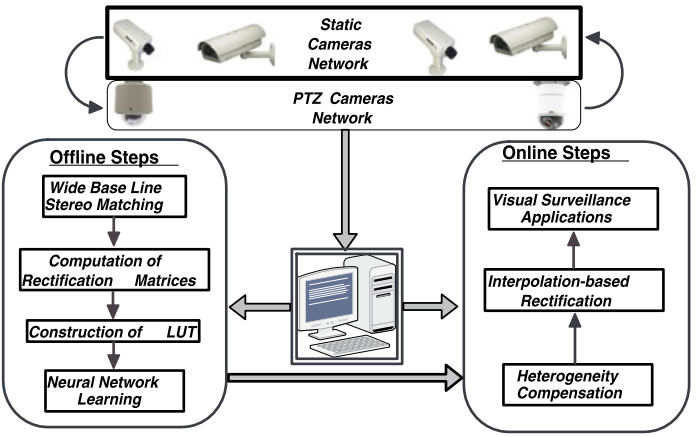


Fig. 1 A virtual design of the proposed stereo system

protocol [17] is used for the cooperation within these cameras' network. This communication system is designed in such a way that it requires low bandwidth. A logic architecture of the proposed system is shown in Fig. 1, where, the top layer of the cameras represents the SCU, while the ASU is shown in the second layer.

The properties of the PTZ camera deployment make the stereo vision problem more difficult when compared to classical stereo systems. Like the captured images from the pair of PTZ cameras are not homogeneous in terms of the intrinsic parameters (resolution and distortion). If we perform rectification on these pairs of images, it introduce errors (distortion effect) in the rectified images. The effect of these unequal intrinsic parameters must be compensated before rectification. Here, the resolution of the images are equalized using the focal-lengths of the two PTZ cameras. The focal lengths are estimated directly from the zoom value. The ratio between the zoom values of two cameras is used to compensate the unequal resolution effect. Once the frames are homogeneous, the rectification transformations are interpolated using a neural network.

3 Correspondence Using Chain of Homographies

SIFT matching [14] is a popular tool for extracting matching points from a pair of stereo images. However, this method is not very accurate in a case when images do not share sufficient common FOV. It can happen when objects are close to both the cameras placed at a wide baseline (see Fig. 2). To sidestep such a problem, concept from the planar homography correspondence can be used [10]. Here, a method based on a chain of homographies is adopted used for establishing the correspondence between the pair of stereo images. Our idea is to initialize the correspondence between the images of two PTZ cameras captured for a far scene and then subsequently use it for other pair of images. To do this, we require the correspondence

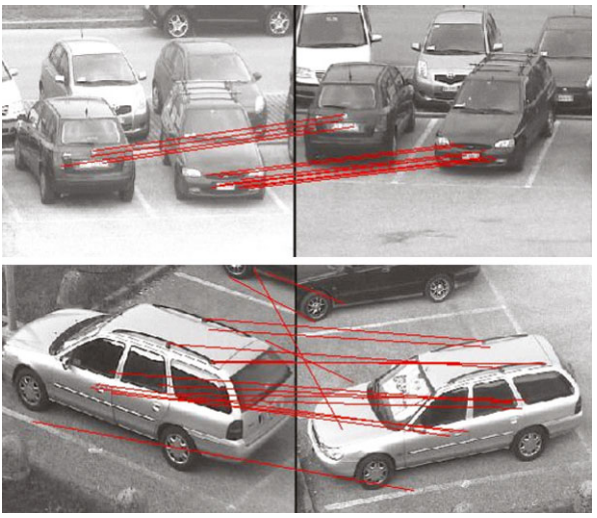


Fig. 2 SIFT matching between wide baseline stereo images of a far (*top*) and near (*bottom*) scenes along the optical axis of camera

Table 1 Descriptions of some symbols and parameters used in the different equations

Symbols	Descriptions
$(\mathbf{I}_l, \mathbf{I}_r)$	Pair of images captured from left and right cameras, respectively
$\mathbf{H}$	Homography
$(\mathbf{J}_l, \mathbf{J}_r)$	Rectified pair of stereo images
$(\mathbf{A}_l, \mathbf{A}_r)$	Rectification matrices for left and right cameras, respectively
$(\mathbf{m}_l, \mathbf{m}_r)$	Set of corresponding points between left and right stereo images
$\mathbf{F}$	Fundamental matrix
$(\mathbf{P}_l, \mathbf{P}_r)$	Left and right cameras' projection matrices
$\mathbf{K}$	Camera intrinsic matrix
(p, t)	pan and tilt settings of PTZ camera
$\mathbf{D}$	Disparitymap

between different overlapped images captured at different pan and tilt settings in case of each PTZ camera which can be obtained using SIFT.

A description of the symbols used in this chapter is given in Table 1. Let $(\mathbf{I}_l^1, \mathbf{I}_r^1)$ be a pair of images of a far scene and $\mathbf{H}^1$ be the homography obtained from the SIFT based matching points between these two images. Let $(\mathbf{I}_l^n, \mathbf{I}_r^n)$ be a pair of images of a scene/object near to the cameras along their optical axis. The problem is to autonomously establish the correspondence between the images $(\mathbf{I}_l^n, \mathbf{I}_r^n)$. Such a correspondence can be established by capturing n images between the scenes those are in $\mathbf{I}_l^1$ and $\mathbf{I}_l^n$ from the left PTZ camera. A similar image grabbing process is required for right PTZ camera to capture the images between the scenes those are in $\mathbf{I}_r^1$ and $\mathbf{I}_r^n$. Let these two sets of images be $(\mathbf{I}_l^1, \mathbf{I}_l^2, \ldots, \mathbf{I}_l^n)$ and $(\mathbf{I}_r^1, \mathbf{I}_r^2, \ldots, \mathbf{I}_r^n)$. The required correspondence between the images $(\mathbf{I}_l^n, \mathbf{I}_r^n)$ in terms of a homography $\mathbf{H}^n$ can be achieved in the following steps:

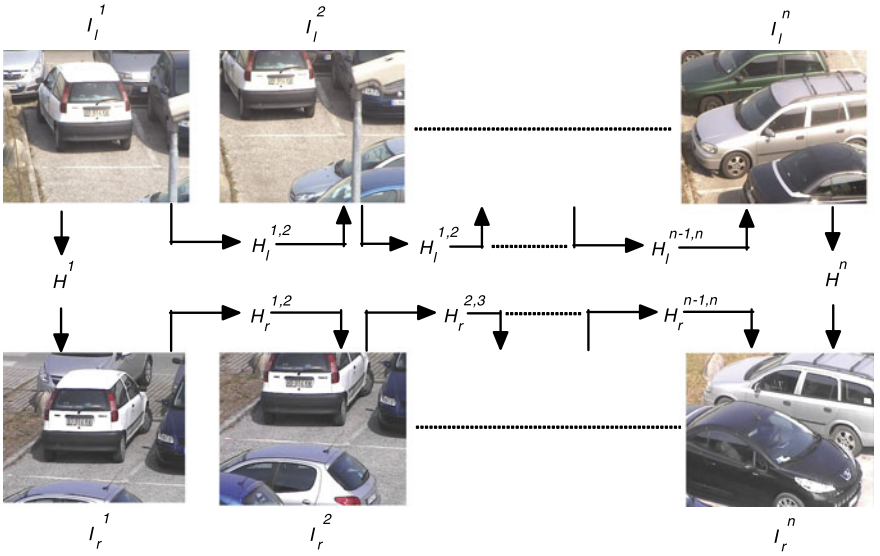


Fig. 3 Wide baseline stereo matching using a chain of homographic matrices

1. Establish the correspondence between image pairs $(\mathbf{I}_l^1, \mathbf{I}_l^2), (\mathbf{I}_l^2, \mathbf{I}_l^3), \ldots, (\mathbf{I}_l^{n-1}, \mathbf{I}_l^n)$ in terms of their respective homographies $\mathbf{H}_l^{1,2}, \mathbf{H}_l^{2,3}, \ldots, \mathbf{H}_l^{n-1,n}$ such that $\mathbf{I}_l^i = \mathbf{H}_l^{i,i+1}\mathbf{I}_l^{i+1}$ for $i = 1, \ldots, n-1$.
2. Repeat the procedure given in the above step to compute $\mathbf{H}_r^{1,2}, \mathbf{H}_r^{2,3}, \ldots, \mathbf{H}_r^{n-1,n}$ for the images captured with the right camera.
3. Compute the homographies $\mathbf{H}_l$ and $\mathbf{H}_r$ as

$$\mathbf{H}_l = \prod_{i=0}^{n-2} \mathbf{H}_l^{n-(i+1),n-i}, \qquad \mathbf{H}_r = \prod_{i=0}^{n-2} \mathbf{H}_r^{n-(i+1),n-i}. \tag{1}$$

4. Compute the required homography matrix $\mathbf{H}^n$ for the pair the images $\mathbf{I}_l^n$ and $\mathbf{I}_r^n$ as

$$\mathbf{H}^n = \mathbf{H}_r \mathbf{H}^1 (\mathbf{H}_l)^{-1}. \tag{2}$$

The homography $\mathbf{H}^n$ can be used to establish correspondence between the images $\mathbf{I}_l^n$ and $\mathbf{I}_r^n$ which is not easy to obtain directly in case of wide baseline stereo systems. Figure 3 gives an intuitive interpretation of the above described procedure. The final homography matrix $\mathbf{H}^n$ can be computed for any value of n; however, the above procedure can accumulate errors in the final homography due to the multiplication of several matrices. In order to minimize this error: (1) we keep the sampling step (i.e., the difference in pan and tilt angles) as low as possible, with a constraint that any pair of images (e.g., $\mathbf{I}_{l/r}^i, \mathbf{I}_{l/r}^{i+1}$) has to share at least 30 % of the FOV; (2) outliers from matching points should be removed before applying a robust approach for the homography estimation.

4 Offline Steps

It is necessary to perform an offline initialization for deriving all the information necessary to determine the rectification transformations during online operations. This includes the computation of the rectification transformations for image pairs captured at different pan and tilt sampling from two PTZ cameras. The rotation parameters related to these rectification transformations are stored in the LUT corresponding to the respective pan and tilt values of the PTZ cameras. The LUT data is used for the training of a set of neural networks that are used for sigmoid interpolation of these transformations in real time.

4.1 Computation of Rectification Transformations and Look-Up Table

A rectification transformation is a linear one-to-one transformation of the projective plane, which is represented by a 3×3 non-singular matrix. For a pair of stereo images $\mathbf{I}_l$ and $\mathbf{I}_r$, the rectification can be expressed as:

$$\mathbf{J}_l = \mathbf{A}_l \mathbf{I}_l, \qquad \mathbf{J}_r = \mathbf{A}_r \mathbf{I}_r$$

where, $(\mathbf{J}_l, \mathbf{J}_r)$ are the rectified images and $(\mathbf{A}_l, \mathbf{A}_r)$ are the rectification matrices. In case of uncalibrated cameras based stereo system, a quasi epipolar rectification [5] has been proposed for computing these rectification transformations by minimizing the following function.

$$\sum_i \left[\left(\mathbf{m}_l^i\right)^T \mathbf{A}_r^T \mathbf{F}_\infty \mathbf{A}_l \mathbf{m}_l^i \right]^2 \tag{3}$$

where, $(\mathbf{m}_l, \mathbf{m}_r)$ are pairs of matching points between images $\mathbf{I}_l$ and $\mathbf{I}_r$. $\mathbf{F}_\infty$ is the fundamental matrix for the rectified pair of images. Generally, the minimization of (3) is time-consuming and therefore it is not easy to compute the rectification transformations in real time. Here, we use this scheme [5] for computing rectification transformations offline for the image pairs captured at different pan and tilt sampling. In real time, this information can be used for computing rectification transformations for a given pan and tilt setting by using sigmoid interpolation. Recently, in [11] such an interpolation based method is adopted to make rectification of stereo pairs in real time. An offline LUT containing rectification matrices corresponding to various image pairs captured at predefined pan and tilt angles is constructed. Then, the rectification transformations can be interpolated in real-time for any arbitrary orientation of both PTZ cameras by using LUT data. However, the interpolation of eighteen parameters (nine elements for each rectification transformation) is again computationally expensive. Here, our effort is to reduce the number of these interpolated parameters into six instead of eighteen by using some suitable assumptions of camera projection matrix.

The meaning of stereo image rectification is that for any given pair of the original camera projection matrices $\mathbf{P}_l$ and $\mathbf{P}_r$, two new virtual projection matrices $\widehat{\mathbf{P}}_l$ and $\widehat{\mathbf{P}}_r$ can be obtained to rotate the cameras around their optical centers until the focal

planes become coplanar. Therefore, the rectification transformations $\mathbf{A}_l$ and $\mathbf{A}_r$ can be decomposed as

$$\mathbf{A}_l = \widehat{\mathbf{P}}_l \mathbf{P}_l^{-1}, \qquad \mathbf{A}_r = \widehat{\mathbf{P}}_r \mathbf{P}_r^{-1}. \tag{4}$$

A camera matrix $\mathbf{P}$ can be decomposed into the intrinsic and extrinsic matrices $\mathbf{P} = \mathbf{KD}$, where, $\mathbf{K}$ is the intrinsic matrix and $\mathbf{D} = [\mathbf{R}\ \mathbf{t}]$ denotes the extrinsic matrix containing rotation matrix $\mathbf{R}$ and translation vector $\mathbf{t}$. Since there is no translation involved in the rectification process, (4) can be rewritten as

$$\mathbf{A}_l = \widehat{\mathbf{K}}_l \overline{\mathbf{R}}_l \mathbf{K}_l^{-1}, \qquad \mathbf{A}_r = \widehat{\mathbf{K}}_r \overline{\mathbf{R}}_r \mathbf{K}_r^{-1} \tag{5}$$

where, $\overline{\mathbf{R}}_l = \widehat{\mathbf{R}}_l \mathbf{R}_l^{-1}$ and $\overline{\mathbf{R}}_r = \widehat{\mathbf{R}}_r \mathbf{R}_r^{-1}$ are the rotation matrices involved in rectification process. Here, the original intrinsic parameter matrices $(\mathbf{K}_l, \mathbf{K}_r)$ and the rotation matrices $(\mathbf{R}_l, \mathbf{R}_r)$ are unknown, whereas the new intrinsic matrices $(\widehat{\mathbf{K}}_l, \widehat{\mathbf{K}}_r)$ can be set arbitrarily, provided that the focal lengths and the coordinates of the principal points must be equal. During the rectification process, the unknown intrinsic parameters can be reduced by considering the zero skew, square pixel and principal point in the center of the image assumptions. Then the intrinsic matrices can be written as:

$$\widehat{\mathbf{K}}_l = \begin{pmatrix} f_l & 0 & w/2 \\ 0 & f_l & h/2 \\ 0 & 0 & 1 \end{pmatrix}; \qquad \widehat{\mathbf{K}}_r = \begin{pmatrix} f_r & 0 & w/2 \\ 0 & f_r & h/2 \\ 0 & 0 & 1 \end{pmatrix} \tag{6}$$

where, w and h are the width and the height of the image. The focal lengths f_l and f_r can be computed directly by reading the zoom parameter of the two PTZ cameras. Thus, the problem of computing a pair of rectification transformations is converted into the computation of only two rotation matrices $(\overline{\mathbf{R}}_l, \overline{\mathbf{R}}_r)$. Hence, for any pan and tilt combination, only three rotation parameters has to be stored in the LUT instead of nine entries of a rectification transformation.

Thus, the main steps to construct the LUT are:

1. The overall monitoring wide-area is divided into a number of subarea in such a way that each subarea is covered in the FOV of each PTZ camera just by changing the pan and tilt angles setting $(p_l^i, t_l^j)_{i=1:1:n_p}^{j=1:1:n_t}$.
2. Capture $n_{\text{tot}} = (n_p \times n_t)^2$ pairs of images of all these local subarea with the two PTZ cameras at equal zoom.
3. Compute the possible k $(> n_{\text{tot}})$ pairs of rectification transformation $(\mathbf{A}_l^k, \mathbf{A}_r^k)$ for the different combination of stereo images. The used images pairs should have images sharing at least the 30 % of their FOV.
4. Decompose rectification transformations as per earlier described scheme and compute their corresponding rotation parameters.
5. Store the rotation parameters in a LUT as dependent variables corresponding to their four independent variables (p_l, t_l, p_r, t_r).

The main problem to be addressed in the creation of the LUT is the establishment of the correspondence between wide baseline stereo images. This has been solved by exploiting the earlier described chain of homographies based approach.

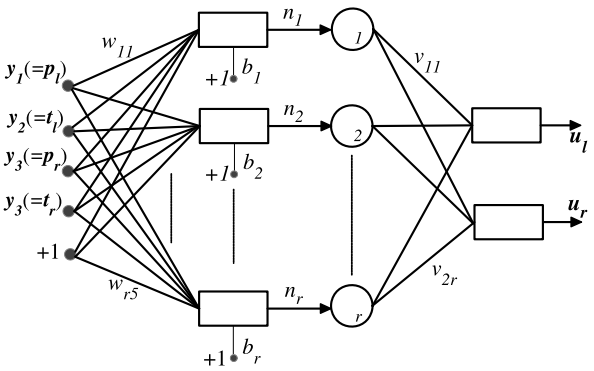


Fig. 4 Architecture of employed neural network

4.2 Training a Neural Network Using LUT

Sigmoid interpolation via a set of neural networks is used for computing the rectification transformations in real time corresponding to any arbitrary orientation of two PTZ cameras. The data stored in LUT is used to train the neural networks. The neural network based interpolation has been chosen due to its strong function approximation property with respect to highly non-linear data. A supervised learning scheme [12] using LUT data has been adopted for the off-line training of the neural networks.

The network considers the pan and tilt angles as input and returns the parameters of the rotation matrices corresponding to the required rectification transformations $(\mathbf{A}_l, \mathbf{A}_r)$ as output. The sets of input and output data are related by a non-linear mapping $\mathbf{U} = f(p_i, t_i)$. For a known set of input-output values, the problem is to find the function $F(\cdot)$ that approximates $f(\cdot)$ over all inputs. That is,

$$\left\| F(p,t) - f(p,t) \right\| < \varepsilon \qquad \text{for all } (p,t), \tag{7}$$

where, ε is a small error. The architecture of the proposed neural network is shown in Fig. 4, where two output nodes are corresponding to the angles for left and right rotation matrices. Three different networks are trained for yaw, pitch and roll elements of the rotation matrices. A detailed learning process for the proposed network is given in [12], where back-propagation algorithm is used with gradient information.

4.3 Zoom to Focal Length Fitting

As aforementioned, the proposed framework is based on a zoom compensation process in case of heterogeneous image-pairs [13]. The effect of this unequal zoom is compensated by using a focal ratio information which requires the focal lengths corresponding to both images. For a static camera, the focal length can be estimated offline once considering that the image parameters (specifically focal length) will remain constant for the whole process. In case of PTZ cameras, the focal length changes as the zoom level is changed to zoom in/out. Thus, the determination of

the accurate focal length associated to any acquired frame is a fundamental even though not an easy task. Moreover, if its computation is not precise enough, the rectification accuracy of the proposed algorithm could be significantly affected. To overcome such a problem, the focal length is computed in two steps: (a) offline fitting of focal lengths corresponding to zoom settings and (b) online estimation given a particular zoom level.

Concerning the first step, the aim is to find out a mapping between the zoom value and the corresponding focal length. For such a purpose, the whole zoom range is sampled and the focal length is estimated by using a calibration process for every sampled zoom tick. In the case of motorized lenses [22], the relation between a given zoom tick z and corresponding focal length f is

$$f(z) = \frac{a_0}{1 + a_1 z + a_2 z^2 + a_3 z^3 + \cdots + a_n z^n} \tag{8}$$

where, the order n and the unknown $a_0, \ldots, a_n$ are camera dependent. For the adopted camera, following the methodology in [22], the estimated optimal value of n is 2. Therefore, a_0, a_1 and a_2 can be estimated by minimizing the following nonlinear function

$$C(a) = \sum_{i=1}^{K} \left[f(z_i) - \frac{a_0}{1 + a_1 z + a_2 z^2} \right]^2. \tag{9}$$

However from (9), the estimation of the focal length is not reliable for small values of zoom, then (9) can be written as

$$C(b) = \sum_{i=1}^{K} \left[p(z_i) - \left(b_0 + b_1 z + b_2 z^2 \right) \right]^2 \tag{10}$$

where, $b_0 = 1/a_0$, $b_1 = a_1/a_0$, $b_2 = a_2/a_0$ and $p(z_i) = 1/f(z_i)$ denotes the lens power. The minimization of (10) is reliable for lower as well as higher zoom settings. The values b_0, b_1 and b_2 corresponding to the minimum value of $C(b)$ are chosen to define the optimal values of a_0, a_1 and a_2. In the real time, the focal length f for any given zoom level z is estimated as

$$f(z) = \frac{a_0}{1 + a_1 z + a_2 z^2}. \tag{11}$$

The above method has been tested on various zoom samples and it has been found reliable for estimating the focal length corresponding to a given zoom.

5 Online Steps

During tracking, stereo tasks can be performed by applying a zoom compensation followed by the rectification of the resulting images. This section contains a detailed description of these two steps.

Algorithm 1 Compensation of unequal zoom settings in PTZ stereo

Read (z_l, z_r)
$Calculate\{f_l, f_r\} = Interpolation(z_l, z_r)$
if $f_l = f_r$ **then**
  STOP
else if $f_l > f_r$ **then**
  $R = f_l / f_r$
  $I_l' = Shrink(I_l, R)$
  $I_l^h = Zeropad(I_l', Size\{I_r\})$ *and* $I_r^h = I_r$
else
  $I_r' = Shrink(I_r, 1/R)$
  $I_r^h = Zeropad(I_r', Size\{I_l\})$ *and* $I_l^h = I_l$
end if

5.1 Unequal Zoom Compensation

The proposed framework allows to operate with couples of PTZ camera acquiring images with different zoom levels. This introduces a heterogeneity between internal imaging parameters of both cameras. However, equivalent zoom values have been used for the two PTZ cameras during the construction of the LUT containing rectification transformations. Therefore, a compensation is required to deal with this heterogeneity with real time performance. A novel approach based on the focal lengths of the two cameras is used to tackle such heterogeneity. In a perspective projection model, the position of any pixel is always proportional to the focal length for the respective camera. Therefore, if the two images are acquired with different zoom levels, then this heterogeneity can be compensated by shrinking the higher zoom image with a focal ratio information.

Let $\mathbf{I}_l$ and $\mathbf{I}_r$ of size $w \times h$ be the two images captured at different zoom levels z_l and z_r from the dual PTZ cameras. Let the corresponding focal lengths be f_l and f_r obtained from earlier described scheme. The idea behind the process of heterogeneity compensation is achieved by shrinking the image having longest focal length by mean of a focal ratio. The overall compensation algorithm is given in Algorithm 1, where the function $Shrink(I_l, R)$ represents that the image I_l is shrunk by a factor of R. The function $Zeropad(I_l', Size\{I_r\})$ denotes that the zero padding is performed around image I until its size becomes equal to the size of I_r. The image pair (I_l^h, I_r^h) is homogeneous in terms of the intrinsic image parameters which is necessary to rectify the stereo images correctly.

5.2 Rectification of Images

Once the zoom compensation is completed, the new pair of images has to be rectified for further stereo processing. This operation can be achieved in the following steps:

1. Interpolate the parameters for generating rotation matrices $\overline{\mathbf{R}}_l^c$ and $\overline{\mathbf{R}}_r^c$ from the trained neural network by giving the pan and tilt angles as input for current pair of frames.
2. Calculate rectification transformations for current frames as

$$\mathbf{A}_l^c = \widehat{\mathbf{K}}_l^c \overline{\mathbf{R}}_l^c \left(\mathbf{K}_l^c\right)^{-1}, \qquad \mathbf{A}_r^c = \widehat{\mathbf{K}}_r^c \overline{\mathbf{R}}_r^c \left(\mathbf{K}_{or}^c\right)^{-1} \tag{12}$$

   where, $(\mathbf{K}_l^c, \mathbf{K}_r^c)$ and $(\widehat{\mathbf{K}}_l^c, \widehat{\mathbf{K}}_r^c)$ are the pairs of intrinsic matrices in the original and the rectified cameras' geometries.
3. Warp the current pair of frames as a rectified pair of images using $\mathbf{A}_l^c$ and $\mathbf{A}_r^c$.

$$\mathbf{J}_l^c = \mathbf{A}_l^c \mathbf{I}_l^c, \qquad \mathbf{J}_r^c = \mathbf{A}_r^c \mathbf{I}_r^c.$$

The above procedure is performed in real time. In this way, rectified pairs of frames can be obtained by using the orientation information of left and right PTZ cameras.

6 High Resolution Depth Map Estimation and Mosaic Construction

The application of the proposed system is given for scene understanding in the case of a large environment. Depth obtained from the stereo images can be a very crucial cue in scene understanding. In a scene having large variations in depths at various positions (like parking lot or a hill), it is necessary to use higher resolution images for obtaining depth map. In case of a flat region (like empty ground) where depths at different points have smooth variation, lower resolution images can be used to obtain the depth map. Such a multiresolution depth map based strategy is useful for establishing a trade off between accuracy and computational cost. Finally, the depth map of the whole environment can be obtained by making the mosaic of several overlapped and multiresolution depth maps.

6.1 Depth Map Estimation

A multi step process is proposed for selecting the optimal zoom values of the two cameras according to the earlier described strategy. When an event of interest has taken place in the FOV of any static camera (let say S_1), this camera delivers the information to the dual PTZ cameras for focusing on the region of interests. Let S_1 delivers the information to the PTZ cameras C_l and C_r. First, the initial resolutions for C_l and C_r is set in such a way that it covers the whole scene, that is, a low resolution more or less equivalent to the static camera. In the second step, the resolution for C_l is refined to acquire the selected region with maximum resolution. Then the disparity map is calculated between the high resolution image from left camera and

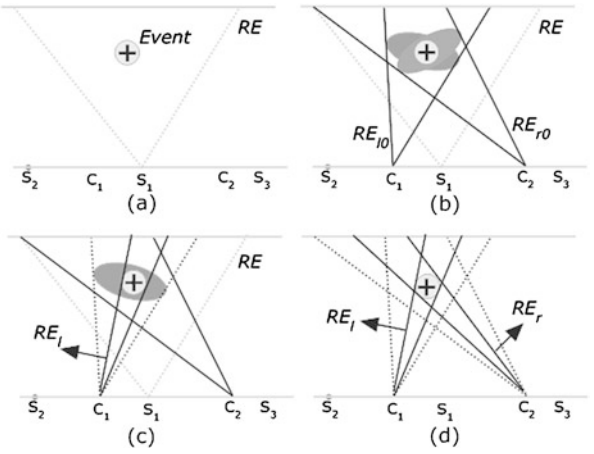


Fig. 5 Procedure for acquainting region of interests by PTZ cameras with maximal resolution

a low resolution image of right camera. The variation of depths are checked from the disparity map to classify the associated region as flat or complex. In the former case, the process is stopped and the computed disparity map is used to compute depths. In the latter case, further processing is required. The following steps are proposed to obtain high resolution depth maps:

1. Detect the event of interest in the static camera S_1 (see Fig. 5(a)).
2. Deliver the information to both PTZ cameras (C_l, C_r) system for focusing towards the regions of interest. Let the region visible in FOVs of these cameras initially be RE_l^0 and RE_r^0 and the corresponding images be $\mathbf{I}_l^0$ and $\mathbf{I}_r^0$ (see Fig. 5(b)).
3. Change the resolution of the left camera in such a way that the event of interest is acquired with best possible resolution for acquiring the image $\mathbf{I}_l$. Some priori information about the left camera are used to adopt the zoom setting for best possible resolution (see Fig. 5(c)). This can be done just by utilizing a Look-Up Table containing details about the various zoom settings and corresponding FOV details.
4. Compute the disparity map $\mathbf{D}^0$ (having disparities $d^0(x, y)$ for all (x, y)) from the images $\mathbf{I}_l$ and $\mathbf{I}_r^0$.
5. Check whether the disparities $d^0(x, y)$ have large variations for all (x, y) in the disparity map $\mathbf{D}^0$. If not, stop the algorithm and use $\mathbf{D}^0$ for computing its corresponding depth map $\widetilde{\mathbf{D}}$. Otherwise, proceed to the next step.
6. Compute an image $\mathbf{I}_r^c$ as $\mathbf{I}_r^c = \mathbf{I}_l + \mathbf{D}^0$.
7. Change the resolution of the right camera based on the image $\mathbf{I}_r^c$ and acquire a new image $\mathbf{I}_r$ with such a resolution (see Fig. 5(d)).
8. Find the higher resolution disparity map $\mathbf{D}$ from the images $\mathbf{I}_l$ and $\mathbf{I}_r$.
9. Compute the depth map $\widetilde{\mathbf{D}}$ from the disparity map $\mathbf{D}$.

Figure 5 provides a graphical representation of the process described in the above steps.

6.2 Construction of Depth Map Mosaic

In general, two approaches can be used to obtain a depth map mosaic of a large scene. The first approach [26] works by stitching the overlapped images for each camera separately to obtain the two stereo panoramic images and then performing the disparity estimation. The second way foresees to compute a depth map for each stereo image pair and then mosaic all the depth maps to construct the panoramic depth map. The main difficulty in the later one is the estimation of matching points between the depth maps of overlapped images, since it is very difficult to apply feature matching between depth maps. To cope this problem, we use the same transformation matrices which are used for stitching the images of left camera. However, the second approach has the following advantages when compared to the earlier one.

- Multiresolution depth cues can be easily maintained in final depth map mosaic.
- We obtain the final depth map mosaic (for a large region) by stitching several depth maps (of various small regions). In this context, the depth value for each pixel belonging to the overlapped regions in consecutive images is calculated by fusing two depth cues, so the robustness and accuracy of the final depth mosaic can be maintained.
- The final depth map can be updated anytime for a new image pair.

The use of disparity drift [24] compensates the uncertainty in the reading of pan, tilt and zoom parameters which is required for correct interpolation of rotation parameters associated with their corresponding rectification transformations. Assuming that there are n rectified pairs $(\mathbf{I}_l^i, \mathbf{I}_r^i)$ of stereo images captured at different pan, tilt and zoom settings. The following steps are adopted to construct the final depth map mosaic.

1. Perform stereo matching between all image pairs $(\mathbf{I}_l^i, \mathbf{I}_r^i)$, and obtain their corresponding disparity maps $\mathbf{D}^i$ for $1 = 1, 2, \ldots, n$.
2. Normalize the gray-level values between consecutive disparity maps. The process starts from the maps used to specify the reference panoramic image coordinate system. This process can be done by finding the linear regression parameters (α_i, β_i) between each consecutive pairs of disparity maps for all matching pixels (x_m, y_m).

$$\mathbf{D}^{(i+1)} = \alpha_i \mathbf{D}^i + \beta_i \tag{13}$$

   where, $i = 1, 2, \ldots, n-1$.
3. Calculate the disparity drift ρ^i for each disparity map $\mathbf{D}^i$.
4. Compute the modified disparity maps as

$$\mathbf{D}_r^i = \mathbf{D}^i + \rho^i \mathbf{I}_d$$

   where, $\mathbf{I}_d$ represents an identity matrix having the same size as the disparity map $\mathbf{D}$.
5. Compute the depth maps $\widetilde{\mathbf{D}}^i$ from their corresponding disparity maps $\mathbf{D}_r^i$.
6. Construct the depth map mosaic **DMM** by stitching all depth maps $\widetilde{\mathbf{D}}^i$ for $i = 1, 2, \ldots, n$, into the reference panoramic image coordinate system.

Fig. 6 SIFT 'x' and Chain of Homographies '+' based correspondence between wide baseline stereo images

Sometimes for a complex scene, a fusion of several depth cues is required for a better representation of the depths for scene understanding. A weighted average method as in [24] can be used for fusing several depth cues together.

7 Results and Discussions

For the experimental validation of the proposed framework, a network of static cameras composed by AXIS 221 network cameras has been adopted. For the stereo unit, two different PTZ cameras (i.e., Axis 213 and Axis 233D) are used. Three different types of experiments have been performed to: (1) evaluate the correspondence between stereo images captured with the two cameras placed far away from each other (i.e., wide baseline stereo), (2) evaluate the proposed interpolation based rectification algorithm for various pairs of stereo images having unequal zoom, (3) evaluate the computation of high resolution depth map mosaic for large scene understanding. Different criterions have been used for comparing the performance of the proposed framework in each case.

7.1 Correspondence Between Wide Baseline Stereo Images

To show the importance of the chain of homographies based matching algorithm in case of wide baseline stereo images, correspondence between a pair of images has been considered. First, a homography H^d has been computed by using the matching points extracted with SIFT method between this pair of images. Then, the homography has been evaluated by using the proposed chain of homographies based approach. To do this, a pair of stereo images has been captured of a far scene where SIFT can be implemented accurately. Then, two different chains of homographies have been computed using five different tilt positions in case of each cameras separately. Finally, the final homography H^n has been computed using (2). Corresponding points in the right image have been computed for 12 selected points in left image using the homographies (H^d and H^n). Figure 6 shows the results for this experiment and it can be observed that the corresponding points obtained from proposed chain of homographies based approach are accurate enough, while the corresponding points obtained from direct method are erroneous.

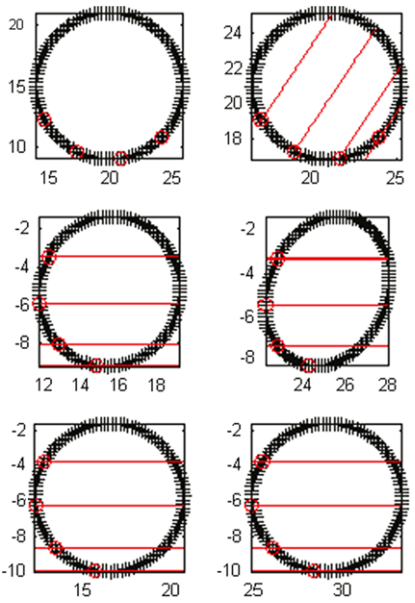


Fig. 7 Rectification of synthetic image pairs: original pair (*first row*); direct rectification (*middle row*); proposed rectification (*in bottom row*)

The unequal zoom settings between the two PTZ cameras produce a distortion error in the rectified images. The distorted images produce the error in the final stereo based 3D localization. To show the effectiveness of the adopted rectification strategy in case of non-homogeneous images, experimental study has been conducted for reconstructing 3D points generated synthetically. To do this, 120 points have been generated in a circular pattern having different coordinates. Later, all these points have been projected on two different planes with two different projection matrices with different intrinsic as well as extrinsic parameters (see Fig. 7). Different projected points in these two planes can be considered as images of two cameras differ with intrinsic parameters (non-homogeneous). These, two circles have been rectified using the traditional rectification algorithm (i.e., without unequal zoom compensation) as well as after using the presented zoom compensation algorithm. After rectification, by using the stereo triangulation on the rectified image planes, 3D points are reconstructed and compared with the original one. The error in reconstruction is reported in Table 2 in terms of different Statistics measures. The error in 3D reconstruction obtained with the rectified images without zoom compensation is increasing drastically, while, it is tolerable in case of rectified images after zoom compensation.

7.2 Depth Map

High resolution depth maps are estimated for a far and large scene in a complex environment. Figure 8 shows the depth map obtained for a building by adopting a coarse-to-fine strategy in two successive iterations. The top row represents the dis-

Table 2 Results for 3D reconstruction error for 120 synthetic points without and after zoom compensation. The focal ratio between two camera is calculated by fixing the projection matrix of left camera and varying it for right camera (see [13])

Focal ratio	Without zoom compensation		After zoom compensation	
	Mean error	Standard deviation	Mean error	Standard deviation
1.0	0.0	0.0	0.0	0.0
0.90	1.57	0.16	0.23	0.05
0.80	2.79	0.27	1.45	0.11
0.70	3.21	0.32	1.74	0.10
0.60	5.30	0.39	1.89	0.13
0.50	6.11	0.54	2.20	0.16

Fig. 8 Experiment for high resolution depth estimation in two successive iterations (*top to bottom*). In each row left camera image, right camera image and corresponding depth map image. Three chosen points (marked with '+') appear to be at different depths in the high resolution depth map

parity map results obtained from a pair of images captured at low resolution of both cameras, that is, in the first iteration of the process when both PTZ cameras are directed towards this region. Then the zoom level of left PTZ camera is selected with the proposed resolution strategy and a corresponding disparity map is obtained. From the obtained disparity map, it is found that the variation in depths are larger for the selected region of interests. In this context, the FOV of the right PTZ camera is refined to acquire high resolution image. The high resolution images for the selected region are given in the second row. Finally, the high resolution depth map (right most in the second row) is obtained with this pair of images. The higher zoom difference between these two pairs of images results in a better depth information for the selected region.

To judge the accuracy of high resolution depth map over the low resolution depth map, three points at different depths have been selected within the region of interest.

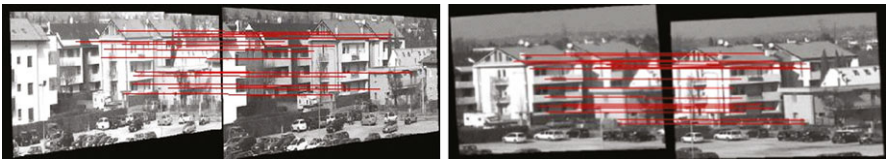


Fig. 9 Testing points with matching lines in two successive iterations (*left to right*)

Table 3 Comparison of depth estimation in two successive iterations

Parameters	Itr-1	Itr-2
Mean relative error (%)	3.29	0.86
Standard dev. error (%)	3.50	1.04
Relative depth uncertainty δ (m)	0.797	0.307
Disparity drift ρ (m)	−0.0088	−0.0064

In the initial depth map (low resolution), the points appear to have the same depths, while these three points appear to be at different depths in the high resolution depth map. To give a quantitative evaluation of the results obtained with the iterative procedure, twenty points with ground truth depths information have been selected. The matching is performed for finding corresponding points in all three rectified image pairs (see Fig. 9) with good pixel precision. Five points have been randomly selected to compute the disparity drift for both pairs of images. Let the calculated and ground truth disparities for these five points be d_j and d'_j, respectively. The disparity drift has been calculated as

$$\rho = \frac{1}{5}\sum_{i=1}^{5} \left|\left(d'_j - d_j\right)\right| \quad \text{for } j = 1, 2, \ldots, 5.$$

For the other 15 points, we have estimated the depths $\tilde{d}_j = f(b/(d_j + \rho))$ and computed the mean and standard deviation of the absolute differences between ground truth and estimated depths in the two successive iterations. Moreover, we have compared the depth uncertainty $\delta = \tilde{d'}^2(u/b)$ in support of our claim that the high resolution depth map is more accurate for the region having more depth variations. Here, $\tilde{d'}$ represents the average depth in a depth map image $\widetilde{\mathbf{D}}$ and u is the horizontal resolution of the rectified images. Table 3 shows the comparison results based on the above mentioned criterion. It is important to notice that all the measures are improving with further iterations, that is, the depth errors are very high in the first iteration while these reduce significantly in the final iteration. In the similar way, relative depth uncertainty and disparity drift iteratively improve. Finally, the depth mosaics from several low and high resolution depth maps have been generated. For this, 12 different pairs of images captured from both PTZ cameras at various zoom settings have been used. The zoom settings are automatically adapted by both cameras using the proposed scheme. All depth map images have been stitched in the

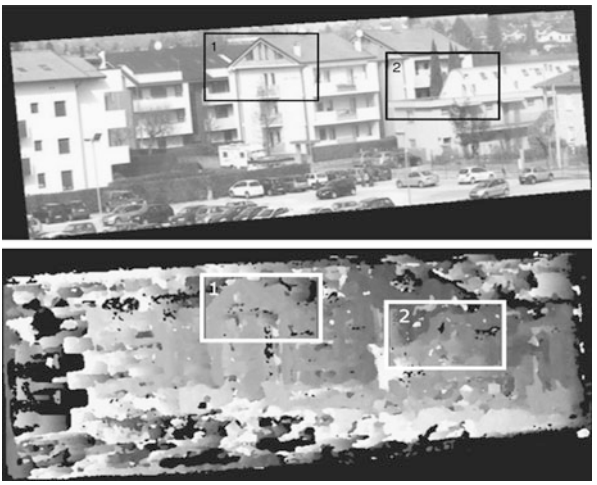


Fig. 10 Depth map mosaic: the mosaic of gray level images for reference camera (*top*) and the mosaic of corresponding depth map images

coordinate frame of a priori selected image. Perspective transformation is used to align such depth maps for generating the mosaic. Figure 10 shows the generated depth map mosaic, in which the gray value linearly reveals the magnitude of the depth value. The visual quality of the obtained depth map mosaic represents that the proposed method works well for a large and complex environment.

8 Conclusions

A dual PTZ camera based stereo system has been presented for video surveillance applications. First, a new real-time rectification algorithm has been proposed. The real-time rectification transformations have been achieved by interpolating the rotation parameters for given orientations of the PTZ cameras. A process for compensating the unequal zoom effects between the images of stereo pairs has been given to generate more accurate rectified images. The rectified frames have been used for constructing a depth map mosaic. The proposed framework is able to obtain high resolution depth maps for regions having larger variation in depths and low resolution depth maps for flatter regions. Moreover, this process requires only limited a priori information. In the near future, the proposed framework will be used to develop a multispectral stereo active system (using visible and thermal PTZ cameras). This will allow us to perform stereo tasks in environmental conditions (like foggy, rainy etc.), where visible cameras do not perform well.

Acknowledgements One of the authors, Sanjeev Kumar is thankful to IIT Roorkee for providing partially financial support through grant number FIG-100550 under New Faculty Initiation Grant Scheme-A to carry out this research work. The authors are also thankful to the AVIRES Lab, University of Udine, Italy for providing access to their image grabbing system.

References

1. Abidi, B., Koschan, A., Kang, S., Mitckes, M., Abidi, M.: Automatic target acquisition and tracking with cooperative static and PTZ video cameras. In: Multisensors Surveillance Systems: The Fusion Perspective, pp. 43–59. Kluwer Academic, Dordrecht (2003)
2. Brown, M., Burschka, D., Hager, G.: Advances in computational stereo. IEEE Trans. Pattern Anal. Mach. Intell. **25**(8), 993–1008 (2003)
3. Chen, C.H., Yao, Y., Page, D., Abidi, B., Koschan, A., Abidi, M.: Heterogeneous fusion of omnidirectional and PTZ cameras for multiple object tracking. IEEE Trans. Circuits Syst. Video Technol. **18**(8), 1052–1063 (2008)
4. Foresti, G., Micheloni, C., Piciarelli, C.: Detecting moving people in video streams. Pattern Recognit. Lett. **26**, 2232–2243 (2005)
5. Fusiello, A., Israra, L.: Quasi epipolar unclaibrated rectification. In: IEEE Int. Conf. on Image Processing (ICPR), pp. 1–4 (2008)
6. Gallup, D., Frahm, J., Mordobhai, P., Pollefeys, M.: Variable baseline/resolution stereo. In: IEEE Int Conf. on Computer Vision and Pattern Recognition (CVPR), pp. 1–8 (2008)
7. Haritaoglu, S., Harwood, D., Davis, L.: Real-time surveillance of people and their activities. IEEE Trans. Pattern Anal. Mach. Intell. **22**(8), 809–830 (2000)
8. Hart, J., Scassellati, B., Zucker, S.: Epipolar geometry for humanoid robotic heads. In: International Cognitive Vision Workshop, pp. 24–36 (2008)
9. Jain, A., Kopell, D., Kakligian, K., Wang, Y.: Using stationary-dynamic camera assemblies for wide-area video surveillance and selective attention. In: IEEE Int. Conf. of Computer Vision and Pattern Recogination (CVPR), pp. 537–544 (2006)
10. Kannala, J.T., Salo, M., Heikkila, J.: Algorithms for computing a planar homography from conics in correspondence. In: Proceedings of the British Machine Vision Conference (BMVC 2006), pp. 9.1–9.10 (2006)
11. Kumar, S., Micheloni, C., Piciarelli, C.: Stereo localization using dual PTZ cameras. In: Computer Analysis of Images and Patterns. LNCS, pp. 1061–1069. Springer, Berlin (2009)
12. Kumar, S., Micheloni, C., Piciarelli, C., Foresti, G.: Stereo localization based on network's uncalibrated camera pairs. In: Sixth IEEE International Conference on Advanced Video and Signal Based Surveillance, pp. 502–507 (2009)
13. Kumar, S., Micheloni, C., Piciarelli, C., Forestia, G.: Stereo rectification of uncalibrated and heterogeneous images. Pattern Recognit. Lett. **31**, 1445–1452 (2010)
14. Lowe, D.: Distinctive image features from scale-invariant keypoints. Int. J. Comput. Vis. **2**(60), 91–160 (2004)
15. Meltzer, J., Soatto, S.: Edge descriptors for robust wide-baseline correspondence. In: IEEE Int. Conf. on Computer Vision and Pattern Recogination, pp. 1–8 (2008)
16. Micheloni, C., Foresti, G., Snidaro, L.: A network of co-operative cameras for visual surveillance. IEE Proc., Vis. Image Signal Process. **152**(2), 205–212 (2005)
17. Micheloni, C., Lestuzzi, M., Foresti, G.: Adaptive video communication for an intelligent distributed system: tuning sensors parameters for surveillance purposes. Mach. Vis. Appl. **19**(5–6), 1432–1769 (2008)
18. Morris, B., Trivedi, M.: A survey of vision-based trajectory learning and analysis for surveillance. IEEE Trans. Circuits Syst. Video Technol. **18**(8), 1114–1127 (2008)
19. Piciarelli, C., Foresti, G.: On-line trajectory clustering for anomalous events detection. Pattern Recognit. Lett. **27**, 1835–1842 (2006)
20. Piciarelli, C., Micheloni, C., Foresti, G.: Trajectory-based anomalous event detection. IEEE Trans. Circuits Syst. Video Technol. **18**(11), 1544–1554 (2008)
21. Qureshi, F., Terzopoulos, D.: Planning ahead for PTZ camera assignment and handoff. In: Third ACM/IEEE International Conf. on Distributed Smart Cameras (ICDSC'09), Como, Italy (2009)
22. Trajkovic, M.: Interactive calibration of a PTZ camera for surveillance applications. In: Asian Conference on Computer Vision (ACCV) (2002)

23. Wan, D., Zhaou, J.: Stereo vision using two PTZ cameras. Comput. Vis. Image Underst. **112**(2), 184–194 (2008)
24. Wan, D., Zhou, J.: Multiresolution and wide-scope depth estimation using a dual-PTZ-camera system. IEEE Trans. Image Process. **18**(3), 677–682 (2009)
25. Yang, J., Arifb, O., Velab, P., Teizerc, J., Shia, Z.: Tracking multiple workers on construction sites using video cameras. Adv. Eng. Inform. **24**, 428–434 (2010)
26. Zhu, Z., Hanson, A.R.: Mosaic-based 3D scene representation and rendering. Signal Process. Image Commun. **21**(9), 739–754 (2006)

Performance Evaluation in Video-Surveillance Systems: The EventVideo Project Evaluation Protocols

Juan C. SanMiguel, Álvaro García-Martín, and José M. Martínez

Abstract During recent years, automatic video-surveillance systems have experienced a great development driven by the growing need for security. Many approaches exist whose performance is not clear for a large variety of available scenarios. To precisely identify which ones operate better for each scenario, empirical performance evaluation has been widely used for determining their strengths and weaknesses through their results. This approach requires defining two aspects (usually named as the evaluation protocol): the dataset (representative sequences) and the metrics (performance estimators). Common empirical approaches use metrics based on ground-truth data that define an ideal result, but there are also some novel approaches that do not require such data. Furthermore, the existence of several metrics and the growing availability of video data increase the complexity of the protocol design as well as require us to automate the whole evaluation process. In this chapter, considering the main analysis stages of a typical video-surveillance system (video object segmentation, people detection, video object tracking and event recognition), we introduce their evaluation protocols within the scope of the EventVideo project.

1 Introduction

During recent years, automatic video-surveillance systems have experienced a great development driven by the need for security in private and public places. Many approaches are available whose effectiveness is not clear [10]. They deal with a huge variety of environments that might change over time (e.g., lighting conditions) or present a substantial difference (e.g., sunny or rainy day). Hence, the performance

J.C. SanMiguel (✉) · Á. García-Martín · J.M. Martínez
Video Processing and Understanding Lab, Escuela Politécnica Superior, Universidad Autónoma de Madrid, Madrid, Spain
e-mail: Juancarlos.SanMiguel@uam.es

Á. García-Martín
e-mail: Alvaro.Garcia@uam.es

J.M. Martínez
e-mail: JoseM.Martinez@uam.es

P.K. Atrey et al. (eds.), *Intelligent Multimedia Surveillance*,
DOI 10.1007/978-3-642-41512-8_9,

of such systems can degrade significantly in these scenarios [17]. As these systems are composed of several analysis stages [35], a performance analysis for each one is required before examining the entire system. To precisely identify which approaches operate better in certain scenarios, performance evaluation has been proposed in the literature as a way to determine their strengths and weaknesses. The widely used empirical approach is based on evaluation through the analysis of the obtained results. For such analysis, two components have to be specified: the dataset (a set of sequences covering the situations that the algorithm might face being large enough to represent real world conditions) and the metrics (which allow us to quantify the performance of algorithms or systems). These two aspects are also known as the evaluation protocol [4, 22]. Traditional performance evaluation approaches use metrics based on ground-truth data that represents a manual annotation of the ideal result. The generation of ground-truth is usually a time consuming task and, therefore, limits the dataset size. Although there are other approaches not focused on ground-truth data [30, 38], most of the current literature assumes the availability of such data. Furthermore, the existence of several metrics increases the complexity of designing an evaluation protocol. Another point to be taken into account is the increasing quantity of video data available, which generates a new need to automate and optimize the whole evaluation process. In this chapter, we present the evaluation protocols (dataset and metrics) for the main analysis stages that compose a typical video-surveillance system (video object segmentation, people detection, video object tracking and event recognition) within the scope of the EventVideo project.[1]

The remainder of this chapter is organized as follows. First, the selected stages and evaluation scenarios of the EventVideo project are described in Sect. 2. Then, the related work on performance evaluation is discussed in Sect. 3. After that, Sect. 4 presents the evaluation protocols of the EventVideo project. Finally, Sect. 5 summarizes the chapter with some conclusions and future work.

2 Evaluation Scenarios

The EventVideo project considers the most common analysis stages of video-surveillance systems and evaluates them under different scenarios. In this section, we describe these stages and the classification criteria for the scenarios.

2.1 Selected Analysis Stages

The stages that compose a typical video-surveillance system are (see Fig. 1) [35]:

- *Video object segmentation*: extracts the foreground objects by applying analysis steps to the video sequence such as foreground analysis [5] and shadow removal [27]. Its output is a binary mask indicating the foreground objects.

[1] http://www-vpu.eps.uam.es/eventvideo/

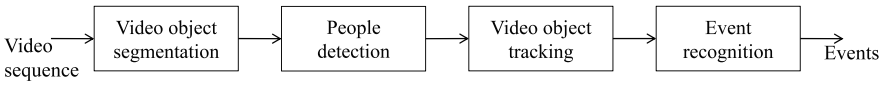


Fig. 1 Typical processing chain for a video-surveillance system

Table 1 Proposed classification for the evaluation scenarios

Scenario	Complexity	Density
S1	Low	Low
S2	High	Low
S3	Low	High
S4	High	High

- *People detection*: assigns a confidence of being people for each candidate region (that could either a frame region or a blob extracted from the foreground binary mask) by computing their similarity with a trained person model [13]. Its outputs are the score (confidence) and location of each analyzed candidate.
- *Video object tracking*: consists on locating the objects of interest (i.e., targets) in the sequence frames [22]. Its output is the location of each tracked target.
- *Event recognition*: detects events using the output of the previous stages [1]. An event is defined as an action performed by one or multiple persons (e.g., walking, handshaking). For each detection, the output includes a descriptor with its spatio-temporal location (frame span and position) and score (detection confidence).

2.2 Scenario Classification

For each stage of the video-surveillance system, the evaluation process should consider different scenarios to appropriately represent real world conditions. For understanding the limitations of current approaches, each scenario is classified according to two criteria: complexity and density. The former describes whether the visual data represents situations that can be easily characterized or not. For example, video object segmentation is an (relatively) easy task for static cameras and scene backgrounds but its complexity highly increases when dealing with moving cameras or motion in the background. The latter considers the number of moving objects in the sequence. Independently of the stage, an increasing number of objects affects its performance. This criterion is particularly interesting in video-surveillance where crowded places are common scenarios (e.g., airports, mass sport events). For example, abandoned object detection presents variable difficulty depending on the moving people density (fewer people, less complexity). Finally, we consider two levels for each criterion (low and high) to define four evaluation scenarios (see Table 1). Sample frames of the evaluation scenarios are depicted in Fig. 2.

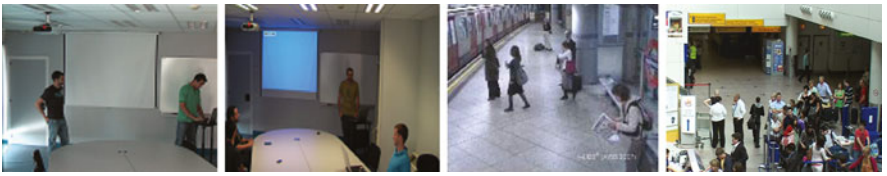


Fig. 2 Sample frames of the evaluation scenarios for event recognition. (*From left to right*): simple event *standing* (S1), complex event *UseObject* (S2), *abandonedObject* detection (S3) and complex event *bag stealing* (S4)

3 Related Work

In this section, we briefly review the state of art for each selected stage with respect to its datasets (see Table 2) and metrics, the two components of evaluation protocol.

3.1 Video Object Segmentation

Video object segmentation also known as foreground/background detection is a critical task in video-surveillance that presents many challenges related with, among others, shadows, camouflage, static objects and background motion [5]. For evaluating the existing approaches under such conditions, several datasets are available:

- VSSN2006:[2] provided within the VSSN Workshop 2006, this dataset consists of 14 sequences with artificial foreground objects introduced into real backgrounds for representing illumination changes, shadows and background motion (ground-truth data is provided for 10 sequences at pixel-level for every frame).
- IPPR06:[3] the IPPR contest motion segmentation dataset includes three different sequences of walking persons (with ground-truth at pixel-level for every frame) that model shadows, illumination changes and image noise.
- CVSG:[4] this dataset [34] consists of 14 sequences that represent the critical segmentation factors for foreground (appearance, size, velocity) and background (appearance, motion, multimodality) by artificially combining real foreground objects and backgrounds (with ground-truth at pixel-level for every frame).
- SABS:[5] this dataset [5] is an artificial dataset that represents nine common challenges of background subtraction for video-surveillance. It consists on nine sequences with isolated challenges which are divided into training and test data (with ground-truth at pixel-level for every frame).

[2] http://imagelab.ing.unimore.it/vssn06/

[3] http://media.ee.ntu.edu.tw/Archer_contest/

[4] http://www-vpu.eps.uam.es/DS/CVSG/

[5] http://www.vis.uni-stuttgart.de/index.php?id=sabs

Table 2 Categorization of existing datasets according to the scenarios of Table 1

	Covered scenario			
	S1	S2	S3	S4
Video object segmentation				
VSSN2006	X	X		
IPPR06	X			
CVSG	X	X		
SABS	X	X		
CDW2012	X	X		
People detection				
ETHZ			X	
TUD-Pedestrians			X	
DCII			X	
Caltech Pedestrian			X	
PDds	X	X	X	
Video object tracking				
PETS	X	X	X	
VISOR	X			
EPFL	X	X		
SOVTds	X		X	
Event detection				
CAVIAR	X	X		
ETISEO	X	X	X	
PETS 2006	X	X		
PETS 2007	X	X		X
I-LIDS	X		X	
VISOR	X	X		
CANDELA	X	X		
CANTATA	X			
ASODds	X	X		
EDds	X		X	

- CDW2012:[6] the IEEE Workshop on Change Detection 2012 proposed a rigorous benchmarking effort for representing well-known segmentation challenges captured in indoor and outdoor settings. In total, it has 31 sequences grouped into six categories (with ground-truth at pixel-level for every frame).

For ground-truth based metrics, video object segmentation can be evaluated at the lowest semantic level, that is, pixel-level, or at higher semantic levels, that is, region-

[6] http://www.changedetection.net

level, object-level, etc. In the literature, the pixel-level evaluation strategy is the most popular [5, 18]. It considers foreground detection as a binary classification of each pixel, resulting in a segmentation mask. The accuracy of this classification is expressed by means of recall (R), precision (P) and their harmonic mean, the F-score (F):

$$P = TP/(TP + FP), \tag{1}$$

$$R = TP/(TP + FN), \tag{2}$$

$$F = 2 \cdot P \cdot R/(P + R), \tag{3}$$

where TP, FP and FN indicate, respectively, the number of correct detections, false alarms and missed detections at pixel-level. For high-level evaluation, [7] used the center of the segmented objects whereas [24] focused on the splits and merges of foreground regions for composing the objects. In addition, [8] introduced spatio-temporal metrics derived from geometrical properties of the segmented objects.

Although non ground-truth based metrics are less popular, according to [29], they can be roughly classified into region (study the segmented regions), model (use available object models) or assisted (use complementary algorithms). Among them, the most relevant is [14] that defined the motion and color contrast along the boundaries of object regions and its adaptation for video object segmentation [29].

3.2 People Detection

The complexity of people detection is mainly related with the difficulty of modeling persons because of their huge variability in appearance, poses, movements, points of views and object-person interactions. This complexity is even higher in crowded video-surveillance scenarios which often include multiple persons, occlusions and background variability. Several datasets are available for its evaluation:

- ETHZ:[7] this dataset [15] consists of four stereo-sequences recorded in a real street walking scenario. For each one, it provides the sequences for both cameras, the camera calibration, the precomputed depth maps using the stereo images, and the ground-truth annotations (at bounding box level).
- TUD-Pedestrians:[8] this dataset [2] consists of 250 images (311 fully visible people) and two complex sequences (highly overlapped people showing significant variation in clothing and articulation), including the bounding box ground-truth.
- DCII:[9] the Daimler Mono Pedestrian Detection Benchmark Data Set II [13] consist of a sequence captured from a moving vehicle in a 27-minute drive through urban traffic and its associated ground-truth at bounding box level.

[7] http://www.vision.ee.ethz.ch/~aess/iccv2007/

[8] http://www.d2.mpi-inf.mpg.de/andriluka_cvpr08

[9] http://www.gavrila.net/

- Caltech Pedestrian Dataset:[10] this dataset [11] consists of approximately 10 hours of video (∼250000 frames divided into clips of 135 minutes) taken from a vehicle driving in an urban environment. In total, around 350000 bounding boxes and 2300 unique pedestrians were annotated. The annotation includes temporal correspondence between bounding boxes and detailed occlusion labels.
- PDds:[11] the PDds corpus [16] consists of 90 sequences for evaluation in video-surveillance covering the most common challenges with variable complexity. For each person, ground-truth is provided for each frame at bounding box level.

Regarding the metrics, people detection performance can be evaluated using ground-truth data at two levels: sequence sub-unit (frame, window, etc) or global sequence. Sub-unit performance is usually measured in terms of Detection Error Tradeoff (DET) [9, 12] or Receiver Operating Characteristics (ROC) [13, 23] curves. Global sequence performance is estimated through Precision-Recall (PR) curves [2, 21, 37]. The first level gives information of the classification stage, while the second one provides the overall system performance. In both cases the detector's output is a confidence score for each person detection, where larger values indicate higher confidence. Both evaluation methods compute progressively the respective parameters such as the number of false positives, Recall rate or Precision rate iterating from the lowest possible score to the highest possible score. Each score threshold iteration provides a point on the curve. On one hand, ROC curves represent the fraction of matched annotations with the detections (true positive rate, *TPR*, Recall or Sensitivity) vs. the fraction of wrong detections out of the negatives (non-people image samples) (false positive rate, *FPR* or 1-Specificity). On the other hand, PR curves represent also the *TPR* but in this case vs. the proportion of positive detections that are true positives (positive predictive value, *PPV* or Precision).

3.3 Video Object Tracking

Video object tracking is a complicated task due to high variability of the data to analyze as well as the many steps involved in the tracking process (feature extraction, target representation and propagation of the target model over time). For evaluating performance of tracking algorithms, several datasets are available:

- PETS:[12] the PETS Workshop series have been releasing a tracking-related dataset almost every year since 2000. As the dataset sizes are large and they cover real situations, these datasets are widely used in the research community. Among the existing datasets, the most important ones related to tracking are the PETS2000 (outdoor people and vehicle tracking for single camera), PETS2001 (outdoor people and vehicle tracking for single camera using two synchronized views) and

[10] http://www.vision.caltech.edu/Image_Datasets/CaltechPedestrians/

[11] http://www-vpu.eps.uam.es/DS/PDds/

[12] http://www.cvg.cs.rdg.ac.uk/slides/pets.html

PETS2009 (outdoor people tracking in crowded environments with multicamera setup). However, PETS datasets have two limitations: there is no ground-truth available and the challenges proposed are focused on event recognition (i.e., without describing the specific tracking problems for each video).

- VISOR:[13] this video repository has been conceived as a support tool for different video-surveillance projects [36]. Related to tracking, it includes six sequences (without ground-truth data) covering common problems such as occlusions, scale changes and complex movements.
- EPFL:[14] this dataset is oriented to multicamera settings for outdoor and indoor video-surveillance. It contains five scenarios with around 30 sequences showing occlusions and scale changes. Although camera calibration is provided for all the scenarios, ground-truth data is only available for some sequences.
- SOVTds:[15] this dataset is provides an extensive coverage of the common tracking-related problems in video-surveillance. For each problem, it is designed with four complexity levels including both real and synthetic sequences carefully selected from other datasets (related and non-related with video tracking). It contains 125 sequences and the associated ground-truth for every frame.

For video object tracking evaluation, metrics based on ground-truth can be divided into frame or sequence level. Frame-level considers the information within the frame being similar to an estimation of classification performance. Hence, standard Precision and Recall (Eqs. 1 and 2) are used for computing the spatial similarity between estimations and ground-truth locations of targets at pixel [25] or object-level [3]. Sequence-level measures the accuracy of the target trajectories such as the temporal accumulation of frame-level pixel accuracy [25] or the trajectory fragmentation [19] (i.e., the number of generated segments).

Approaches for tracking evaluation without ground-truth can be grouped into trajectory-based, feature-based and hybrid categories [30]. Trajectory-based approaches analyze the generated trajectories in which the time-reversibility of object motion is commonly used [38]. Feature-based approaches analyze target feature variation [30] or compute statistics for checking model consistency such as the covariance of the target state [26]. Finally, hybrid category describes the combinations of the previous approaches such as the use of the time-reversibility and the covariance analysis [32].

3.4 Event Recognition

As event recognition considers all the outputs of the stages that compose the video-surveillance system and therefore, its performance is influenced by all the factors affecting each stage. For evaluating its performance, several datasets are available:

[13] http://www.openvisor.org/

[14] http://cvlab.epfl.ch/data/pom/

[15] http://www-vpu.eps.uam.es/DS/SOVTds

- CAVIAR:[16] this dataset includes 17 sequences of human activities for indoor video-surveillance. It covers several events (with ground-truth data) such as people walking alone, meeting with others, window shopping, entering and exiting shops, fighting and passing out and leaving a package in a public place.
- ETISEO:[17] this dataset [25] contains 86 indoor and outdoor video-surveillance sequences (corridors, streets, building entries, subway, ...) with different types of complexity levels. Several events are annotated considering person-object interactions as well as person movement.
- PETS 2006:[18] this dataset is focused on multicamera sequences for *abandoned luggage* detection with increasing scene complexity in terms of nearby people. It contains 28 sequences (~1–2 minutes long) with 24 annotated events.
- PETS 2007:[19] this dataset considers the events *loitering*, *stolen luggage* and *abandoned luggage* in a crowded scenario. A four-camera setting is employed to record, 32 sequences (~2–3 minutes long) containing 36 events in total.
- I-LIDS:[20] this dataset has three sequences (~3.5 minutes long) for abandoned object detection at an underground station classified into three complexity levels (easy, medium, and hard), which are defined considering the crowd density.
- ViSOR:[21] this dataset is classified in different categories including outdoor and indoor events (human actions, traffic monitoring, cast shadows, ...). A total of 140 sequences with variable length is available for events related with human-object interactions (*abandoned object*, *Leave car, Enter Car,* ...).
- CANDELA:[22] this dataset contains 16 indoor sequences (~30 secs long) for *abandoned object*, including interactions between object owners. Despite the simplicity of the scenario, the low resolution and the relatively small size of objects present challenges for detecting the events.
- CANTATA:[23] this dataset is focused on *abandoned* and *stolen object*s in non-crowded outdoor scenarios. A total of 31 sequences (~2 minutes long) are available from two different views (leaving and removing objects in the sequences).
- ASODds:[24] this dataset provide a representative test-set for discriminating previously detected stationary regions in video-surveillance systems able to detect abandoned and stolen objects. Annotations of both events are also provided. Sequences (over 100) have been extracted from related public datasets.
- EDds:[25] this dataset contains 17 sequences (~3–4 minutes long) focused on human-related events for indoor video-surveillance considering interactions be-

[16] http://homepages.inf.ed.ac.uk/rbf/CAVIAR

[17] http://www-sop.inria.fr/orion/ETISEO/intro_presentation.htm

[18] http://www.cvg.rdg.ac.uk/PETS2006/data.html

[19] http://www.cvg.rdg.ac.uk/PETS2007/data.html

[20] http://www.eecs.qmul.ac.uk/~andrea/avss2007_d.html

[21] http://www.openvisor.org/

[22] http://www.multitel.be/~va/candela/abandon.html

[23] http://www.multitel.be/~va/cantata/LeftObject/

[24] http://www-vpu.eps.uam.es/DS/ASODds

[25] http://www-vpu.eps.uam.es/DS/EDds

Table 3 Critical factors in video object segmentation specified in the CVSG dataset

Foreground		Background	Camera
Single objects	Groups		
Textural complexity, apparent velocity, object structure, uncovered extent, object size	Largest difference, object interactions	Textural complexity, multimodality	Motion

tween persons and environmental objects and activities without involving physical contact. In particular, two activities (*HandUp* and *Walking*) and three person-object interactions (*Leave*, *Get* and *Use object*) have been annotated.

For event recognition, the common evaluation scheme is to optimally determine the match between ground-truth annotations the event detections. This one-to-one mapping can be done temporally or spatio-temporally [25]. The former only considers the duration of the detection and the annotation whereas the latter extends it by including a constraint for similar spatial locations. Moreover, an additional constraint can be imposed considering the confidence of the detected event [31].

4 Evaluation Protocols

In this section, we introduce the proposed protocols for performance evaluation of the selected video-surveillance stages within the scope of the EventVideo project.

4.1 Video Object Segmentation

4.1.1 Selected Dataset

For this stage, the Chroma Video Segmentation Ground-truth (CVSG) dataset [34] is selected as it covers the main problems of video object segmentation. It consists of a set of video sequences obtained according to a thorough study of the critical factors affecting segmentation performance (summarized in Table 3). As specific values of these factors can significantly increase or decrease the complexity of the segmentation task (and therefore, the expected algorithm accuracy), they are convenient for designing multiple sequences with variable complexity. Foreground objects have been recorded in a chroma studio, in order to automatically obtain pixel-level high quality segmentation masks with different foreground factors. Then, real scene backgrounds are also recorded with different camera and background factors. Finally, the resulting corpus consists on the composition of the foreground and background sequences obtaining a total of 14 sequences ($\sim$7000 frames). Some examples are shown in Fig. 3. As it can be observed, they present low density scenarios with variable complexity thus covering the S1–S2 scenarios defined in Table 1.

Fig. 3 Sample frames for the sequences of the CVSG dataset

4.1.2 Metrics Based on Ground-Truth Data

As a first approach, we have selected the pixel-wise evaluation based on ground-truth data [18]. In order to evaluate and compare the segmentation techniques, we have selected the precision and recall measures for foreground ($P1$, $R1$) and background ($P0$, $R0$) detection:

$$P0 = TN/(TN + FN), \qquad R0 = TN/(TN + FP), \tag{4}$$

$$P1 = TP/(TP + FP), \qquad R1 = TP/(TP + FN), \tag{5}$$

where TP indicates the number of foreground pixels correctly detected, TN the number of background ones correctly detected, FP the number of foreground pixels wrongly detected as background and FN the number of background ones wrongly detected as foreground. Additionally, the F-Score measure has been selected to combine P and R measures for foreground ($F1$) and background ($F0$) results:

$$F0 = 2 \cdot P0 \cdot R0/(P0 + R0), \tag{6}$$

$$F1 = 2 \cdot P1 \cdot R1/(P1 + R1). \tag{7}$$

In order to achieve the objective of evaluating and finding the optimal parameters of the algorithms, it have been maximized the average of the F-score measures for foreground and background, $F0$ and $F1$.

4.1.3 Metrics not Based on Ground-Truth Data

We also evaluate segmentation performance without ground-truth data by means of the color-based metric $DC1$ proposed by [29]. It relies on comparing the boundaries

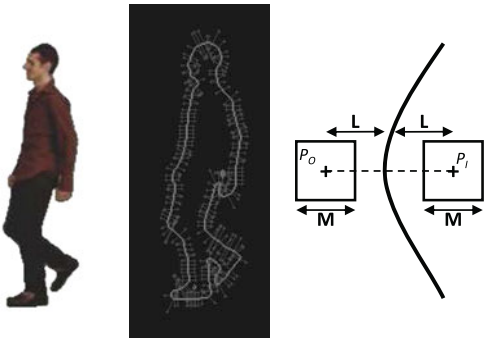


Fig. 4 Boundary-based contrast scheme proposed by [29]. (**a**) Segmented object, (**b**) its boundary with the normal lines and (**c**) a zoom on a boundary pixel location

of the segmented objects against the color boundaries extracted from each frame. The scheme is depicted in Fig. 4. For each boundary pixel, a normal line of length $2L+1$ is defined and the color differences between the initial (P_I) and ending (P_O) points of this line are obtained in a $M \times M$ patch as follows:

$$CD(t;i) = \frac{\|P_O^i(t) - P_I^i(t)\|}{\sqrt{3 \cdot 255^2}}, \tag{8}$$

where $P_O^i(t)$ and $P_I^i(t)$ are the mean colors of the $M \times M$ patches centered at P_I and P_O points (using the RGB color space quantified into 256 levels) extracted from each ith boundary pixel of the foreground region at time t. $CD(t;i)$ ranges from 0 to 1 if both points belong to, respectively, the same or different color regions.

Then, the evaluation of the foreground segmentation for each region, O_j is performed and combined for multiple foreground regions as follows:

$$DC1_{O_j}(j) = \frac{1}{K_t} \sum_{i=1}^{K_t} CD(t;i,j), \tag{9}$$

$$DC1(t) = \min_j \left(DC1_{O_j}(t)\right), \tag{10}$$

where K_t is the number of boundary pixels, $CD(t;i,j)$ is the color difference of the ith boundary pixel of the jth analyzed foreground region. Its value ranges from 0 (lowest segmentation quality) to 1 (highest segmentation quality). Finally, the mean of $DC1(t)$ is taken over all the sequence frames to get an evaluation score.

4.2 People Detection

4.2.1 Selected Dataset

For this stage, the Person Detection dataset (PDds) [16] is selected as it covers the main problems affecting people detection in video-surveillance. It consists of a set of sequences with different levels of complexity and their associated ground-truth

Table 4 Critical factors in people detection

Background		Classification	
Textural complexity	Variability	Appearance variability	People-object interactions
Low, medium, high	Lighting changes, view changes, multimodal	Pose variations, different clothes, carry objects	Objects, people, objects & people

Table 5 Description of the PDds dataset and their associated critical factors

Sequence	Category	Subcategory	Background		Classification	
			Textural complexity	Variability	Appearance variability	People/object interactions
1–4	C1	C1-a	Low	Low	Low	Low
5–6	C1	C1-b	Low	Medium	Low	Low
7–8	C2	C2-a	Low	Low	Medium	Low
9–10	C2	C2-b	Low	Low	Medium	Medium
11–12	C2	C2-c	Low	Medium	Low	Medium
13	C3	C3-a	Medium	Medium	Medium	Low
14–16	C3	C3-b	Medium	Medium	Medium	Medium
17–18	C4	C4-a	Low	Low	Medium	High
19–20	C4	C4-b	Low	Low	High	Medium
21	C4	C4-c	Low	Low	High	High
22–24	C5	C5-a	Medium	High	Medium	High
25	C5	C5-b	Medium	High	High	Medium
26	C5	C5-c	High	High	Medium	High
27–33	C5	C5-d	High	High	High	Low
34–65	C5	C5-e	High	High	High	Medium
66–90	C5	C5-f	High	High	High	High

(bounding box annotations for each frame). Sequences have been classified into different complexity categories depending on previously identified critical factors for people detection performance. Table 4 summarizes such factors and Table 5 lists the video sequences and their complexity. Sample frames are shown in Fig. 5. The resulting corpus contains 91 sequences (∼28000 frames) exceeding other public pedestrian datasets in the amount of data and its complexity variability. As it can be observed, they present low density scenarios with variable complexity thus covering the S1, S2 and S3 scenarios defined in Table 1.

4.2.2 Metrics Based on Ground-Truth Data

For evaluating people detection performance based on ground-truth, we aim to compare the overall performance of different detection systems, so we have chosen the

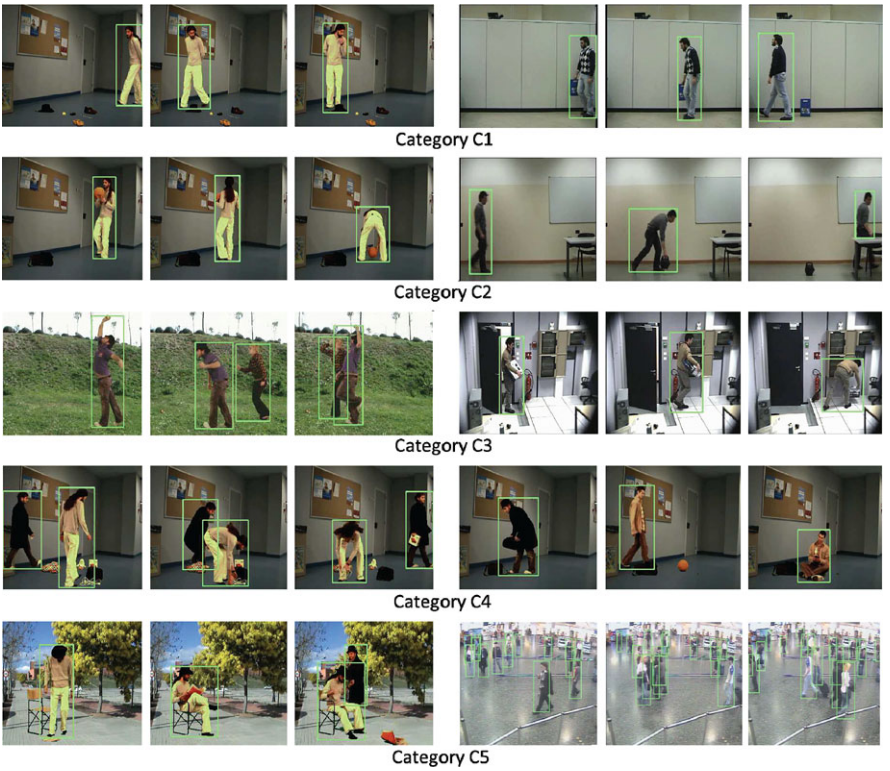
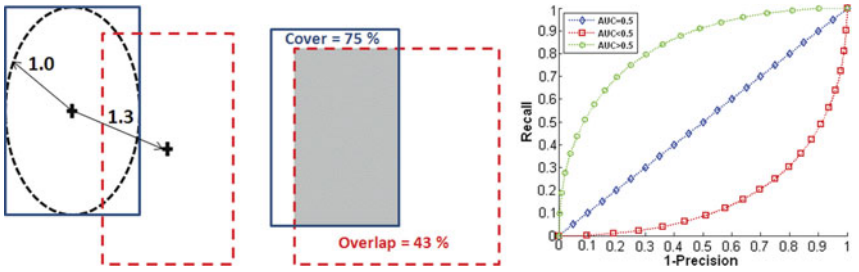


Fig. 5 Sample frames for the categories of the PDds dataset

PR evaluation method (see Sect. 3.2). For each value of the detection confidence, PR curves compute Precision and Recall as shown in Eqs. 1 and 2.

In order to evaluate not only the (binary) yes/no detection but also the precise pedestrians locations and extents, we use three criteria, defined by [20], that allow comparing hypotheses at different scales: the relative distance, cover, and overlap. The relative distance dr measures the distance between the bounding box centers in relation to the size of the annotated bounding box (see Fig. 6a). Cover and overlap measure how much of the annotated bounding box is covered by the detection hypothesis and vice versa (see Fig. 6b). A detection is considered true if $dr \leq 0.5$ (corresponding to a deviation up to 25 % of the true object size) and cover and overlap are both above 50 %. Only one hypothesis per object is accepted as correct, so any additional hypothesis on the same object is considered as a false positive.

We usually use the integrated Average Precision (AP) to summarize the overall performance, represented geometrically as the area under the PR curve (AUC-PR), in order to express more clearly the results we have chosen the representation Recall vs 1-Precision (see Fig. 6c). In addition, focusing on the people detection evaluation in video security systems, we want also to evaluate the detector at the operating point, that is, at the predefined optimal decision threshold for each algorithm. Thus,



(a) Relative distance criterion for comparing bounding boxes.

(b) Cover and overlap criteria for comparing bounding boxes.

(c) (1-Precision)-Recall curve and area under the curve.

Fig. 6 Performance evaluation metrics for people detection

Table 6 Complexity factors for the video tracking dataset

Problem	Criteria (factors)
Complex movement	The target changes its speed (pixels/frame) abruptly in consecutive frames
Gradual illumination	The average intensity of an area changes gradually with time until a maximum intensity difference is reached
Abrupt illumination	The average intensity of an area changes abruptly with respect to its surroundings (maximum intensity difference)
Noise	It includes natural (snow) or white Gaussian noise which is manually added with varying deviation value
Occlusion	Objects in the scene occlude a percentage of the target
Scale changes	The target changes its size with a maximum relative change regarding its original size
Similar objects	An object with similar color to the target appears in the neighborhood of the target

we can compare the final operational performance and not just its overall performance.

4.3 Video Object Tracking

4.3.1 Selected Dataset

For this stage, the Single Object Video Tracking dataset (SOVTds) is selected to evaluate single-object tracking algorithms for video-surveillance. SOVTds covers seven common tracking problems in video-surveillance by identifying its critical factors (see Table 6). Then, it organizes the sequences into four situations: synthetic, real laboratory, simple real and complex real data. For the first two situations, the

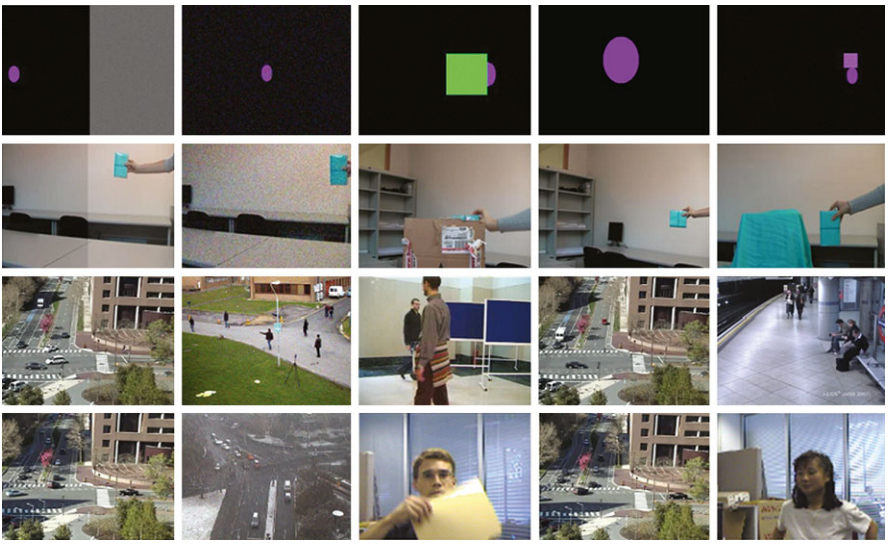


Fig. 7 Sample frames for the situations of the proposed dataset (*from top row to bottom row*): synthetic, laboratory, Simple real and Complex real. Samples of some tracking problems are also presented for each column (*from left to right*): abrupt illumination change, noise, occlusion, scale change and (color-based) similar objects

sequences were recorded trying to isolate the tracking problems whereas the last two situations contain carefully selected clips from existing datasets. In total, the corpus has 125 sequences (∼23000 frames). Sample frames are shown in Fig. 7. Moreover, the complexity of the tracking problems is estimated for each sequence through the factors. As this dataset represents simple and complex problems in nonhighly crowded situations, it covers the S1, S2 and S3 scenarios defined in Table 1.

4.3.2 Metrics Based on Ground-Truth Data

In order to evaluate the tracking accuracy, the SFDA (Sequence Frame Detection Accuracy) metric was chosen which calculates for each frame the spatial overlap between the estimated target location and the ground-truth annotation.

$$SDFA = \frac{\sum_{t=1}^{N_{\text{frames}}} FDA(t)}{\sum_{t=1}^{N_{\text{frames}}} \exists(N_{GT}^t + N_P^t)} \tag{11}$$

$$FDA(t) = \frac{OverlapRatio}{\frac{N_{GT}^t + N_P^t}{2}} \tag{12}$$

where N_{frames} is the number of frames, N_{GT}^t and N_P^t represent the number of ground-truth and estimated locations in the th frame, $\exists(\cdot)$ indicates if ground-truth or estimation data exist for the th frame and *OverlapRatio* is the pixel-level spatial overlap between both locations divided by their area sum.

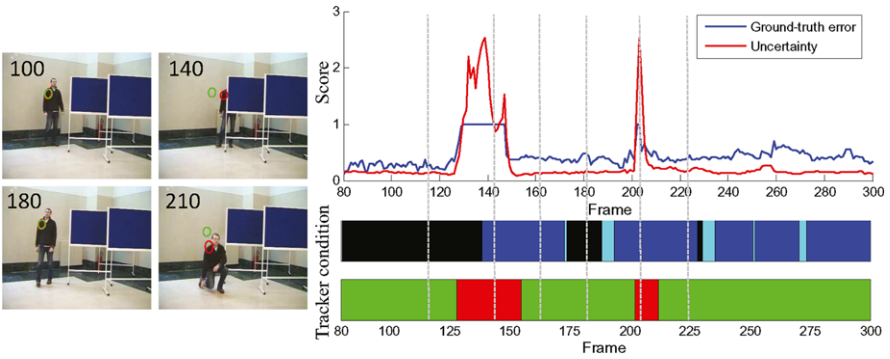


Fig. 8 Tracking results, tracker condition estimation and temporal segmentation for target H5 (occlusion_1 sequence; frames shown are 100, 140, 180 and 210) [32]. Tracking results and ground-truth annotations are represented as *green* and *red* ellipses, respectively. (*Green*: successful tracking; *Red*: unsuccessful tracking; *Black*: scanning; *Cyan*: locking in; *Blue*: locked on.)

4.3.3 Metrics not Based on Ground-Truth Data

For estimating tracking performance without ground-truth data, we use [32] which is based on estimating the uncertainty of the tracking algorithm (i.e., tracker) and then, analyzing its values to decide whether it is successful or not. Such uncertainty, S_t, can be used as indicator of periods of unstable output data (e.g., wrong target estimation) allowing the tracker evaluation. It can be measured by analyzing the state-space representation of particle-filter based approaches [22] or by adapting the output of deterministic trackers such as for Mean-shift tracking [33].

Then, we identify when the tracker is stable (i.e., following the target) by detecting changes of S_t within a window of length λ. We compute two relative variations of uncertainty for the change of $S_{t-\lambda}$ with respect to S_t and vice versa, using two lengths for short and long term changes (λ_1 and λ_2) as defined in [32]. The former change indicates low-to-high uncertainty changes whereas the latter represents high-to-low uncertainty changes. As a result, four signals are computed by combining the two variations and the two lengths. Then, changes on the four signals are detected by using a three-threshold scheme and combined in a finite-state machine for estimating the tracker condition: focused on the target, scanning the video frame or locking on the target after a failure [32]. Finally, we use time-reversed analysis to check the tracker recovery when it focuses on an object after failure (transition from third to first tracker condition) as it might be on a distractor (background objects with features similar to those of the target). A tracker in reverse direction from this recovery instant until a reference point (the last time instant when the tracker was successful) [32] and the spatial overlap between the reverse and the forward trackers (the one to evaluate) is computed for determining if the tracker has recovered or not. Figure 8 shows an example of tracker condition and successful estimation.

Category 1 Category 2 Category 3

Fig. 9 Sample frames for the available categories in the ASODds dataset

Table 7 ASODds dataset description

Category	Number of annotations (blobs)				Complexity
	Annotated sequences		Real sequences		
	Abandoned	Stolen	Abandoned	Stolen	
C1	771	442	756	863	Low
C2	666	316	794	397	Medium
C3	595	174	852	660	High
All	2032	932	2402	1920	

4.4 Event Detection

4.4.1 Selected Datasets

For event detection, two datasets have been selected: the Abandoned and Stolen Discrimination dataset (ASODds) and the Event Detection dataset (EDds).

Abandoned and Stolen Object Discrimination Dataset—ASODds The ASODds dataset [6] consists of two annotation sets of the foreground binary masks for abandoned and stolen objects. The first one has been obtained by manually annotating the objects of interest in the video sequence (annotated data). The second one represents real data has been obtained by running [28] over the sequences to get inaccurate masks (real data). Then, the sequences have been grouped into three categories according to a subjective estimation of the background complexity that consists on the presence of edges, multiple textures, lighting changes, reflections, shadows and objects belonging to the background. Currently, three categories have been defined considering low (C1), medium (C2) and high (C3) background complexity. According to the criteria proposed in Sect. 2, the categories C1 and C2 present low complexity and few number of objects (situation S1) whereas the C3 covers low complex and crowded scenarios (situation S3). Sample frames of such categories are shown in Fig. 9 and a summary of the annotated events in the dataset and the associated complexity of each category is available in Table 7.

Table 8 EDds dataset description. The complexity estimation codes are Low (L), Medium (M), High (H) and Very High (V). The events are Leave-object (LEA), Get-object (GET), Use-object (USE), Hand Up (HUP) and Walking (WLK)

Sc1	Events occurrences					Complexity estimation			
	Iterations			Activities		S1	S2	S3	S4
	LEA	GET	USE	HUP	WLK				
1	18	13	9	9	54	M	L	M	M
2	7	7	10	14	44	M	M	M	H
3	14	14	22	20	10	V	H	V	V

Category 1 Category 2 Category 3

Fig. 10 Available categories in the EDds dataset

Event Detection Dataset—EDds Currently, the dataset EDds [31] contains 17 sequences recorded using a stationary camera at resolution of 320×240 at 12 fps. It is focused on two types of human-related events: interactions and activities. In particular, two activities (*HandUp* and *Walking*) and three human-object interactions (*Leave*, *Get* and *Use object*) have been annotated. Moreover, all the test sequences have been grouped into three categories according to a subjective estimation of the analysis complexity according to the criteria defined in the previous subsections for the foreground, tracking, feature and event stages that compose a typical event detection system. A summary of the annotated events in the dataset and the associated complexity of each category is available in the Table 8. Sample frames of such categories are shown in the Fig. 10.

4.4.2 Metrics Based on Ground-Truth Data

For matching event annotations and detections, we use the following conditions:

$$Match\left(E^{GT}, E^{D}\right) = \begin{cases} 1 & \text{if } score > \rho \quad \wedge \\ & |T^{D}_{\text{start}} - T^{GT}_{\text{start}}| < \tau_1 \quad \wedge \\ & |T^{D}_{\text{end}} - T^{GT}_{\text{end}}| < \tau_2 \quad \wedge \\ & \frac{2|A^{GT} \cap A^{D}|}{|A^{GT}|+|A^{D}|} > \sigma \\ 0 & \text{otherwise} \end{cases} \tag{13}$$

Table 9 Classification of datasets according to criteria defined in Sect. 2.2. The (–) indicates that the dataset partially fulfills the requirements of such criterion

		Density	
		Low	High
Complexity	Low	CVSG, PDds, SOVTds, ASODds, EDds	PDds (–), ASODds (–)
	High	PDds (–), SOVTds (–), EDds (–)	

where E^{GT} and E^{D} are the annotated and detected events; score is the detection probability; (T^{D}_{start}; T^{D}_{end}) and (T^{GT}_{start}; T^{GT}_{end}) are the frame intervals of the annotated (GT) and detected (D) events; A^{GT} and A^{D} represent the average area (in pixels) of each event; $|A^{GT} \cap A^{D}|$ is their average spatial overlap (in pixels); ρ, τ_1, τ_2 and σ are positive thresholds (heuristically set to the values $\rho = 0.75$, $\tau_1 = \tau_2 = 100$, and $\sigma = 0.5$).

Then, we use the Precision (P) and Recall (R) measures for evaluating the performance of the matching process. Precision is the ratio between the correct and the total number of detections. Recall is the ratio between the correct detections and the total number of annotations. We also use the F-score measure, β, to combine Precision and Recall as shown in Eqs. 1 and 2.

5 Conclusions

In this chapter, we have presented the material for performance evaluation within the EventVideo project. In particular, we have selected some stages: video object segmentation, people detection, video object tracking and event detection. Then, we have described the employed datasets and protocols for their evaluation in Sect. 4 (CVSG, PDds, SOVTds, ASODds y EDds; all of them available at http://www-vpu.eps.uam.es/webvpu/en/recursos-publicos/datasets/).

In addition, a novel methodology that does not follow the traditional ground-truth based approach has been presented in Sects. 4.1.3 and 4.3.3 for, respectively, the video object segmentation and tracking stages. Moreover, according to the scenario classification of Sect. 2.2 (with the variables complexity and density), the datasets used in the EventVideo project are categorized as listed in Table 9.

As future work, the selected datasets will be used for comparing the most recent approaches for evaluating the current status of the state-of-the-art (and which of the criteria in Table 9 could be considered as achieved). Moreover, we will consider the extension of the datasets to cover the highest levels of the defined situations and the inclusion of additional information to help visual analysis (such as depth and laser).

Acknowledgements Work supported by the Spanish Government under Project TEC2011-25995 EventVideo.

References

1. Aggarwal, J., Ryoo, M.: Human activity analysis: a review. ACM Comput. Surv. **43**(3), 16:1–16:43 (2011)
2. Andriluka, M., Roth, S., Schiele, B.: People-tracking-by-detection and people-detection-by-tracking. In: IEEE Int. Conf. on Computer Vision and Pattern Recognition, Anchorage, Alaska, USA, pp. 1–8 (2008)
3. Bashir, F., Porikli, F.: Performance evaluation of object detection and tracking systems. In: Proc. IEEE Int. Workshop Perform. Eval. Track. Surveill, New York, USA, pp. 7–14 (2006)
4. Baumann, A., Boltz, M., Ebling, J., Koenig, M., Loos, H.S., Merkel, M., Niem, W., Warzelham, J.K., Yu, J.: A review and comparison of measures for automatic video surveillance systems. EURASIP J. Image Video Process., 1–30 (2008)
5. Brutzer, S., Heferlin, B., Heidemann, G.: Evaluation of background subtraction techniques for video surveillance. In: Proc. of IEEE Conf. on Computer Vision and Pattern Recognition, Colorado Springs, CO, USA, pp. 1937–1944 (2011)
6. Campos, L.C., SanMiguel, J.C., Martínez, J.M.: Discrimination of abandoned and stolen object based on active contours. In: Proc. of the IEEE Int. Conf. on Advanced Video and Signal based Surveillance, Klagenfurt, Austria, pp. 101–106 (2011)
7. Charles, J.J., Kuncheva, L.I., Wells, B., Lim, I.S.: An evaluation measure of image segmentation based on object centres. In: Proc. of the International Conference on Image Analysis and Recognition, pp. 283–294 (2006)
8. Correia, P., Pereira, F.: Objective evaluation of video segmentation quality. IEEE Trans. Image Process. **12**(2), 186–200 (2003)
9. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: IEEE Int. Conf. on Computer Vision and Pattern Recognition, San Diego, CA, USA, pp. 886–893 (2005)
10. Dee, H., Velastin, S.: How close are we to solving the problem of automated visual surveillance? Mach. Vis. Appl. **19**(5), 329–343 (2008)
11. Dollar, P., Wojek, C., Schiele, B., Perona, P.: Pedestrian detection: a benchmark. In: IEEE Int. Conf. on Computer Vision and Pattern Recognition, Miami, FL, USA, pp. 304–311 (2009)
12. Dollar, P., Wojek, C., Schiele, B., Perona, P.: Pedestrian detection: an evaluation of the state of the art. IEEE Trans. Pattern Anal. Mach. Intell. **34**(4), 743–761 (2012)
13. Enzweiler, M., Gavrila, D.M.: Monocular pedestrian detection: survey and experiments. IEEE Trans. Pattern Anal. Mach. Intell. **31**(12), 2179–2195 (2009)
14. Erdem, C., Sankur, E., Tekalp, A.: Performance measures for video object segmentation and tracking. IEEE Trans. Image Process. **13**(7), 937–951 (2004)
15. Ess, A., Leibe, B., Gool, L.V.: Depth and appearance for mobile scene analysis. In: IEEE Int. Conf. on Computer Vision, Rio de Janeiro, Brazil, pp. 1–8 (2007)
16. García-Martín, A., Martínez, J.M., Bescós, J.: A corpus for benchmarking of people detection algorithms. Pattern Recognit. Lett. **33**(2), 152–156 (2012)
17. Greoris, B., Bremond, F., Thonnat, M.: Real-time control of video surveillance systems with program supervision techniques. Mach. Vis. Appl. **18**(3), 185–205 (2007)
18. Herrero, S., Bescós, J.: Background subtraction techniques: systematic evaluation and comparative analysis. In: Proc. of the Advanced Concepts for Intelligent Vision Systems, Bordeaux, France, pp. 33–42 (2009)
19. Kasturi, R., Goldgof, D., Soundararajan, P., Manohar, V., Garofolo, J., Bowers, R., Boonstra, M., Korzhova, V., Zhang, J.: Framework for performance evaluation of face, text, and vehicle detection and tracking in video: data, metrics, and protocol. IEEE Trans. Pattern Anal. Mach. Intell. **31**(2), 319–336 (2009)
20. Leibe, B., Seemann, E., Schiele, B.: Pedestrian detection in crowded scenes. In: IEEE Int. Conf. on Computer Vision and Pattern Recognition, San Diego, CA, USA, pp. 878–885 (2005)
21. Leibe, B., Leonardis, A., Schiele, B.: Robust object detection with interleaved categorization and segmentation. Int. J. Comput. Vis. **77**(1–3), 259–289 (2008)

22. Maggio, E., Cavallaro, A.: Video Tracking: Theory and Practice. Wiley, New York (2011)
23. Munder, S., Gavrila, D.M.: An experimental study on pedestrian classification. IEEE Trans. Pattern Anal. Mach. Intell. **28**(11), 1863–1868 (2006)
24. Nascimento, J., Marques, J.: Performance evaluation of object detection algorithms for video surveillance. IEEE Trans. Multimed. **8**(4), 761–774 (2006)
25. Nghiem, A.-T., Bremond, F., Thonnat, M., Valentin, V.: Etiseo, performance evaluation for video surveillance systems. In: Proc. of the IEEE Int. Conf. on Advanced Video and Signal based Surveillance, London, UK (2007)
26. Nickels, K., Hutch, S.: Estimating uncertainty in ssd-based feature tracking. Image Vis. Comput. **20**(1), 47–58 (2002)
27. Prati, A., Mikic, I., Trivedi, M., Cucchiara, R.: Detecting moving shadows: algorithms and evaluation. IEEE Trans. Pattern Anal. Mach. Intell. **25**(7), 918–923 (2003)
28. SanMiguel, J.C., Martínez, J.M.: Robust unattended and stolen object detection by fusing simple algorithms. In: Proc. of IEEE Int. Conf. on Advanced Video and Signal based Surveillance, Santa Fe, USA, pp. 18–25 (2008)
29. SanMiguel, J.C., Martínez, J.M.: On the evaluation of background subtraction algorithms without ground-truth. In: Proc. of the IEEE Int. Conf. on Advanced Video and Signal Based Surveillance, Boston, USA, pp. 180–187 (2010)
30. SanMiguel, J.C., Cavallaro, A., Martínez, J.M.: Evaluation of on-line quality estimators for object tracking. In: Proc. of the IEEE Int. Conf. on Image Processing, Hong Kong, China, pp. 825–828 (2010)
31. SanMiguel, J.C., Escudero-Viñolo, M., Martínez, J.M., Bescós, J.: Real-time single-view video event recognition in controlled environments. In: Proc. of the Int. Workshop on Content-Based Multimedia Indexing, Madrid, Spain, pp. 91–96 (2011)
32. SanMiguel, J.C., Cavallaro, A., Martínez, J.M.: Adaptive online performance evaluation of video trackers. IEEE Trans. Image Process. **21**(5), 2812–2823 (2012)
33. SanMiguel, J.C., Cavallaro, A., Martínez, J.M.: Standalone evaluation of deterministic video tracking. In: IEEE Int. Conference on Image Processing, Orlando, FL, USA, pp. 1353–1356 (2012)
34. Tiburzi, F., Escudero, M., Bescós, J., Martínez, J.M.: A ground-truth for motion-based video-object segmentation. In: Proc. of IEEE Int. Conf. on Image Processing, San Diego, USA, pp. 17–20 (2008)
35. Valera, M., Velastin, S.: Intelligent distributed surveillance systems: a review. IEE Proc., Vis. Image Signal Process. **152**(2), 192–204 (2005)
36. Vezzani, R., Cucchiara, R.: Video surveillance online repository (ViSOR): an integrated framework. Multimed. Tools Appl. **50**(2), 359–380 (2010)
37. Wojek, C., Walk, S., Schiele, B.: Multi-cue onboard pedestrian detection. In: IEEE Int. Conf. on Computer Vision and Pattern Recognition, Miami, FL, USA, pp. 794–801 (2009)
38. Wu, H., Sankaranarayanan, A., Chellappa, R.: Online empirical evaluation of tracking algorithms. IEEE Trans. Pattern Anal. Mach. Intell. **32**(8), 1443–1458 (2010)